Conditional Average Treatment Effects and Decision Making*

Mario Samano†

April 21, 2014

Abstract

This paper develops a decision making empirical method to evaluate welfare programs accounting for heterogeneity of impacts. We find outcome predictive distributions for different subgroups of the population and use a characterization of second order stochastic dominance to give a policy recommendation conditional on covariates with minimal requirements on the social planner's utility function. Further, we can estimate quantile treatment effects within subgroups. We apply this method to the Connecticut's Jobs First program and find subgroups for which the program may have been welfare enhancing even though some statistics may suggest the opposite and vice-versa.

JEL: C11, D31, I38, J31

Keywords: conditional average treatment effects, heterogeneous impacts, hierarchical Bayesian model, decision making, stochastic dominance

*Some of the data used in this paper are derived from data files made available by the Manpower Demonstration Research Corporation (MDRC). The author remains solely responsible for how the data have been used or interpreted. The paper benefited from comments at seminar presentations at HEC Montreal, the Canadian Economics Association Annual Meeting, the University of Toronto, and from Martin Burda, Emmanuel Flachaire, and Daniel Parent. I thank Jonah Gelbach, Kei Hirano, Ron Oaxaca, and participants at a seminar presentation at the University of Arizona for their comments on an earlier version of the paper. All errors are my own.

†HEC Montreal: Institute of Applied Economics. email: mario.samano@hec.ca

1 Introduction

Studies evaluating social programs usually concentrate on overall population measures that miss the possible heterogeneity in the program impacts. Knowledge of the degree of heterogeneity is relevant to policy makers when extending the program to another jurisdiction or when reducing the number of beneficiaries since that may or may not bring potential welfare losses. This paper considers an empirical method to infer the heterogeneous impacts of social programs and how to map those into decisions using expected utility theory.

The problem can arise in the following situation, before implementing the social program to a larger population, a pilot phase is run in order to get insight about its benefits. Using the data collected in that phase, the social planner has to decide whether to implement the program or not on a larger sample, also the planner could decide to restrict the application of the program to a smaller population. The classical approach to answer this question is to find an overall treatment effect of the program and if it is statistically significant different from zero and positive, the program is likely to be preserved or extended. Here we propose a method to make inference on the treatment effect conditional on covariates even if the sample is very small within the subgroups, which brings a curse of dimensionality problem, and how to use that information in a decision making process.

Specifically, we find the probability distributions of potential outcomes for treated and controls for different subgroups of the population defined by covariates using a hierarchical Bayesian model. Then we can determine whether a social planner with a strictly increasing and concave utility function would prefer to assign a similar individual to the treatment or not by taking the expectation over the potential outcomes probability distributions. This can be done without having to specify the functional form of the planner's utility function[1]. Moreover, with knowledge of the outcome distributions we can also calculate quantile treatment effects within the subgroups. Thus the method gives not only information on the heterogeneity across subgroups but also within.

[1]Specific conditions on the utility function are discussed later in the paper. The method is agnostic in two cases: a risk-neutral social planner and positive conditional average treatment effect, and in the absence of second order stochastic dominance.

A few papers[2] have considered the program evaluation assessment as a statistical decision problem: by proposing a measure of the value of covariate information and how the higher the variation in the treatment response as a function of covariates, the higher that value (Manski [2001], Manski [2004]); by looking at predictive outcome distributions for treated and controls and then using first order stochastic dominance to rank those distributions (Dehejia [2005]); by extending and analyzing Manski's approach to what the rules of assignment converge to in distribution when a loss function evaluates the decision rule and its interaction with risk rather than pure statistical rules (Hirano and Porter [2009]); by constructing a framework to evaluate policies using quasi-experimental data (Kasy [2013]); and by deriving asymptotic frequentist confidence intervals for welfare gains in randomized assignment experiments when there are budget constraints (Bhattacharya and Dupas [2012]).

On the other hand, another rather small branch of the literature has looked into the empirical problems to measure the heterogeneity of treatment effects (Abrevaya et al. [2012], Bitler et al. [2006], Bitler et al. [2010]); and by considering the theoretical issues of the hierarchical Bayesian models for evaluation programs when there is heterogeneity and the use of an empirical Bayesian approach (Chamberlain [2011]). The use of Bayesian methods to estimate treatment effects dates back to Rubin [1978][3].

More recently, a related branch of the literature has focused on extending Bayesian inference methods for the heterogeneity of endogenous treatment effects. Hu et al. [2011] and Hu [2011] for example use a Dirichlet process mixture to determine the number of components of heterogeneity based on a three-equation selection model. One equation for the selection, one for the treated, and one for the controls. Chib et al. [2009] propose a four-equation model to simultaneously allow for sample selection, endogeneity, and nonparametric covari-

[2]Imbens and Wooldridge [2009] describe how the literature has concentrated on the overall treatment effect and argue that this would only make sense if we were to mandate exposure to the treatment to either everyone in the population or to no one, which in practice that is not usually the case.

[3]He studies the estimation of causal effects by finding the predictive distribution of the outcomes that correspond to the treatments that are not available in the experiment. The idea is that the unobservable outcomes (because an individual can be assigned to either the treatment or to the control group but not both) are missing data and those values can be imputed by giving a prior distribution on the potential outcomes. After this paper, little had been done on estimating treatment effects using Bayesian techniques (Imbens [2004]), perhaps because of the intensive computational requirements.

ate effects. In their Bayesian implementation of the model there is no need to augment the data to impute missing values due to the selection mechanism as opposed to the general non-frequentist approach to latent variable models. Li and Tobias [2011] propose a three-equation model and its Bayesian inference to obtain heterogenous causal effects. The model allows for joint inference on the determinants of the outcome, the endogenous treatment variable, and an individual-specific component of the causal effect. But instead of keeping these individual effects throughout the inference, individuals are exogenously put in different groups leading to determining heterogeneity using a mixture model over these groups, not across the individuals[4]. In general, estimation of treatment effects for small samples or for very restrictive subsets of covariates has been avoided because of the potential lack of statistical significance under classical approaches[5].

