Leveraging Machine Learning to Estimate Effect Modification[1]

Ang Yu[1,*], Chan Park[2], Hyunseung Kang[2], and Jason Fletcher[1,3]

[1]Department of Sociology, University of Wisconsin-Madison

[2]Department of Statistics, University of Wisconsin-Madison

[3]La Follette School of Public Affairs, University of Wisconsin-Madison

[*]ayu33@wisc.edu

[Last edited: August 18, 2021]

Abstract:

Sociologists are often interested in estimating and testing whether some causal effect varies by a modifier of interest. The conventional regression estimator for effect modification is inflexible in functional form and prone to misspecification bias. Machine Learning (ML) algorithms can aid the estimation of effect modification in observational studies by controlling for confounders in a highly flexible, automated, yet principled way. Therefore, leveraging ML for effect modification helps reduce misspecification bias and enhance the credibility of causal identification. We introduce a novel estimator that estimates effect modification in a familiar regression framework after using ML algorithms to fit nuisance components of the model. We show that this estimator is more flexible than the conventional regression model while more efficient and suitable for theory-driven sociological research than other ML-based methods. We use the new estimator to study the modification in the effect of a college degree on adult family income by gender and

[1] This paper has been presented at the annual meeting of the American Sociological Association and the Stanford conference on computational sociology, both in 2020. We would like to thank Felix Elwert, Xiang Zhou, and Eric Grodsky for helpful suggestions.

family income in adolescence in the United States. Along these two dimensions, the benefits of a college degree are rather equally distributed.

1. Introduction

Machine Learning (ML) has been making its way into empirical sociology in recent years. For example, methodological adaptations of a wide variety of ML methods have appeared in life course research (Billari, Fürnkranz, and Prskawetz 2006), criminology (Baćak and Kennedy 2019), survey design (Fu, Guo, and Land 2018), text and image analysis (Rona-Tas et al. 2019; Zhang and Pan 2019), causal inference (Brand et al. 2021; Liu 2021; Torrats-Espinosa 2021), log-linear modelling (Bucca and Urbina 2019), and general evaluation of the predictability of life and physical outcomes (Daoud, Kim, and Subramanian 2019; Salganik et al. 2020).

For causal inference tasks based on observational data, ML can provide flexible yet principled ways to control for confounders, hence buttressing the credibility of causal estimates. Effect heterogeneity, as a specific causal inference problem, is at the core of sociological research (Xie 2013). And some recent developments in statistics and econometrics have started to harness the strength of ML to address effect heterogeneity. Molina and Garip (2019, p34, p37-38), in a general review of machine learning for sociologists, have briefly introduced some developments in using ML for effect heterogeneity.

However, the existing ML algorithms for effect heterogeneity target the approximation of individual-level treatment effects, taking an optimal treatment assignment regime as the eventual goal (Athey and Imbens 2019). It is rare that sociological research is guided by that goal. Indeed, most sociological research involving treatment heterogeneity focuses instead on theoretically motivated dimensions along which the treatment effects may vary, such as gender (Legewie and

DiPrete 2012), race (Gorry 2019), socioeconomic status (Xie et al. 2020), occupational characteristics (Yu and Kuo 2017) age (Wodtke, Elwert, and Harding 2016), and genotype (Fletcher 2012). In other words, the dominant approach to effect heterogeneity in sociology is effect modification, which is about whether and how some causal effect varies with specified modifiers of interest. The distinction between effect modification and other approaches of studying effect heterogeneity has not been adequately addressed in the methodological literature, potentially hindering optimally leveraging ML for effect modification in sociology.

This paper is structured as follows. First, we will review how ML-based methods can, in general, facilitate causal inference with observational data and introduce, in particular, the effect modification estimand. Second, we will also demonstrate, in more detail, differences between the effect modification estimand and other approaches to examine effect heterogeneity. Third, we will introduce a new ML-based Sample Splitting (ML-SS) regression estimator which is an extension of the Groupwise Inference Method developed by Park and Kang (2019). Compared with the Groupwise Inference Method, the ML-SS regression can accommodate arbitrary parametrizations of the treatment effect function and various auxiliary adjustments for survey data. Fourth, we will demonstrate that the ML-SS is particularly suitable for the effect modification estimand in a simulation study. Finally, we will present an empirical study that applies the ML-SS regression to investigate the modification roles of gender and family income in adolescence in the treatment effect of college degree on adult family income.

2. ML-based Methods and Observational Studies

In observational studies on causal effect, conditioning on observable confounders is a common strategy for identification of both average treatment effect (ATE) and effect modification.

For instance, one conventional way of estimating and testing effect modification is using the following generalized linear regression:

$$g[E(Y|W,M,\boldsymbol{X})] = \alpha + \beta_1 W + \beta_2 M + \beta_3 W \cdot M + \beta_4 \boldsymbol{X},$$

where W is the treatment, M is the modifier of interest, $g(\cdot)$ is a link function such as logit or probit. And β_3, the coefficient for the product term between the treatment and modifier, is taken as evidence of effect modification by M, or its lack thereof. Importantly, in this conventional model, the vector of control variables, $\boldsymbol{X}$, enters the model linearly or some other researcher-specified way.

In the conditioning-on-observables framework, the identifiability of the causal estimand depends first on whether the confounders are available to the researcher. If some confounders are simply not measured at all, then the researcher must either select an alternative identification strategy or use sensitivity analysis. However, even in a case where all confounders are observed in some form, there is still a considerable degree of uncertainty in how to practically implement the conditioning. Indeed, researchers often have access to a large number of potential confounders, but do not have a strong theory about which covariates (and which interaction terms between them) need to be conditioned on and in what functional form. However, misspecification in any manner will likely lead to bias in estimates and false conclusions. In particular, in the conventional regression for effect modification, the linearity assumed for the relationship between control variables and the outcome is a highly restrictive functional form. If there is deviation in the true function from linearity regarding the confounders, then bias will arise in the estimation of effect modification (see Breen, Choi, and Holm 2015).

In the face of uncertainty about model form and the peril of misspecification bias, sociologists have proposed to characterize the uncertainty in terms of a distribution and incorporate it in the analysis. Muñoz and Young (2018) propose to obtain estimates across all plausible model forms and assess the entire distribution of these estimates. The approach of Winship and Western (2016) is to designate a Bayesian prior distribution for a parameter that represents misspecification in OLS and integrate that distribution into inference. In a comment on Muñoz and Young (2018), Western (2018) also describes another distributional solution that amounts to averaging coefficients in all possible models, weighted by the posterior probability of each model.

In this paper, we focus on a different and complementary approach, where we strive to find an optimal model, instead of just recognizing the uncertainty as a distribution. One key to obtain the model that best approximates the true functional form is to allow for more flexibility in modelling. ML-based methods are generally nonparametric or semiparametric, fitting the model in a data-driven way without imposing researcher-specified parametrization. Consequently, ML-based causal inference has the advantage of flexibility over traditional parametric models.

Flexibility in estimation can be trivially achieved by aggressively adding more covariates in the model in ever more flexible forms. However, doing so will also increasingly cause overfitting, reduce efficiency, and weaken replicability. Hence, as we endeavor to avoid misspecification, we also need to guard against overfitting and efficiency loss. This is what ML methods are designed to do, as they regularize (limit) model complexity for best out-of-sample performance via cross-validation instead of greedily fitting the current sample.

When integrated with causal inference, ML methods are often also combined with cross-fitting, which amounts to using one part of the sample to fit the function that is used to predict values for another part of the sample. Cross-fitting also helps avoid overfitting and is crucial for valid inference after applying ML algorithms as part of the estimation. An early version of cross-fitting appeared in Chernozhukov et al., (2018) and has been adopted widely for causal inference using ML for nuisance functions (eg. Knaus, Lechner, and Strittmatter 2021; Nie and Wager 2020).

The "honest" estimation of Causal Tree in Athey and Imbens (2016) is also a type of cross-fitting. The ML-SS estimator that we introduce in this paper also uses cross-fitting, which is premised on sampling-splitting.

In summary, ML methods can provide principled flexibility in controlling for confounders, resulting in functional forms that are neither overly restrictive nor overly complex.

Furthermore, ML algorithms are automated procedures that, when applied to controlling for confounders, enable researchers to be free of ad-hoc decisions with regard to the confounders and focus on specifying and interpreting parameters of substantive interest. Automating the modelling of control variables can also help with preventing the bias-inducing behavior of model refinement or specification searching, i.e. p-hacking on the part of the researcher(Christensen, Freese, and Miguel 2019; Muñoz and Young 2018). [2] ML-based causal inference, thus, facilitates research practices that are both convenient and of integrity.

3. Effect Modification and Various ML Methods for Effect Heterogeneity

We first define the *effect modification* estimand formally. Next, we explain why existing ML methods are designed instead for *individual-level effect prediction* hence not directly and optimally suitable for sociological research targeting effect modification as the estimand. We will also briefly discuss the possibility of using ML methods to *discover previously unknown modifiers*, which is yet another goal of research different from the effect modification estimand. Figure 1 summarizes the relationships between the estimands or goals of research appearing in this paper.

Let W_i be a binary treatment variable, Y_i be the measured outcome and $Y_i(W_i = w)$ be the potential outcome that would be observed if i's treatment W_i was set to w by external intervention.

[2] Correspondingly, we argue that with regard to the parameters of interest, variables and functional forms should be pre-specified based on theory, preferably pre-registered.

Finally, let M_i be a low dimensional and pre-specified vector of pre-treatment variables. The effect modification estimand can then be defined as such:

$$E[Y_i(W_i = 1) - Y_i(W_i = 0)|M_i]. \tag{1}$$

Importantly, the effect modification estimand is different from the causal interaction estimand based on the potential outcomes $E[Y_i(W_i, M_i)]$. The difference is that the effect modification estimand only seeks to identify the causal effect of W_i, but not that of M_i (Keele and Stevenson 2020; VanderWeele 2009).

