Bachelor's Thesis

Can Machine Learning Aid Econometrics? A Methodological Study on Heterogeneous Treatment Effects Estimation

Matilde Dolfato

Supervisor: Prof. Jerome Adda

Economic and Social Sciences
Bocconi University

September 18, 2024

Abstract

The study of policy effects is core in empirical economic research. The effect of a policy is likely to vary across different units of a population. Recently, many approaches to estimate heterogeneous treatment effects have been proposed in econometrics, also touching upon the machine learning field. This has triggered the emergence of causal machine learning, and empirical economists have diverging opinion on whether it should be used. The main contribution of this study is to provide an exhaustive methodological review on the estimation of heterogeneous causal effects. Moreover, it investigates the impact of causal machine learning on econometrics, attempting to answer the question of whether and how it can be beneficial. It is found that the main methods to estimate heterogeneous effects are: including interactions in a linear regression; the marginal treatment effect estimator; matching on propensity scores; local regressions with differencing; double machine learning; and causal forests. Moreover, I argue that machine learning can be beneficial to econometrics, especially as it flexibly models data and can be adapted to accurately estimate causal effects, and therefore it should be integrated in this discipline.

A special thank to professor Jerome Adda, who inspired me everyday in class, and taught me something that is found in this thesis, and goes way beyond it.

I thank professor David Rossell, who first sparked in me the curiosity for machine learning and helped crafting the idea for this thesis, and Lorenzo M, who gave me precious suggestions on these matters and much more.

I am also deeply grateful to professors Massimo Marinacci, Maristella Botticini and Chiara Fumagalli: I am honoured to say that you gave me some advice I will always keep with me.

A heartfelt thank to my friends Lucrezia, Bianca, Ilaria, Lorenzo G, Francesco S, Francesco D, Luigi, Federico, Gabriele N, Alessandro, Zoe, Michelle and Gabriele P. You made these three years better than I could have ever imagined. Thank you for keeping me going every time I don't see how or don't know where to go; I found in you another home.

Thank you also to my best friends Chiara, Roberta, Celeste and Anna, who will always remain such, no matter where life brings us.

My last thanks to Elisa and Paola, without whom I sincerely think I would be completely lost. You are my role models and I feel incredibly grateful to have you by my side and be by yours for all the life to come.

This thesis is dedicated to my parents Giovanna and Fabio, the best team I know. You give me an incredible strength to navigate life, while reminding me not to take it too seriously.

Contents

Introduction **7**

Preliminary Concepts 10

1 Estimating Heterogeneous Treatment Effects **17**

1.1 An Overview 17

1.2 The Methods 19

1.2.1 Including Interactions in a Linear Regression 19

1.2.2 Marginal Treatment Effect Estimator 20

1.2.3 Matching on Propensity Scores 22

1.2.4 Local Regression with Differencing 23

1.2.5 Double Machine Learning 23

1.2.6 Causal Forests 25

2 The Impact of Causal Machine Learning on Econometrics **27**

2.1 Are They Even Different? 27

2.2 The Debate 30

2.3 Empirical Evidence 31

2.3.1 Discussion 35

Conclusions **37**

Introduction

The study of policy effects is core in empirical economic research. A major part of this discipline is concerned with investigating how a certain policy, called treatment, affects a population (of individuals, firms, countries, etc.), by measuring the causal relationship between such treatment and a certain variable of interest. The econometric literature provides a wide range of methods to estimate the treatment effect on a population on average (some collections are Angrist and Pischke, 2009; Imbens and Rubin, 2015; Miguel, Hernan, and James, 2023). The effect of a policy, however, is likely to vary across different units of a population, probably at the individual level. This variation is due to individuals differing in many characteristics, that determine how they react to a policy, and that they may self-select into treatment. Estimating the *heterogeneous* effects of the policy, at least for some subgroups of the population, is crucial for social policy. In fact, if policymakers have information on which subgroups benefit the most from a treatment, they can choose more effectively to which part of the population and to which extent the treatment should be applied, allocating resources more efficiently (Heckman and Vytlacil, 2005; Xie, Brand, and Jann, 2012; Zhou and Xie, 2020).

The econometric literature on heterogeneous causal effects estimation, however, has not a long history, and methods to perform this task have been proposed only quite recently. This may be due to the inferior availability of data and precise individual-level information that there was in the last century, compared to today (Wager and Athey, 2018). In fact, the pioneer of these methods is Angrist, Imbens, and Rubin [1996]'s Local Average Treatment Effect (LATE) estimator. It is an Instrumental Variables (IV)-based method, which estimates the policy effect only for a subgroup of the population, those who comply with a change in the instrument. The Nobel Prize in Economic Sciences [1] to G. Imbens and J. Angrist in 2021 (together with D. Card) particularly honours this significant breakthrough in econometrics (Lechner, 2023). Indeed, the LATE estimator paved the way for the development of other methods to estimate heterogeneous treatment effects, and since then novel advanced approaches have been proposed. Among these, there are both structural models developing on the IV approach (e.g. Heckman and Vytlacil, 2005), and others using trending statistical techniques like non-parametric methods (e.g. Nie and Wager, 2021; Xie et al., 2012).

These advancements in econometrics have also touched upon the computer science and machine learning fields. Machine learning is a subfield of artificial intelligence, that develops algorithms to be applied to data, mainly for tasks of prediction and classification. These disciplines have come to the forefront in the last decades, bringing profound technological improvements, also tied to the access to larger datasets and greater computational capabilities of recent times. Indeed, economists have been using methods that typically belong to machine learning for economic research. Most recently, they are also adapting them to the specific case of causal inference, developing the new area of research of *causal* machine learning (Athey, 2019; Efron, 2020; Jordan and Mitchell [2015]; Lechner, 2023).[2] In fact, the estimation of heterogeneous causal effects appears to be one of the most suitable econometric tasks for the application of causal machine learning tools (Curth, Peck, McKinney, Weatherall, and van Der Schaar, 2024; Marginal Revolution University, 2022).

However, the use of causal machine learning is not embraced by all empirical economists. Some of them strongly support it for the prediction accuracy and computational capacity of machine learning algorithms, while others are more cautious and skeptical about its suitability for economic studies.

[1] More specifically, the "Sveriges Riskbank Prize in Economic Sciences in Memory of Alfred Nobel".

[2] Throughout this thesis, I use "causal machine learning" and "machine learning for causal inference" or "for empirical economic research" as synonyms.