The method we develop in this paper connects these two branches of the literature. We use a model that captures correlations across the subgroups of the population through a hierarchical Bayesian model and then second order stochastic dominance to map the within subgroups probability distributions of potential outcomes into expected utility values[6]. In our case we do not model any self-selection issue in the data, but focus on the randomized experiment on those who signed up for cash assistance. Moreover, our approach keeps parametric assumptions at a minimum. An element in common with this literature is that in order to implement the methods, it is needed to exogenously assume a number of subgroups in the population over which heterogenous effects can be determined; Bayes factors or other criteria could be used to choose the number of the subgroups.

Perhaps the closest papers to ours are Bitler et al. [2010] and Dehejia [2005]. In the former they measured heterogenous impacts of the Connecticut's Jobs First program by

[4]There is another branch of the literature developing tests to isolate the covariates that have an impact on an outcome out of a large collection of covariates. This is the typical situation in studies attempting to identify the genes associated with a specific disease, as in Efron [2011].

[5]Graham and Hirano [2011] discuss the performance of different estimators when there are no observations for particular combinations of covariates values.

[6]At a broader scale, methods to estimate treatment effects can be classified as follows: (i) regressions of the outcome on the covariates and a treatment dummy, (ii) matching on covariates, (iii) using the propensity score, (iv) combinations of those methods, and (v) Bayesian methods. Imbens [2004] Heckman [2010] gives an in-depth analysis of the state-of-the-art on the structural versus the reduced form approaches to do empirical work on program evaluation.

looking at the differences in income between treated and controls at each different quantile of the income distribution for different subgroups of the population. Then they developed a test for the null that conditional treatment effects across subgroups are the same. This is similar to our work in that by using completely different methods we find a high degree of heterogeneity. The main difference is that we can rank distributions. Dehejia [2005], using data from a randomized experiment, estimates posterior densities for the treatment effects and evaluates under different scenarios the maximization of welfare using the posterior densities. His contribution is that he models program evaluation as a decision problem. For example, in order to decide whether all the individuals should be assigned to the treatment. Another scenario is the one of a caseworker that has to decide if a person should be assigned to the treatment or to the control group. In order to carry out these decisions, he uses classical expected utility theory since the predictive distributions for the outcomes are known in a Bayesian framework. The main difference with respect to that paper is that we also condition on covariates so that the heterogeneity problem overlaps with expected utility rankings.

We apply the method to the Connecticut's Jobs First program. This arose as one of the welfare reform experiments in the U.S. caused by the elimination of the Aid to Families with Dependent Children (AFDC) program in 1996. The Federal government required the states to replace AFDC with programs that had a time limit to participate and that would enhance job training. Thus we focus in this paper on the transitional program Jobs First implemented in Connecticut, specifically in Manchester and New Haven. The welfare experiment consisted in randomly assigning 4,803 people with at least one dependent children to either AFDC or to Jobs First[7]. The outcome of interest is the average quarterly amount of earnings over the first seven quarters after inception to the experiment since positive earnings directly reflect employment, which is the main program's objective. One of the main differences of Jobs First with respect to AFDC is that the former has a limit of time of cash assistance of 21 months plus 6 months if an extension is requested and approved. This feature coupled with the requirement of following a job training program instead of just general education programs,

[7]Of these, 96% were women.

were expected to make a difference in the level of the impacts[8]. It is however unclear how the program impacts vary with covariates, and even less so how to make decisions on extensions of the program based on expected utility arguments. The impacts here are the differences between the outcome level if the individual was in the experimental group and the outcome level if she was in the control group.

Our results suggest that even though classical statistical measures of conditional treatment effects on earnings are negative for 9 out of 24 subgroups, only 5 of them are welfare-decreasing. When using total average income -earnings plus transfers-, the conditional average treatment effect for only one subgroup in our main specification is negative, when using the expected utility approach there are two other subgroups for which exposure to the treatment does not maximize utility. This difference is due to the curse of dimensionality when trying to estimate expected values using very small sample sizes, and the lack of use of information across subgroups. There is also a high degree of heterogeneity in quantile treatment effects within the subgroups. A second specification in which pre-treatment covariates are discarded when forming the subgroups wipes out most of the heterogeneity.

We believe it is relevant to understand the impacts of social programs not just as an overall measure for the entire population but rather how large the heterogeneity of these impacts is and how to make assignment decisions based on robust theoretical expected utility results that do not depend on particular functional forms but only minimal conditions. The rest of the paper is organized as follows. Section 2 presents the definitions and formal statements of the problem. Details on Jobs First and some data issues are explained in Section 3. Section 4 presents a hierarchical Bayesian model and a Gibbs sampler. Finally, in Section 5 we present the connection between predictive income distributions and expected utility theory as well as the main results. Section 6 concludes.

[8]For more details on Jobs First see the MDRC's Final Report (Adams-Ciardullo et al. [2002]).

2 Conditional Average Treatment Effects

A social planner is trying to decide whether an individual should be assigned to a treatment or not to maximize an outcome. In this paper, the main outcome is earnings and we do robustness checks using total income. Equivalent outcomes in other applications are health quality, level of education, and in general, any outcome of a social program that can be quantified. Our social planner has data on N individuals. The data for each individual consist of individual characteristics, whether the individual received the treatment or not, and the outcome. Individuals are randomly assigned to receive the treatment or not. Each individual is characterized by a vector of covariates X_i of size $1 \times K$. Denote by $T_i = 1$ if individual is exposed to treatment and by $T_i = 0$ if the individual is not. The outcome will be denoted by

$$Y_i \equiv T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0).$$

Since treatment is completely at random, the unconfoundedness assumption which captures the idea that treatment assignment is exogeneous conditioned on X_i is automatically satisfied, $(Y_i(0), Y_i(1)) \perp T_i | X_i$[9]. Thus, conditioning on X, treatment assignment is independent of the potential outcome. However, the realization of T does affect which outcome is observed.