However, existing methods leverage ML for effect heterogeneity in a different way. The estimands of these methods can be similarly written in potential outcomes as

$$E[Y_i(W_i = 1) - Y_i(W_i = 0)|\boldsymbol{C_i}]. \tag{2}$$

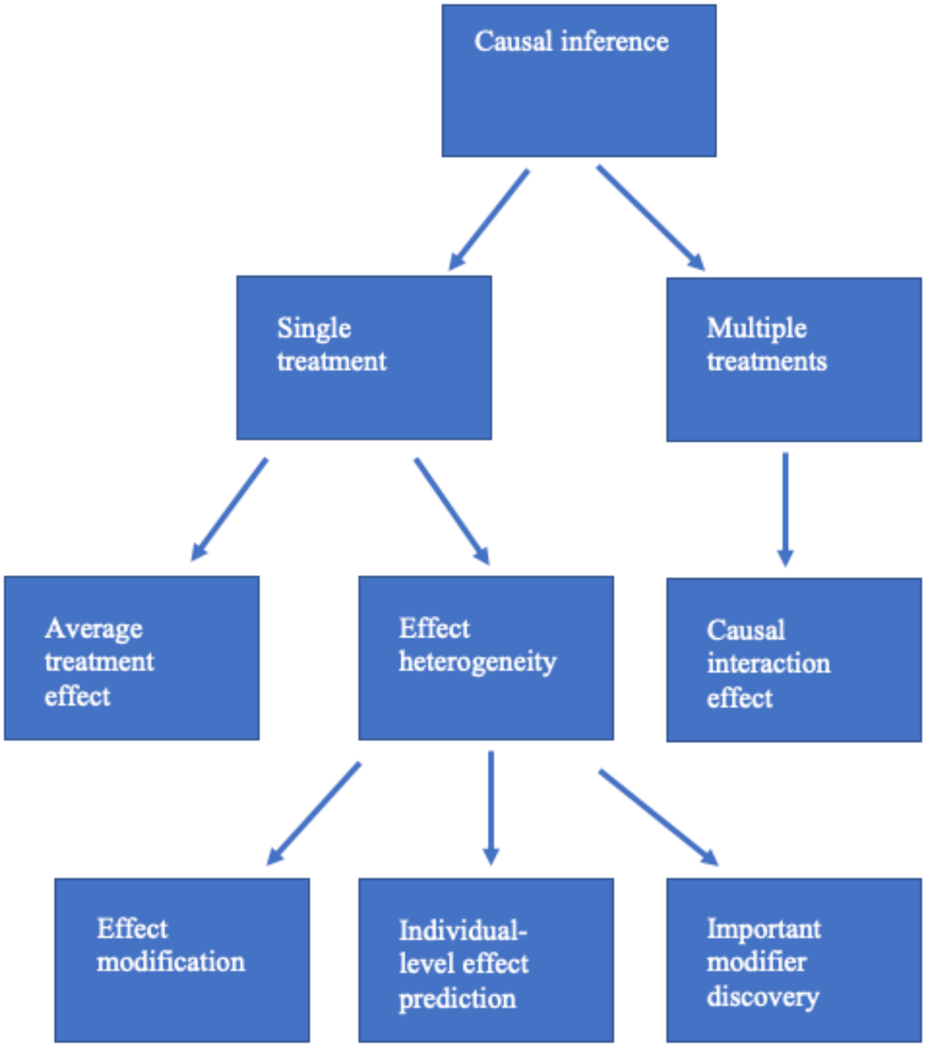
This can be called the individual-level effect prediction estimand, and the substantive goal of this estimand is to predict the individual-level treatment effect $Y_i(W_i = 1) - Y_i(W_i = 0)$ for each out-of-sample individual i as accurately as possible. [3] The key difference between individual-level effect prediction estimand in (2) and the effect modification estimand in (1) lies in that $\boldsymbol{C_i}$ is a high-dimensional vector of modifiers that are not pre-specified. Hence, while (2) aims to capture the entire generating process of treatment effects using as many modifiers $\boldsymbol{C_i}$ as possible, (1) summarizes the effect heterogeneity in a substantively important or interesting way with respect to just one modifier of interest, M_i. Formally, $M_i \in \boldsymbol{C_i}$, and the relationship between (1) and (2) can be expressed in terms of an iterated conditional expectation:

$$E[Y_i(W_i = 1) - Y_i(W_i = 0)|M_i] = E\{E[Y_i(W_i = 1) - Y_i(W_i = 0)|\boldsymbol{C_i}]|M_i\},$$

[3] Equivalently, the goal is also to approximate what is called the Conditional Average Treatment Effect function that includes all modifiers there are (Künzel et al. 2019:4157).

meaning that the effect modification estimand is a conditional expectation of the individual-level effect heterogeneity estimand, aggregating the latter over all modifiers except for M_i.

Figure 1.



There has been a fast growing number of methods that target the estimand in (2). They optimize the accuracy of individual-level predictions and directly output predictions at the individual level as opposed to the aggregate level of substantive interest. The causal forests, a variant of Generalized Random Forests (GRF), (Athey, Tibshirani, and Wager 2019; Athey and Wager 2019) is a prominent example in this category and has been applied in multiple social science settings (Brand et al. 2021; Daoud and Johansson 2019; Knittel and Stolper 2019; Tiffin

2019). Other examples include the R-learner (Nie and Wager 2020), the X-learner (Künzel et al. 2019), and the Modified Covariate Method (Chen et al. 2017; Tian et al. 2014). In fact, there were already at least 23 types of ML methods that target treatment effects at the individual level as of 2018 (Knaus et al. 2021) and the literature is still growing very rapidly. These methods tend to perform well for the task they explicitly target, i.e., predicting treatment effects for individuals, hence it is obvious that they are very promising for practical applications such as personalized recommendation and precision medicine. However, when the goal is to estimate and test effect modification, they become practically cumbersome and statistically inefficient, as they do not directly summarize effect heterogeneity along key dimensions of interest, nor do they allow for directly comparing treatment effects at different values of the modifier. In order to obtain effect modification estimates using these methods above, one has to manually average over modifiers in $\boldsymbol{C_i}$ but not in M_i, which is an unnecessary detour step and essentially wastes statistical power at the more granular level that is not of theoretical interest. This type of averaging procedure is indeed proposed by Lechner (2018) and Athey and Wager (2019) based on different specific implementations (for a sociological application, also see Liu 2021). However, as our simulation study will demonstrate, using the Augmented Inverse Propensity Weighting (AIPW) procedure in Athey and Wager (2019) to aggregate individual-level estimates to the group level is less efficient than the ML-SS regression that we propose and that directly estimates effect modification.

Apart from effect modification and individual-level effect prediction, there is yet another goal that may interest empirical researchers. Researchers have also employed ML methods to discover important effect modifiers that are previously ignored and may lead to novel insights about effect heterogeneity (Brand et al. 2021). For example, causal forests output a variable importance metric that can be used for this purpose.

In summary, of the three approaches for heterogeneity in treatment effect, individual-level effect prediction and important modifier discovery are exploratory analyses and do not follow the hypothesis-testing logic. The effect modification estimand, on the other hand, is fundamentally confirmatory, hence requiring a pre-specified modifier of interest and statistical testing for it. For effect modification, we need a more well-tailored method.

4. An ML-based Sample Splitting Regression Model for Effect Modification

We introduce a new ML-based Sample Splitting (ML-SS) regression estimator that targets effect modification by design. This new estimator extends the Groupwise Inference Method recently developed by Park and Kang (2019) to broader types of modifiers and integrates the Groupwise Inference Method with familiar practices in sociology with survey data. The strength of ML-SS regression lies in leveraging ML to nonparametrically control for confounders while focusing on parametrized modifiers based on theoretical hypotheses in the familiar and interpretable structure of a regression model.

Causal identification of effect modification using ML-SS regression depends on three identifying assumptions that routinely appear in causal inference on the familiar ATE (eg. Morgan and Winship 2014):

1) Conditional ignorability assumption:

$$Y_i(W_i = 1), Y_i(W_i = 0) \perp W_i | \boldsymbol{X_i}.$$

The conditional ignorability assumption states that conditional on a vector of pre-treatment control variables $\boldsymbol{X_i}$, W_i is statistically independent of the two potential outcomes of Y_i under assignment of $W_i = 1$ and $W_i = 0$. Roughly speaking, the conditional ignorability assumption is satisfied when there is no unobserved confounder that influences both treatment assignment and potential outcomes.

2) Consistency assumption:

$$Y_i = Y_i(W_i = 0) + [Y_i(W_i = 1) - Y_i(W_i = 0)]W_i.$$

The consistency assumption states that the observed outcome equals the potential outcome of the treatment that was actually assigned. Substantively, this means the hypothetical intervention does not change the causal mechanism linking W_i and Y_i.

3) Positivity assumption:

$$0 < P(W_i = 1|\boldsymbol{X_i} = \boldsymbol{x}) < 1 \; for \; all \; \boldsymbol{x}.$$

The positivity assumption requires that for all values of the confounder, the DGP does not mechanically rule out the possibility of being assigned to either treatment or control. This makes sure the expected potential outcomes $E[Y(W_i = 1)|\boldsymbol{X_i} = \boldsymbol{x}]$ and $E[Y(W_i = 0)|\boldsymbol{X_i} = \boldsymbol{x}]$ are sensible for all $\boldsymbol{x}$.[4]

Under these identifying assumptions, we can apply ML-SS regression to effect modification estimation, which involves two steps. In the first step, the original sample is split into two subsamples. In the first subsample, ML-SS fits two nuisance functions, $E(Y_i|\boldsymbol{X_i})$ and $E(W_i|\boldsymbol{X_i})$,[5] using any ML algorithm whose fitted function converges to the true conditional expectation function (CEF) in large sample. These two CEFs are called nuisance functions because they are only necessary for controlling confounders $\boldsymbol{X_i}$ and do not capture what is of interest, namely effect modification. Next, our estimator predicts $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$ for each individual in the second subsample using the model fitted in the first subsample. Then the roles of first and second subsamples are switched so that every data point is used to both fit the nuisance functions and to predict $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$.

[4] Technically, the positivity assumption is also needed for the design matrix in our final regression model to be invertible however we define $h(M_i)$.

[5] $E(W_i|\boldsymbol{X_i})$ is the propensity score.