For instance, in an interview by I. Andrews in May 2022, the Nobel Laureates that contributed so much to the advancement in econometric methodology themselves, G. Imbens and J. Angrist, expose sharply diverging opinions on the impact of machine learning on economics (Marginal Revolution University, 2022). Another example are J. Heckman and R. Pinto, who in a paper comparing some methods from econometrics and computer science, conclude that "*Economics has a rich body of theory and tools to address policy problems. Applied economists today would do well to use the impressive body of tools embodied in modern structural econometrics*" (Heckman and Pinto, 2022). S. Athey and G. Imbens, on the other hand, present an overview of machine learning methods relevant to economics, and state that the list of tools they discuss "*should be part of the empirical economist's toolkit and should be covered in the core econometrics graduate courses*" (Athey and Imbens, 2019).

As it is evident, a vivid debate is currently underway among the economic academia, and although it is not possible to go back in time and completely deny the use of machine learning, the way in which it can affect empirical economic research is still an open question.

In light of this scenario, the aim of this study is to clarify the recently developed array of methods available to economists for estimating heterogeneous causal effects, combining econometrics and machine learning. Furthermore, this thesis seeks to understand what are the roots of the two opposing views on the use of causal machine learning. By doing so, it intends to contribute to answering the question of whether and how machine learning can actually benefit empirical economic research, especially in the case of heterogeneous treatment effects estimation.

Among similar methodological surveys there are, for the econometric part, Xie et al. [2012] and Zhou and Xie [2020], and for the machine learning side, Gong, Hu, Basu, and Zhao [2021]; Knaus, Lechner, and Strittmatter [2020] and part of Athey [2019]. Moreover, there are works comparing the two approaches, like Dorie, Hill, Shalit, Scott, and Cervone [2019] and Heckman and Pinto [2022], but concerning the estimation of the average treatment effect and considering rather specific algorithms. Hence, the main contribution of this study is to combine the two fields and provide an exhaustive overview of all the fundamental methods economists have to estimate heterogeneous causal effects, and also to assess whether the machine learning ones can be advantageous for this task.

The rest of this thesis is organized as follows. First, the next section explains some basic concepts and algorithms to ease the understanding of the following technical analysis and discussion. Chapter 1 concerns the methodology review and explains each method in detail. Chapter 2 investigates how causal machine learning methods relate to econometric ones and what are their potential gains, and looks at some empirical evidence from the literature to assess how they perform for heterogeneous effects estimation.

Preliminary Concepts

The general setting considered throughout this thesis is that of an observational study on a sample of n individuals, each one denoted by i. Some of these are exposed to a treatment D, distinguishing the "treatment" group ($D = 1$) and the "control" group with ($D = 0$), to which no treatment (or in a more general setting, for other values of D, a different one) is applied. We study the effect of the treatment on an outcome variable of interest Y given a set of J individual characteristics X. We further denote Y_1 the potential outcome for the treated individuals ($D = 1$), and Y_0 the potential outcome for control individuals ($D = 0$). We indicate in lowercase the realizations of a variable, and with a hat the estimates. For instance, y is a realization of Y, and $\hat{y}$ is an estimate.

In general, treatment effects are estimated as the difference in outcomes between treated and non treated units, after controlling for the characteristics X to ensure comparability between the groups (Imbens and Wooldridge, 2009).

Furthermore, some approaches to studying policy effects assume that the treatment is exogenous. The exogeneity (or ignorability) assumption states that the treatment is independent of potential outcomes given X, i.e.

$$(Y_1, Y_0) \perp\!\!\!\perp D|X \text{ }^{3} \tag{1}$$

Violations of this assumption are given, for instance, by individuals self-selecting into the treatment, that is one of the source of heterogeneity in policy effects, besides individuals reacting differently to the treatment (Zhou and Xie, 2020).

Propensity Scores Many methods to estimate heterogeneous treatment effects (HTE henceforth) presented in this study are based on propensity scores. Let us define propensity scores as the probability of an individual in the sample of receiving the treatment, given their characteristics X, i.e.

$$P(X) \equiv Pr(D = 1|X) \text{ }^{4} \tag{2}$$

Rosenbaum and Rubin [1983] crucially found that under exogeneity, it is sufficient to condition on propensity scores $P(X)$ to estimate HTE. So, the assumption in equation 1 becomes

$$(Y_1, Y_0) \perp\!\!\!\perp D|P(X) \tag{3}$$

In fact, propensity scores allow to summarize the relevant information in the possibly high-dimensional covariates X $(n \times J)$ in a lower dimensional variable $P(X)$ $(n \times 1)$, to more easily balance covariates and ensure comparability of the units between the treatment and control group.

Kernel Regression The kernel regression is a specific type of local regressions. Local regressions are, in general, non-parametric statistical methods that estimate the

[3] I refer to the notation used by Xie et al. [2012].

[4] I refer to the definition used by Xie et al. [2012], even though this definition of propensity scores is generally used and accepted.

outcome Y *locally*, allowing for a non-linear relationship between the covariates and the outcome. The idea is to estimate the outcome y_i as a weighted average of the outcomes of the observations "close" to i, meaning with values of X in a neighbourhood of x_i.[5] The kernel regression uses a specific weighting function, the kernel, in this algorithm.

More specifically, following the lines of Aman [2016], say we want to estimate the function $\mu(X)$, where

$$Y = \mu(X) + U \tag{4}$$

with U being the random error component. We do so by modelling $\mu(x)$ locally at a given x, applying a linear regression to the data *close* to x. This means that we consider the linear regression

$$y_i = \alpha(x) - (x_i - x)\beta(x) + u_i \tag{5}$$

for each x_i with $i = 1, ..., n$ to determine the local regression function $\alpha(x)$ for $Y = y$ at $X = x$. The local linear regression method leads to the minimization problem

$$\min \sum_{i=1}^{n} (y_i - \alpha(x) - (x_i - x)\beta(x))^2 \, K\left(\frac{x - x_i}{h}\right) \tag{6}$$

where $K(\cdot)$ is the kernel weighting function. It is a decreasing function of the distances of x_i from the point x, with h being the bandwidth that determines the amount of local information used to determine the estimated outcome y. By solving the minimization problem, we obtain the estimated regression function for Y at $X = x$,

$$\hat{\mu}(x) = \hat{\alpha}(x) \tag{7}$$

and $\hat{\beta}(x)$ is the local slope. Iterating this process for each x in X we estimate a smoothed curve of the outcomes Y.