We are interested in estimating the effect of being under the treatment for a given X. As one observes either $Y_i(0)$ or $Y_i(1)$ but not both at the same time for individual i, we always have some uncertainty in one of the two outcomes. The statistic of interest is the difference between the outcomes had the individual been in both cases. We call this quantity the *average treatment effect* (ATE), $\beta := E(Y(1) - Y(0))$. A cell is each of the elements of the discretization over all the covariates. If we are interested in the treatment effect for individuals in the cell $X = x$, we define the *conditional average treatment effect* (CATE), $\beta(x) := E(Y_i(1) - Y_i(0)|X = x)$. Note that this can be interpreted as an average treatment effect for a group of individuals with characteristics $X = x$. We can rewrite $\beta(x)$ as $\beta(x) = E(Y_i|T = 1, X = x) - E(Y_i|T = 0, X = x)$ where $Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$. The overall

[9]First proposed by Rosenbaum and Rubin [1983].

ATE can be obtained by taking the expectation with respect to X,

$$\beta = E(\beta(x)).$$

This suggests that estimation for $\beta(x)$ consists of two parts, $\beta_1(x) = E(Y|T = 1, X = x)$ and $\beta_0(x) = E(Y|T = 0, X = x)$. A simple estimator for the CATE can be constructed using sample means,

$$\begin{aligned}\hat{\beta}_1(x) &= \frac{1}{N_1(x)} \sum_{i:T_i=1,X=x} Y_i \\ \hat{\beta}_0(x) &= \frac{1}{N - N_1(x)} \sum_{i:T_i=0,X=x} Y_i\end{aligned}$$

where $N_1(x)$ is the number of individuals with covariate $X = x$ and under the treatment. An estimator for the CATE is the difference

$$\hat{\beta}(x) = \hat{\beta}_1(x) - \hat{\beta}_0(x). \tag{1}$$

It is a consistent estimator[10] and its variance is $Var(\hat{\beta}(x)) = \frac{1}{N_1(x)}\sigma_1^2(x) + \frac{1}{N-N_1(x)}\sigma_0^2(x)$, where $\sigma_1^2(x)$ and $\sigma_0^2(x)$ are the variance of the outcomes for the treated and the controls respectively when $X = x$. By replacing the variances with their sample analogs we have $\widehat{Var(\hat{\beta}(x))} = \frac{1}{N_1(x)}\hat{\sigma}_1^2(x) + \frac{1}{N-N_1(x)}\hat{\sigma}_0^2(x)$.

Our problem is to find a rule to decide whether a new individual $N+1$ should be assigned to the treatment or not. Suppose that the social planner wants to maximize the population mean outcome of the treatment if she has an increasing and concave utility function. This is equivalent to maximize the outcome of each individual. That is, conditional on X, she wants to choose T_{N+1} such that

$$T_{N+1} = \arg \max_{t \in \{0,1\}} \{E(Y_{N+1}(t)|X_{N+1})\}$$

[10]By using the unconfoundedness assumption to get the third equality,

$$\begin{aligned}\operatorname{p\,lim} \hat{\beta} &= E(Y_i|T = 1, X = x) - E(Y_i|T = 0, X = x) \\ &= E(Y_i(1)|T = 1, X = x) - E(Y_i(0)|T = 0, X = x) \\ &= E(Y_i(1)|X = x) - E(Y_i(0)|X = x) = \beta.\end{aligned}$$

so that the population mean outcome is $E(\max_{t\in\{0,1\}}\{E(Y(t)|X)\})$, where the outside expectation is with respect to X (see Manski [2001]). In order to compute the last expectation we need the entire distribution of $Y(t)|X$ and not just one point statistic. It is clear that the function $E(Y(t)|X)$ is identified given the random assigment, but as the partition becomes finer we get to a curse of dimensionality problem and there might not be in the data enough individuals to estimate this function. We can use some non-parametric approach to infer the distribution, or a Bayesian model that as an output gives us the desired distribution. We opt for the latter since it easily handles correlations across different subgroups. Those correlations should not arise given the random assignment of the treatment, however the small sample sizes of the subgroups might induce across-subgroups correlation.

We use a characterization of second order stochastic dominance. This criterion for ranking distributions relies on the inspection of the cumulative areas under each distribution function. If for every point in the domain the cumulative area of the distribution function of one random variable is lower than for another random variable's, we say that the former second order stochastic dominates the latter. The well-known result we use in this paper is that for two random variables X and Z, $E(u(X)) \geq E(u(Z))$ for *any* strictly increasing and concave function u if and only if X second order stochastically dominates Z. The latter statement can be easily evaluated using the predictive income distributions obtained from the data.

The assigning rule would now be to take the $(N+1)$-th individual, and according to her characteristics $X_{N+1} = x$, she would belong to one of the subgroups, and she would be assigned to the treatment if the random variable $Y|T = 1, X = x$ second order stochastically dominates $Y|T = 0, X = x$.

The social planner has some information to decide how to discretize each covariate. In the empirical application it is assumed that this criterion is exogenous to the problem of maximization of the expected value of the outcome and the level of discretization is left for future research. For example, if the covariate is age the social planner might discretize the variable into young, adults and seniors according to specific ranges for the covariate age. If the sample is large enough the level of discretization can be less coarse. In the example, young, adults and seniors are the three cells obtained by the discretization. For identification

purposes, we require that in each of the cells there is at least one individual under the treatment and at least one individual in the control group. That is, in each segment of each covariate there exists at least one pair $(X_i, Y_i(0))$ and at least one pair $(X_i, Y_i(1))$. Once we have the discretization for each covariate we can form subgroups of individuals according to all the possible combinations of the segments. This means that we take the cartesian product over the segments of the covariates. In this way, every individual belongs to exactly one subgroup and only one and each subgroup contains at least one pair $(X_i, Y_i(0))$ and one pair $(X_i, Y_i(1))$.

3 Connecticut's *Jobs First* Program

Connecticut's Jobs First was put in place just before the passage of the Federal welfare reform through the Personal Responsibility and Work Opportunity Act (PRWORA) signed by President Bill Clinton in 1996. One of the main consequences of PRWORA was the substitution of AFDC with the Temporary Assistance for Needy Families (TANF) program. Jobs First was one of the very first programs with all the TANF main features to be implemented, and it is thus one of the few examples available to get an insight on the performance of TANF programs versus the previous AFDC format. The main differences between the two are that TANF-type programs have time limits, different implied rates of taxation, and work requirements where as AFDC did not.[11]

The time limit consisted of a maximum allowable period of 21 months of cash assistance, with the possibility of an extension of 6 months if the recipient complied with requirements for an exemption. AFDC stops providing assistance when all children are 18 or older. TANF now allows for up to 60 months of cash assistance. We should keep in mind though that about two thirds of those in our sample who reached the time limit under Jobs First were approved the extension of 6 months. This is why we look at the quarterly average of earnings over the first seven quarters after inception to the program.