After estimating $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$ for every data point in the sample, the estimator proceeds to the second step, which is fitting a regression model with outcome and treatment residualized against the nuisance functions. The regression model in the second step is as follows:

$$Y_i - \hat{E}(Y_i|\boldsymbol{X_i}) = \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]h(M_i) + \epsilon_i, \tag{3}$$

which is a model regressing the residualized Y_i on the interaction term between $h(M_i)$ and residualized W_i, where $h(M_i)$ is a parametrized characterization of effect modification.[6] In practice, the researcher needs to specify a parametric form for $h(M_i)$, then the regression model can be simply fitted using any standard statistical package, such as the "regress" command in Stata or the "lm" command in R. If M_i is a discrete variable, it is natural to use a groupwise parametrization:

$$h(M_i) = \sum_{g=1}^{G} \tau_g I_g(M_i),$$

where $I_g(M_i)$ is an indicator for each value of M_i and τ_g captures the groupwise ATE corresponding to group g. Therefore, the vector of τ_g, and the difference between the τ_g for different groups g, directly capture groupwise effect modification. In the regression model in (3), τ_g are just the coefficients on the interaction terms between $I_g(M_i)$ and $\left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]$.

If the modifier M_i is originally a continuous variable, it will be necessary to parametrize $h(M_i)$. Depending on one's belief about the true DGP, one can opt to discretize M_i into, for example, quantile groups and use the groupwise parametrization as above. Alternatively, it is possible to retain the continuous nature of M_i and use, for example, a linear parametrization for $h(M_i)$:

$$h(M_i) = \tau_c + \tau_s M_i,$$

[6] Residualization eliminates the intercept in the regression.

where τ_c is a treatment effect constant indicating the baseline treatment effect when $M_i = 0$ and τ_s is a treatment effect slope on M_i showing show the treatment effect linearly increases or decreases with M_i. Concretely, τ_c is the coefficient on the "main effect" of $[W_i - \hat{E}(W_i|\boldsymbol{X_i})]$ and τ_s is the coefficient on the interaction term between M_i and $[W_i - \hat{E}(W_i|\boldsymbol{X_i})]$ in the regression model in (3). Of course, for a continuous modifier, there are other possible parametrizations that may involve recentering or polynomials. The original Groupwise Inference Method (Park and Kang 2019) only focuses on the groupwise parametrization and in this paper, we extend the estimator to accommodate any arbitrary parametrizations.

It is important to parametrize $h(M_i)$ in a way that captures the true treatment effect as a function of M_i. ML-SS regression will exactly achieve its theoretically guaranteed properties (detailed below) when $h(M_i)$ is the same as the true conditional treatment effect function. Although this may appear a strong requirement, one can always conduct diagnostics for the chosen parametrization using the residuals from the regression model in a manner analogous to the diagnostics in the conventional regression setting. This is due to the regression formulation of the ML-SS regression. We will showcase the diagnostics in the empirical application.

The formulation in (3) reveals the nature and strength of ML-SS regression, which takes a unique approach to estimate effect modification. ML-SS regression estimates the nuisance components of the model nonparametrically, hence providing much more flexibility and ensuing credibility in the controlling strategy for confounders than the conventional regression model. At the same time, by subtracting the predicted values of the nuisance components, this estimator isolates the quantity of interest in $h(M_i)$ and estimates it in the form of a pre-specified parametric function. [7] As a result, ML-SS regression gains advantageous statistical power relative to ML-

[7] This maneuver is called Robinson (1988) Transformation.

based methods that do not prioritize the pre-specified M_i by design and need to aggregate individual level predictions to the group level. In fact, when ϵ_i in (3) is homoscedastic, the estimates reach the semiparametric efficiency bound (Park and Kang 2019). Therefore, ML-SS regression is a unique combination of the theory-driven and data-driven approaches, finding a middle ground between the conventional regression estimator and ML-based methods for individual effect prediction.

ML-SS regression has desirable theoretical properties including consistency and asymptotic normality. Park and Kang (2019) originally prove these properties specifically for the case of groupwise parametrization. But they be extended to cover arbitrary parametrization of $h(M_i)$. In Appendix A, we sketch a proof of the consistency of the estimator with any parametrization of effect modification. Specifically, under aforementioned identifying assumptions and correct specification of $h(M_i)$, $\hat{h}(M_i)$ estimated using the OLS model in (3) is consistent for $E[Y_i(W_i = 1) - Y_i(W_i = 0)|M_i]$, the effect modification estimand.

As for inference, the estimates of $\hat{\tau}$, namely $\hat{\tau}_g$, $\hat{\tau}_c$ for groupwise parametrizations or $\hat{\tau}_s$ for linear parametrization are asymptotically normally distributed (Park and Kang 2019). Therefore, standard errors in large samples are analytically derived, and the construction of confidence intervals is straightforward using the standard normal quantile function. The standard errors for the coefficients in our estimator can be consistently estimated using Huber-White heteroskedasticity-robust standard errors for linear regression models:

For groupwise parametrization:

$$\widehat{s.e.}_{\tau_g} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2 \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2 I_g(M_i)}{\left[\sum_i \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2 I_g(M_i)\right]^2}} \qquad \text{for each } g.$$

For linear parametrization:

$$\widehat{s.e.}_{\tau_c} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2 \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2}{\left[\sum_i \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2\right]^2}} \qquad \text{for treatment effect constant } \gamma_l,$$

$$\widehat{s.e.}_{\tau_s} = \sqrt{\frac{\sum_i \hat{\epsilon}_i^2 \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2 M_i^2}{\left[\sum_i \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]^2 M_i^2\right]^2}} \qquad \text{for treatment effect slope } \tau_l,$$

where $\hat{\epsilon}_i = Y_i - \hat{E}(Y_i|\boldsymbol{X_i}) - \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]\hat{h}(M_i)$. The $1 - \alpha$ confidence interval for the point estimates can be easily constructed as

$$\hat{\tau} \pm z_{\alpha/2}\widehat{s.e.},$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Correspondingly, the p value for the null hypothesis that the point estimate equals to zero is $2 \times \left[1 - \Phi\left(\frac{|\hat{\tau}|}{\widehat{s.e.}}\right)\right]$, where Φ is the cumulative distribution function of standard normal distribution. To construct confidence interval for the difference between two groupwise ATEs when $h(M_i)$ is groupwise-parametrized, the following can be used

$$\left(\hat{\tau}_{g=1} - \hat{\tau}_{g=2}\right) \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{s.e.}^2{}_{g=1} + \widehat{s.e.}^2{}_{g=2}},$$

and the p value under the null hypothesis that the groupwise difference equals to zero is $2 \times \left[1 - \Phi\left(|\hat{\tau}_g| / \sqrt{\widehat{s.e.}^2{}_{g=1} + \widehat{s.e.}^2{}_{g=2}}\right)\right]$.

There are some other methods that provide flexibility via non- or semi-parametric estimation and also privilege a low dimensional set of modifiers chosen based on substantive reasons. Blackwell and Olson (2021) apply the post-double-selection Lasso developed by Belloni, Chernozhukov, and Hansen (2014b, 2014a) to the study of effect modification. Zeldow et al., (2019) adapt Bayesian Additive Regression Trees (BART) to effect modification and propose to parametrically estimate a linear interaction effect between the treatment and the modifier and nonparametrically fit the baseline outcome as a function of control variables. Hainmueller et al., (2019) introduce a kernel estimator developed by Li and Racine (2010), which automates the fitting of functions in a smooth manner. The bandwidths of the kernels are automatically selected via cross-validation. Abrevaya et al., (2015) also develop a kernel estimator to flexibly fit the relationship between a continuous modifier and treatment effects using nonparametric or semiparametric estimation in a first step.

The methods of Blackwell and Olson (2021) and Zeldow et al., (2019) are in the same spirit as ML-SS regression since they also require researcher-specified parametrization for the treatment effect function. But these methods are respectively based on one specific algorithm (LASSO and BART, respectively), hence not enabling the researcher to leverage the full range of ML algorithms like our estimator does. Different ML methods are adept at fitting different underlying functions in terms of the presence of irrelevant features, nonlinearities, and interactions (Athey and Imbens 2019). ML-SS regression allows the researcher to choose any ML method for the estimation of the nuisance components in the model as they see fit. The kernel methods of Hainmueller et al., (2019) and Abrevaya et al., (2015), on the other hand, take a data-driven approach regarding the specification of the effect modification function for continuous modifiers. Consequently, their methods do not allow succinct summarization of effect modification in a parametric form, nor could statistical tests and confidence intervals be constructed. Nevertheless, they could complement the ML-SS regression as exploratory tools in preparation for the confirmatory analysis performed by the ML-SS regression.

5. Simulation Study

We conduct a simulation study to compare the ML-SS regression relative to both conventional regression models and causal forests that target individual-level effect prediction.

We also investigate whether the relative advantage of the ML-SS regression varies by the covariance between confounders.