Regularization, Over Fitting and Cross-Validation In general, machine learning algorithms intrinsically perform *regularization*. Regularization is the auto-

[5] For simplicity, assume that here X contains only one variable.

matic selection of relevant covariates performed by statistical methods, reducing errors caused by *over fitting*. Over fitting, on the other hand, occurs when a model fits the data too closely, hence it is unable to make accurate out-of-sample predictions and leaves little to the interpretation of causal relationships. Selecting covariates implicates reducing the variance of the estimates and also the risk of over fitting, but possibly biasing the results (as some information is "missing"). Hence, this creates a tradeoff between bias (regularization) and variance (over fitting), which are in general inherently balanced by machine learning algorithms. The next two paragraphs provide technical examples of how this mechanism works.

Furthermore, *cross-validation* is a technique used in machine learning to evaluate the performance of a model and provide for over fitting. In brief, it consists in dividing the sample in multiple subsamples, or folds, and using some of them to train the model, and the rest of them (generally, only one) to test it. This procedure is repeated many times, each time with a different test set (Refaeilzadeh, Tang, and Liu, 2009).

LASSO Regression The Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996) is a regularization technique typically used in machine learning. It adds a penalization term to the residual sum of squares of the regression, to inherently select the variables.

More specifically, to visualize how it works we look at the logit version of the LASSO, as presented by Goller, Lechner, Moczall, and Wolff [2020] for propensity scores estimation (see method 1.2.3).

Consider the minimization function

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left[-d_i x_i \ + \log(1 + \exp(x_i \tau)) \right] + \lambda \sum_{j=1}^{J} |\tau_j| \right\} \tag{8}$$

where d is an observation of the treatment variable D, and τ is the coefficient measuring the relationship between the covariates X and the treatment. The first summation term corresponds to the residual sum of squares, while the second one is the penalization term, controlled by λ. This corresponds to the bias that will result in the estimates.

By adding this term, the absolute value of the coefficients τ_j is shrank towards 0, reducing their importance or altogether eliminating them from the model, resulting in an automatic model selection.

Indeed, as λ increases, the penalization term will be higher and the regularization will be more stringent. Hence, the estimates will have a lower variance and a larger bias. So, the value of λ controls the tradeoff between variance and bias, and its choice is crucial. The optimal choice of the λ is based on data, generally with a cross-validation technique.

RANDOM FORESTS Random forests (Breiman, 2001) are a machine learning nonparametric and non-linear estimation technique. They are constructed on regression trees (Breiman, 2017), which are a specific version of decision trees where the variable to be predicted is numerical. Decision trees, in general, are a common machine learning algorithm that recursively splits the covariate space into intervals, while minimising the error in the prediction of the outcome variable. The final structure is a rotated tree, where the initial trunk contains all the data, and each branch corresponds to a split of the observations defined by a certain threshold or value (for categorical variables) of one or more covariates. At the end of the splitting procedure, in the final nodes (leaves), the predictions are given as the average of the outcomes of all the observations included in that leaf. [6]

A random forest is an ensemble of regression trees, meaning that it combines the output of multiple trees, built on different random subsets of the covariate space. In fact, the final model is obtained through an average of the many trees, to increase the accuracy of the prediction.[7]

In random forests, the balance between bias and variance is controlled by the deep to which each tree is developed - in other words, by the number of observations in the

[6] In simple words, an example of two branches with X being age and gender is: *for male individuals (branch 1), if they are older than 25 (branch 1.1), we predict $\hat{y_1}$; if they are younger than 25, $\hat{y_2}$ (branch 1.2); for female individuals (branch 2), if they are older than 40 (branch 2.1), we predict $\hat{y_3}$; if they are younger than 40, $\hat{y_4}$ (branch 2.2).*

[7] Note that the accuracy of a prediction is defined as the combination of (the opposite of) both its bias and its variance (Walther and Moore, 2005). This concept is crucial and should be kept in mind for the subsequent discussions in this study.

final nodes used for each prediction. A higher number of observations in the leaves determines a lower variance and a higher bias of the prediction. In fact, it indicates that less splits have been made, and thus that intuitively the prediction is more “coarse”. This mechanism, together with averaging between many trees to obtain the final forest, entails the automatic selection of the variables and controls over fitting (Athey, 2019; Goller et al., 2020).

Chapter 1

Estimating Heterogeneous Treatment Effects

This chapter pursues the first aim of this study, that of providing a survey of the main methods used by economists to estimate heterogeneous causal effects, combining classic econometrics and newly proposed causal machine learning methods. Before delving into the technical review in Section 1.2, Section 1.1 provides an overall idea of the methodological scenario.

1.1 An Overview

There is a diverse group of methods to estimate heterogeneous causal effects, that differ substantially from one another yet share some basic features and are related to each other.

First, there are regression-based methods (1.2.1, 1.2.2, 1.2.4), which either assume to include all the relevant individual characteristics (under the exogeneity assumption) or consider unobservables that could in part drive the heterogeneity, using an IV approach. This is the case of the marginal treatment effect (MTE) estimator by Heckman and Vytlacil [2005], which is based on a specific structural model. Alternatively, the idea behind regression-based methods is that we need to allow for non-linearities in the relationship between the response variable and the treatment, to obtain heterogeneous causal effects. Some econometric ways to do so are including interactions in a linear regression (for instance, discussed by Zhou and Xie, 2020) and the commonly used non-parametric local regressions (Ferwerda, Hainmueller, and Hazlett, 2015; Hastie, Tibshirani, Friedman, and Friedman, 2009; Huber, 2023; Imbens and Wooldridge, 2009; Shim and Lee, 2009).

Then, other methods consider matching the units between the treatment and control groups based on similar characteristics X (1.2.3), to ensure the comparability between the units and estimate treatment effects based on this comparison (e.g. Abadie and Imbens, 2016; Xie et al., 2012).

In addition to this, some of these methods use propensity scores, as proposed by Xie et al. [2012] (1.2.3, 1.2.4). Propensity scores replace the covariates X as balancing measure (for instance, in a regression), because they summarize the relevant information of the individuals in a lower-dimensional parameter easier to use, as explained in the related paragraph above.