The second key feature of Jobs First is that the earnings from working are completely

[11] Adams-Ciardullo et al. [2002].

disregarded when calculating the food stamps benefits as long as these earnings were below the FPL[12]. This was not the case under AFDC where the earnings disregarded were a function of the length of the time in the program[13].

Jobs First required recipients to sign up for job training programs, not just any type of education. Lack of compliance with this requirement imposed grant reductions of 20-35% up to 100% reductions if in extended period. AFDC on the other hand removed non-compliant individual from transfers calculations, which is equivalent to about a $100 per month reduction. Other features of Jobs First are that the asset limit for cash assistance is $3,000 whereas under AFDC was $1,000. The benefit per child conceived while mother was receiving cash assistance was $50 per month under Jobs First and about $100 under AFDC. Jobs First was a state-wide program, however the data we use here come from only two locations: New Haven and Manchester which altogether account for about one fourth of the entire caseload. Among all these recipients, 96% were women. All recipients are eligible only if they have at least one child at the moment of inception to either program. Figure 1 compares the benefits between the two programs for typical households.

Estimates on the mean impact of Jobs First versus AFDC have been found to be $453 on income per quarter according to the Final Report (Adams-Ciardullo et al. [2002])[14]. Other studies such as Bitler et al. [2006] have found a mean impact of $294 per quarter over the first seven quarters in the program. Moreover, Bitler et al. [2006] found significant heterogeneous effects when looking at the impacts conditional on quantiles of the income distribution. These impacts range from 0 to $800. The impacts increase with the quantile and then decrease towards the upper quantiles. This may be attributed to an opt-in effect as some individuals might lower their income in order to be eligible for the program. The type of heterogeneity they looked at is the difference between the cumulative distribution of income for the treated and for the controls at each different quantile[15]. These large variations in

[12]Federal Poverty Line, which was $1,138 per month for a family of three people in 1998.

[13]AFDC disregarded $120 plus 33% of earnings per month during the first four months in the program. During months 5-12 only $120 per month were disregarded. After 12 months only $90 were disregarded per month when calculating the food stamps benefits.

[14]Income is defined as the sum of earnings from working and the food stamps benefits.

[15]The horizontal difference between the graphs of the cumulative distribution functions for any given

the impacts conditional on income quantiles can be further decomposed into heterogeneous impacts conditional on covariates. This is possible using the method proposed in this paper.

The sample consists of 2,396 Jobs First recipients and 2,407 AFDC recipients, making a total of 4,803 families. In this paper we focus on four covariates: age, whether the individual was employed before inception to the program, number of children, and high school degree. Summary statistics for the entire sample are shown in Table 2. The partition of the sample consists of the cartesian product of the segments created in each variable as follows: the covariate age is segmented in two parts depending on whether the parent is at least 30 years old or not; a dummy variable for employment before the program; dummy variables for being pregnant or with one kid, with two kids, or with three or more kids; and whether the parent has a high school degree or not. This gives a total of $J = 24$ subgroups and a total of 4,463 observations due to missing values in the high school degree covariate. Table 1 describes the subgroups and Table 2 shows summary statistics for the entire sample. As explained before, our main outcome of interest is the quarterly average of earnings since it reflects the program's main objective: employment. Over the first seven quarters, 27% of the sample has zero earnings, 5.5% no transfers, and 1.2% neither. This produces a mass point over zero on the empirical earnings distributions due to the individuals who are unemployed, an issue we further discuss in the next section.

The particular choice of the subgroups is certainly exogenous in the model. This however has been done using covariates that the Final Report uses as explanatory variables in their estimation of impacts. Bitler et al. [2010] use the same covariates as well to diagnose subgroup heterogeneity in Jobs First. In general this is a good approach to form the subgroups because it is realistic to think that a social planner extending or constraining the program to a different population will be able to sustain her criteria if they are based on such covariates because age can determine your expertise in your job and the number of kids your actual economic needs. Having being previously employed is also used as a criterion for welfare aid, such as in unemployment benefits. Even though we have data on race, we believe its use in reality

number between 0 and 1. For example, at the q-th quantile, the difference $Y_{1,q} - Y_{0,q}$ is the quantile treatment effect, where $F_i(Y_{i,q}) = q$ is the cumulative distribution function of income Y for group i.

could bring problems as it could be taken as racial profiling. In the robustness section we discuss two other different specifications for the subgroups, using the race information and using only non-historical covariates. We also discuss the results when the outcome is the average total income -earnings plus transfers- over the first seven quarters of inception to the program.

As a benchmark for the results, Table 3 and Table 4 report the estimates for the CATEs using the estimator (1). We can see the tremendous amount of variation between the subgroups' CATEs. The goal is to find which subgroups have a high value for their CATE, not only by taking the within subgroup sample means because of the lack of statistical significance we could get in some of the subgroups, but from a decision making standpoint. Column 10 in Table 3 and Table 4 shows the t-statistic for the null hypothesis that the CATE is zero[16] and how in most subgroups the null hypothesis cannot be strongly rejected. This paper proposes how to calculate predictive probability distributions for income and how to use them in an expected utility maximization setting.

4 Estimation and Results

Once the subgroups are formed, the sample can be described by pairs (T_{ij}, Y_{ij}) where i represents the individual observation and j is the index for different subgroups in the sample. Each subgroup can have a different number of observations, so $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$ where J is the number of subgroups. One could estimate the CATEs by taking the difference of the sample means in each subgroup but this might not give any meaningful results since the size of the subgroup could be very small. That would not be taking into account the correlation with other cells either. Table 3 and Table 4 show the results. There are two observations to make, first, statistical significance of these differences as measured by the t-statistic is rarely found, and second, among the CATEs that exhibit certain degree of significance it is evident that the size of the program impacts substantially differs from one

[16] These are calculated to test whether the two means are statistically significantly different. The statistic is $t = \frac{\bar{Y}(1)-\bar{Y}(0)}{\sqrt{\frac{(n_1-1)S_1^2+(n_0-1)S_0^2}{n_1+n_0-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}}$ where S_i^2 are the sample variances for the treated and the controls. See Dehejia [2008] for more details on rules-of-thumb for inference of treatment effects.

subgroup to another. The former is just a consequence of the lack of statistical power of this simple estimator on small sample sizes. The second suggests that in effect, the impact of the program is not constant across subgroups. A different approach, which also relies on asymptotics, is to create dummy variables for each subgroup and then estimate a linear model on these dummies and their interactions with the treatment indicator variable. The results are available from the author upon request. Once again, it is rare to come across with a statistically significant coefficient.