In each iteration, we first generate 5500 independent samples of X_1, X_1, X_3, X_4, X_5, ϵ_Y, and ϵ_τ following multivariate normal distribution,

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ \epsilon_Y \\ \epsilon_\tau \end{pmatrix} \sim N\left(\begin{bmatrix} -1 \\ 3 \\ 1 \\ -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & Q & Q & Q & Q & 0 & 0 \\ Q & 1 & Q & Q & Q & 0 & 0 \\ Q & Q & 1 & Q & Q & 0 & 0 \\ Q & Q & Q & 1 & Q & 0 & 0 \\ Q & Q & Q & Q & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.3 \end{bmatrix} \right).$$

The sample size is to imitate the empirical application in this paper, and we vary Q to be 0.2, 0.4, 0.6, and 0.8 in order to capture variation in the performance of the estimators. Then, we generate the outcome, Y, and the treatment, W, according to the rules below. We make the data generating process for confounding highly nonlinear to showcase the flexibility of ML-based estimation compared with OLS with restrictive functional form.

$$Y(W = 0) = \arctan(X_1) + X_1 X_2 + I(X_2 < 2) + I(X_2 >= 4) - X_3^2 + \sqrt{|X_4|} +$$

$$\exp(-\,exp(X_5)) + \epsilon_Y.$$

$W \in (0,1)$ is drawn from a binomial distribution with

$$p = 1/\{1 + \exp[1 - 0.2 \cdot (X_1 X_2 + I(X_2 < 2) + I(X_2 >= 4) + X_3^2 + \sqrt{|X_4|} +$$

$$\exp(-\,exp(X_5)))]\}.$$

Finally, the treatment effect and the observed outcome are generated so that the treatment effect varies systematically only with X_3 and randomly across individuals:

$$\tau = I(X_3 \geq 1.5) - I(X_3 < 1.5) + \epsilon_\tau,$$

$$Y = Y(W = 0) + W\tau.$$

For effect modification by X_3, we fit four models. All of them have the correct treatment effect function, i.e., the group membership specification is correct in all models. Two of them are conventional OLS models, one with linear additive specification for confounders (all five X variables), as is often the case in practice, and the other with the correct specification for confounders, which is implausible in actual research. The OLS model with linear specification takes this form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \gamma I(X_3 \geq 1.5) X_3 + \xi.$$

And the OLS model with correct specification is:

$$Y = \arctan(X_1) + X_1 X_2 + I(X_2 < 2) + I(X_2 >= 4) - X_3^2 + \sqrt{|X_4|} +$$
$$\exp(-\,exp(X_5)) + \gamma I(X_3 \geq 1.5) X_3 + \xi.$$

We also fit ML-SS regression and causal forests. For the nuisance functions in ML-SS regression, we use an ensemble of Random Forests and Single Layer Neural Networks, both of which are used in the empirical application of this paper and will be explained in that section. And the ensemble weights are estimated using a generalized linear model. In the case of causal forests, we aggregate individual-level effect predictions to group level via AIPW.

In Table 1, we report the average bias, the empirical standard error, and the empirical coverage rate of 95% confidence interval for each of the four models over 500 iterations of Monte Carlo simulation. Bias in each simulation iteration is measured as the square root of the mean squared errors for the two treatment effect groups defined in terms of X_3, namely, $\left[\frac{(\hat{\tau}_{X_3<1.5}-\tau_{X_3<1.5})^2+(\hat{\tau}_{X_3\geq 1.5}-\tau_{X_3\geq 1.5})^2}{2}\right]^{\frac{1}{2}}$. As for empirical standard error, we report the square root of the mean empirical variances of the two treatment effect groups, $\left[\frac{V\widehat{ar}(\hat{\tau}_{X_3<1.5})+V\widehat{ar}(\hat{\tau}_{X_3\geq 1.5})}{2}\right]^{\frac{1}{2}}$. And the rate of confidence interval coverage is the average (over groups) proportion of simulation iterations where the estimated confidence interval by each method covers the true treatment effect.

Table 1. Performance of the models in the simulation study

	Confounders' covariance	ML-SS regression	OLS linear	OLS correct	causal forests
Average bias	0.2	0.043	1.130	0.018	0.061
	0.4	0.040	0.819	0.018	0.052
	0.6	0.038	0.402	0.019	0.047
	0.8	0.037	0.288	0.019	0.065
Empirical standard error	0.2	0.034	0.082	0.017	0.033
	0.4	0.031	0.081	0.016	0.033
	0.6	0.032	0.070	0.017	0.039
	0.8	0.032	0.060	0.017	0.086
Confidence interval coverage	0.2	0.956	0	0.92	0.928
	0.4	0.954	0	0.922	0.906
	0.6	0.926	0	0.916	0.882
	0.8	0.938	0.042	0.912	0.872

In Table 1, the two OLS models are two extremes on the spectrum of performance. The OLS model with linear specification for confounders always produces very large biases, leading to confidence intervals that miss the true effects in all or most iterations, depending on the value of confounders' covariance. This reflects how serious the misspecification bias can become even if the researcher controls for all the necessary confounders in some way. On the other hand, the OLS model with correct specification for confounders has the lowest biases and standard errors. However, the correct model is infeasible in practice as researchers rarely know what the true functional forms for confounders are.

Estimates of ML-SS regression always have small biases relative to the true ATEs, which are -1 and 1 in the two groups. And the coverage rate of its estimated confidence interval is close to the nominal coverage rate of 95% with all covariance values, fulfilling the estimator's theoretical property. Despite the presence of complicated nuisance functions, ML-SS regression still manages to fit the functions well enough and reduce bias substantially relative to the OLS model with linear specification. In practice, the flexible fitting of nuisance functions in the ML-SS approach allows researchers to remain agnostic about confounders' functional forms while achieving consistent estimation.

We also compare the ML-SS method with the popular causal forests method (Athey et al. 2019). With only one exception where causal forests have very slightly lower empirical standard error, our method outperforms causal forests in having 19% to 43% lower biases, 6% to 63% lower empirical standard errors, and confidence intervals that are 65% to 91% closer to the theoretical coverage rate. Using a different simulation design, Park and Kang (2019) also show that the ML-SS regression has higher statistical power in testing the groupwise differences in treatment effect compared with causal forests. These results confirm that by targeting a prespecified low-dimensional treatment effect function, the ML-SS regression is more efficient than methods aiming at individual-level predictions.

Comparing the ML-SS regression with causal forests across the confounder covariances, we notice that the empirical standard errors and interval coverage rates of causal forests deteriorate as the covariance increases, while ML-SS regression maintains stably good performance. Intuitively, we suspect this is because non-focal confounders/modifiers become more predictive of individual-level treatment effects as the covariance increases, simply by virtue of their stronger covariance with the focal modifier (X_3, in this case). Hence, targeting individual-level treatment effects, the statistical power of causal forests gets more diluted by non-focal modifiers when they become more strongly correlated. The practical implication of this simulation study is that the ML-SS regression should generally be preferred for effect modification estimation, but especially when the confounders are highly correlated with the focal modifier.

6. Empirical application

We use an empirical application to showcase how researchers can apply the ML-SS regression in a typical study in sociology. Using National Longitudinal Survey of Youth 1979 (NLSY79), we investigate modification of the effect of a college degree on adult income along the dimensions of gender and income origin. Studying college effect heterogeneity provides insights about higher education and inform the influence of potential policies that could affect the college admission process (Brand et al. 2021; Brand and Xie 2010; Forster, van de Werfhorst, and Leopold 2021; Heckman, Humphries, and Veramendi 2018). In addition, the pattern of effect modification also implies whether college education serves as an equalizer that reduces inequality in terms of the focal modifier. In particular, if college is an equalizer, then its effect should be higher for more disadvantaged students. The equalizer role of college with regard to income origin has been extensively examined with mixed findings (Fiel 2020; Hout 2012; Karlson 2019; Torche 2011; Zhou 2019).

We measure the outcome, adult income, by the percentile rank, in order for it to reflect stabilized and overall economic well-being. [8] of family income averaged over five waves of survey when the respondent is 35 to 44 years old and divided by the square root of family size. The treatment variable is a binary indicator of whether the respondent graduates from college by the age of 31. We constructed 34 control variables [9] for the nuisance functions $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$, which are expanded to 68 after factor variables are binarized and variables with very low variances are dropped. All control variables originally have less than 20% missing values.

For effect modification by gender, we use a groupwise parametrization, estimating groupwise ATE for men and women. For income origin, we constructed a percentile rank measure based on the average of the respondent's family incomes in the first three waves of the survey (1979, 1980, and 1981) divided by the square root of the family size. And for effect modification by income origin, we use both a linear parametrization using the percentile rank measure and a groupwise parametrization by quartile groups. In linear modification by income origin, since income origin and income destination are both percentile ranks, the modification estimates have the interpretation of changes in the rank-rank slope. We will test the appropriateness of these parametrizations using residual diagnostics.

We only use the subset of the data where respondents are between 14 to 17 years old at the time of the baseline survey so that we can reasonably assume income origin is measured prior to

[8] Both income origin percentile rank and destination percentile rank are constructed on the basis of weighted sample and income values in each survey year are harmonized to the dollar value in 2019 using personal consumption expenditures index.

[9] These control variables are drawn with the goal of covering as many plausible confounders as possible and in reference to prior publications using NLSY79 to study the effect of college education.

treatment, i.e., the earliest age the vast majority of respondents might have a chance to graduate from college. [10] The resulting sample size is 5583.

We use three ML algorithms to fit the nuisance functions for the ML-SS regression, Elastic Net, Random Forests, and Neural Networks. Elastic Net [11] is a data-adaptive synthesis of LASSO and Ridge regressions. It operates in a generalized linear setting and selects important covariates based on regularizing factors that penalize the number and size of coefficients. Random Forests is an ensemble method based on pooling many "trees", each of which fits the data by recursively splitting the covariate space into "leaves". And tree pruning is used to prevent over-fitting. Neural Networks flexibly combine the original covariates into many hidden nodes and further combine the generated hidden nodes into another set of nodes. We use Single Layer Neural Networks which is often enough to approximate fairly complicated functions, although multilayer neural networks may further improve the performance at expense of computational time. Introductions to these methods aiming at a social science audience can be found in Zeng (1999), Montgomery and Olivella (2018), and Athey and Imbens (2019). All these methods can be used for both regression and classification tasks, corresponding to estimating $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$, respectively. In our application, the method applied to $\hat{E}(Y_i|\boldsymbol{X_i})$ and to $\hat{E}(W_i|\boldsymbol{X_i})$ are always the same, but it is entirely at researchers' discretion to use either the same or different methods for each of them.

ML-SS regression can be combined with various axillary tools common in data analysis in sociology, such as weighting, clustering, and multiple imputation. To illustrate how to integrate ML-SS regression with these tools, we use multiple imputation with five imputed datasets to deal

[10] Income origin is measured by a three-year average and the last survey year contributing to that average is 1981, recording family income in 1980, and the oldest people in our sample are 18 years old in 1980.

[11] Since Elastic Net does not generate interaction terms itself (unlike Random Forest and Neural Network), we manually create all pairwise interactions between all covariates and squared terms of all continuous covariates and then drop those with very low variances, resulting in over one thousand covariates being fed into the Elastic Net (the number of covariates ranges from 1035 to 1045 over multiply imputed samples).

with missing values in our analysis. We also weight the sample with custom survey weights of NLSY79. Similarly, empirical practitioners may adjust for clustering and other structures in residuals straightforwardly when using the ML-SS regression.