Finally, there are two main uses of machine learning for HTE estimation. First, Cannas and Arpino [2019]; Goller et al. [2020]; Lee, Lessler, and Stuart [2010] and McCaffrey, Ridgeway, and Morral [2004], among others, propose to use algorithms from machine learning to perform the first-stage estimation of propensity scores, thus as an auxiliary, possibly complementary tool to econometric methods (like methods 1.2.3 and 1.2.4). Secondly, economists have been using off-the-shelf causal machine learning algorithms (methods 1.2.5, 1.2.6), that entirely substitute traditional ones (some examples are Athey and Imbens, 2015; Athey and Imbens, 2016; Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018; Bargagli-Stoffi, Cadei, Lee, and Dominici, 2020). Among others, Athey [2019], Huber [2023] and Lechner [2023] discuss the use of these methods for HTE estimation.

1.2 The Methods

In brief, the methods to estimate HTE presented in this study are: a linear regression with interaction terms; the MTE estimator by Heckman and Vytlacil [2005]; a matching algorithm that then smooths the estimated effects; a local regression applied to the treatment and control group, with differencing; the double machine learning algorithm by Chernozhukov et al. [2018]; and the causal forests algorithm by Wager and Athey [2018].

Also some insights on how each method relates to the others and when it is most suitable are given.

1.2.1 Including Interactions in a Linear Regression

The most simple way of estimating HTE is to include interactions terms between the treatment and the covariates in a basic linear regression. Indeed, if we linearly regress Y on the treatment D and the individual characteristics X, the coefficient of the treatment variable will give us an estimate of the average treatment effect (ATE), as it is equal to $\hat{Y}_1 - \hat{Y}_0$. We shall include in the regression an interaction term $X_j \cdot D$, with $j \in [0, ..., J]$, where X_j is a categorical variable or a discretized one with K groups or intervals, across which the treatment effect is expected to vary. The resulting regression is

$$Y = \alpha + \beta X + \delta_0 D + \delta_j X_j D + \epsilon$$

Estimating it will give us an estimate of the treatment effect for each category in X_j. The k-th estimate of the HTE will be equal to $\delta_0 + \delta_{j,k}$, with $k = 1, ..., K$.

The main drawback of this method is that it assumes linearity in the relationship between the outcome Y and the treatment and covariates, which could bias the results (Zhou and Xie, 2020). Moreover, it requires to know a priori which is the individual characteristic responsible for the heterogeneity, and this needs to be categorical or to be discretized. So, the method is a good solution for small and simple datasets, providing very clear results, but it is not suitable for more complex studies.

1.2.2 Marginal Treatment Effect Estimator

The marginal treatment effect (MTE) estimator proposed by Heckman and Vytlacil [2005] is a novel and advanced version among the IV approaches, implementing a *local* IV. It generalizes the ATE and LATE estimators, as they can be obtained as a weighted average of the MTE. The model on which the MTE is based combines treatment effect literature with structural estimation. Indeed, it allows for the selection into treatment to be determined by a decision rule (structural part), and this choice, as well as potential outcomes, can be influenced also by unobservables, relaxing the exogeneity assumption. Here, the MTE method is explained on the lines of Carneiro, Heckman, and Vytlacil [2011].

First, consider the regression

$$Y = \alpha + \delta D + \varepsilon \tag{1.1}$$

where δ is the effect of the treatment on the outcome. Define potential outcomes as

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1, \\ Y_0 &= \mu_0(X) + U_0 \end{aligned} \tag{1.2}$$

where U_1, U_0 are the unobservable components influencing potential outcomes and $Y_1 - Y_0 = \delta$. Furthermore, consider the latent variable choice model

$$I_D = \mu_D(Z) - V \tag{1.3}$$

and the decision rule

$$D = 1 \text{ if } I_D \geq 0; \quad D = 0 \text{ otherwise} \tag{1.4}$$

Notice that the unobservable component in the choice model V collects unobserved effects that make the individual *less* inclined to be treated, and define

$$U_D \equiv F_V(V) \tag{1.5}$$

where F_V is a strictly increasing function. Then the propensity scores $P(Z)$ are estimated with the instrument Z, getting

$$P(z) \equiv Pr(D = 1|Z = z) = F_V(\mu_D(z)) \tag{1.6}$$

Finally, the MTE, defined as

$$\text{MTE}(x, u_D) \equiv E(\delta|X = x, U_D = u_D) \tag{1.7}$$

can be estimated as a function of propensity scores and the unobservables. Indeed, we consider the case in which the propensity score equals the value of the unobservable component, $p = u_D$, i.e. when the individual is indifferent between being treated or not. So, we are able to estimate the MTE for the indifferent individuals at different values of the propensity score, as

$$\text{MTE}(x, p) = \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p} \tag{1.8}$$

for $P(Z) = p$ and $X = x$. Indeed, $P(Z)$ is referred to as a local IV. The group of individuals who are indifferent between being treated or not, for which we estimate the MTE at different values of the propensity score, are the relevant ones policy-wise, because they are the ones who, at the margin, would change their mind if treated, and let us estimate the optimal extent to which the policy should be applied.

Compared to the others, this method is a more structural approach and it relaxes the exogeneity assumption, which most likely does not hold in reality. This allows for a greater economic interpretability of the results and to apply the method to more datasets and studies. Alongside this, Zhou and Xie [2016] discuss in detail the differences between the following propensity scores-based methods (1.2.3 and 1.2.4) and the MTE, also finding that the former two are valid even under some violations of exogeneity, partly decreasing the distance between the applicability of the two approaches.

1.2.3 Matching on Propensity Scores

Matching is a non-parametric statistical technique that directly compares (groups of) units to estimate heterogeneous treatment effects as the difference in their outcomes, for each match. Xie et al. [2012] propose to use propensity scores to match the units and then smooth the resulting estimates of the treatment effects.

An intuitive idea of the rationale behind this method is provided by D'Agostino Jr [1998]. Individuals in the two groups with equal propensity scores have equal probability of being treated or not, as if they were randomized into either of the two groups. Randomized experiments, in fact, entail to randomly assign units to the treatment and control group, ensuring that on average there are no systematic differences between them. Thus, matching units on their propensity scores allows to directly compare the outcomes of each match of treated and non treated individuals and estimate the treatment effect.

The algorithm works as follows. First, it requires to estimate $P(X)$ and to match treated and non-treated units with the 1:1 matching framework. [1] Then, it computes estimates of the HTE as the difference in the outcomes of each pair of units, $\Delta y = y_1 - y_0$, getting individual-level information on the treatment effect. Finally, we fit a local regression of Δy on $P(\hat{X})$, to obtain a smoothed curve of the treatment effect heterogeneity.