4.1 A Hierarchical Bayesian Model

To overcome issues of estimators that rely on asymptotics we use a Bayesian approach instead. We assume that each realization of the random variable Y_{ij} is an independent draw from a normal distribution,

$$Y_{ij} \sim N(\mu_j, \tau_{\mu_j}).$$

This way, each subgroup has its own mean μ_j and own precision τ_{μ_j}[17]. We want to capture that these means are not independent, equivalently, there is some correlation across the subgroups[18]. Assume an underlying common distribution for all the subgroups from which we draw these means,

$$\mu_j \sim N(\mu, h_1),$$

where μ is the overall outcome mean. If we independently draw from that distribution we have that the subgroup means $\mu_j|\mu$ are conditionally independent. However, without conditioning on μ, the subgroup means are correlated. To see why, we complete this model with two more layers and then compute the correlation across subgroups.

Assume the prior on μ and on ν,

$$\mu \sim N(\nu, h_2) \tag{2}$$

$$\nu \sim N(0, h_3)$$

[17]We assume the traditional notation in the Bayesian literature in which the second parameter of the normal distribution specification refers to the precision not the variance.

[18]As discussed earlier, in principle there should not be any correlation since assignment to treatment is random but the curse of dimensionality may bring a non-negligible correlation across subgroups.

and a gamma density prior on τ_{μ_j} for each $j = 1, \ldots, J$,

$$\tau_{\mu_j} \sim G(\alpha, \beta).$$

The parameters h_3, α and β are fixed. Since we do not know how the subgroup means μ_j are allocated with respect to μ, we give a prior on the precision h_1 that captures whether the subgroups share more or less information. A flexible prior is a gamma distribution with most of its mass close to the origin which is equivalent to have a difuse prior on μ_j[19] [20]. Suppose that the subgroups are very similar to each other, then the subgroup means μ_j will be similar as well. On the other hand, if the subgroups do not share much information, we expect the subgroup means μ_j to be very different and the posterior distribution should be using little information from μ to estimate the subgroup means. This completes the model, a hierarchical Bayesian model.

Alternatively, we can rewrite the model as $\mu_j = \mu + \epsilon_j$ with $\epsilon_j \sim N(0, h_1)$, and $\mu = \nu + \epsilon_\mu$ with $\epsilon_\mu \sim N(0, h_2)$. This implies that

$$\mu_j \sim N(\nu, 1/(h_1^{-1} + h_2^{-1})).$$

Thus the variance of a subgroup and the covariance between two given subgroups are respectively,

$$Var(\mu_j) = h_1^{-1} + h_2^{-1}$$

$$Cov(\mu_j, \mu_k) = E(\epsilon_\mu^2) = h_2^{-1}.$$

If the subgroups are very different from each other, the covariance between two subgroups should be small, large h_2, and thus the variance of one single subgroup mean primarily depends on the size of h_1^{-1} and not on the size of h_2^{-1}. Although it may seem restrictive at first to have a constant value for the subgroup means correlation, this greatly simplifies the number of parameters in the model. If each pair of subgroups had its own correlation parameter, this would require $\frac{J(J-1)}{2}$ prior distributions and an equal number of posterior distributions to determine. Notice also that the constant correlation property of this model

[19] See details in the Appendix.

[20] Gamma distributions will be denoted as $G(a, b)$ with mean a/b and variance a/b^2.

in its specification does not mean it cannot provide information about the heterogeneity of such correlation. The Bayesian model outputs a posterior distribution on h_2, and even though we cannot distinguish where the pairwise correlations are found with respect to that density, it gives information on the dispersion of the different correlations could they be modelled separately. Other properties about the across subgroups correlation arise in the posterior distributions below.

The normality assumption for the outcome is not ideal since there is a probability mass point on the left of the income distribution at zero as explained in the previous section. The censoring at the left of the distribution occurs because of the people in the sample that are unemployed, or because they did not receive any transfers in the form of food stamps benefits in the case of the total income distribution. For this reason we choose a Tobit likelihood for our most general model[21]. For the ease of exposition, we first present the model as if there was no mass point at zero and then present the full model.

The full conditional density is proportional to

$$\begin{aligned} p(\mu_1,\ldots,\mu_J,\mu,\tau_{\mu_1},\ldots,\tau_{\mu_J}|Y_{ij},h_1,h_2,h_3,\alpha,\beta) \quad \propto \quad & \left(\prod_{i,j}\tau_{\mu_j}^{1/2}\exp\left(-\frac{1}{2}\tau_{\mu_j}(Y_{ij}-\mu_j)^2\right)\right)\times \\ & \left(\prod_{j=1}^{J}h_1^{1/2}\exp\left(-\frac{1}{2}h_1(\mu_j-\mu)^2\right)\right)\times \\ & h_2^{1/2}\exp\left(-\frac{1}{2}h_2\mu^2\right)\times \\ & \prod_{j=1}^{J}\tau_{\mu_j}^{\alpha-1}\exp(-\beta\tau_{\mu_j}) \end{aligned}$$

where each term of the product above is the kernel of the densities for Y_{ij}, $\{\mu_j\}_{j=1,\ldots,J}$, μ and $\{\tau_{\mu_j}\}_{j=1,\ldots,J}$ respectively.

