What Is the Value Added by Using Causal Machine Learning Methods in a Welfare Experiment Evaluation?

Anthony Strittmatter

University of St. Gallen

First Draft: September 17, 2018

March 19, 2019

Abstract

Recent studies have proposed causal machine learning (CML) methods to estimate conditional average treatment effects (CATEs). In this study, I investigate whether CML methods add value compared to conventional CATE estimators by re-evaluating Connecticut's Jobs First welfare experiment. This experiment entails a mix of positive and negative work incentives. Previous studies show that it is hard to tackle the effect heterogeneity of Jobs First by means of CATEs. I report evidence that CML methods can provide support for the theoretical labor supply predictions. Furthermore, I document reasons why some conventional CATE estimators fail and discuss the limitations of CML methods.

1   Introduction

Conditional average treatment effects (CATEs) estimate heterogeneous policy effects by means of exogenous covariates. Recently, causal machine learning (CML) methods have been proposed to estimate CATEs (see, e.g., Athey, 2018, Chernozhukov, Demirer, Duflo, and Fernández-Val, 2018, for comprehensive summaries). Compared to more conventional estimation methods, CML methods have three potential advantages. First, they make it convenient to incorporate many covariates that are potentially responsible for effect heterogeneity. Several conventional estimation approaches also make it possible to incorporate many covariates, but CML methods can avoid the potential risk of overfitting and are computationally feasible even when the covariate space is very large. Some CML methods allow for more covariates than observations. Second, CML methods are relatively flexible when dealing with the covariates. Some CML methods incorporate nonlinear and interaction terms automatically. Third, the systematic CML algorithms make it less likely to overlook important effect heterogeneity. However, judiciously conducted conventional approaches may also find important heterogeneity margins. Furthermore, CML methods are largely black-box approaches, which is certainly a disadvantage. Accordingly, it is unclear how much value CML methods can add to economic applications compared to more conventional estimation methods.

In this study, I revisit the effects of Connecticut's Jobs First welfare experiment on the labor supply. Well-established labor supply theory provides clear predictions about the heterogeneity margins of this experiment (see, e.g., Kline and Tartari, 2016, for a comprehensive summary). However, Bitler, Gelbach, and Hoynes (2017) document the limitations of a conventional CATE estimator in terms of its ability to provide evidence for these theoretical predictions. This is puzzling because Bitler, Gelbach, and Hoynes (2006) show that quantile treatment effects (QTEs) can uncover evidence for the theoretical labor supply model. It appears that the Jobs First data contain relevant information that can support labor supply theory, but how to uncover the appropriate heterogeneity by means of conventional CATE estimators is not straightforward.

This study contributes to the aforementioned literature in at least three ways. First, I investigate whether CATEs estimated with CML methods can provide evidence supporting the theoretical labor supply predictions of the Jobs First program, which would clearly represent value added compared to conventional CATE estimators. Second, I reveal modeling restrictions that prevent the conventional CATE estimators from revealing more effect heterogeneity. Third, I test whether the estimates of the CATEs and QTEs are nested.

Bitler, Gelbach, and Hoynes (2017) consider local constant models, which are one

of the workhorse methods used to estimate CATEs in empirical economics. Local constant models stratify the sample into different groups defined by the covariates and report subgroups' average treatment effects. Local constant models uncover effect heterogeneity across groups but report constant effects within groups. There are three potential reasons why local constant models fail to support labor supply theory. First, the choice of the subgroups could be suboptimal. Second, constant effects within groups do not accurately approximate the possibly continuously distributed treatment effects. Third, the covariates used in the local constant models may be insufficient for explaining effect heterogeneity.[1] Suitable CML methods can overcome all three possible disadvantages of local constant models. It is essential to understand why local constant models fail to support theoretical predictions in the Jobs First case because these models are widely used, hence, the disadvantages may carry over to other applications.

My main analysis is based on the double machine learning approach proposed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). This is a generic approach that can incorporate many different machine learning estimators. Accordingly, it is suitable to compare machine learning estimators with different modeling restrictions. I consider off-the-shelf machine learning estimators that can be used with standard personal computers. These machine learning estimators can be accessed by a widespread audience. Furthermore, no special IT infrastructure is required, in contrast to machine learning methods that require cloud computing infrastructure (e.g., TensorFlow). This could be useful when dealing with confidential data, such as data from the Manpower Demonstration Research Corporation (MDRC), because it is often easier to comply with data security regulations and laws using a personal computer than using applications on a cloud.

In particular, I consider the "tree" and the "random forest" machine learning estimators (see, e.g., Hastie, Tibshirani, and Friedman, 2009). Tree estimators split the data into mutually exclusive groups defined by the covariates and report effect heterogeneity as the subgroups' average treatment effects. Similar to local constant models, tree estimators uncover effect heterogeneity across groups but report constant effects within groups. However, tree estimators employ data-driven algorithms to select subgroups, whereas for local constant models, subgroups must be manually selected. Both methods could, in principle, incorporate many covariates. However, it is more convenient to use tree estimators when the covariate space is large. In particular, trees can automatically incorporate different subgroup definitions (based on covariates and interactions between covariates) without precoding. Random for-

---

[1]Obviously, local constant models would also fail if the theoretical predictions are not good approximations of the labor supply effects or if measurement error or other data problems prevent me from finding empirical support for the labor supply predictions.

est estimators are ensemble methods. They estimate many trees based on different subsets of the data and covariates and then report the average of the different tree estimates. These features imply that random forest estimators additionally relax the restriction of constant effects within subgroups. Accordingly, the tree and random forest estimators are suitable for relaxing, in a stepwise fashion, the modeling restrictions of local constant models.

The results suggest that random forest estimators can provide evidence for the theoretical labor supply model when they incorporate many covariates. This suggests that including many relevant covariates and allowing for continuously distributed treatment effects are important ingredients for establishing a match between theoretical predictions and empirical results in the Jobs First application. Using solely data-driven methods of selecting subgroups, without incrementing the covariates and flexibility of the model, does not seem to improve the matches.

Furthermore, the results provide evidence that the CML estimates of the CATEs and QTEs provide disparate sets of information for evaluating the Jobs First case, suggesting that the estimated CATEs do not uncover all of the inherent effect heterogeneity of the Jobs First experiment. Both CATEs and QTEs have advantages and disadvantages. CATEs could be useful to design welfare schemes that optimize the labor supply response of specific target groups or to create assignment rules. QTEs enable the study of the responses of the entire labor supply distribution (in a fully flexible way), but it is difficult to assign these responses to specific groups.

Using case studies to demonstrate the capabilities of new methods is common in economics. Recently, Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) illustrate how predictive machine learning methods can improve human decisions, using the example of bail decisions made by judges. Doriey, Hill, Shalit, Scott, and Cervone (2018) launched a causal inference data analysis challenge. Contributors received real-world data to estimate the effects of birth weight on child's IQ. The real-world data were slightly calibrated such that the ground truth was known by the organizers of the challenge. Their main conclusion is that flexible methods with fewer modeling restrictions perform better, which is coherent with my findings for the Jobs First application.

Economic applications using CML methods to predict CATEs are still rare. Davis and Heller (2017) estimate the heterogeneous effects of summer jobs on the probability of committing a violent crime. Taddy, Gardner, Chen, and Draper (2016) investigate the heterogeneous effects of A/B experiments in online-auctions (eBay) on customer responses. Bertrand, Crépon, Marguerie, and Premand (2017) estimate the heterogeneous effects of a work experience program in Côte d'Ivoire on post-participation employment and wages. Knaus, Lechner, and Strittmatter (2018a) estimate the heterogeneous employment effects of a job search program in Switzer-

land.

In the next section, I provide some background information about the Jobs First welfare experiment. In Section 3, I introduce the MDRC data. In Section 4, I describe the empirical framework of this study. In Section 5, I document the empirical results. The final section concludes. The Online Appendices A-F provide supplementary material and descriptive statistics.

2   The Jobs First welfare program

In 1996, Connecticut replaced the Aid for Families with Dependent Children (AFDC) program with the Jobs First program. The Jobs First program created financial work incentives for people on assistance that differed from those offered by the AFDC program.

Figure 1 shows the earnings and welfare transfers in a stylized way. The maximum welfare payment W is similar under both welfare schemes, but additional earnings lead to different welfare payment deductions. AFDC disregarded all earnings below a fixed amount B, which was $120 per month during the first 12 months of employment while on assistance and $90 per month afterward. Furthermore, 51% of any additional earnings was disregarded during the first four months of employment and 27% of any additional earnings afterward. By contrast, the Jobs First program disregards all earnings below the federal poverty line (FPL). Earnings above the FPL terminate all welfare benefit payments from the Jobs First program (which is like a cliff in the benefits payment scheme). The Jobs First and AFDC programs differ in other aspects besides the financial work incentives. The additional changes are summarized in the Online Appendix A.

Figure 1 around here

Bitler, Gelbach, and Hoynes (2006) use a static labor supply model to develop four hypotheses. First, the Jobs First program deducts fewer earnings from welfare payments than the AFDC program. Thus, Jobs First should have a positive effect on the extensive margin of the labor supply. Second, Jobs First recipients with relatively low earnings (between B and E in Figure 1) can keep more of their additional income than AFDC recipients. This should create positive work incentives when the substitution effects dominate the income effects. Third, the FPL is considerably higher than the earnings amount E at which participants lose their eligibility for AFDC welfare benefit payments. Accordingly, this provides a lump-sum transfer to Jobs First welfare recipients with earnings between E and the FPL, which reduces the optimal earnings in the presence of negative income effects. Furthermore, the cliff

construction of the Jobs First payment scheme creates incentives to reduce earnings to just below the FPL, which might increase income or leisure time. Fourth, for recipients sufficiently above the FPL, the AFDC and Jobs First programs provide the same labor supply incentives.

The theoretical predictions and an additional data restriction lead to three empirically testable hypotheses:

H1: The Jobs First program has positive earnings effects for some individuals with zero earnings under AFDC. Furthermore, the earnings effects cannot be negative for this group because the earnings outcome cannot be negative (additional data restriction).

H2: There is a mix of positive and negative earnings effects in the group of individuals with positive earnings below the FPL under AFDC.

H3: Jobs First has non-positive earnings effects for individuals with earnings above the FPL under AFDC.

3 Experimental data

The Connecticut Department of Social Services required the MDRC to conduct a randomized controlled trial to evaluate the Jobs First program. Experimental participants were single-parent welfare applicants and recipients who lived in Manchester or New Haven. Between January 1996 and February 1997, 4,803 experimental participants were randomly assigned to either the AFDC (control group) or Jobs First (treatment group) programs.[2]

3.1 Variable definitions

The MDRC's public use files for the Jobs First program contain baseline data on demographic and family composition variables merged with longitudinal administrative information on welfare and food stamps payments and earnings provided by the state unemployment insurance system. The outcome variable is earnings per quarter in US dollars.[3] The treatment is assignment to the Jobs First program. I follow 4,802 experimental participants for seven quarters after random assignment (RA) to the Jobs First program or AFDC program. The total sample contains 33,614 observations.

---

[2]I drop one experimental participant who had extraordinarily high earnings. This does not change the point estimates much but makes the estimation of the confidence intervals more stable.