Pooling results from multiply imputed datasets, the final point estimates and their standard errors are calculated as follows:

$$\hat{\tau}_{k,MI} = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}_k^{(m)}$$

, $\hat{\tau}_k^{(m)}$ is the point estimate of τ_g, τ_c, or τ_s using ML method k in the mth imputed sample. M is the number of imputed datasets. The final estimate is just an average of the groupwise estimates over imputed samples. The post-imputation standard error estimate is

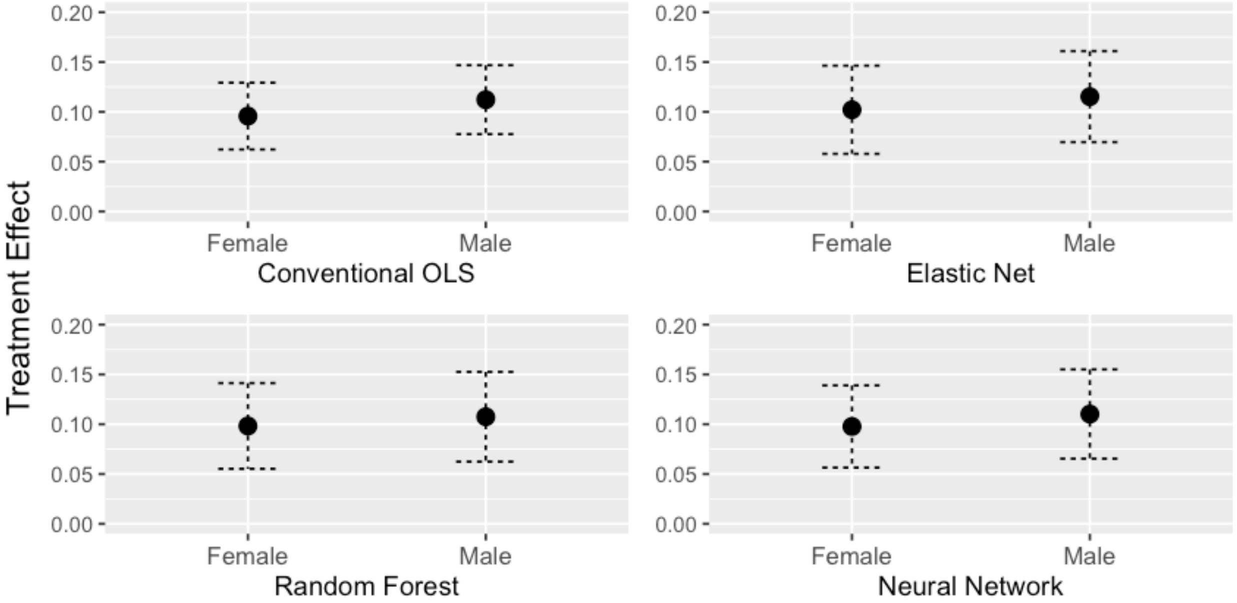
$$\widehat{s.e.}_{MI} = \sqrt{\frac{Var_{within} + (\frac{1+M}{M})Var_{between}}{N}}$$

, where $Var_{within} = \frac{1}{M}\sum_{m=1}^{M} N \cdot \widehat{s.e.}^{2\,(m)}_{k}$ and $Var_{between} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\tau}_k^{(m)} - \hat{\tau}_{k,MI}\right)^2$. $\widehat{s.e.}^{2\,(m)}_{k}$ is the squared standard error of the point estimate using ML method k in the mth imputed sample. $\hat{\tau}_k^{(m)}$ and $\hat{\tau}_{k,MI}$ are defined as above.

The estimation of point estimates and standard errors can also easily accommodate survey weights. Combining weight specification with robust standard error in off-the-shelf statistical packages will easily enable researchers to do so. If multiple imputation is employed in conjunction with, $\hat{\tau}_k^{(m)}$ and $\widehat{s.e.}^{2\,(m)}_{k}$ will be replaced by weighted estimates in each imputed sample.

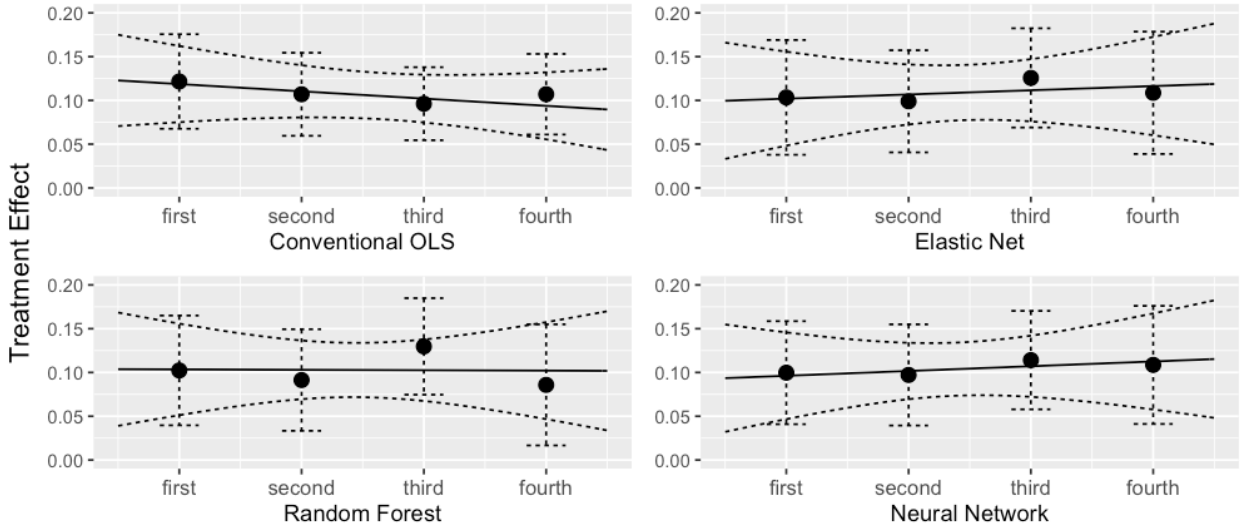
We present results on effect modification by gender in Figure 2. Apart from the ML-SS regression implemented via the three chosen ML algorithms, we also include results from a conventional OLS model where control variables are conditioned on linearly and an interaction term between gender and college completion is used to calculate groupwise ATEs. All implementations of our estimator, as well as the conventional model, agree that the college effect on adult income is slightly higher for men than for women. In this case, the conventional model appears to produce fairly similar results to those of our estimator. However, as we have shown using simulation, this is not always true, and one would not know how much the results of the conventional model are tied to its restrictive functional form had they not also made use of the flexible ML-based method.

Figure 2. Groupwise Effect Modification by Gender



Note: Solid points are point estimates, dotted lines indicate 95% confidence intervals.

Figure 3. Groupwise and Linear Effect Modification by Income Quartile Group



Note: Solid points and lines are point estimates, dotted lines indicate 95% confidence intervals.

Figure 3 show effect modification by income origin in both groupwise parametrization and linear parametrization. [12] The three implementations of ML-SS regression agree that the third income origin quartile group has a slightly stronger causal effect of college than other quartile groups and correspondingly, there is a slightly upward slope on income origin percentile. On the contrary, the conventional model has the third quartile group ranked as the lowest in effect and hence producing a slightly downward slope on income origin percentile. Therefore, there is some evidence that the conventional model may suffer from misspecification bias due to its restrictive functional form. In other words, although the variables as input are the same for conventional regression and ML-SS regression, the use of ML algorithms does make a difference by adaptively specifying how they are modelled.

As can be seen in both Figure 2 and Figure 3, college completion has a strong positive effect for all groups we consider. Across groups and ML algorithms employed in the ML-SS

[12] The confidence interval for the linear parametrization is calculated as $\hat{\tau}^T \boldsymbol{M}_i \pm z_{\alpha/2}\sqrt{\boldsymbol{M}_i^T \hat{\Sigma} \boldsymbol{M}_i}$, where $\hat{\Sigma} = \left\{\sum_i [W_i - \hat{E}(W_i|\boldsymbol{X_i})]^2 [\boldsymbol{M}_i \boldsymbol{M}_i^T]\right\}^{-1} \left\{\sum_i \hat{\epsilon}_i^2 [W_i - \hat{E}(W_i|\boldsymbol{X_i})]^2 [\boldsymbol{M}_i \boldsymbol{M}_i^T]\right\} \left\{\sum_i [W_i - \hat{E}(W_i|\boldsymbol{X_i})]^2 [\boldsymbol{M}_i \boldsymbol{M}_i^T]\right\}^{-1}$, and $\boldsymbol{M}_i = (1, M_i)^T$, $\hat{\tau} = (\tau_c, \tau_s)^T$.

regression, the lowest groupwise estimate is an 8.6 percentile increase in adult income and the highest estimate is 13 percentiles. The linear parametrization of effect modification by income origin also illustrates that across the distribution of income origin, everybody benefits from college degree by about 10 percentiles in adult income. In particular, even for the neural network estimates, where the slope is the steepest, individuals of the highest income origin only have a treatment higher by 2.2 percentiles than that of individuals of the lowest income origin.

Table 2 shows all statistical tests we perform based on the estimates of our estimator. Since for each ML algorithm, we perform 15 tests, we adjust the threshold of statistical significance to be $.05/15 = .0033$ by a Bonferroni correction in order to obtain a conventional .05 Family-Wise Error Rate. All the groupwise ATEs and the constant treatment effect for linear effect modification by income origin are significantly different from 0 across algorithms even after the Bonferroni correction, confirming the result that college degree causally boosts income across modifier values considered in this study. Meanwhile, all groupwise differences and the treatment effect slope on income percentile are not significantly different from zero, suggesting that evidence is scarce for effect modification along these dimensions. The non-significance, however, is highly informative, given the nonconclusive findings of previous research (see Abadie 2020). Substantively, our findings lend strong support to the absence of modification, tipping the balance in the literature.