The estimation is done using a Gibbs sampler. In general, this method produces, when the number of iterations is large enough, a random draw from some density $p(\theta|A)$, where θ is the random vector of interest and A is the set of fixed parameters. Assume that θ can be decomposed in blocks of parameters $\theta=(\theta_1,\theta_2)$, so that $p(\theta|A)$ can be written as $p(\theta_1|\theta_2,A)$

[21] A third suitable likelihood is a mixture of a mass point on zero and a normal density on the $\log(income)$ for the positive observations.

and as $p(\theta_2|\theta_1, A)$ and it is possible to get random draws from these two densities. In this setting, a Gibbs sampler consists of getting a random draw $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(0)}, A)$ using some initial value $\theta_2^{(0)}$ where the superscripts denote the number of the iteration, then getting a random draw $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(1)}, A)$. Note that an initial value was needed in order to sample from $p(\theta_1|\theta_2, A)$, and also note that we update the values of θ_i in the next iteration. If the iteration is done N times, we get a random sample of size N of draws for θ_1 and for θ_2. Under some general conditions, it can be shown that[22]

$$E(g(\theta)) \approx \frac{1}{N}\sum_{j=1}^{N} g(\theta_1^{(j)}, \theta_2^{(j)}).$$

for large values of N.

Going back to the estimation of CATEs, it is of interest to write down the joint distribution as a set of full conditional distributions. Because of the form of the joint distribution in this particular setting, we can write the full conditional distributions as[23]

$$\mu_j|\mu, \mu_{-j}, \tau_\mu, Y_{ij}, h_1, h_2, \alpha, \beta \sim N\left(\frac{\sum_{i=1}^{n_j} Y_{ij} + h_1\mu/\tau_{\mu_j}}{n_j + h_1/\tau_{\mu_j}}, \tau_{\mu_j} n_j + h_1\right),\ j = 1, \ldots, J,$$

$$\mu|\{\mu_j\}_{j=1,\ldots,J}, \{\tau_{\mu_j}\}_{j=1,\ldots,J}, Y_{ij}, h_1, h_2, \alpha, \beta \sim N\left(\frac{\sum_{j=1}^{J}\mu_j}{J + h_2/h_1}, Jh_1 + h_2\right),$$

$$\tau_{\mu_j}|\mu, \{\mu_j\}_{j=1,\ldots,J}, \tau_{\mu\ -j}, Y_{ij}, h_1, h_2, \alpha, \beta \sim G\left(\alpha + n_j/2, \beta + \frac{1}{2}\sum_{i\in j}(Y_{ij} - \mu_j)^2\right),\ j = 1, \ldots, J.$$

where μ_{-j} is the collection of the μ_k with $k \neq j$. These three densities can be used in a Gibbs sampler. We can see that the posterior mean of the subgroup means[24] $\mu_j|A$ is a weighted average of the outcomes within the subgroup and the overall outcome mean μ. The precision is a weighted sum of the precision of the income for subgroup j and the precision of μ_j. The overall mean μ is weighted by the precision of subgroup j. Also, the precision of the overall distribution is weighted by the precision of group j. Observe that if n_j increases -there are more individuals in subgroup j-, the posterior precision $\tau_{\mu_j} n_j + h_1$ increases, which is a good attribute of the model.

[22] See Geweke [2005].

[23] See Appendix for details.

[24] The letter A henceforth will denote all the parameters and data we are conditioning on at that particular step of the sampler.

From the posterior distribution of μ we see that its posterior mean is almost the average of the subgroup means, but the denominator contains the term h_2/h_1. This suggests that h_2/h_1 should be small compared to J, the number of subgroups, in order to let the data dominate and not the initial values of the parameters. Then, in the specification of the priors we will require $h_2 < h_1$. As shown in the Appendix, the posterior distributions for these hyperparameters do not explicitly depend on the value of the observations. The posterior precision suggests that when the number of subgroups increases, the precision is higher and that h_2, the precision in the prior for μ, should be very small to have an uninformative prior on μ.

The third posterior distribution is for the subgroup precisions. Observe that these hyperparameters do not depend on h_1 nor h_2 in the conditional density, but in the number of total observations and the overall sum of the squares of the deviations from the subgroup means. The parameters α and β should take on values that do not dominate the data. Let $\epsilon_j \equiv \frac{1}{2}\sum_{i\in j}(Y_{ij} - \mu_j)^2$. The mean of τ_{μ_j} is clearly not dominated by the parameters if α is much smaller than half the size of the smallest subgroup, and if β is much smaller than ϵ. However, when there is not much variation in the outcomes within the subgroup, i.e. when $\epsilon \to 0$, the expected value should be large enough to reflect the fact that the subgroup variance is close to 0. Therefore $(\alpha + n_j/2)/\beta$ cannot be too small. At the same time, the variance for this conditional density is $(\alpha+n_j/2)/(\beta+\epsilon)^2$ and it should be relatively larger in the $\epsilon \to 0$ case since there is not much information to accurately infer the subgroup variance. A good compromise on these parameter values is as shown in Table 5 and robustness checks are discussed at the end of the next section.

Finally, a discussion on the prior for h_1. As mentioned before, the precision of the distribution of μ_j is $h_1 \sim G(\gamma, \gamma)$, and a similar prior for h_2 applies. Another possibility is to use a generalized inverse gamma prior, but since our model is parameterized with the precisions, not with the variances, it is more tractable to use gamma densities. Remember that the posterior distribution for μ suggests that $h_2 < h_1$ is a good selection for the initial value of these parameters. The posterior distribution for $h_1|A_{-h_1}$, where A_{-h_1} represents the rest of the parameters, is $h_1|A_{-h_1} \sim G(\gamma + J/2, \gamma + \frac{1}{2}\sum_j(\mu_j - \mu)^2)$.[25] Note that if the

[25]See Appendix for details.

subgroup means μ_j are very far from each other then the posterior mean for h_1 is close to zero, which implies a large variance in the distribution of the subgroup means. The posterior density for h_2 is $h_2|A_{-h_2} \sim G(\gamma + 1/2, \gamma + \frac{1}{2}\mu^2)$ so the dispersion of the subgroup means depends on their overall mean and initial parameters, however the influence of the latter is negligible compared to the size of the other arguments.