The first-step estimation of propensity scores can be performed with many algorithms. A traditional approach is to use a logistic regression, as in Xie et al. [2012]. Alternatively, more advanced algorithms like a LASSO regression or random forests can be used, as in Goller et al. [2020] and Knaus et al. [2020]. The implications of the

[1] See Xie et al. [2012] for more detail on the many matching types and algorithms that can be used.

choice of the algorithm to predict propensity scores are discussed in the next chapter, in Section 2.3.

1.2.4 Local Regression with Differencing

Another option economists have is to use local non-parametric regressions on propensity scores, as presented by Xie et al. [2012] as *"Smoothing - Differencing method"*. The fundamental concept behind this method is to use a local regression to allow for a non-linear relationship between the covariates and the response variable, as explained in the paragraph on the kernel regression.

In fact, the first step consists in estimating $P(X)$, with the algorithms discussed for method 1.2.3. Then, we fit two local regressions of the outcome Y on $\hat{P(X)}$ for the treatment and control groups separately, getting two smoothed curves of the fitted values. Finally, we take the difference between the two regression lines over the common support of $\hat{P(X)}$, to get the estimates of HTE for each value of the propensity score.

This approach can be considered as an extension of the interactions method 1.2.1. Indeed, including interactions in a linear regression is a way of allowing for a piece-wise linear relationship between D and Y, where the slope (the estimated coefficient) can change at different values of X_j (the variable giving the heterogeneity). Local regressions extend this idea in a non-parametric form, allowing for a non-linear relationship over the whole range of X, so they relax further the functional assumptions on the relationship between Y and D. A further development of this approach are machine learning algorithms that are based on interactions, like causal forests by Wager and Athey [2018] that are presented below (method 1.2.6).

1.2.5 Double Machine Learning

The first off-the-shelf causal machine learning method is double machine learning (DML), first proposed by Chernozhukov et al. [2018].

The concept at the heart of this method is not far from the idea explained at the end of Section 1.2.3, i.e. combining machine learning and econometric methods, using the former to estimate nuisance parameters. [2] However, these estimates may be biased, as explained in the paragraph on regularization and further discussed below (Section 2.3). Indeed, DML is a more complete algorithm that uses a *doubly robust* approach and an orthogonal estimator of treatment effects to provide unbiased treatment effect estimates. Doubly robust methods (see for reference Huber, 2023; Kennedy, 2023) propose estimators that are a function of both propensity scores and potential outcomes, thus requiring that either one of the two models is correctly specified, being more robust to model mis-specification (and biased estimates) than previously discussed methods.

Technically, the DML method consists in a two-step estimation of treatment effects. First, we estimate the (potentially) high-dimensional propensity scores and potential outcomes using algorithms from machine learning (again, like a LASSO regression or random forests). Then, we plug the residuals from these estimations in a residual-on-residual regression, obtaining a score function that is Neyman orthogonal, thus it is not sensitive to the possible biases in the nuisance parameters estimates. Lastly, it uses cross-validation techniques to correct for over fitting.

Finally, for what regards the second-stage model to estimate the treatment effect, the DML approach has been analyzed in multiple papers in the literature. The original paper Chernozhukov et al. [2018] considers the case of estimating the ATE or a low dimensional LATE. Instead, in the specific case of HTE estimation, the online article Ahmed [2022] proposes to use an interaction model in the second step to allow for non-linear relationships, so it would be an evolution of method 1.2.1. Huber [2023] suggests to regress the score function defining the ATE on X or on a subset of it to get an estimate for HTE. Nie and Wager [2021] considers the case in which the function identifying treatment effects can be modeled through a kernel regression, and thus augment method 1.2.4 above. Lastly, Wager and Athey [2018] models treatment effects

[2] Nuisance parameters are a subset of the parameters of a model (here, propensity scores and potential outcomes) that are needed for the model specification, but only in order to estimate other parameters, the ones which we are interested to inference (here, the treatment effects).

with a causal forest, which we present in the next section.

1.2.6 Causal Forests

Another causal machine learning method to estimate heterogeneous treatment effects are causal forests, proposed by Wager and Athey [2018]. Causal forests are an adaptation of the common machine learning algorithm of random forests to causal inference. In causal forests, briefly, the "leaves" of each tree give the estimates of causal effects rather than estimates of the outcome Y. The splitting algorithm, in fact, maximizes the heterogeneity in the treatment effect, instead of minimizing the prediction error as in random forests. Thus, the estimated treatment effect varies across different values of the covariates in X. Moreover, the algorithm is "honest", in the sense that for each *causal* tree, the sample used for prediction is different from the samples used to build the trees and split the covariate space.

Indeed, causal forests can be seen as a machine learning version of allowing for non-linearities in the relationship between X and D, and, in this view, a development of the interaction method 1.2.1 and to some extent the local regression method 1.2.4. As Wager and Athey [2018] explains, in fact, this algorithm is close to kernels and matching in that they estimate the parameters as a weighted average of "nearby" observations, but they determine which observations receive more weight based on data, that is particularly important in complex and high-dimensional settings.

Chapter 2

The Impact of Causal Machine Learning on Econometrics

Some of the methods presented come from the emerging area of causal machine learning. As suggested above, empirical economists have diverging opinions on its use, and what is the impact of (causal) machine learning on empirical economic research is still debated (Athey, 2019; Desai, 2023; Dorie et al., 2019; Heckman and Pinto, 2022; Marginal Revolution University, 2022).

This chapter aims at shedding light on whether and how machine learning methods can be used beneficially in empirical economic research.

First, Section 2.1 presents how machine learning tools relate to econometric ones; then Section 2.2 dives into the potential gains and losses of machine learning for empirical economics, outlining the ongoing debate; lastly, Section 2.3 discusses in this regard the machine learning methods considered in the previous survey, using evidence from the empirical literature.

2.1 Are They Even Different?

S. Athey, while discussing the impact of machine learning on economics, questions if machine learning really needed a new name other than statistics (Athey, 2019).

In fact, the basic algorithms behind machine learning are not new to empirical economists, because they share the same root as statistical regression methods (Efron, 2020). Consequently, it is not always possible - nor necessary - to draw a line and classify each technique as either machine learning or econometric, as also arises from this study. For instance, the LASSO regression is widely recognized as a machine learning tool, but received some attention in the econometric literature before machine learning arrived in this field. Even more strikingly, the cross-validation technique is viewed as a fundamental part of machine learning algorithms, but has been historically used in economics, for example to determine the bandwith for a kernel regression (Athey, 2019).