[3]I do not observe the earnings reported by the experimental participants to the welfare agency, which could matter when misreporting is a major practice (see discussion, e.g., in Kline and Tartari, 2016).

Being above or below the FPL is one of the major factors driving effect heterogeneity according to the theoretical considerations. Similar to previous Jobs First studies, I cannot observe the administrative assistance unit size, which determines the FPL of the experimental participant. Following the suggestions of Kline and Tartari (2016), I calculate the assistance unit size based on the number of children at RA.[4] This approach may lead to the underestimation of the FPL because mothers may have more children during the seven quarters after RA. To account for this potential disadvantage, I inflate the number of children by one for all mothers as a robustness check (in the following, I call this calculation of the FPL "extra child").

I distinguish between three sets of covariates that I use for the heterogeneity analysis. I label them "baseline", "decent", and "kitchen sink" covariates. Table 1 summarizes the different covariate categories. The baseline covariates contain the elapsed quarters since RA and the earnings in the seven quarters prior to RA. The selection of the baseline covariates follows Bitler, Gelbach, and Hoynes (2017), who use these two covariates in some of their main specifications.

<div align="center">Table 1 around here</div>

The decent covariates include 13 variables. In addition to the baseline variables, they include age, education, information about children, and more information about earnings and welfare history. This set of covariates is still relatively small; however, it may be difficult to consider even these relatively few variables with a standard CATE estimator when including many non-linearities and interactions between the covariates.

The kitchen sink covariates include 68 variables. I include all exogenous covariates for which data are available from the MDRC and that might affect effect heterogeneity. The kitchen sink covariates include different measures of the variables that are already included in the decent covariates. The additional variables are ethnicity, marital status, information about the residence, information on previous participation in education or labor market programs, and more information on earnings and welfare history.

3.2 Descriptive statistics

Table 2 reports the descriptive statistics of the main variables. The mean earnings do not differ greatly between the Jobs First and AFDC participants. Thus, the average effects of the Jobs First experiment do not differ significantly from zero (similar to previous findings, e.g., in Bitler, Gelbach, and Hoynes, 2006). However,

---

[4]I cannot calculate the assistance unit size for 160 experimental participants because the number of children is not reported.

fewer participants in the Jobs First program are unemployed compared to the participants in the AFDC program. More participants in the Jobs First program have employment with earnings below the FPL than participants in the AFDC program. These descriptive findings are in line with the theoretical labor supply predictions. The group of participants with earnings above the FPL is relatively small regardless of whether I inflate the number of children used to calculate the FPL or not. The share of participants with earnings above the FPL does not differ much between the Jobs First and AFDC participants.

Table 2 documents the standardized difference in the baseline and decent covariates between the Jobs First and AFDC participants.[5] Table B.1 in the Online Appendix B shows the descriptive statistics of the kitchen sink covariates. If the RA to the Jobs First and AFDC programs was appropriately random, then we expect all pre-RA covariates to be balanced. Table 2 shows that there are no large differences between the pre-RA covariates. However, there are some small differences. Jobs First participants have slightly more children, less previous earnings, and received more welfare than AFDC recipients.

Table 2 around here

4 Empirical approach

4.1 Estimation target

The treatment dummy D_i equals one when an experimental participant is assigned to Jobs First and zero when she is assigned to the AFDC welfare scheme. Following Rubin's (1974) potential outcome framework, $Y_{it}(1)$ denotes the potential earnings outcome under Jobs First for individual i in quarter t (for $i = 1, \ldots, N$ and $t = 1, ..., 7$). Correspondingly, $Y_{it}(0)$ denotes the potential earnings outcome under AFDC for individual i in quarter t. Each individual can be assigned to either the Jobs First or AFDC but not to both welfare schemes simultaneously. Thus, only one potential outcome is observable. Under the stable unit treatment value assumption (SUTVA), the observed outcome equals

$$Y_{it} = Y_{it}(1)D_i + Y_{it}(0)(1 - D_i). \tag{1}$$

---

[5]The standardized difference in variable X between samples A and B is defined as

$$SD = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{1}{2}\left(Var(X_A) + Var(X_B)\right)}} \cdot 100,$$

where $\bar{X}_A$ denotes the mean of sample A and $\bar{X}_B$ denotes the mean of sample B. Rosenbaum and Rubin (1983) consider an absolute standardized difference higher than 20 to be "large."

Individual i's causal effect of being assigned to Jobs First instead of the AFDC welfare scheme on earnings is

$$\delta_{it} = Y_{it}(1) - Y_{it}(0).$$

We cannot identify individual causal effects without assumptions that are implausible in many applications (e.g., the assumption of effect homogeneity). Nevertheless, group averages of δ_{it} may be identifiable under plausible assumptions. For example, the identification of the average treatment effect (ATE), $\rho = E[\delta_{it}]$, and the average treatment effect on the treated (ATET), $\theta = E[\delta_{it}|D_i = 1]$, is standard in policy evaluations (see, e.g., Imbens and Wooldridge, 2009). CATEs can potentially uncover effect heterogeneity based on exogenous pre-treatment variables X_{it}. The CATEs are

$$\bar{\delta}(x) = E[\delta_{it}|X_{it} = x].$$

Under the random treatment assignment and SUTVA, the CATEs

$$\bar{\delta}(x) = E[Y_{it}|D_i = 1, X_{it} = x] - E[Y_{it}|D_i = 0, X_{it} = x],$$

are identified from observable data on (Y_{it}, D_i, X_{it}).

The CATEs are often labeled as individualized or personalized treatment effects. To some extent, this is misleading because these labels might suggest that the CATEs closely approach the individual causal effects. However, to achieve this, the individual causal effects must be (almost) deterministic, and all relevant determining variables must be observed. In many applications, these requirements are too strong. Nevertheless, even when the CATEs are not equal to the individual causal effects, they have the potential to provide a more complete picture of the effect heterogeneity than the ATEs can.

4.2 Local constant model

The local constant model partitions the sample into mutually exclusive groups. Let $\pi = \{g_1, ..., g_{\#(\pi)}\}$ be a specific sample partition, let $g_j \equiv g_j(x, \pi)$ be the respective group (for $j = 1, \ldots, \#(\pi)$), and let $\#(\pi)$ be the number of groups in the partition π. The group $g_j(x, \pi)$ of partition π is a function of the covariate space of X_{it}. For an explicit example, consider that X_{it} contains only a binary indicator for gender. Then, we can choose between two possible sample partitions; either keep men and women together, $\pi' = \{g_1\} = \{\text{men}, \text{women}\}$, or we partition men and women into two separate groups, $\pi'' = \{g_1, g_2\} = \{\{\text{men}\}, \{\text{women}\}\}$.

Local linear models can be estimated with a linear model that incorporates the

interaction terms with D_i

$$Y_{it} = \sum_{j=1}^{\#(\pi)} \gamma_j \cdot 1\{X_{it} \in g_j(x, \pi)\} + \sum_{j=1}^{\#(\pi)} \delta_j \cdot D_i \cdot 1\{X_{it} \in g_j(x, \pi)\}. \qquad (2)$$

The CATEs are $\bar{\delta}(X_{it}) = \sum_{j=1}^{\#(\pi)} \delta_j \cdot 1\{X_{it} \in g_j(x, \pi)\}$.

The empirical challenge is to find a useful sample partition π. Often the choice of π is a rather discretionary decision. Not accounting for the selection of π (which may include many manual model selection steps) can lead to invalid inference procedures.

4.3 CML approach

Given the discussion in the last section, it would be useful to select π in a data-driven way based on some optimality criteria. In the ideal case, these optimality criteria would be based on the individual causal effect δ_{it}. However, the fundamental problem of causal analysis is the unobservability of δ_{it}.

To overcome the fundamental identification problem, a popular CML approach is to modify the outcome. For example, I could replace the outcome with the orthogonal score

$$Y_{it}^* = \mu_1(X_{it}) - \mu_0(X_{it}) + \frac{D_i(Y_{it} - \mu_1(X_{it}))}{p(X_{it})} - \frac{(1 - D_i)(Y_{it} - \mu_0(X_{it}))}{1 - p(X_{it})},$$

which goes back to Robins and Rotnitzky (1995). It includes the three nuisance parameters $\mu_1(X_{it}) = E[Y_{it}|D_i = 1, X_{it}]$, $\mu_0(X_{it}) = E[Y_{it}(0)|X_{it}] = E[Y_{it}|D_i = 0, X_{it}]$, and $p(X_{it}) = E[D_i|X_{it}]$ that model the potential outcomes and the selection into treatment. Each of the nuisance parameters can be estimated with methods suited to making predictions, such as machine learning estimators. The expected value of the efficient score is the ATE, $\rho = E[Y_{it}^*]$, and the conditional expectations are the CATEs, $\bar{\delta}(x) = E[Y_{it}^*|X_{it} = x]$ (see the proof in the Online Appendix C). The orthogonal score has the advantage that the causal effect estimates remain consistent even when either $\mu_0(x)$ and $\mu_1(x)$ or $p(x)$ is misspecified.

Chernozhukov et al. (2018) call this approach the double machine learning approach because it combines first-step auxiliary predictions of nuisance parameters to estimate the causal effects in the second step. These authors discuss how it is possible to obtain $\sqrt{N}$-consistent and asymptotically normal estimates of the ATE and other low-dimensional causal parameters. An important finding is that, even if the estimates of the nuisance parameter have a slow convergence rate (e.g., $\sqrt[4]{N}$), the ATE estimates can still converge with $\sqrt{N}$. However, much less is known about the asymptotic properties of the modified outcome approach regarding the functions of the efficient score, such as the CATEs (see discussion, e.g., in Chernozhukov,

Demirer, Duflo, and Fernández-Val, 2018). Lee, Okui, and Whang (2017) discuss the asymptotic properties of CATEs estimated with the orthogonal score, but they assume that the covariates used to model effect heterogeneity are low-dimensional and that the nuisance parameters are estimated at a parametric rate.

Another popular CML approach is the causal forest estimator (see Athey, Tibshirani, and Wager, 2018, Wager and Athey, 2018). Causal forest estimators are consistent and asymptotically normal for the estimation of CATEs, but the convergence rates are below $\sqrt{N}$. The modified outcome and causal forest approaches have similar good finite sample properties for estimating CATEs (see, e.g., Knaus, Lechner, and Strittmatter, 2018b, and references therein). I use the modified outcome approach in the main specifications. Additionally, I show that the results do not change considerably when using causal forest estimators.

Chernozhukov et al. (2018) suggest using a cross-fitting procedure to break through the correlation structure between the estimated nuisance parameters and the causal effect estimation. To implement this procedure, I partition the data into two random samples. I use the first sample to estimate the nuisance parameters and extrapolate the fitted values of the nuisance parameters to the second sample. Then, I use the second sample to estimate the CATEs.[6]

4.4 Machine learning estimators

Many different machine learning estimators can be combined with the modified outcome approach (see, e.g., Hastie, Tibshirani, and Friedman, 2009, for an overview of different machine learning estimators). I focus on the regression tree and random forest estimators (e.g., Breiman, 2001) because they mimic the modeling restrictions of the local constant model. I use the R packages `rpart` and `grf` to implement those estimators.