The posterior densities above form a Gibbs sampler. Given some initial values for (μ_j, τ_{μ_j}) for $j = 1, \ldots, J$ and μ, the posterior distribution for μ_j is known and we can get a draw from that distribution and update those values to draw a value for μ, then we use that value to draw values for the τ_{μ_j}s from their posterior distribution. We repeat this process until getting convergence for all the posterior distributions. The estimation is made separately for the treated and the controls. The CATE for subgroup j is defined as $\mu_j^{treated} - \mu_j^{control}$. The link between $\mu_j^{treated}$ and $\mu_j^{control}$ is the prior from which the overall mean for the treated and the overall mean for the controls are drawn as in (2). The specification also allows for different variances for the overall mean of treated and controls, respectively.

As pointed out above, the data show evidence of censoring close to the zero earnings point. To alleviate this problem we include, on top of the Gibbs sampler explained above, a data augmentation step for Tobit models first proposed by Chib [1992]. Once the mass point at the extreme left of the distribution has been replaced with hypothetical negative observations, the Gibbs sampler above is applied in the exact same way as before[26].

4.2 Posterior CATEs

After an amount of iterations of the Gibbs sampler[27] and a burn-in phase[28] the histograms for all the parameters' posterior distributions show convergence in distribution, see Figure 2 which shows the time series of one of the Markov chains for selected parameters from the treated subsample. We can now calculate posterior means for the parameters, in particular we are interested in the posterior means for μ_j and the τ_{μ_j}. Table 6 presents the results

[26] See Appendix for details.

[27] 2,000 iterations for the outer loop and 2,000 iterations for the data augmentation step at each iteration of the outer loop.

[28] The first 1,000 draws from the Gibbs sampler are discarded to avoid any influence from the initial parameters of the priors in the final results.

by subgroup. Columns 2-5 from the top part of that table show the posterior means and the posterior standard deviations for the treated. The last four columns do the same for the controls. The bottom part of the table shows posterior means for the hyperparameters. Notice that even though the posterior mean values for the precisions τ_{μ_j} are in the order of 10^{-7}, the implied subgroup standard deviations are in the order of 10^3 dollars, which is in the same order of magnitude as our data on outcomes[29]. The bottom part of that table shows the posterior means for the rest of the hyperparameters. In particular, the values for h_2 imply non-negligible correlations, although this value is much larger for treated than for controls in the case of average earnings. For the case of total income both controls and treatment exhibit about the same level of correlation and it is not concentrated on a single point.

When looking at the distributions of predicted earnings, we can see how the mass point near zero has been spread out over negative values of the outcome[30]. This is an essential step to tackle the censoring problem in the data. With posterior densities of the subgroup means in hand we can calculate posterior densities for the CATEs for each subgroup. For every draw of the joint posterior distribution of parameters we find the predicted outcome for treated and for controls and their difference, this is the conditional average treatment effect posterior distribution. Figure 3 and the last two columns of Table 3 show these results. Figure 3 plots the kernel density estimates for the distribution of the posterior of CATEs for earnings. These densities can be used as well for calculating the probability of a positive treatment impact. Most subgroups' CATEs distributions span over the negative values range, leading to unclear evidence of any gains from the program.

We also run the model when the outcome is total average income. The posterior means for this case are shown in the last two columns of Table 4. The sign of the treatment effect does not change in the posterior means with respect to the sample means results. However, the level of the effects slightly changes in most cases. We can also see from the posterior densities of the CATEs the heterogeneity of the impacts within each of the subgroups. For example, subgroups 9, 10, 12, and 15 have a lot of heterogeneity if measured by the spread

[29]We further discuss sensitivity to different initial parameter values in the section on robustness checks.
[30]The next section explains how to construct predictive earnings distributions.

of the distribution. The first three correspond to people who were previously employed and no high school degree, whereas the last one corresponds to people with high school degree and three or more children[31].

5 Welfare

We can form predictive distributions for the outcome in the following way.

1. Get a random draw from the posterior distributions $p(\mu_j|A_{-\mu_j})$ and $p(\tau_{\mu_j}|A_{-\tau_{\mu_j}})$.

2. Get a random draw from $p(Y_{ij}|A)$.

Where A represents all the other parameters and data respectively updated at each step. Since this posterior distribution is the one of a normal, we censored the negative random draws to get back to the original format of the data, that is, non-negative earnings or income. Had we not done this censoring and just taken sample means over the non-censored random draws, those means would be negative in some cases, that is why in Table 6 some of the subgroup control posterior means are equal to zero. The same censoring applies when computing the CATE posterior means in Table 3 and Table 4 but not for the CATE densities shown in Figure 3.

The social planner is concerned with the optimal choice of subgroups to be assigned to or exempted from the treatment in order to maximize welfare. Using the definition of second order stochastic dominance we can get a natural welfare interpretation of the relationship between the posterior predictive distributions within each group.

Define $G_X(a)$ as the integral of the cumulative distribution function,

$$G_X(a) = \int^a F_X(t)dt$$

for a random variable X with density dF_X and a real number a. Using this notation, a random variable X *second order stochastically dominates* (SOSD) a random variable Y if

$$G_X(a) \leq G_Y(a) \tag{3}$$

[31]The complete results on average income are available upon request.

for all a in the support of the distributions, where G_i is defined as above.

A well-known result on stochastic dominance relates expected utility theory and SOSD. This result allows us to translate our problem into an expected utility inequality. A random variable X SOSD the random variable Y if and only if $E_X(u(X)) \geq E_Y(u(Y))$ for all increasing and concave functions u, where the operator $E_X(\cdot)$ represents the expected value with respect to the distribution of X[32].

Thus, if the social planner has a utility function that is increasing and concave she will always prefer X to Y if X SOSD Y[33]. From the predictive posterior distributions we can get the CDFs and compute the quantity $G_X(a)$ for the treated and for the controls for any given range of values for a. Specifically, we choose a to be positive and no larger than the maximum quarterly income observation in the data. This guarantees that we are ranking the two distributions for income values within the appropriate domain of this random variable. The ranking may not exist or change for very large values of a, but since we do not observe individuals getting such large values, the ranking is irrelevant at those points. By doing this we can examine whether the relation (3) holds, if that is the case we can determine whether the social planner prefers one outcome over the other, all this *without* having to specify any particular functional form for the utility function, other than requiring it to be increasing and concave. Furthermore, there is no need to run any kind of statistical test since we integrated out all the parameter uncertainty. There are cases however, in which the method is agnostic as to what ranking is welfare maximizing. For example, if $u(Y) = Y$ and $CATE(X = x) > 0$ then a risk-neutral social planner prefers treatment assignment even though there may not be SOSD. Thus for the following results to hold, we need to restrict our attention to strictly increasing and concave utility functions.