Having said that, there are some criteria on which machine learning methods can be distinguished from traditional econometric ones, and which justify the debate on their use.

First of all, methods from machine learning are tailored to handle very complex high-dimensional datasets, with a huge number of covariates and observations, compared to smaller datasets traditionally used in economic research (Desai, 2023; Efron, 2020).

Then, machine learning algorithms are focused on prediction, rather than on estimation (see Efron, 2020 for a detailed explanation of prediction vs. estimation in general). This means that the concern of these methods is the goodness of fit of the model, and they aim at obtaining an accurate prediction of the outcomes (very high explanatory power). On the contrary, as causal relationships are the main interest of empirical economic research, classic econometric methods are focused on obtaining unbiased estimates of the causal effects between variables, at the cost of the goodness of fit of the model and of predictive accuracy of the outcomes (possibly low explanatory power). In fact, machine learning models generally include myriads of weak predictors, as opposed to econometric models that include only few strong explanatory variables. Related to this, econometric models are based on statistical theory and the estimators they provide have asymptotic properties, meaning they are asymptotically normal and their results can be inferred to a whole population. This is crucial for economic causal inference, because for instance it allows to obtain confidence intervals and perform hypothesis testing. True values of real causal effects are generally not available, hence this property is particularly important to assess the quality of the estimation. Traditional (predictive) machine learning algorithms do not have such features, and therefore they are more suitable for simpler classification purposes, where a ground truth (the observed outcomes) is available (Athey, 2019; Efron, 2020).

The last crucial point is that machine learning methods perform a data-based model selection, after a (large) set of covariates is provided by the researcher. This, intuitively, relates to the very high goodness of fit typical of these models. Also, this determines a tradeoff between over fitting and regularization, which machine learning algorithms intrinsically balance. Please refer to the paragraph on regularization and over fitting for a more detailed explanation of this mechanism. On the contrary, empirical economists generally specify only one model, then use all the data for the estimation, and inference the results based on statistical theory (Athey, 2019). This is also related to the typically smaller size of datasets used with econometric methods, compared to machine learning ones.

Recently a new movement is emerging, combining machine learning algorithms with causal inference. The so-called causal machine learning is a new research area that is adapting machine learning tools to estimate causal effects (some leading examples are Athey and Imbens, 2015; Athey and Imbens, 2016; Athey et al., 2019; Bargagli-Stoffi et al., 2020; Chernozhukov et al., 2018; Wager and Athey, 2018). In doing so, the objective is to harness the strengths of machine learning, while providing for some of the shortcomings these algorithms present when applied to economic research. For instance, Wager and Athey [2018] adapt random forests providing causal forests with a tractable asymptotic theory and allowing statistical inference. In fact both causal forests and the DML method presented above belong to this movement (methods 1.2.6 and 1.2.5). Therefore, causal machine learning is likely to further narrow the distance between the econometric and machine learning traditions (Athey, 2019; Lechner, 2023).

2.2 The Debate

Based on the aspects that distinguish machine learning methods from econometric ones, some potential gains together with some limits of using machine learning in economic research arise; these outline the vivid debate going on among economists on this matter.

One of the arguments of economists that support the use of machine learning for empirical economic research is the data-driven model selection. In fact, it is more solid than classic procedures and it avoids economists' custom of checking alternative models "behind the scenes" (Athey, 2019). Moreover, this property has many implications. It avoids imposing functional model assumptions, as the functional form is at least in part derived from the real distribution of data. Indeed, these algorithms flexibly model the (possibly complex) relationships between the variables. This also results in capturing non-linearities or other complex structures that could not be discovered in advance, and that traditional methods could fail to detect. This, for instance, is particularly relevant to detect heterogeneity in the policy effects (Desai, 2023; Mullainathan and Spiess, 2017). In fact, S. Athey precisely maintains that regularization and automated model selection have several advantages on traditional methods and will likely become part of standard empirical practice in economics (Athey, 2019).

On the other hand, economists who are reluctant about the use of machine learning bring the main, strong argument of interpretability. Being able to interpret the algorithm with which some predictions are obtained is crucial for economic research, that tackles questions on causal effects rather than prediction. Machine learning algorithms, however, are traditionally not interpretable, as the algorithm they use is not known in detail to the researcher and as they do not provide causal effects estimates. Moreover, the lack of a structural component in machine learning models impairs further interpretability of causal relationships (Heckman and Pinto, 2022; Marginal Revolution University, 2022).

However, causal machine learning algorithms overcome this hurdle, at least in part. This is because they are tailored to estimate causal effects and adapted to present asymptotic theory and perform causal inference, as explained at the end of the previous section. Therefore, they leverage the accuracy of the prediction and the flexibility typical of machine learning algorithms to produce accurate estimates of causal effects and aid econometrics in finding the correct functional form to achieve unbiasedness (Athey, 2019).

Finally, as explained above, machine learning algorithms are tailored to handle huge datasets, and especially for a number of covariates that is much larger than the number of observations. This feature could be void for economic studies, which are generally based on few covariates. For instance, in the specific example of estimating policy effects, getting effects estimates for very specific groups of the population identified by many characteristics could be useless and not actionable upon (Marginal Revolution University, 2022). In fact, with simpler datasets classic econometric methods are sufficient, and machine learning ones do not appear to be useful (Desai, 2023). At the same time, nowadays there is an easier and easier access to datasets that are many orders of magnitude larger than before. Indeed, machine learning tools combined with newly available data sets could change economic research providing new questions, new approaches, and more interdisciplinary works (Athey, 2019).

2.3 Empirical Evidence

The final question addressed now is how the specific methods considered in this study relate to the discussion on machine learning applied to economic research, and thus whether they improve heterogeneous causal effects estimation.

First of all, the machine learning methods considered overcome at least in part the non-interpretability argument discussed in Section 2.1. Causal forests do so by presenting asymptotic properties. Also using machine learning for first-stage estimation of nuisance parameters, either with DML or as part of the matching or local regression methods, overcomes this limit. As explained by Lechner [2023] and Lee et al. [2010], in fact, in the first step we are not interested in the causal relationship between controls X and the treatment or the outcome. In the second step, instead, we want to be able to understand the causal effect of the treatment on the outcome, and thus we implement traditional more interpretable econometric methods to estimate treatment effects.

Furthermore, there is some empirical literature that compares machine learning algorithms to a logistic regression to estimate propensity scores.