4.4.1 Tree estimator

Similar to the local constant model, regression trees partition the sample into mutually exclusive groups g_j, which are now called leaves. For a specific sample partition π, which is now called a tree, I can estimate the CATEs by

$$\hat{\hat{\delta}}(x,\pi) = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{7} 1\{X_{it} \in g_j(x,\pi)\}} \sum_{i=1}^{N}\sum_{t=1}^{7} 1\{X_{it} \in g_j(x,\pi)\} \cdot \hat{Y}_{it}^{*},$$

---

[6]For the forest estimators, I additionally switch the first and second samples and repeat the cross-fitting procedure. Then, I report the average CATEs obtained from the first and second samples. In the clustered bootstrap procedure that I use to compute the p-values and confidence intervals, I ensure that each individual can enter only one cross-fitting sample and never both.

where $\hat{Y}_{it}^*$ is the estimate of Y_{it}^*, which I extrapolate from the retained cross-fitting sample.

Regression trees select the partition π with a greedy algorithm, i.e., by adding recursive sample splits to the tree without anticipating later splits (e.g., Breiman, Friedman, Stone, and Olshen, 1984). Using the modified outcome approach, regression trees seek to minimize the mean squared error (MSE) with regard to $\hat{Y}_{it}^*$. Accordingly, they select the splits that fit the approximated CATEs best since the group average of $\hat{Y}_{it}^*$ approximates the CATEs. The first splits contribute more effect heterogeneity than the last splits, because of the hierarchical partition structure. I select the optimal tree π^* and $\bar{\delta}_{\text{tree}}(x) = \delta(x, \pi^*)$ based on the out-of-sample MSE, which I calculate with a 10-fold cross-validation procedure. To stabilize the trees, I impose the restriction that each leaf should contain at least 50 observations.

Following the suggestions of Athey and Imbens (2016), I use the so-called honest inference procedure, which means that I split the sample into two parts of equal size. Then, I use the first partition to build the tree (training sample) and the second partition to estimate the CATEs (estimation sample).[7] This separation between the training and estimation samples avoids overfitting of the estimated CATEs.

4.4.2 Generalized random forest estimator

Generalized random forests are assembled from H decorrelated honest trees $\delta(x, \pi_h)$ (for $h = 1, ..., H$). The decorrelated honest trees are estimated using different subsamples of the data and subsets of the covariates. Decorrelation is necessary because, without it, each tree would have a similar structure and I would not be able to gain much from assembling the trees. The random forest estimator of the CATEs is the average of these honest trees:

$$\hat{\bar{\delta}}_{RF}(x) = \frac{1}{H} \sum_{h=1}^{H} \hat{\bar{\delta}}_{\text{tree}}(x, \pi_h).$$

The honest trees of a random forest are built deep (i.e., with small terminal leaves). Thus, I no longer try to optimize the leaf size of the trees with the cross-validation procedure. Instead, I build many deep honest trees that have a small bias but a large variance. Averaging across different honest trees reduces the variance (which is often called "bagging"). Athey, Tibshirani, and Wager (2018) explore the consistency and asymptotic normality of generalized random forests.

I build random forests with 1,000 decorrelated trees, each with a minimum leaf size of 10 observations.[8] In each subsample, I randomly select 50% of the individuals

---

[7]In the clustered bootstrap procedure that I use to compute p-values and confidence intervals, I ensure that each individual can enter either the training or estimation sample, but never both.

[8]The number of trees H is an important tuning parameter for random forests. Table E.1 in the

and two-thirds of the covariates.

4.5   Testing the theoretical hypotheses

To provide evidence for the theoretical predictions, I want to test whether the estimated CATEs are non-positive or non-negative in specific subsamples. From the forest estimators, I obtain separate CATE estimates for each individual. Imposing single hypothesis tests on each CATE would cause the multiple hypothesis testing problem. To avoid this problems, I employ discrete versions of first-order stochastic dominance tests and compute p-values using a clustered bootstrap procedure (e.g., Anderson, 1996, Barrett and Donald, 2003). First, I test the null hypothesis H_0^+ that all estimated CATEs are non-negative. Second, I test the null hypothesis H_0^- that all estimated CATEs are non-positive. I provide the details of the distributional tests in the Online Appendix D.

For the tests, I restrict the sample to the AFDC participants, which does not alter the expected CATE estimates because of the random treatment assignment. However, it enables the identification of CATEs at specific earnings levels under AFDC. According to the theoretical labor supply predictions, the estimated CATEs should

H1: reject H_0^- and not reject H_0^+ in the subsample of unemployed AFDC participants,

H2: reject H_0^+ and H_0^- in the subsample of AFDC participants with positive earnings below FPL, and

H3: reject H_0^+ and not reject H_0^- in the subsample of AFDC participants with earnings above the FPL.

These are necessary but not sufficient conditions of the labor supply predictions.

4.6   Comparison between CATEs and QTEs

In contrast to the CATEs, QTEs do not identify heterogeneity by subgroups of the population but rather by the potential earnings distributions.[9] The potential

---

Online Appendix E shows that the out-of-sample MSE improves when I increase the number of trees. However, with 100 trees, the prediction power of random forests is already almost saturated, and the MSE improvements are marginal.

[9]Under very strong assumptions, the effects on the potential outcome distribution coincide with the individual causal effects. These assumptions imply that individuals do not systematically change their ranks in the potential outcome distributions as a result of treatment status (see the discussion in, e.g., Chernozhukov and Hansen, 2005, Firpo, 2007).

outcome distributions are defined by

$$
\begin{aligned}
F_{Y(1)}(y) &= Pr(Y_{it}(1) \leq y) = Pr(Y_{it}(0) + \delta_{it} \leq y), \text{ and} \\
F_{Y(0)}(y) &= Pr(Y_{it}(0) \leq y) = Pr(Y_{it}(1) - \delta_{it} \leq y),
\end{aligned}
\tag{3}
$$

with $Y_{it}(1) = Y_{it}(0) + \delta_{it}$. The potential quantile $Q_{Y(d)}(\tau)$ is the minimum value of $Y_{it}(d)$ such that, at minimum, the share τ of the earnings distribution lies below this value. QTEs are defined as

$$
\delta_{QTE}(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau),
$$

the difference between the potential quantiles.

Bitler, Gelbach, and Hoynes (2017) propose the simulated potential outcome distributions,

$$
\begin{aligned}
F_{Y(1)}^{S}(y) &= Pr(Y_{it}(0) + \bar{\delta}(X_{it}) \leq y), \text{ and} \\
F_{Y(0)}^{S}(y) &= Pr(Y_{it}(1) - \bar{\delta}(X_{it}) \leq y).
\end{aligned}
\tag{4}
$$

When

$$
F_{Y(1)}^{S}(y) = F_{Y(1)}(y), \text{ and} \tag{5}
$$

$$
F_{Y(0)}^{S}(y) = F_{Y(0)}(y), \tag{6}
$$

then CATEs and QTEs carry the same information and differ only in how they report the effects.[10] I exploit conditions (5) and (6) to determine whether CATEs and QTEs are nested. Moreover, (5) and (6) are necessary (but not sufficient) conditions for $\delta_{it} = \bar{\delta}(X_{it})$, as can be observed when comparing (3) and (4).

5 Results

5.1 Local constant model

To create a benchmark, I first document some results for the local constant model. Following one of the main specifications in Bitler, Gelbach, and Hoynes (2017), I stratify the data by previous earnings and quarters elapsed since RA. I classify previous earnings seven quarters before RA into zero earnings ($earn_0$), earnings below the median among those with positive earnings ($earn_1$), and earnings above the median ($earn_2$). Furthermore, I create dummies for each elapsed quarter since RA (q_1, ..., q_7). Then, I fully interact these dummy variables and the treatment

---

[10]When CATEs and QTEs are nested, it is always possible to calculate the QTEs from the CATEs using the simulated potential outcome distributions. However, it is not necessarily possible to calculate CATEs from QTEs.

dummy (D), such that the model

$$Y = \sum_{k=0}^{2} \sum_{t=1}^{7} \gamma_{kt} \cdot earn_k \cdot q_t + \sum_{k=0}^{2} \sum_{t=1}^{7} \delta_{kt} \cdot D \cdot earn_k \cdot q_t,$$

is fully stratified. The CATEs are δ_{kt}.

Table 3 reports the percentage of positive and negative CATEs. The first column reports the results for the full sample. 82% of the CATEs are positive, and 18% are negative. The second column reports the share of positive and negative CATEs in the sample of AFDC participants. The shares do not change compared to the full sample, which provides reassurance that the predicted CATEs are balanced.

Table 3 around here

Column three of Table 3 reports the CATE results for unemployed individuals under AFDC. The theoretical labor supply model predicts that I find some positive and no negative CATEs (H1). Indeed, the share of negative CATEs (10%) is relatively low, and the null hypothesis H_0^+ for non-negative CATEs cannot be rejected. Thus, there is empirical support for the theoretical hypothesis H1.

Column four of Table 3 reports the CATE results for the AFDC participants with positive earnings below the FPL. For this subsample, the theoretical labor supply model predicts a mix of positive and negative CATEs (H2). However, the null hypothesis H_0^+ for non-negative CATEs cannot be rejected. I find empirical evidence only for positive CATEs (H_0^- rejected), which is insufficient to provide evidence for theoretical hypothesis H2.

Columns five and six of Table 3 report the CATE results for the AFDC participants with earnings above the FPL. Column five uses the exact number of children to measure the FPL and column six uses the inflated FPL measure considering one "extra child". No matter which FPL measure is used, the null hypothesis H_0^- is rejected, suggesting that some individuals with earnings above the FPL have positive CATEs, contradicting theoretical hypothesis H3.

To summarize, the local constant model provides evidence for theoretical hypothesis H1 but rejects H2 and H3.

5.2 CML results

5.2.1 Estimation of the nuisance parameters

To estimate the nuisance parameters, I always use the random forest estimator with the kitchen sink covariates. In this way, the results of the different CATE estimators do not depend on the specification of the nuisance parameters.

Figure B.1 in the Online Appendix B documents the histogram of the estimated Jobs First assignment probability. Under RA, I would expect no variation in the assignment probability. The assignment probability varies between 41% and 61%, with an average assignment probability of 50%. Even though the range of the estimated treatment probability is narrow, it is far from homogeneous. However, the modified outcome approach accounts for the differences in the treatment probabilities. Furthermore, I do not have to worry about common support problems because the propensity score is far from zero and one (see discussion in, e.g., Lechner and Strittmatter, 2017).

Figure B.2 in the Online Appendix B reports the densities of the estimated earnings under the AFDC and Jobs First programs. The estimated earnings are always higher than zero, i.e., they do not capture the mass points at zero earnings (see Figure B.3 in the Online Appendix B).

5.2.2 Results of the modified outcome approach

Table 4 reports the results of the modified outcome approach. A comparison of columns (1) and (2) suggests that restricting the sample to AFDC participants does not alter the results strongly, no matter which estimator is used.