In summary, the treatment effect of college degree on adult earnings does not appear to be modified by gender and income origin. From gender and income origin perspectives, everybody gains rather equally from attending and graduating college in the United States. In agreement with Zhou (2019), Fiel (2020), and Forster et al. (2021), our finding also casts doubt on the great equalizer role of college, which would require higher treatment effect among disadvantaged groups (cf. Brand et al. 2021; Torche 2011). If anything, high-income-origin is associated with a slightly higher return to college education. However, it is important that our findings do not rule out the possibility that college effect may be modified by some other variables we do not examine here. It is also noteworthy that the specific choice of ML algorithm for the nuisance components of the model do not appear to matter substantially for the results. In addition, as suggested by the confidence intervals in Figure 2 and Figure 3, the much higher level of functional flexibility afforded by our ML-based estimator does not come at much cost of efficiency and statistical power compared with the conventional regression model. [13] Finally, the second step of ML-SS regression is just a regression model with residualized Y_i and W_i, which enables us to use residual plots to detect incorrect specification of the treatment effect function $h(M_i)$. Based on the plots we present in Appendix B, both the linear and groupwise parametrizations of the effect modification by income origin seem to be valid. [14]

Table 2. P Values for Statistical Tests for the Results of the ML-SS Regression

H0	Elastic Net	Random Forest	Neural Network	H0	Elastic Net	Random Forest	Neural Network
Male=0	**<.0001**	**<.0001**	**<.0001**	Q1-Q2=0	.921	.801	.951
Female=0	**<.0001**	**<.0001**	**<.0001**	Q1-Q3=0	.614	.520	.729
Male-Female=0	.685	.772	.687	Q1-Q4=0	.914	.729	.846
Origin Q1=0	**.0020**	**.0014**	**.0009**	Q2-Q3=0	.545	.367	.683

[13] The conventional OLS model has smaller standard errors because it imposes more structure on the functional form. However, as is clear in the case of modification by income origin, the functional form imposed by the conventional model leads to some misspecification bias. Briefly, the way ML-SS regression tackles the bias variance dilemma makes it have smaller specification error while losing a little efficiency, when compared with the conventional OLS.

[14] We do not show the plots for effect modification by gender because there is not a continuous raw measure against which the residuals can be plotted. But for any modifier, the same diagnostics can also be used to detect omitted modifiers when the plot is against suspected modifiers that are not included in the model.

Origin Q2=0	**.0009**	**.0020**	**.0010**	Q2-Q4=0	.843	.907	.801
Origin Q3=0	**<.0001**	**<.0001**	**<.0001**	Q3-Q4=0	.711	.330	.903
Origin Q4=0	**.0023**	**.0151**	**.0016**	Origin Constant=0	**.003**	**.002**	**.003**
				Origin Slope=0	.754	.975	.706

Note: bold values are smaller than the threshold of .05 Family-Wise Error Rate under Bonferroni correction.

7. Conclusion

ML methods may help sociologists who are interested in the estimation of effect modification by controlling for confounders in an automated manner, providing principled flexibility in model specification. However, despite the presence of numerous ML-based methods that address effect heterogeneity, they in fact target the estimation of individual-level treatment effects rather than effect modification by substantively important or interesting variables that dominates the sociological approach to effect heterogeneity. In this article, we introduced the ML-based Sample Splitting regression, which is adopted from the Groupwise Inference Method (Park and Kang 2019), which corresponds to the groupwise parametrization in this article. And we extend the use of it to other parametrizations of effect modification and particularly show that the linear parametrization which, in some cases, lead to familiar interpretation of the results such as changes in rank-rank slope. We also further enhance the applicability of the estimator by integrating it with multiple imputation and survey weights.

As both the simulation study and the empirical application show, our estimator improves upon the conventional regression for effect modification by controlling for the confounders nonparametrically without compromising much efficiency. Compared with another ML-based method for effect heterogeneity, causal forests, the ML-SS regression also has superior performance, especially when the covariance between other confounders and the focal modifier is high. Thus, the ML-SS regression is flexible, efficient, easy to combine with common tools in empirical analysis, and directly suitable for research on effect modification. On the other hand, although the ML-SS regression substantially relaxes functional form assumptions regarding the confounder-outcome relationship relative to the conventional regression, it still imposes a functional form to the relationship between the modifier and the treatment effect. Therefore, its optimal use is when the researcher is willing to estimate the effect modification pattern as a parametrized function.

Appendix A

We sketch a proof for the consistency of the final OLS model

$$Y_i - \hat{E}(Y_i|\boldsymbol{X_i}) = \left[W_i - \hat{E}(W_i|\boldsymbol{X_i})\right]h(M_i) + \epsilon_i. \qquad (A1)$$

For a more rigorous and complete proof of the consistency and asymptotic normality of the estimator, see the appendix of Park and Kang (2019).

Consider the infeasible regression model $(A2)$ where the unknown $E(Y_i|\boldsymbol{X_i})$ and $E(W_i|\boldsymbol{X_i})$ are present. We will first show the unbiasedness of the infeasible model for $h(M_i)$.

$$Y_i - E(Y_i|\boldsymbol{X_i}) = [W_i - E(W_i|\boldsymbol{X_i})]h(M_i) + \epsilon_i. \qquad (A2)$$

By consistency assumption,

$$Y_i = Y(0)_i + \tau_i W_i, \qquad (A3)$$

where $\tau_i = Y_i(1) - Y_i(0)$. Taking expectation conditional on the confounder vector on both sides of $(A3)$, we get

$$E(Y_i|\boldsymbol{X_i}) = E[Y(0)_i|\boldsymbol{X_i}] + E(\tau_i W_i|\boldsymbol{X_i}).$$

By the conditional ignorability assumption,

$$E(Y_i|\boldsymbol{X_i}) = E[Y(0)_i|\boldsymbol{X_i}] + E(\tau_i|\boldsymbol{X_i})E(W_i|\boldsymbol{X_i}). \qquad (A4)$$

Note that by the positivity assumption, $E(\tau_i|\boldsymbol{X_i})$ is well defined for all $\boldsymbol{X}$ values.

Subtracting $(A4)$ from $(A3)$ on both sides,

$$Y_i - E(Y_i|\boldsymbol{X_i}) = Y(0)_i - E[Y(0)_i|\boldsymbol{X_i}] + \tau_i W_i - E(\tau_i|\boldsymbol{X_i})E(W_i|\boldsymbol{X_i})$$

$$= Y(0)_i - E[Y(0)_i|\boldsymbol{X_i}] + E(\tau_i|\boldsymbol{X_i})W_i - E(\tau_i|\boldsymbol{X_i})E(W_i|\boldsymbol{X_i}) + [W_i\tau_i - W_iE(\tau_i|\boldsymbol{X_i})]$$

$$= E(\tau_i|\boldsymbol{X_i})[W_i - E(W_i|\boldsymbol{X_i})] + Y(0)_i - E[Y(0)_i|\boldsymbol{X_i}] + [W_i\tau_i - W_iE(\tau_i|\boldsymbol{X_i})].$$

Then we parametrize $E(\tau_i|\boldsymbol{X_i})$ by $h(M_i)$, where $M_i \in \boldsymbol{X_i}$ is one single modifier of interest. In the case of groupwise parametrization, $h(M_i) = \sum_{g=1}^{G} \tau_g I_{gi}(M_i)$, and in the case of linear parametrization for a continuous M_i, $h(M_i) = \tau_c + \tau_l M_i$. Note that, since τ_g, τ_c, and τ_l represent parametrized $E(\tau_i|\boldsymbol{X_i})$, they are all causal estimands defined by counterfactuals. Then,

$$\begin{aligned}
&Y_i - E(Y_i|\boldsymbol{X_i}) \\
&= [W_i - E(W_i|\boldsymbol{X_i})]h(M_i) \\
&+ \underbrace{Y(0)_i - E[Y(0)_i|\boldsymbol{X_i}] + W_i\tau_i - W_iE(\tau_i|\boldsymbol{X_i}) + [W_i - E(W_i|\boldsymbol{X_i})][E(\tau_i|\boldsymbol{X_i}) - h(M_i)]}.
\end{aligned}$$

And the underbraced part becomes the error term ϵ_i in the structural model behind $(A2)$. As the regressor(s) in $(A2)$, $[W_i - E(W_i|\boldsymbol{X_i})]h(M_i)$, is a function of W_i and $\boldsymbol{X_i}$, we need $E(\epsilon_i|W_i, \boldsymbol{X_i}) = 0$ to show that coefficients in $h(M_i)$ in $(A2)$ are unbiased for the corresponding causal quantities they are meant to estimate.

$$\begin{aligned}
E(\epsilon_i|W_i, \boldsymbol{X_i}) &= E[Y(0)_i|W_i, \boldsymbol{X_i}] - E\{E[Y(0)_i|\boldsymbol{X_i}]|W_i, \boldsymbol{X_i}\} + E(W_i\tau_i|W_i, \boldsymbol{X_i}) \\
&\quad - E[W_iE(\tau_i|\boldsymbol{X_i})|W_i, \boldsymbol{X_i}] + E\{[W_i - E(W_i|\boldsymbol{X_i})][E(\tau_i|\boldsymbol{X_i}) - h(M_i)]|W_i, \boldsymbol{X_i}\} \\
&= E[Y(0)_i|W_i, \boldsymbol{X_i}] - E[Y(0)_i|\boldsymbol{X_i}] + W_iE(\tau_i|W_i, \boldsymbol{X_i}) - W_iE(\tau_i|\boldsymbol{X_i}) \\
&\quad + [W_i - E(W_i|\boldsymbol{X_i})][E(\tau_i|\boldsymbol{X_i}) - h(M_i)].
\end{aligned}$$

Again, by the conditional ignorability assumption,

$$= E[Y(0)_i|\boldsymbol{X_i}] - E[Y(0)_i|\boldsymbol{X_i}] + W_i E(\tau_i|\boldsymbol{X_i}) - W_i E(\tau_i|\boldsymbol{X_i})$$
$$+ [W_i - E(W_i|\boldsymbol{X_i})][E(\tau_i|\boldsymbol{X_i}) - h(M_i)]$$
$$= [W_i - E(W_i|\boldsymbol{X_i})][E(\tau_i|\boldsymbol{X_i}) - h(M_i)].$$