Goller et al. [2020] find that the LASSO regression performs particularly well compared to a logistic regression. Random forests, on the other hand, appear to lead to misleading results especially when the share of treated is low, possibly because the algorithm does not split deep enough to properly balance covariates. [1] Moreover, they conclude that since knowing a priori which of the many methods works best, implementing off-the-shelf causal machine learning algorithms like DML or causal forests could be more convenient. Lee et al. [2010], instead, find that ensemble methods, like random forests or an advanced version of regression trees, show the best performance in balancing covariates to then estimate effects, recommending them for propensity scores estimation. Cannas and Arpino [2019] provide similar results, preferring random forests to other machine learning algorithms and recommending them in place of logistic regressions to estimate propensity scores. Lastly, Lee et al. [2010] find that the random forest algorithm performs better than others also for smaller datasets (with $n = 500$).

[1] With a small share of treated observations, splitting much the covariates space may easily lead to overfitting, as few (treated) observations would remain in each leaf

Having said that, a matter to be addressed is the presence of a regularization bias in propensity scores estimates. Indeed, as explained in the paragraph on regularization and over fitting, the inherent variable selection of machine learning algorithms enhances prediction accuracy, but possibly produces biased estimates. Goller et al. [2020] maintains that it is necessary to correct for this bias in propensity scores as it can translate to a higher bias in the treatment effect estimates. To this end, they implement a matching estimator with bias adjustment, but in the case of average treatment effect estimation. So, further work should be conducted to obtain HTE bias adjusted estimators, possibly on the lines of the matching or local regression algorithms (methods 1.2.3 and 1.2.4). At the same time, Lee et al. [2010] consider that also econometric methods can lead to biased estimates if the modelling assumptions on which they are based are incorrect. Indeed, they compare a logistic regression without including interactions or non-linearities (powers of X) to many machine learning algorithms. They find that when the real data presents non-linearity, the logit model gives the highest bias; also when there is linearity, its performance is comparable to that of some of the other methods. So, they implement the machine learning algorithms without correcting for the bias when estimating treatment effects. Thus, machine learning algorithms could possibly be implemented for first-stage estimation, jointly with the matching 1.2.3 or local regression 1.2.4 estimators for HTE, without the need of bias adjustment.

For what regards the DML and causal forests methods, there is not much empirical literature on comparisons to classic econometric approaches, especially for heterogeneous effects in economic studies. This may be because they are off-the-shelf causal machine learning algorithms substituting econometric ones, and also because they have been proposed very recently.

However, there are some papers that implement these algorithms for HTE estimation, with promising results.

Two of the few examples for the DML algorithm are Fuhr, Berens, and Papies [2024] and Wang, Huang, and Zhang, 2023. Both papers find that it gives more accurate results and that it fits non-linear relationships better than econometric methods. In the first paper, moreover, they state that structural analysis and assumptions are needed behind the application of this method, because it doesn't account for unobserved confoundness (as, for comparison, the MTE method 1.2.2 does). Indeed, as discussed by Lechner [2023], the upside of this algorithm is that it can be used as an helpful tool to human decision making, especially when combined with decision trees (or, equivalently, random forests), that clearly provide an idea of the criteria used by the machine learning algorithm, so that the human can understand it.

Among the studies implementing causal forests, Davis and Heller [2017] find that causal forests identify treatment heterogeneity that other interaction approaches would have missed. Bonander and Svensson [2021] conclude that the method is suitable for estimating heterogeneity, but that one of its limits is that it performs poorly in low-dimensional datasets. Additionally, there are some papers in the medical literature comparing the performance of causal forests to simple regressions. An example is Elek and Bíró [2021] [2], that maintains that, compared to a full interaction linear regression model, the automatic selection of which variables give the heterogeneity performed by causal forests improves statistical power, and they are therefore preferred. Venkatasubramaniam, Mateen, Shields, Hattersley, Jones, Vollmer, and Dennis [2023], instead, in a similar study conclude that causal forests should not be used alone, but always compared to classic regression methods, that in their evaluation performed better.

On the whole, it is found that the methods from the machine learning literature considered in this study allow for interpretability of the causal treatment effects, while harnessing the higher predictive accuracy of machine learning algorithms. Moreover, the empirical evidence shows that estimating nuisance parameters with machine learning gives more accurate results and better covariate balance, and both LASSO regression

[2] It is a medical study, but given the limited number of covariates included (around 10) its results are relevant also for economic applications.

and random forests are preferred to a logistic regression for this task. Their performance, however, varies depending on the studies, and using off-the-shelf causal machine learning methods may be recommendable. These, in fact, appear to perform well for heterogeneous effects estimation, also detecting heterogeneity patterns that classic econometric methods miss.

2.3.1 Discussion

From this study it arises that causal machine learning can bring fascinating improvements to empirical economic research. The main benefit it brings is the ability to model the data with highly flexible functional forms. This enables the researcher to discover hidden patterns of the relationships between variables, that are the basis of empirical economics questions. This is especially true for the investigation of heterogeneous causal effects, as the empirical results support. Furthermore, causal machine learning provides empirical economists with a solid model selection procedure based on data and very accurate estimations of causal effects.

Alongside this, machine learning may also open new horizons for empirical economics, introducing new approaches and questions, possibly exploiting the abundance of very precise data available today.

It remains true that the methodological choice depends largely on the characteristics of the data and the study that is being conducted. Hence, the best choice is to integrate the two disciplines to leverage machine learning tools when they are most suitable, and use econometric ones when they are sufficient. Indeed, economic and statistical theory must remain the foundation of the model choice procedure.

As a last consideration, it is evident that more work should be conducted in the direction of embodying machine learning in econometrics and empirical economics, as this integration presents some frictions. Nevertheless, even from the early steps taken so far, it is evident that causal machine learning can greatly contribute to econometrics, suggesting that this direction should be pursued further.

Conclusions

Investigating heterogeneous policy effects is highly relevant for empirical economic research, and many methodological developments have been done in the last decades concerning this task (Zhou and Xie, 2020). Not only, but heterogeneous effects estimation is also closely related to the recent advent of machine learning techniques in econometrics, which has triggered a vivid debate among experts on the use of *causal* machine learning (Athey, 2019; Marginal Revolution University, 2022).

In light of this, this study has first reviewed the main methods to estimate heterogeneous causal effects, combining the two disciplines, and then it has inquired whether and how causal machine learning can benefit empirical economic research.