Table 4 around here

The results for trees and forests with baseline controls are qualitatively similar to the findings from the local constant model, suggesting that using a data-driven approach to select subgroups, without incrementing the covariates, does not improve the match between the empirical and theoretical results. I can provide evidence for H1 but must reject H2 and H3.

Using the decent selection of covariates and the tree estimator enables me to detect evidence for the theoretical hypothesis H2 (Table 4, column (4)). I find evidence of positive and negative CATEs for the group with positive earnings below the FPL under AFDC. However, I still reject H3. Furthermore, I find evidence for negative CATEs for the group of unemployed AFDC participants. This is not possible because of the limited support for earnings. Using the kitchen sink controls and the tree estimator allows me to provide evidence for H2 and H3, but I still have to reject H1 because of negative CATEs for the unemployed, suggesting that increasing the number of covariates while maintaining the within-group constant effect modeling restriction does not allow me to provide empirical evidence for all theoretical hypotheses.

In Figures E.1, E.2, and E.3 in the Online Appendix E, I document the relative MSE of the cross-validation samples. For the tree with the baseline controls, the

relative MSE is almost flat. Regardless of how the tree estimator stratifies the sample, it never significantly outperforms the benchmark of homogeneous effects. The selected tree minimizes the MSE with 26 final leaves or 25 splits (as opposed to 21 groups in the local constant model). Figure E.4 in the Online Appendix E shows the structure of the tree with baseline controls. Of the 25 splits, 24 are based on previous earnings, and only one is based on the quarters elapsed. The relative MSE of the trees with the decent and kitchen sink covariates is never saturated (see Figures E.2 and E.3 in the Online Appendix E). Eventually, the tree estimators do not generate additional leaves because I impose the restriction that each leaf must have 50 observations, which further indicates that the modeling restriction of the within-group constant effects is not appropriate in the Jobs First case. Figures E.5 and E.6 in the Online Appendix E show the complex structure of the trees with the decent and kitchen sink covariates. Earnings and welfare history, as well as information about children, are important split variables.

The forest and tree results are qualitatively similar when using the decent controls. However, using the forest estimator and the kitchen sink controls enables me to find empirical evidence supporting the theoretical hypotheses H1, H2, and H3. Accordingly, this is the only CML specification that provides evidence for the theoretical labor supply model. For the unemployed under AFDC, the estimated CATEs of the forests with kitchen sink covariates are sometimes positive and otherwise non-negative (Table 4, column (3)). This result is empirical evidence supporting the theoretical hypothesis H1. For the group with positive earnings below the FPL under AFDC, the estimated CATEs are sometimes positive and sometimes negative (Table 4, column (4)). This result represents empirical evidence for the theoretical hypothesis H2. For the groups with earnings above the FPL under AFDC, the estimated CATEs are sometimes negative and otherwise non-positive (Table 4, columns (5) and (6)). This result is empirical evidence for the theoretical hypothesis H3.

Kline and Tartari (2016) report that between 20% and 100% of women who do not receive welfare under AFDC reduce their labor supply under Jobs First. The results of the forest estimator with kitchen sink controls suggest 85-90% negative CATEs for individuals with earnings above the FPL under AFDC. Furthermore, 29% of the individuals with positive earnings below the FPL have negative CATEs (in this group, not all AFDC recipients are off welfare). This result suggests that the point estimates are roughly in the range of Kline and Tartari (2016), even without having confidence intervals for labor supply response probabilities.

Figure 2 shows the aggregated CATEs of the forest estimator with kitchen sink controls on the ordinate and the difference between earnings and FPL under AFDC on the abscissa. Figure 2 documents positive effects far below the FPL, negative effects slightly below and above the FPL, and insignificant effects far above the

FPL. This finding nicely summarizes the theoretical predictions. Accordingly, CML methods can provide empirical evidence supporting the theoretical predictions of the Jobs First experiment, but I must incorporate many covariates and allow for continuously distributed treatment effects. This finding is consistent with the results of Doriey et al. (2018), who show with their data challenge that as more flexible algorithms model the response surface, their performance is more promising.

Figure 2 around here

5.2.3 Additional results

Table 5 reports the results obtained from the causal forest estimator (e.g., Athey, Tibshirani, and Wager, 2018) with the kitchen sink covariates. The results do not differ strongly from the estimates of the modified outcome approach with the random forest and kitchen sink covariates. Both CML methods provide empirical evidence supporting the theoretical labor supply predictions.

Table 5 around here

The early CML literature suggests that the modified outcome method without confounder adjustments

$$Y_{it}^{**} = \frac{D_i - Pr(D_i = 1)}{Pr(D_i = 1)Pr(D_i = 0)} Y_{it}$$

is sufficient to estimate the CATEs, $\delta(X_{it}) = E[Y_i^{**}|X_{it}]$, in randomized experiments. Table 6 documents that the modified outcome method without covariate adjustments fails to provide evidence for the theoretical labor supply predictions, even for the random forest estimator and kitchen sink controls. This finding is consistent with previous studies documenting the poor properties of the modified outcome method without covariate adjustments (e.g., Athey and Imbens, 2016). The shares of positive and negative CATES differ greatly between the first and second columns of Table 6, suggesting that the estimated CATEs are not balanced between the Jobs First and AFDC participants. The adjustment for confounders appears crucial (especially when many covariates are incorporated) because even small covariate imbalances could be picked up and misused by the machine learning algorithms.

Table 6 around here

5.3 Relation between CATEs and QTEs

Bitler, Gelbach, and Hoynes (2006) use a QTE approach to evaluate the Jobs First

program. Figure 3 replicates their main results. QTE can also provide empirical evidence supporting the theoretical labor supply predictions, which raises the following question: do CATEs and QTEs contain the same information?

<center>Figure 3 around here</center>

Figure 4 reports the simulated earnings distributions using the random forest estimator with the kitchen sink covariates. The simulated earnings distributions are sometimes above and sometimes below the potential earnings distributions. However, the simulated distributions cannot detect the mass point at zero. Heckman, Smith, and Clements (1997) point out, that distributions with differing mass points cannot be equal. Thus, the hypothesis that the potential and simulated earnings distributions are equal can be formally rejected, which suggests that the QTEs and CATEs contain different information. Furthermore, this finding is evidence that the CATEs are not equal to the individual causal effects.

<center>Figure 4 around here</center>

In addition, I report Kolmogorov-Smirnov tests for equality of the positive part of the potential and simulated earnings distributions in Online Appendix F. For all estimation approaches, the tests reject that the QTEs and CATEs are nested.

Whether it is more appropriate to use CATEs or QTEs depends on the concrete research questions. For example, CATEs are more appropriate for developing assignment rules for programs (see, e.g., Athey and Wager, 2018, for a discussion). QTEs are more appropriate for investigating earnings inequalities when we are not concerned about the exact locations of specific individuals in the earnings distributions.

6 Conclusions

I study the value added by using CML methods in a case study of Connecticut's' Jobs First randomized welfare experiment. In this application, conventional CATE estimators fail to find supporting evidence for the theoretical labor supply predictions. I provide evidence that CML methods can overcome this disadvantage. Accordingly, CML methods can add value to a Jobs First evaluation in the sense that they can provide evidence supporting the theoretical labor supply predictions. However, this strategy works only when the CML methods incorporate many important heterogeneity variables and allow for continuously distributed treatment effects.

CML methods cannot uncover the entire effect heterogeneity of the Jobs First program. The estimates of the CATEs and QTEs do not contain the same infor-

<center>19</center>

mation. Furthermore, CML methods cannot detect mass points in the earnings distributions. Because of the case study style of this research, it is difficult to make statements about the external validity of the results.

References

ANDERSON, G. (1996): "Nonparametric Tests of Stochastic Dominance in Income Distributions," *Econometrica*, 64(5), 1183–1193.

ARNOLD, T. A., AND J. W. EMERSON (2011): "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions," *R Journal*, 3(2), 34–39.

ATHEY, S. (2018): "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*, ed. by A. Agrawal, J. Gans, and A. Goldfarb. NBER, Proceedings of Economics AI Conference.

ATHEY, S., AND G. IMBENS (2016): "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Science of the United States of America (PNAS)*, 113(27), 7353–7360.

ATHEY, S., J. TIBSHIRANI, AND S. WAGER (2018): "Generalized Random Forests," *Annals of Statistic*, forthcoming.

ATHEY, S., AND S. WAGER (2018): "Efficient Policy Learning," *arXiv:1702.02896v4*.

BARRETT, G. F., AND G. D. DONALD (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71(1), 71–104.

BERTRAND, M., B. CRÉPON, A. MARGUERIE, AND P. PREMAND (2017): "Contemporaneous and Post-Program Impacts of a Public Works Program: Evidence from Côte d'Ivoire," *World Bank Working Paper*.

BITLER, M., J. GELBACH, AND H. HOYNES (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96(4), 988–1012.

——— (2017): "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment," *Review of Economics and Statistics*, 99(4), 683–697.

BLOOM, D., S. SCRIVENER, C. MICHALOPOULOS, P. MORRIS, R. HENDRA, D. ADAMS-CIARDULLO, J. WALTER, AND W. VARGAS (2002): *Jobs First: Final Report on Connecticut's Welfare Reform Initiative.* Manpower Demonstration Research Corporation, New York.

BREIMAN, L. (2001): "Random Forests," *Machine Learning*, 45(1), 5–32.

BREIMAN, L., J. FRIEDMAN, C. STONE, AND R. OLSHEN (1984): *Classification and Regression Trees.* CRC Press.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21(1), C1–C68.

CHERNOZHUKOV, V., M. DEMIRER, E. DUFLO, AND I. FERNÁNDEZ-VAL (2018): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," *arXiv:1712.04802v3*.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

DAVIS, J. M. V., AND S. B. HELLER (2017): "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs," *American Economic Review*, 107(5), 546–550.

DONALD, S. G., AND Y.-C. HSU (2016): "Improving the Power of Tests of Stochastic Dominance," *Econometric Reviews*, 35(4), 553–585.

DORIEY, V., J. HILL, U. SHALIT, M. SCOTT, AND D. CERVONE (2018): "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition," *arXiv:1707.02641v5*.

FIRPO, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

GINÉ, E., AND J. ZINN (1990): "Bootstrapping General Empirical Measures," *Annals of Probability*, 18(2), 851–869.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *Elemants of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2nd edn.

HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 63(4), 487–535.

IMBENS, G., AND J. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

KLEINBERG, J., H. LAKKARAJU, J. LESKOVEC, J. LUDWIG, AND S. MULLAINATHAN (2018): "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133(1), 237–293.

KLINE, P., AND M. TARTARI (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106(4), 972–1014.

KNAUS, M., M. LECHNER, AND A. STRITTMATTER (2018a): "Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach," *arXiv:1709.10279*.

——— (2018b): "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," *arXiv:1810.13237*.

LECHNER, M., AND A. STRITTMATTER (2017): "Practical Procedures to Deal with Common Support Problems in Matching Estimation," *Econometric Reviews*, forthcoming.

LEE, S., R. OKUI, AND Y.-J. WHANG (2017): "Doubly Robust Uniform Confidence Band for the Conditional Average Treatment Effect Function," *Journal of Applied Econometrics*, 32(7), 1207–1225.