Under the assumption that $h(M_i)$ correctly characterizes $E(\tau_i|\boldsymbol{X_i})$,

$$E(\tau_i|\boldsymbol{X_i}) - h(M_i) = 0,$$

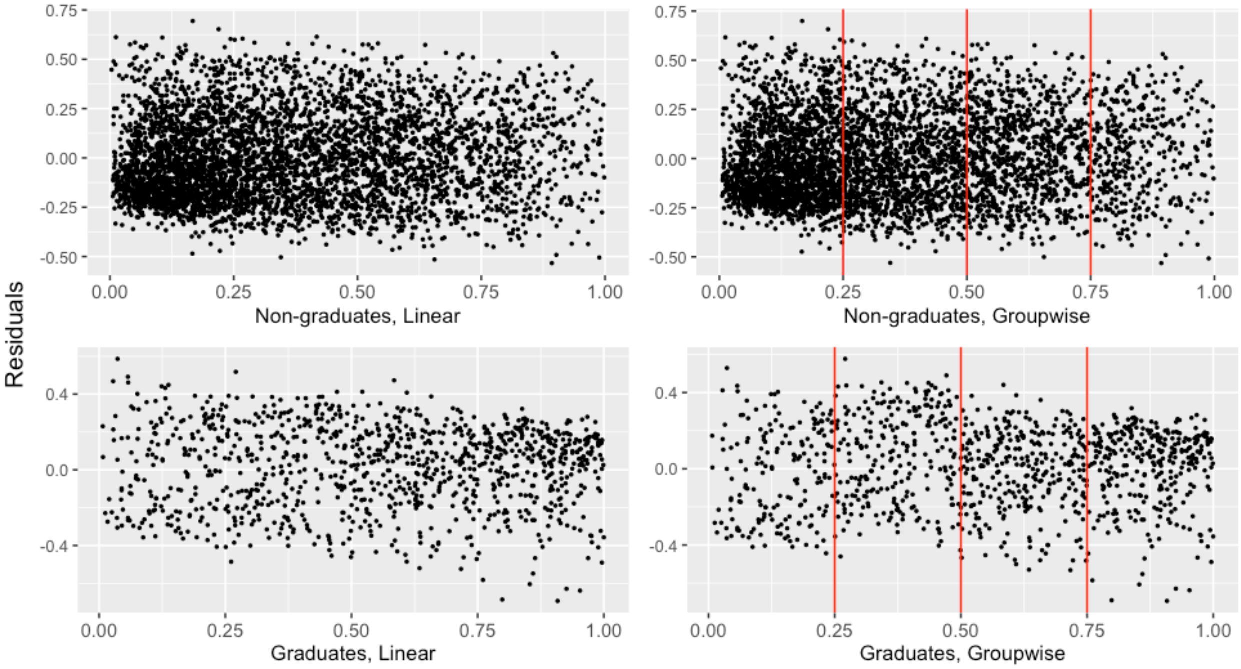
hence $E(\epsilon_i|W_i, \boldsymbol{X_i}) = 0$. Therefore, coefficients in the infeasible model $(A2)$ are unbiased for the causal quantities τ_g, τ_c, and τ_l that we use to characterize groupwise and linear effect modification.

Further, using the fact that $\hat{E}(Y_i|\boldsymbol{X_i})$ and $\hat{E}(W_i|\boldsymbol{X_i})$ are estimated using sample splitting and under the assumption that our ML estimates of these conditional expectation functions converge to their true values at a certain rate, it can be shown that the plug-in feasible estimator as in $(A1)$ is consistent. The asymptotic normality of estimates by $(A1)$ and the fact that they reach the semiparametric efficiency bound also follow (Park and Kang 2019).

Appendix B

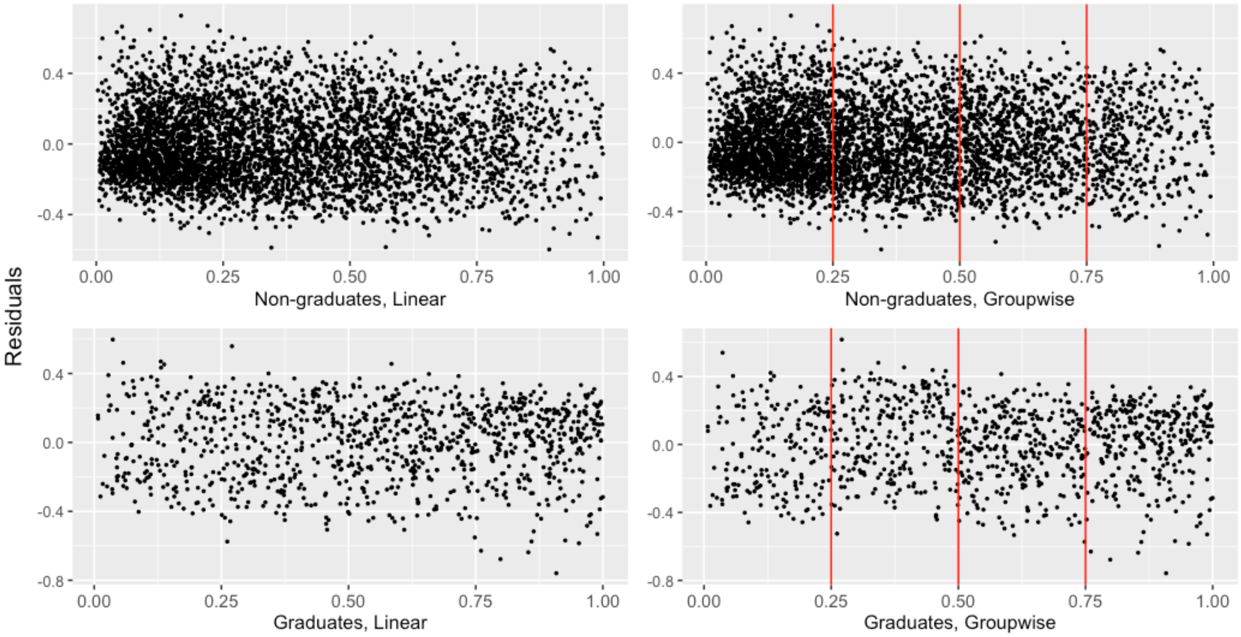
Below we show some residual plots as a diagnostic tool. Conditional on treatment status, i.e., college graduation status in the current study, the residuals from the regression model in the second step should not have any trend that is not paralleled with the x-axis when plotted against the underlying continuous measure of income percentile. In other words, similar to the residual plots for conventional regression models, misspecification of the treatment effect function will show in these plots as some unexpected pattern. To save space, we only show the plots for one of the five multiple imputation samples.

Figure A1. Diagnostic Plots For Elastic Net by Treatment Status and Parametrization



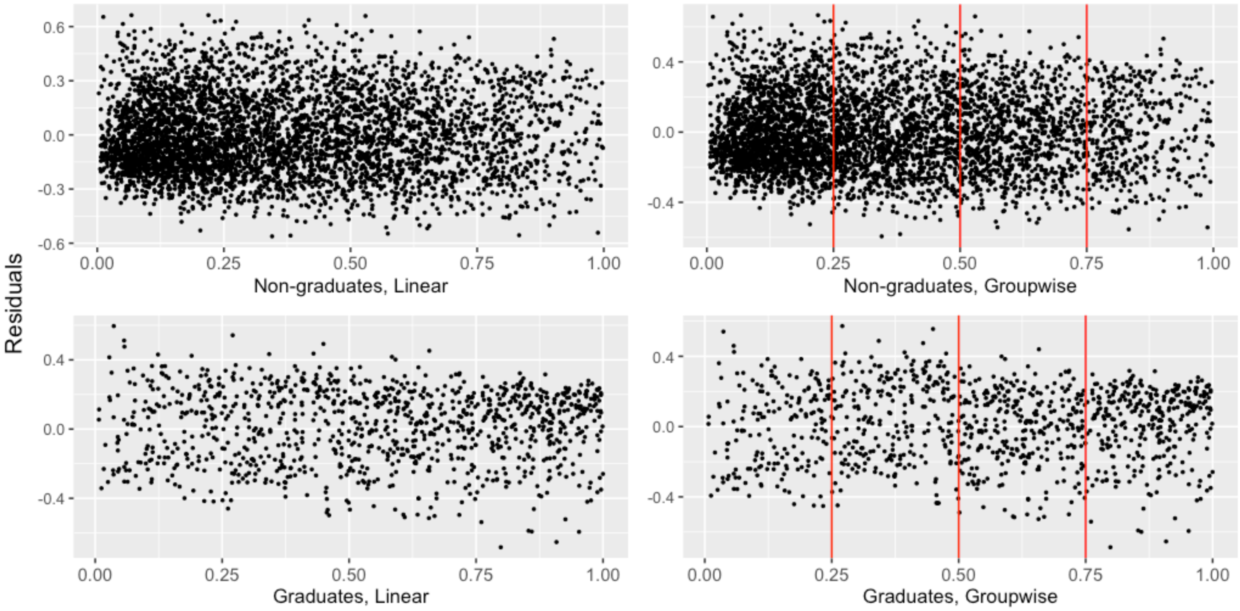
Note: These plots are for the first multiple imputation sample. Verticle lines indicate quartile groups.

Figure A2. Diagnostic Plots For Random Forest by Treatment Status and Parametrization



Note: These plots are for the first multiple imputation sample. Verticle lines indicate quartile groups.

Figure A3. Diagnostic Plots For Neural Network by Treatment Status and Parametrization



Note: These plots are for the first multiple imputation sample. Verticle lines indicate quartile groups.

References

Abadie, Alberto. 2020. “Statistical Nonsignificance in Empirical Economics.” *American Economic Review: Insights* 2(2):193–208. doi: 10.1257/aeri.20190252.

Abrevaya, Jason, Yu-Chin Hsu, and Robert P. Lieli. 2015. “Estimating Conditional Average Treatment Effects.” *Journal of Business & Economic Statistics* 33(4):485–505. doi: 10.1080/07350015.2014.975555.

Athey, Susan, and Guido Imbens. 2016. “Recursive Partitioning for Heterogeneous Causal Effects.” *Proceedings of the National Academy of Sciences* 113(27):7353–60. doi: 10.1073/pnas.1510489113.