In summary, the main methods economists have to estimate heterogeneous causal effects are: including interactions in a linear regression, mostly suitable for simple datasets; the marginal treatment effect estimator, a structural approach that allows for unobserved resistance to treatment (by Heckman and Vytlacil, 2005); a matching algorithm and a local regression method based on propensity scores, which are flexible statistical approaches that can be combined with machine learning tools (by Xie et al., 2012); the causal machine learning algorithms of causal forests, an ensemble of regression trees adapted to causal inference (by Wager and Athey, 2018), and double machine learning, entailing machine learning estimation of nuisance parameters and the use of econometric models for the second-stage estimation of treatment effects (by Chernozhukov et al., 2018).

Moreover, it results that causal machine learning can greatly aid econometrics.

Indeed, the main benefits of these methods are to flexibly model especially high-dimensional data and to capture hidden patterns of heterogeneity. Also, their high prediction accuracy can be leveraged for estimating causal effects. This is supported by the empirical literature evidence, especially for causal forests and double machine learning. Hence, although further work is needed to fully combine these disciplines and economic theory must remain at the basis of model choices, integrating causal machine learning in empirical economics is a direction that should be pursued.

A proposed future development of this study is to conduct an empirical application of the methods considered, to compare their performances on the same dataset, and possibly also including other promising methods like bayesian additive regression trees by Hill [2011].

Bibliography

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

Zain Ahmed. Heterogeneous treatment effect using double machine learning, 2022. Available at: `https://towardsdatascience.com/heterogeneous-treatment-effect-using-double-machine-learning-65ab41f9a5dc`.

Ullah Aman. Kernel estimators in econometrics. In *The New Palgrave Dictionary of Economics*, pages 1–4. Palgrave Macmillan UK, London, 2016.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91 (434):444–455, 1996.

Susan Athey. The impact of machine learning on economics. In Joshua Gans Ajay Agrawal and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda*, pages 507–547. University of Chicago Press, 2019.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

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

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Falco J Bargagli-Stoffi, Riccardo Cadei, Kwonsang Lee, and Francesca Dominici. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

Carl Bonander and Mikael Svensson. Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Economics*, 30(8):1818–1832, 2021. doi: https://doi.org/10.1002/hec.4263. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4263`.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

Massimo Cannas and Bruno Arpino. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, 61(4):1049–1072, 2019.

Pedro Carneiro, James J Heckman, and Edward J Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754–2781, 2011.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Alicia Curth, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van Der Schaar. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 2024.

Ralph B D'Agostino Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281, 1998.

Jonathan M.V. Davis and Sara B. Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50, May 2017. doi: 10.1257/aer.p20171000. URL `https://www.aeaweb.org/articles?id=10.1257/aer.p20171000`.

Ajit Desai. Machine learning for economics research: when, what and how, 2023. Available at: `https://www.bankofcanada.ca/2023/10/staff-analytical-note-2023-16/#Machine-learning-for-economic-applications`.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

Bradley Efron. Prediction, estimation, and attribution. *International Statistical Review*, 88:S28–S59, 2020.

Péter Elek and Anikó Bíró. Regional differences in diabetes across europe–regression and causal forest analyses. *Economics & Human Biology*, 40:100948, 2021.

Jeremy Ferwerda, Jens Hainmueller, and Chad Hazlett. Krls: A stata package for kernel-based regularized least squares. *Available at SSRN 2325523*, 2015.

Jonathan Fuhr, Philipp Berens, and Dominik Papies. Estimating causal effects with double machine learning – a method evaluation, 2024. URL `https://arxiv.org/abs/2403.14385`.

Daniel Goller, Michael Lechner, Andreas Moczall, and Joachim Wolff. Does the estimation of the propensity score by machine learning improve matching estimation? the case of germany's programmes for long term unemployed. *Labour Economics*, 65: 101855, 2020.

Xiajing Gong, Meng Hu, Mahashweta Basu, and Liang Zhao. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT: Pharmacometrics & Systems Pharmacology*, 10(11):1433–1443, 2021.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

James J Heckman and Rodrigo Pinto. Causal inference of social experiments using orthogonal designs. *Journal of Quantitative Economics*, 20(Suppl 1):7–30, 2022.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Martin Huber. *Causal analysis: Impact evaluation and Causal Machine Learning with applications in R*. MIT Press, 2023.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. doi: 10.1126/science.aaa8415. URL `https://www.science.org/doi/abs/10.1126/science.aaa8415`.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Michael C Knaus, Michael Lechner, and Anthony Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161, June 2020. ISSN 1368-423X. doi: 10.1093/ectj/utaa014. URL `http://dx.doi.org/10.1093/ectj/utaa014`.

Michael Lechner. Causal machine learning and its use for public policy. *Swiss Journal of Economics and Statistics*, 159(1):8, 2023.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

Marginal Revolution University. How will machine learning impact economics? (guido imbens, josh angrist, isaiah andrews), 2022. Available at: `https://www.youtube.com/watch?v=kM8B2X8pdNA&t=448s`.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

A Miguel, Robins Hernan, and M James. *Causal inference: what if*. CRC PRESS, 2023.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL `https://doi.org/10.1007/978-0-387-39940-9_565`.

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.

Joo-Yong Shim and Jang-Taek Lee. Kernel method for autoregressive data. *Journal of the Korean Data and Information Science Society*, 20(5):949–954, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ashwini Venkatasubramaniam, Bilal A Mateen, Beverley M Shields, Andrew T Hattersley, Angus G Jones, Sebastian J Vollmer, and John M Dennis. Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine. *BMC Medical Informatics and Decision Making*, 23(1):110, 2023.

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.

Bruno A Walther and Joslin L Moore. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28(6):815–829, 2005.

Xia Wang, Xiaoyan Huang, and Yun Zhang. Application of machine learning method based estimation of heterogeneous treatment effects in economics. In *Proceedings of the 7th International Conference on Intelligent Information Processing*, ICIIP '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450396714. doi: 10.1145/3570236.3570262. URL `https://doi.org/10.1145/3570236.3570262`.

Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

Xiang Zhou and Yu Xie. Propensity score–based methods versus mte-based methods in causal inference: Identification, estimation, and application. *Sociological Methods & Research*, 45(1):3–40, 2016.

Xiang Zhou and Yu Xie. Heterogeneous treatment effects in the presence of self-selection: a propensity score perspective. *Sociological Methodology*, 50(1):350–385, 2020.