ROBINS, J. M., AND A. ROTNITZKY (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5), 688–701.

TADDY, M., M. GARDNER, L. CHEN, AND D. DRAPER (2016): "A Nonparametric Bayesian Analysis of Heterogenous Treatment Effects in Digital Experimentation," *Journal of Business and Economic Statistics*, 34(4), 661–672.

WAGER, S., AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

Figures
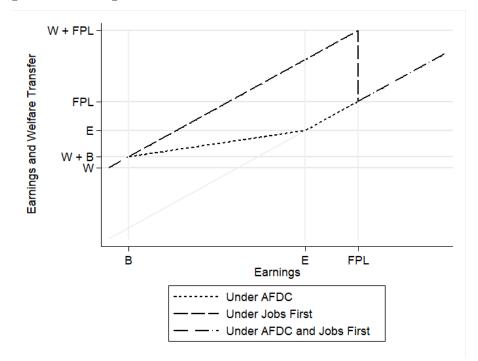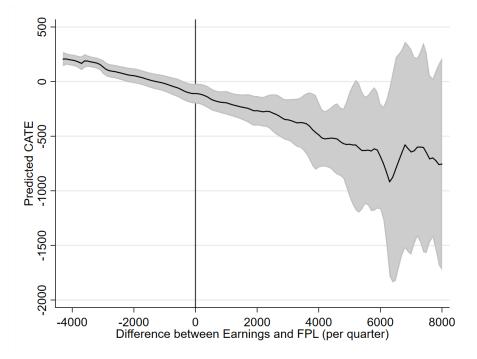
Figure 1: Earnings and welfare transfers under AFDC and Jobs First.
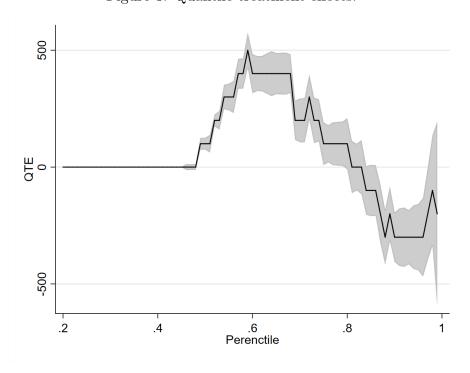

Note: Unemployed persons receive the maximum welfare amount W under AFDC and Jobs First. Under AFDC, all earnings below B are disregarded. Any earnings above B reduce the welfare amount proportionally. Welfare is completely terminated at earnings E. Under Jobs First, all earnings below the FPL are disregarded. Any earnings above the FPL terminate welfare payments.

Figure 2: CATEs by difference between earnings and FPL (per quarter).
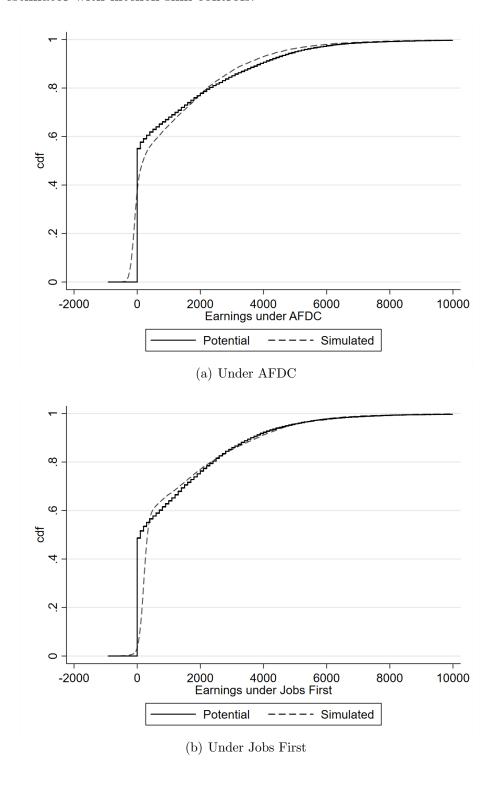
Note: I estimate the aggregated CATEs by the difference between the quarterly earnings and FPL ("extra child") using a local-constant regression. I use Silverman's rule to specify the bandwidth. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size. Figure B.4 in Online Appendix B shows the same figure using the FPL without inflating the number of children by one. The gray area reports the 95% confidence intervals that are estimated using an individual-level clustered bootstrap approach (with 1,999 replications). I control for the kitchen sink covariates. The figure is truncated at 8,000 US dollars.

Figure 3: Quantile treatment effects.

Note: Replication of Bitler, Gelbach, and Hoynes (2006). The gray area shows the 95% confidence interval.

Figure 4: Potential and simulated earnings distributions obtained from the random forest estimator with kitchen sink controls.

(a) Under AFDC

(b) Under Jobs First

Tables

Table 1: List of covariates.

Baseline:	quarters elapsed since RA, earnings seven quarters prior to RA
Decent:	age, education, number of children, age of the youngest child, amount of AFDC assistance received seven quarters prior to RA, amount of food stamps received seven quarters prior to RA, dummy variable indicating a positive amount of earnings in at least one of the seven quarters prior to RA, dummy variable indicating a positive amount of AFDC assistance received in at least one of the seven quarters prior to RA, dummy variable indicating a positive amount of food stamps in at least one of the seven quarters prior to RA
Kitchen sink:	ethnicity, marital status, city of residence, information on residence in a publicly subsidized home, information on relocations, participation in different types of education and labor market programs in the 12 months prior to RA (e.g., English as a Secondary Language (ESL), Adult Basic Education (ABE), General Education Development (GED), job readiness skills, work experience, vocational education, post-secondary education, and high school), earnings for each of the seven quarters prior to the RA, the amount of AFDC assistance received for each of the seven quarters prior to the RA, the amount of food stamps received for each of the seven quarters prior to RA, the number of quarters on AFDC, a dummy variable indicating whether the family received AFDC during childhood, a dummy variable indicating whether work was never recorded, a dummy variable indicating whether work was recorded at RA

Note: The decent covariates also include the baseline covariates. The kitchen sink covariates also include the decent covariates. I include dummies for missing values whenever necessary (see Table B.2 in the Online Appendix B for details).

Table 2: Descriptive statistics of the main variables.

	Jobs First		AFDC		SD
	Mean	St. Dev.	Mean	St. Dev.	
	(1)	(2)	(3)	(4)	(5)
Earnings per quarter (in $)	1173	1789	1125	1868	2.6
Share of participants with					
no earnings	0.49	0.50	0.55	0.50	13.3
earn. below FPL	0.39	0.49	0.31	0.46	17.0
earn. above FPL	0.13	0.33	0.14	0.35	4.1
earn. above FPL ("extra child")	0.09	0.28	0.10	0.30	4.4
Baseline covariates					
Quarters since RA	4.0	2.0	4.0	2.0	0.0
Earnings in pre-Q7 (in $)	682	1552	774	1781	5.5
Decent covariates					
Age categories					
< 20 years	0.09	0.28	0.09	0.28	1.2
20-24 years	0.20	0.40	0.21	0.41	2.7
25-34 years	0.41	0.49	0.42	0.49	1.5
35-45 years	0.25	0.43	0.23	0.42	4.1
> 44 years	0.05	0.22	0.06	0.23	1.2
Education categories					
No degree	0.33	0.47	0.31	0.46	3.8
High school	0.55	0.50	0.57	0.50	3.2
More than high school	0.06	0.24	0.06	0.23	1.9
Age of youngest child (in years)	4.6	4.7	4.5	4.8	2.2
Number of children	1.6	1.0	1.5	1.0	6.0
AFDC pre-Q7 (in $)	920	925	865	896	6.0
Food stamps pre-Q7 (in $)	306	319	293	301	4.4
Any earnings pre-Q1/7	0.33	0.37	0.36	0.38	7.9
Any AFDC pre-Q1/7	0.57	0.45	0.54	0.45	6.5
Any food stamps pre-Q1/7	0.61	0.44	0.60	0.43	2.1
Participants	2,396		2,406		
Observations	16,772		16,842		

Note: The last column reports the standardized difference (SD). Earnings in pre-Q7 refers to earnings in the seven quarters before RA. Any earnings pre-Q1/7 is a dummy variable indicating that earnings were positive in at least one of the seven quarters prior to RA. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size.

Table 3: Results of the local constant model.

| | Full Sample | Under AFDC | | | | |
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL	Earn. above FPL ("extra child")
	(1)	(2)	(3)	(4)	(5)	(6)
Local constant model with baseline covariates						
Positive CATEs	82%	82%	90%	79%	56%	53%
Negative CATEs	18%	18%	10%	21%	44%	47%
p-value H_0^+	0.29	0.29	0.56	0.12	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	0.00	0.00
Observations	33,621	16,842	8,988	4,967	2,313	1,713

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). The details of the distributional tests are provided in the Online Appendix D.

Table 4: Results of the modified covariate approach.

| | Full Sample | Under AFDC | | | | |
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL	Earn. above FPL ("extra child")
	(1)	(2)	(3)	(4)	(5)	(6)
Tree with baseline covariates						
Positive CATEs	88%	87%	93%	85%	68%	66%
Negative CATEs	12%	13%	7%	15%	32%	34%
p-value H_0^+	0.32	0.29	0.45	0.27	0.06	0.04
p-value H_0^-	0.00	0.00	0.00	0.00	0.01	0.01
Tree with decent covariates						
Positive CATEs	53%	53%	60%	50%	35%	33%
Negative CATEs	47%	47%	40%	50%	65%	67%
p-value H_0^+	0.00	0.00	0.00	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	0.01	0.07
Tree with kitchen sink covariates						
Positive CATEs	55%	56%	67%	51%	29%	30%
Negative CATEs	45%	44%	33%	49%	71%	70%
p-value H_0^+	0.00	0.00	0.00	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00	1.00
Forest with baseline covariates						
Positive CATEs	84%	83%	91%	79%	58%	55%
Negative CATEs	16%	17%	9%	21%	42%	45%
p-value H_0^+	0.52	0.47	0.77	0.32	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	0.00	0.00
Forest with decent covariates						
Positive CATEs	69%	70%	86%	63%	20%	15%
Negative CATEs	31%	30%	14%	37%	80%	85%
p-value H_0^+	0.00	0.00	0.00	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00	1.00
Forest with kitchen sink covariates						
Positive CATEs	75%	75%	94%	71%	15%	10%
Negative CATEs	25%	25%	6%	29%	85%	90%
p-value H_0^+	0.00	0.00	0.12	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00	1.00
Observations	33,621	16,842	8,988	4,967	2,313	1,713

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). The details of the distributional tests are provided in the Online Appendix D.

Table 5: Results of the causal forest approach.

| | Full Sample | Under AFDC | | | | |
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL	Earn. above FPL ("extra child")
	(1)	(2)	(3)	(4)	(5)	(6)
Causal forest with kitchen sink covariates						
Positive CATEs	78%	77%	93%	74%	25%	20%
Negative CATEs	22%	23%	7%	26%	75%	80%
p-value H_0^+	0.00	0.00	0.34	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00	1.00
Observations	33,621	16,842	8,988	4,967	2,313	1,713

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). The details of the distributional tests are provided in the Online Appendix D.