Athey, Susan, and Guido W. Imbens. 2019. “Machine Learning Methods That Economists Should Know About.” *Annual Review of Economics* 11(1):685–725. doi: 10.1146/annurev-economics-080217-053433.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. “Generalized Random Forests.” *The Annals of Statistics* 47(2):1148–78. doi: 10.1214/18-AOS1709.

Athey, Susan, and Stefan Wager. 2019. “Estimating Treatment Effects with Causal Forests: An Application.” *ArXiv:1902.07409 [Stat]*.

Baćak, Valerio, and Edward H. Kennedy. 2019. "Principled Machine Learning Using the Super Learner: An Application to Predicting Prison Violence." *Sociological Methods & Research* 48(3):698–721. doi: 10.1177/0049124117747301.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014a. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28(2):29–50. doi: 10.1257/jep.28.2.29.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014b. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81(2):608–50. doi: 10.1093/restud/rdt044.

Billari, Francesco C., Johannes Fürnkranz, and Alexia Prskawetz. 2006. "Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach." *European Journal of Population / Revue Européenne de Démographie* 22(1):37–65. doi: 10.1007/s10680-005-5549-0.

Blackwell, Matthew, and Michael P. Olson. 2021. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 1–20. doi: 10.1017/pan.2021.19.

Brand, Jennie E., and Yu Xie. 2010. "Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Review* 75(2):273–302. doi: 10.1177/0003122410363567.

Brand, Jennie E., Jiahui Xu, Bernard Koch, and Pablo Geraldo. 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning." *Sociological Methodology* 008117502199350. doi: 10.1177/0081175021993503.

Breen, Richard, Seungsoo Choi, and Anders Holm. 2015. "Heterogeneous Causal Effects and Sample Selection Bias." *Sociological Science* 2:351–69. doi: 10.15195/v2.a17.

Bucca, Mauricio, and Daniela R. Urbina. 2019. "Lasso Regularization for Selection of Log-Linear Models: An Application to Educational Assortative Mating." *Sociological Methods & Research* 004912411982615. doi: 10.1177/0049124119826154.

Chen, Shuai, Lu Tian, Tianxi Cai, and Menggang Yu. 2017. "A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring." *Biometrics* 73(4):1199–1209. doi: 10.1111/biom.12676.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21(1):C1–68. doi: 10.1111/ectj.12097.

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press.

Daoud, Adel, and Fredrik Johansson. 2019. “Estimating Treatment Heterogeneity of International Monetary Fund Programs on Child Poverty with Generalized Random Forest.” doi: 10.31235/osf.io/awfjt.

Daoud, Adel, Rockli Kim, and S. V. Subramanian. 2019. “Predicting Women’s Height from Their Socioeconomic Status: A Machine Learning Approach.” *Social Science & Medicine* 238:112486. doi: 10.1016/j.socscimed.2019.112486.

Fiel, Jeremy E. 2020. “Great Equalizer or Great Selector? Reconsidering Education as a Moderator of Intergenerational Transmissions.” *Sociology of Education* 003804072092788. doi: 10.1177/0038040720927886.

Fletcher, Jason M. 2012. “Why Have Tobacco Control Policies Stalled? Using Genetic Moderation to Examine Policy Impacts” edited by B. Mittal. *PLoS ONE* 7(12):e50576. doi: 10.1371/journal.pone.0050576.

Forster, Andrea G., Herman G. van de Werfhorst, and Thomas Leopold. 2021. “Who Benefits Most from College? Dimensions of Selection and Heterogeneous Returns to Higher Education in the United States and the Netherlands.” *Research in Social Stratification and Mobility* 73:100607. doi: 10.1016/j.rssm.2021.100607.

Fu, Qiang, Xin Guo, and Kenneth C. Land. 2018. “Optimizing Count Responses in Surveys: A Machine-Learning Approach.” *Sociological Methods & Research* 004912411774730. doi: 10.1177/0049124117747302.

Gorry, Devon. 2019. “Heterogeneous Consequences of Teenage Childbearing.” *Demography* 56(6):2147–68. doi: 10.1007/s13524-019-00830-1.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. “How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice.” *Political Analysis* 27(2):163–92. doi: 10.1017/pan.2018.46.

Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2018. “Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking.” *Journal of Political Economy* 126(S1):50.

Hout, Michael. 2012. “Social and Economic Returns to College Education in the United States.” *Annual Review of Sociology* 38(1):379–400. doi: 10.1146/annurev.soc.012809.102503.

Karlson, Kristian Bernt. 2019. “College as Equalizer? Testing the Selectivity Hypothesis.” *Social Science Research* 80:216–29. doi: 10.1016/j.ssresearch.2018.12.001.

Keele, Luke, and Randolph T. Stevenson. 2020. “Causal Interaction and Effect Modification: Same Model, Different Concepts.” *Political Science Research and Methods* 1–9. doi: 10.1017/psrm.2020.12.

Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. 2021. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." *The Econometrics Journal* 24(1):134–61. doi: 10.1093/ectj/utaa014.

Knittel, Christopher, and Samuel Stolper. 2019. *Using Machine Learning to Target Treatment: The Case of Household Energy Use*. w26531. Cambridge, MA: National Bureau of Economic Research.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116(10):4156–65. doi: 10.1073/pnas.1804597116.

Langche, Zeng. 1999. "Prediction and Classification with Neural Network Models." *Sociological Methods & Research* 27(4):499–524.

Legewie, Joscha, and Thomas A. DiPrete. 2012. "School Context and the Gender Gap in Educational Achievement." *American Sociological Review* 77(3):463–85. doi: 10.1177/0003122412440802.

Liu, Ran. 2021. "Leveraging Machine Learning Methods to Estimate Heterogeneous Effects: Father Absence in China as an Example." *Chinese Sociological Review* 1–29. doi: 10.1080/21620555.2021.1948828.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45:27–45.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data: TREE-BASED MODELS FOR POLITICAL SCIENCE DATA." *American Journal of Political Science* 62(3):729–44. doi: 10.1111/ajps.12361.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods And Principles For Social Research*. 2nd edition. Cambridge University Press.

Muñoz, John, and Cristobal Young. 2018. "We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness." *Sociological Methodology* 48(1):1–33. doi: 10.1177/0081175018777988.

Nie, X., and S. Wager. 2020. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika*. doi: 10.1093/biomet/asaa076.

Park, Chan, and Hyunseung Kang. 2019. "A Groupwise Approach for Inferring Heterogeneous Treatment Effects in Causal Inference." *ArXiv:1908.04427 [Stat]*.

Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56(4):931. doi: 10.2307/1912705.

Rona-Tas, Akos, Antoine Cornuéjols, Sandrine Blanchemanche, Antonin Duroy, and Christine Martin. 2019. "Enlisting Supervised Machine Learning in Mapping Scientific

Uncertainty Expressed in Food Risk Analysis." *Sociological Methods & Research* 48(3):608–41. doi: 10.1177/0049124117729701.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15):8398–8403. doi: 10.1073/pnas.1915006117.

Tian, Lu, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. 2014. "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates." *Journal of the American Statistical Association* 109(508):1517–32. doi: 10.1080/01621459.2014.951443.

Tiffin, Andrew. 2019. "Machine Learning and Causality: The Impact of Financial Crises on Growth." *IMF Working Papers* 19(228). doi: 10.5089/9781513518305.001.

Torche, Florencia. 2011. "Is a College Degree Still the Great Equalizer? Intergenerational Mobility across Levels of Schooling in the United States." *American Journal of Sociology* 117(3):763–807. doi: 10.1086/661904.

Torrats-Espinosa, Gerard. 2021. "Using Machine Learning to Estimate the Effect of Racial Segregation on COVID-19 Mortality in the United States." *Proceedings of the National Academy of Sciences* 118(7):e2015577118. doi: 10.1073/pnas.2015577118.

VanderWeele, Tyler J. 2009. "On the Distinction Between Interaction and Effect Modification:" *Epidemiology* 20(6):863–71. doi: 10.1097/EDE.0b013e3181ba333c.

Western, Bruce. 2018. “Comment: Bayes, Model Uncertainty, and Learning from Data.” *Sociological Methodology* 48(1):39–43. doi: 10.1177/0081175018799095.

Winship, Christopher, and Bruce Western. 2016. “Multicollinearity and Model Misspecification.” *Sociological Science* 3:627–49. doi: 10.15195/v3.a27.

Wodtke, Geoffrey T., Felix Elwert, and David J. Harding. 2016. “Neighborhood Effect Heterogeneity by Family Income and Developmental Period.” *American Journal of Sociology* 121(4):1168–1222. doi: 10.1086/684137.

Xie, Y. 2013. “Population Heterogeneity and Causal Inference.” *Proceedings of the National Academy of Sciences* 110(16):6262–68. doi: 10.1073/pnas.1303102110.

Xie, Yu, Christopher Near, Hongwei Xu, and Xi Song. 2020. “Heterogeneous Treatment Effects on Children’s Cognitive/Non-Cognitive Skills: A Reevaluation of an Influential Early Childhood Intervention.” *Social Science Research* 86:102389. doi: 10.1016/j.ssresearch.2019.102389.

Yu, Wei-hsin, and Janet Chen-Lan Kuo. 2017. “The Motherhood Wage Penalty by Work Conditions: How Do Occupational Characteristics Hinder or Empower Mothers?” *American Sociological Review* 82(4):744–69. doi: 10.1177/0003122417712729.

Zeldow, Bret, Vincent Lo Re III, and Jason Roy. 2019. “A Semiparametric Modeling Approach Using Bayesian Additive Regression Trees with an Application to Evaluate Heterogeneous Treatment Effects.” *The Annals of Applied Statistics* 13(3):1989–2010. doi: 10.1214/19-AOAS1266.

Zhang, Han, and Jennifer Pan. 2019. “CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media.” *Sociological Methodology* 49(1):1–57. doi: 10.1177/0081175019860244.

Zhou, Xiang. 2019. “Equalization or Selection? Reassessing the ‘Meritocratic Power’ of a College Degree in Intergenerational Income Mobility.” *American Sociological Review* 84(3):459–85. doi: 10.1177/0003122419844992.