Table 6: Results of the modified outcome approach without confounder adjustment.

| | Full Sample | Under AFDC | | | | |
		All	Unempl.	Pos. earn. below FPL	Earn. above FPL	Earn. above FPL ("extra child")
	(1)	(2)	(3)	(4)	(5)	(6)
Forest with kitchen sink covariates						
Positive CATEs	68%	50%	74%	29%	3%	2%
Negative CATEs	32%	50%	26%	71%	97%	98%
p-value H_0^+	0.00	0.00	0.00	0.00	0.00	0.00
p-value H_0^-	0.00	0.00	0.00	0.00	1.00	1.00
Observations	33,621	16,842	8,988	4,967	2,313	1,713

Note: H_0^+ is the null hypothesis that all CATEs are non-negative. H_0^- is the null hypothesis that all CATEs are non-positive. FPL ("extra child") means that I inflate the number of children per mother by one when calculating the assistance unit size. P-values are calculated with an individual-level clustered bootstrap procedure (with 1,999 replications). The details of the distributional tests are provided in the Online Appendix D.

Online Appendix to "What Is the Value Added by Using Causal Machine Learning Methods in a Welfare Experiment Evaluation?"

Anthony Strittmatter

University of St.Gallen

Sections:

A Additional differences between Jobs First and AFDC

The Jobs First and AFDC programs differ in aspects other than financial work incentives. AFDC participants receive welfare benefits for an unlimited amount of time, whereas the lifetime benefit duration of Jobs First recipients is limited to 21 months. However, Bloom, Scrivener, Michalopoulos, Morris, Hendra, Adams-Ciardullo, Walter, and Vargas (2002) document that extensions to this time limit were possible and common. To account for possible time limit differences, I focus on the short-term effects of the Jobs First program on earnings during the first 21 months after random assignment (RA).

The recipients of both the Jobs First and AFDC programs had work requirements. Unemployed AFDC recipients could meet the work requirements by participating in an employment-related program. Bloom et al. (2002) argue that AFDC work requirements were only casually enforced. Unemployed Jobs First recipients could meet the work requirements by participating in employment programs meant to reduce the time taken for job placement. Moreover, Jobs First implemented stricter sanction rules in the case of non-compliance with the work requirement.

Furthermore, Jobs First recipients had different access to child care subsidies and Medicaid than AFDC recipients. However, Bloom et al. (2002) argue that these differences had little impact on actual child care and health care availability because contemporaneous state-level programs compensated for the differences. Finally, the reform affected how alimony payments from fathers were deducted from welfare payments but only when the alimony payments were between $50 and $100. Kline and Tartari (2016) argue that these differences are negligible because they affected only a small fraction of mothers.

B   Additional descriptive statistics

Table B.1: Descriptive statistics of the kitchen sink covariates.

	Jobs First		AFDC		SD
	Mean	St. Dev.	Mean	St. Dev.	
	(1)	(2)	(3)	(4)	(5)
Kitchen sink covariates					
Ethnicity					
Black	0.37	0.48	0.37	0.48	0.6
White	0.36	0.48	0.35	0.48	2.9
Hispanic	0.21	0.41	0.22	0.41	2.3
Other ethnicity	0.06	0.24	0.06	0.25	0.6
Marital status					
Never married	0.62	0.48	0.63	0.48	1.4
Married	0.33	0.47	0.32	0.47	1.4
Separated	0.06	0.23	0.05	0.22	2.5
Divorced	0.11	0.32	0.12	0.33	3.2
Widowed	0.01	0.09	0.01	0.11	3.2
City of residence					
New Haven	0.75	0.43	0.76	0.43	0.9
Manchester	0.25	0.43	0.24	0.43	0.9
Housing information					
Public housing	0.11	0.31	0.10	0.30	1.4
Subsidized housing	0.23	0.42	0.22	0.42	1.3
Temporary housing	0.01	0.11	0.01	0.10	1.2
Other housing	0.65	0.48	0.66	0.47	2.3
Num. moves last 2 years	0.91	1.23	0.95	1.22	3.0
Any program or educ. during last year	0.20	0.40	0.20	0.40	0.8
ESL	0.01	0.08	0.01	0.09	2.9
ABE	0.01	0.08	0.01	0.09	1.4
GED	0.03	0.16	0.03	0.16	0.5
Vocational education	0.03	0.16	0.03	0.17	3.0
Post-secondary educ.	0.06	0.23	0.05	0.21	4.8
Job search program	0.01	0.10	0.01	0.10	0.8
Work experience program	0.06	0.24	0.07	0.25	3.5
High school	0.02	0.14	0.02	0.14	1.2
Job readiness skills	0.01	0.08	0.01	0.08	0.5

< Table continues on next page. >

|	Jobs First		AFDC		SD
	Mean	St. Dev.	Mean	St. Dev.	
	(1)	(2)	(3)	(4)	(5)
Employment history					
Earnings in pre-Q6	727	1932	814	1759	4.7
Earnings in pre-Q5	683	1506	840	1769	9.6
Earnings in pre-Q4	695	1522	845	1776	9.1
Earnings in pre-Q3	682	1440	791	1620	7.1
Earnings in pre-Q2	678	1513	726	1495	3.1
Earnings in pre-Q1	632	1368	689	1399	4.1
Never worked	0.12	0.32	0.11	0.31	2.3
Employed at RA	0.21	0.41	0.23	0.42	5.3
Welfare history					
AFDC pre-Q6	916	924	862	904	5.9
AFDC pre-Q5	899	905	837	895	6.9
AFDC pre-Q4	887	879	820	872	7.7
AFDC pre-Q3	862	845	801	824	7.3
AFDC pre-Q2	869	812	822	802	5.8
AFDC pre-Q1	884	786	837	795	6.0
Food stamps pre-Q6	309	326	301	311	2.3
Food stamps pre-Q5	323	337	312	321	3.3
Food stamps pre-Q4	349	351	333	340	4.7
Food stamps pre-Q3	370	361	357	349	3.7
Food stamps pre-Q2	394	366	383	360	3.3
Food stamps pre-Q1	413	360	396	353	4.8
Quarters on AFDC at RA	4.1	2.3	4.1	2.2	3.6
AFDC during childhood	0.23	0.42	0.24	0.43	2.3
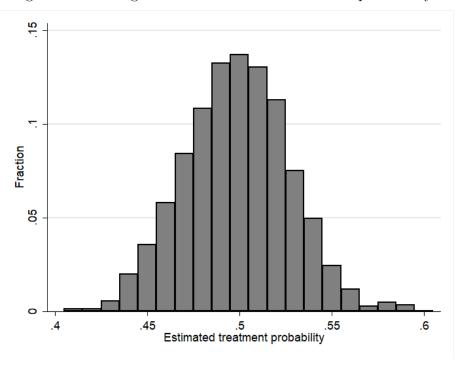
Note: The last column reports the standardized difference (SD). Earnings, AFDC and food stamp payments are measured in US dollars. Earnings in pre-Q6 means earnings six quarters before RA. The abbreviations for the other quarters are similar. ESL is the abbreviation for English Secondary Language. ABE is the abbreviation for Adult Basic Education. GED is the abbreviation for General Education Development.

Table B.2: Balance of missing dummies.

| | Jobs First | | AFDC | | SD |
| | Mean | St. Dev. | Mean | St. Dev. | |
	(1)	(2)	(3)	(4)	(5)
Missing Dummies					
Education	0.06	0.23	0.06	0.24	2.5
Children	0.12	0.32	0.14	0.35	7.7
Marital status	0.05	0.21	0.05	0.21	0.3
Num. moves last 2 years	0.05	0.21	0.04	0.21	0.7
program or educ. during last year	0.06	0.23	0.05	0.22	2.1
Never worked	0.03	0.18	0.03	0.18	0.9
Employed at RA	0.04	0.20	0.05	0.21	1.5
Quarters on AFDC at RA	0.05	0.23	0.05	0.22	1.0
AFDC during childhood	0.07	0.25	0.06	0.25	0.3
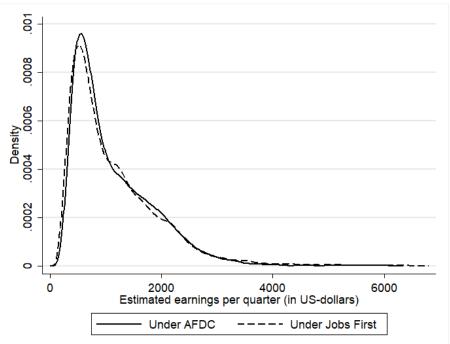
Note: The last column reports the standardized difference (SD).

Figure B.1: Histogram of the estimated treatment probability.

Note: I use the random forest estimator with the kitchen sink covariates to estimate the propensity score.

Figure B.2: Density of estimated potential earnings.

Note: Kernel density plot. I use Silverman's rule to specify the bandwidth. I use the random forest estimator with the kitchen sink covariates to estimate potential earnings.

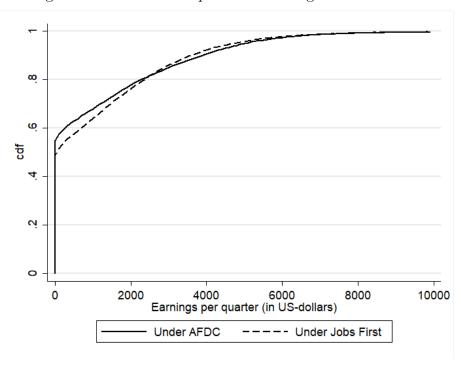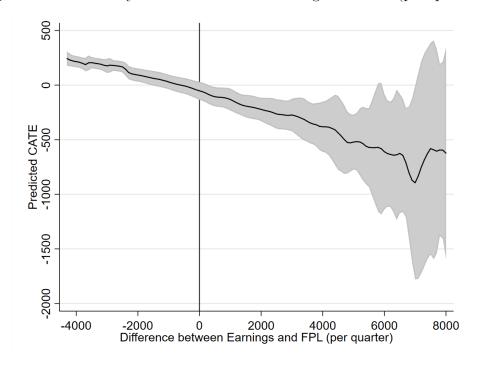Figure B.3: Cumulative potential earnings distributions.

Figure B.4: CATEs by difference between earnings and FPL (per quarter).

Note: I estimate the aggregated CATEs by the difference between the quarterly earnings and FPL using a local-constant regression. I use Silverman's rule to specify the bandwidth. The gray area reports the 95% confidence interval estimated using an individual-level clustered bootstrap approach (with 1,999 replications). I control for the kitchen sink covariates. The figure is truncated at 8,000 US dollars.

C Identification of ATEs and CATEs

In the following, I summarize the assumptions required to identify ATEs and CATEs and proof the identification using the orthogonal score.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)

$$Y_{it} = Y_{it}(1)D_i + Y_{it}(0)(1 - D_i).$$

Assumption 2: Exogeneity of Covariates

$$X_{it}(1) = X_{it}(0).$$

Assumption 3: No Support Problems

$$\epsilon < Pr(D_i = 1 | X_{it} = x) = p(x) < 1 - \epsilon$$

for some small $\epsilon > 0$ and all x in the support of X_{it}.

Assumption 4: Conditional Independence Assumption (CIA)

$$Y_{it}(1), Y_{it}(0) \perp\!\!\!\perp D_i | X_{it} = x$$

for all x in the support of X_{it}.

Assumption 1 excludes spillover and other general equilibrium effects. Assumption 2 requires that all covariates are exogenous. $X_{it}(1)$ is the potential value of the covariates under Jobs First and $X_{it}(0)$ is the potential value of the covariates under AFDC. Assumption 3 excludes individuals having very small or very large treatment probabilities. Figure B.1 shows the distribution of the estimated propensity score, which is far greater than zero and far less than one. Assumption 4 implies that the potential outcomes are independent of the treatment status conditional on X_{it}. This requires conditioning on all confounding variables, which jointly affect the potential outcomes and the treatment probability. In the Jobs First application this assumption is satisfied because of the randomized treatment assignment, even without conditioning on X_{it}. In practice, it might still be useful to condition on X_{it} to avoid the phenomenon in which the machine learning algorithm exploits small covariate imbalances (see the discussion in Section 5.2.3).

The conditional expectation of the orthogonal score Y_{it}^* equals

$$
\begin{aligned}
\bar{\delta}(x) = & E\left[\mu_1(x) - \mu_0(x) + \frac{D_i(Y_{it} - \mu_1(x))}{p(x)} - \frac{(1-D_i)(Y_{it} - \mu_0(x))}{1-p(x)}\,\middle|\, X_{it} = x\right], \\
= & E\left[\frac{D_i - p(x)}{p(x)(1-p(x))}Y_{it}\,\middle|\, X_{it} = x\right] + \underbrace{\frac{E[D_i - p(x)|X_{it} = x]}{p(x)}}_{=0}\mu_1(x) \\
& - \underbrace{\frac{E[D_i - p(x)|X_{it} = x]}{1-p(x)}}_{=0}\mu_0(x), \\
= & E\left[\frac{D_i - p(x)}{p(x)(1-p(x))}Y_{it}\,\middle|\, X_{it} = x\right], \\
= & E\left[\frac{D_i Y_{it}}{p(x)}\,\middle|\, X_{it} = x\right] - E\left[\frac{(1-D_i)Y_{it}}{1-p(x)}\,\middle|\, X_{it} = x\right], \\
\stackrel{Bayes}{=} & E\left[Y_{it}|D_i = 1, X_{it} = x\right] - E\left[Y_{it}|D_i = 0, X_{it} = x\right], \\
= & E\left[Y_{it}(1)|D_i = 1, X_{it} = x\right] - E\left[Y_{it}(0)|D_i = 0, X_{it} = x\right], \\
\stackrel{CIA}{=} & E\left[Y_{it}(1)|X_{it} = x\right] - E\left[Y_{it}(0)|X_{it} = x\right],
\end{aligned}
$$

with $\mu_1 = E[Y_{it}(1)|X_{it} = x]$ and $\mu_0 = E[Y_{it}(0)|X_{it} = x]$. This proves that the CATEs equal the conditional expectation of the orthogonal score. An application of the law of iterative expectations shows that the expectation of the orthogonal score identifies the ATEs,

$$
\rho = E[Y_{it}(1) - Y_{it}(0)] = E[E[Y_{it}(1) - Y_{it}(0)|X_{it} = x]] = E[\bar{\delta}(x)].
$$

D Tests of the CATE distribution

The empirical distribution of the estimated CATEs is

$$\hat{F}_{\bar{\delta}}(z) = \frac{1}{N} \sum_{i=1}^{N} 1\{\hat{\bar{\delta}}(X_{it}\} \le z)$$

and the hypothetical distribution, when all estimated CATEs would be zero, is $F_0(z) = 1\{z \ge 0\}$. Anderson (1996) notes that stochastic dominance tests can also be used for discrete distributions. In particular, they can be used to test the null hypotheses that all estimated CATEs are non-negative, $H_0^+ : F_{\bar{\delta}}(z) \le F_0(z)$ for all z, or all estimated CATEs are non-positive, $H_0^- : F_{\bar{\delta}}(z) \ge F_0(z)$ for all z.

The test statistics are

$$\widehat{D}_+ = \sup_z (\hat{F}_{\bar{\delta}}(z) - F_0(z)) \text{ and}$$
$$\widehat{D}_- = \sup_z (F_0(z) - \hat{F}_{\bar{\delta}}(z)).$$

Note that $\widehat{D}_+$ is equivalent to the probability that the estimated CATEs are negative and $\widehat{D}_-$ is equivalent to the probability that the estimated CATEs are positive. Defining the test statistics by means of distributions instead of probabilities enables recentering of the bootstrap test statistic (see below), which is known to increase the finite sample testing power (e.g., Donald and Hsu, 2016).

To obtain the p-values for the null hypotheses, I compute the bootstrap distributions of the test statistics (see, e.g., Arnold and Emerson, 2011, for discrete null distributions). I construct B bootstrap resamples drawing random observations at the individual level with replacement. Let $\hat{F}_{\bar{\delta}}^{(b)}(z)$ denote the bootstrapped distribution of the estimated CATEs in bootstrap sample b (for $b = 1, ..., B$). The bootstrapped re-centered test statistics are

$$\widehat{D}_+^{(b)} = \sup_z (\hat{F}_{\bar{\delta}}^{(b)}(z) - \hat{F}_{\bar{\delta}}(z)) \text{ and}$$
$$\widehat{D}_-^{(b)} = \sup_z (\hat{F}_{\bar{\delta}}(z) - \hat{F}_{\bar{\delta}}^{(b)}(z)).$$

The estimated p-values that the null hypotheses are true equal

$$\hat{\alpha}(H_0^+) = \frac{1}{B} \sum_{b=1}^{B} 1\left\{ \widehat{D}_+^{(b)} > \widehat{D}_+ \right\} \text{ and}$$
$$\hat{\alpha}(H_0^-) = \frac{1}{B} \sum_{b=1}^{B} 1\left\{ \widehat{D}_-^{(b)} > \widehat{D}_- \right\}.$$

Table D.1 reports the results of a small simulation exercise. I consider two data generating processes (DGPs):

$$\text{DGP 1:} \quad Y_i = D_i \cdot X_i \cdot \beta + X_i \cdot \gamma + U_i \text{ and}$$
$$\text{DGP 2:} \quad Y_i = -D_i \cdot X_i \cdot \beta + X_i \cdot \gamma + U_i,$$

with D_i being the randomized binary treatment dummy, with $Pr(D_i = 1) = 0.5$, X_i has a random uniform distribution between zero and one, and U_i has a random normal distribution with a mean zero and standard deviation of four. The coefficients β and γ are both one. The true CATEs are $\bar{\delta}(X_i) = X_i$ in DGP 1 and $\bar{\delta}(X_i) = -X_i$ in DGP 2. Accordingly, H_0^- is true and H_0^+ is false for DGP 1, whereas H_0^+ is true and H_0^- is false for DGP 2. I estimate both DGPs with the linear model,

$$Y_i = \hat{\gamma}_0 + D_i \cdot \hat{\beta}_0 + X_i \cdot \hat{\gamma}_1 + D_i \cdot X_i \cdot \hat{\beta}_1 + \hat{U}_i.$$

The estimated CATEs are $\hat{\bar{\delta}}(X_i) = \hat{\beta}_0 + X_i \cdot \hat{\beta}_1$. I derive the bootstrapped p-values of H_0^+ and H_0^- as described above (using 1,999 bootstrap replications).

Table D.1 shows the results with and without re-centering the test statistic for sample sizes of 500, 1,000, and 2,000. The results suggest that the re-centered stochastic dominance tests have considerably more power than their uncentered counterparts. Furthermore, the re-centered test rejects H_0^+ and H_0^- at approximately correct rates for both DGPs.

Table D.1: Simulation of first-order stochastic dominance test.

	Without re-centering			With re-centering		
	$N = 500$	$N = 1,000$	$N = 2,000$	$N = 500$	$N = 1,000$	$N = 2,000$
	(1)	(2)	(3)	(4)	(5)	(6)
	DGP 1 – Rejection rates at 5%-level					
H_0^+	0.021	0.006	0.010	0.079	0.049	0.054
H_0^-	0.428	0.444	0.485	0.625	0.769	0.897
	DGP 2 – Rejection rates at 5%-level					
H_0^+	0.425	0.474	0.499	0.822	0.950	0.994
H_0^-	0.012	0.000	0.019	0.060	0.054	0.047

E   Specification of the tuning parameters

Table E.1: RMSE by forest size.

| | Single tree | Tree size of forest | | | | | | |
| | | 10 | 50 | 100 | 500 | 1,000 | 5,000 | 10,000 |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
P-score		0.506	0.502	0.500	0.500	0.500	0.500	0.500
$Y(0)$		1551	1540	1537	1532	1532	1531	1532
$Y(1)$		1563	1537	1536	1531	1531	1530	1531
Y^*, baseline	3059	3190	3013	3026	3002	3001	2999	3000
Y^*, decent	3252	3228	3015	3029	3002	2999	2997	2998
Y^*, kitchen sink	3353	3221	3013	3024	2995	2995	2991	2992

Figure E.1: Relative MSE for the tree with baseline covariates.
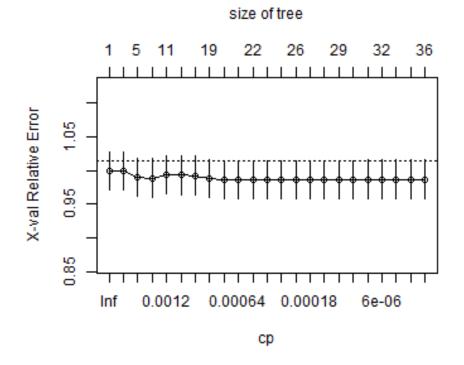
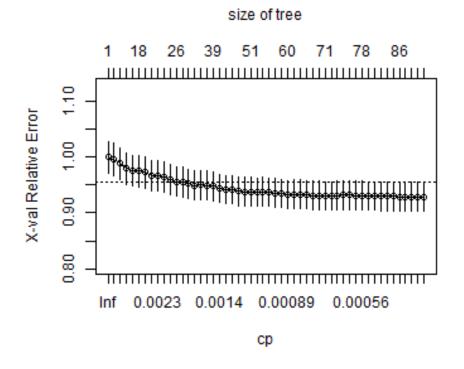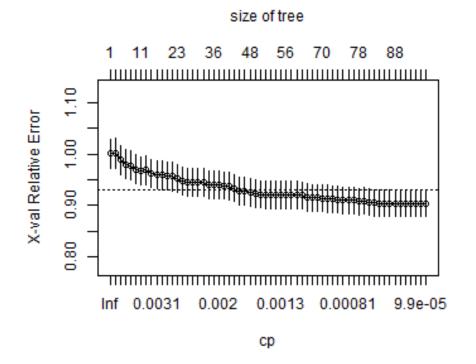Figure E.2: Relative MSE for the tree with decent covariates.

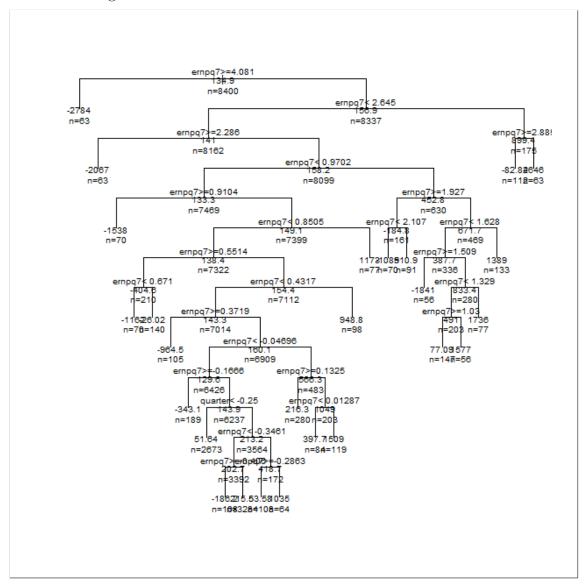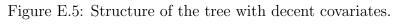Figure E.3: Relative MSE for the tree with kitchen sink covariates.

Figure E.4: Structure of the tree with baseline covariates.

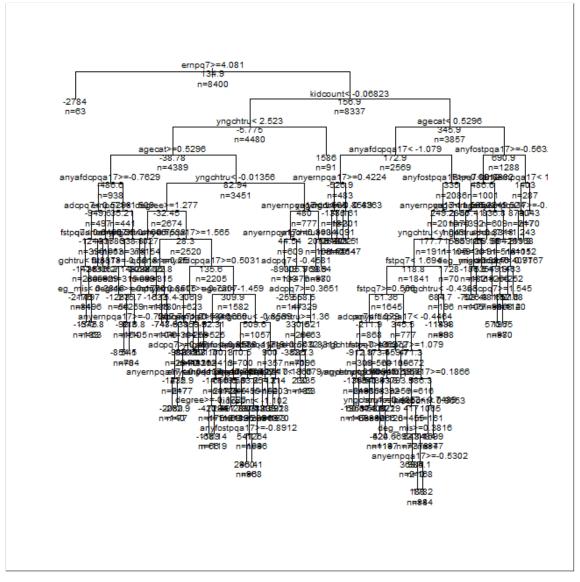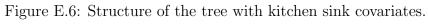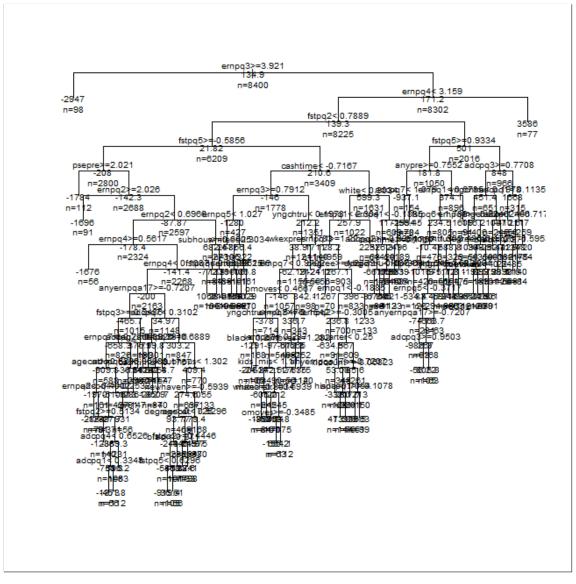Figure E.5: Structure of the tree with decent covariates.

Figure E.6: Structure of the tree with kitchen sink covariates.

F   Comparison of CATEs and QTEs

Bitler, Gelbach, and Hoynes (2017) propose a test for within-group constant treatment effects that vary across a limited number of observable subgroups. They suggest a testing procedure based on within-group permutations. This procedure cannot be applied to the random forest estimator because it does not have a clear group structure.

Therefore, I define the positive potential outcome distributions,

$$
\begin{aligned}
F_{Y(1)}^{+}(y) &= Pr(0 < Y_{it}(1) \leq y) = Pr(0 < Y_{it}(0) + \delta_{it} \leq y), \text{ and} \\
F_{Y(0)}^{+}(y) &= Pr(0 < Y_{it}(0) \leq y) = Pr(0 < Y_{it}(1) - \delta_{it} \leq y),
\end{aligned}
$$

and positive simulated potential outcome distributions,

$$
\begin{aligned}
F_{Y(1)}^{S+}(y) &= Pr(0 < Y_{it}(0) + \bar{\delta}(X_{it}) \leq y), \text{ and} \\
F_{Y(0)}^{S+}(y) &= Pr(0 < Y_{it}(1) - \bar{\delta}(X_{it}) \leq y).
\end{aligned}
$$

When conditions

$$
\begin{aligned}
F_{Y(1)}^{S}(y) &= F_{Y(1)}(y), \text{ and} \\
F_{Y(0)}^{S}(y) &= F_{Y(0)}(y),
\end{aligned}
$$

hold (see Section 4.6), then

$$
F_{Y(1)}^{S+}(y) = F_{Y(1)}^{+}(y), \text{ and} \tag{1}
$$

$$
F_{Y(0)}^{S+}(y) = F_{Y(0)}^{+}(y), \tag{2}
$$

hold too. Accordingly, (1) and (2) are necessary conditions for the equality of the potential and simulated potential outcome distribution. Conditions (1) and (2) can be tested in the data using a Kolgomorov-Smirnov (KS) test, because the underlying distributions are (almost) continuous (especially for the more flexible approaches).

The positive potential and simulated potential outcome distributions

$$
\begin{aligned}
F_{Y(1)}^{+}(y) &= Pr(0 < Y_{it} \leq y | D_i = 1), \\
F_{Y(0)}^{+}(y) &= Pr(0 < Y_{it} \leq y | D_i = 0), \\
F_{Y(1)}^{S+}(y) &= Pr(0 < Y_{it} + \delta(X_{it}) \leq y | D_i = 0), \\
F_{Y(0)}^{S+}(y) &= Pr(0 < Y_{it} - \delta(X_{it}) \leq y | D_i = 1),
\end{aligned}
$$

are identified from observable data because of the randomized treatment assignment. The empirical distribution functions are denoted by $\hat{F}_{Y(1)}^{+}(y)$, $\hat{F}_{Y(1)}^{+}(y)$, $\hat{F}_{Y(1)}^{S+}(y)$, and $\hat{F}_{Y(1)}^{S+}(y)$. Furthermore, let $\hat{F}_{Y(1)}^{+(b)}(y)$, $\hat{F}_{Y(1)}^{+(b)}(y)$, $\hat{F}_{Y(1)}^{S+(b)}(y)$, and $\hat{F}_{Y(1)}^{S+(b)}(y)$ denote the

respective bootstrapped distributions of the bootstrap sample b (for $b = 1, ..., B$). The bootstrap resamples randomly draw observations at the individual-level with replacement (see, e.g., Giné and Zinn, 1990, for two-sided tests).

We can impose three null hypotheses:

$$H_0^{Y(1)} : F_{Y(1)}^{S+}(y) = F_{Y(1)}^+(y),$$
$$H_0^{Y(0)} : F_{Y(0)}^{S+}(y) = F_{Y(0)}^+(y), \text{ and}$$
$$H_0^{Y(1),Y(0)} : F_{Y(1)}^{S+}(y) = F_{Y(1)}^+(y) \text{ and } F_{Y(0)}^{S+}(y) = F_{Y(0)}^+(y)$$

The possible KS test statistics are

$$\widehat{KS}_{Y(1)} = \sup_y |\hat{F}_{Y(1)}^+(y) - \hat{F}_{Y(1)}^{S+}(y)|,$$
$$\widehat{KS}_{Y(0)} = \sup_y |\hat{F}_{Y(0)}^+(y) - \hat{F}_{Y(0)}^{S+}(y)|, \text{ and}$$
$$\widehat{KS}_{Y(1),Y(0)} = \max(\widehat{KS}_{Y(1)}, \widehat{KS}_{Y(0)}),$$

where $\widehat{KS}_{Y(1),Y(0)}$ is a joint test of both conditions. The re-centered KS test statistics are

$$\widehat{KS}_{Y(1)}^{(b)} = \sup_y |\hat{F}_{Y(1)}^{+(b)}(y) - \hat{F}_{Y(1)}^{S+(b)}(y) - (\hat{F}_{Y(1)}^+(y) - \hat{F}_{Y(1)}^{S+}(y))|,$$
$$\widehat{KS}_{Y(0)}^{(b)} = \sup_y |\hat{F}_{Y(0)}^{+(b)}(y) - \hat{F}_{Y(0)}^{S+(b)}(y) - (\hat{F}_{Y(0)}^+(y) - \hat{F}_{Y(0)}^{+S}(y))|, \text{ and}$$
$$\widehat{KS}_{Y(1),Y(0)}^{(b)} = \max(\widehat{KS}_{Y(1)}^{(b)}, \widehat{KS}_{Y(0)}^{(b)}).$$

The estimated p-values for the null hypotheses, that the potential and simulated potential outcome distributions are equal, are

$$\hat{\alpha}(H_0^{Y(1)}) = \frac{1}{B} \sum_{b=1}^B 1\left\{ \widehat{KS}_{Y(1)}^{(b)} > \widehat{KS}_{Y(1)} \right\},$$
$$\hat{\alpha}(H_0^{Y(0)}) = \frac{1}{B} \sum_{b=1}^B 1\left\{ \widehat{KS}_{Y(0)}^{(b)} > \widehat{KS}_{Y(0)} \right\}, \text{ and}$$
$$\hat{\alpha}(H_0^{Y(1),Y(0)}) = \frac{1}{B} \sum_{b=1}^B 1\left\{ \widehat{KS}_{Y(1),Y(0)}^{(b)} > \widehat{KS}_{Y(1),Y(0)} \right\}.$$

Table F.1 reports the test results. The null hypotheses for the equality of distributions must be rejected for all estimation approaches. Accordingly, I have the reject the hypothesis that the predicted CATEs and QTE are nested. Furthermore, the predicted CATEs do not approximate the individual causal effects.

Table F.1: Test for equality between positive potential and simulated earnings distributions.

	Joint test		Under Jobs First		Under AFDC	
	test stat.	p-value	test stat.	p-value	test stat.	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Local constant model						
Baseline	0.461	0.000	0.461	0.000	0.093	0.059
Tree						
Baseline	0.462	0.033	0.462	0.033	0.081	0.066
Decent	0.190	0.000	0.190	0.000	0.148	0.000
Kitchen sink	0.168	0.000	0.168	0.000	0.139	0.000
Random forest						
Baseline	0.434	0.000	0.434	0.000	0.096	0.001
Decent	0.393	0.000	0.393	0.000	0.209	0.000
Kitchen sink	0.434	0.000	0.434	0.000	0.170	0.000
Causal forest						
Kitchen sink	0.439	0.000	0.439	0.000	0.130	0.000
Forest without confounder adjustment						
Kitchen sink	0.400	0.000	0.400	0.000	0.163	0.000

Note: The p-values are bootstrapped with 1,999 replications. Columns (3) and (4) report the test results for condition (1), columns (5) and (6) report the test results for condition (2), and columns (1) and (2) report the test results for both conditions jointly.