Figure 4 shows the CDFs by subgroup for average earnings and Figure 5 and Figure 6 show the graphs for differences between the functions $G_i(\cdot)$ for the treated ($i = 1$) and for the controls ($i = 0$) for average earnings and total income respectively. The CDFs also give a stochastic ranking according to the first order stochastic dominance. However, we could get

[32]For a proof of this result see Green et al. [1995].

[33]Another characterization of SOSD is that X SOSD Y is equivalent to a mean-preserving spread of the distribution of X in order to obtain the distribution of Y, see Green et al. [1995].

intersections of the two CDFs for the same subgroup and there is no natural link with utility theory when this happens[34]. Thus we look for the subgroups where the distribution for the treated SOSD the one from the controls. This would look as negative values for the functions in Figure 5 and Figure 6 and because of the characterization above this would mean that the expected utility of having that subgroup exposed to the treatment is greater than the same subgroup's utility when left as a control. This seems to occur in most cases for total income, although in most of them there is indifference at low income levels. These results suggest that subgroups 3, 10, and 23 would not have been picked to be exposed to the treatment by our social planner if the relevant domain for this random variable is [0, 3000] and only subgroups 3 and 10 if the relevant domain is [0, 5000] dollars per quarter. These subgroups correspond to subjects with no high school degree, not previously employed and 3 or more kids and previously employed with 1 kid or pregnant, respectively. The positive differences appear over the entire interval of observed income. In the cases where the inequality does not hold for the entire interval there is no dominance since the criterion applies when the inequality holds for the entire support of the distribution, which in our case is the range of observed total income. Some of these results contrast with what was obtained by just looking at the CATEs. In that case, only subgroup 3 did not benefit from the program. Subgroups 10 and 23, under the glasses of CATE, have positive mean impacts.

In the case of earnings, 5 out of the 24 subgroups exhibit SOSD from being exposed to the control over being exposed to the treatment. These subgroups have in common that the subjects were previously employed although at the beginning of the program they are unemployed. This might suggest that having employment experience makes an individual more likely to have greater earnings if exposed to the treatment in such a way that the expected utility value is larger in this case. The number of subgroups that exhibit SOSD is in contrast with the 9 out of 24 subgroups that have a negative sample CATE, although none of those is statistically significant. The results for earnings are also different from the ones

[34] X FOSD Y is equivalent to $E_X(u(X)) > E_Y(u(Y))$ for all increasing functions u regardless of their concavity or convexity. So if there are crossings in the CDFs nothing can be said about FOSD, but there could still be SOSD. In other words, FOSD implies SOSD, but SOSD does not necessarily imply FOSD. To summarize, if the CDFs do *not* intersect, for sure there is SOSD, but if they do intersect, there might or might not be SOSD.

when the outcome is total income because in the former we are only capturing employment enrolment, whereas in the latter both employment and transfers are confounded. Moreover, as it will be explained in the robustness checks section, the subgroups for which the program was not welfare enhancing if using the income outcome is a subset of the corresponding subgroups when the outcome is average earnings.

The heterogeneity across subgroups can be theoretically explained by looking at the different possible scenarios of a person facing either JF or AFDC in an income-leisure space[35]. A budget line represents the maximum attainable bundles of income and leisure. JF and AFDC however distort this budget line in different ways. AFDC modifies the budget line only at its right lower corner by shifting it upwards by the maximum amount of benefits attainable and by changing the slope of the original budget line up to a certain number of work hours. JF shifts the budget line upwards in a parallel way over a wider interval than AFDC because JF disregards all earnings below the FPL. Depending on where the person finds herself along the AFDC budget line and the shape of her indifference curves, exposure to JF can have different effects on income. A person at a very low income level, exposure to JF can either increase her income or let it unchanged. A person at a top income level would either remain with the same income or reduce her number of working hours, thus lowering her final income in order to maximize her utility.

That explanation works at two different levels. First, two different subgroups contain each homogenous individuals that behave one way or the other as described in the previous paragraph. Second, within a subgroup there is more than one type of individual and the explanation above works equally for each of them *within* the subgroup. This is an aspect we further discuss in the next section through the use of quantile treatment effects.

5.1 Within-QTEs

Another appealing feature of the method in this paper is that we can recover the quantile treatment effects (QTE) for each subgroup by taking the horizontal difference between the

[35]See Bitler et al. [2006]. Leisure is represented in the horizontal axis as work hours in decreasing order, i.e. 0 is to the right of the axis.

graphs of the two cumulative distribution functions for the outcome. Figure 7 shows the subgroup QTEs when the outcome is quarterly average earnings[36]. This is in some way an extension of the heterogeneity analysis of Jobs First in Bitler et al. [2006] except that here we disaggregated that heterogeneity by subgroups. The first thing to remark is that the increasing and then decreasing behavior of the QTEs is not characteristic to all the subgroups as it has been documented in that paper for the entire sample. Observe also how at the left of the distribution the QTEs are zero because at low income levels there are not any impacts on employment. In some subgroups the QTEs are rather a constant (subgroups 1, 7, and 23) and in a few cases with non-decreasing QTEs (subgroups 2, 13, and 15). The opt-in effect that Bitler et al. [2006] found at the upper quantiles and then confirmed by Kline and Tartari [2013] using partial identification arguments is present in almost all the subgroups. We remain agnostic as to whether the behavior in subgroups 2, 13, and 15 is driven by behavior not captured before by other methods or whether the QTEs at extreme quantiles are not reliable enough. For the case of total income as the outcome, subgroups 19 and 22 show almost constant QTEs and subgroups 8 and 12 non-decreasing QTEs.

These within-QTEs also explain why in some subgroups the CATEs are positive and yet the decision making method suggests the opposite. For example, in subgroups 3, 8, and 15 the QTEs are both positive and negative depending on the specific interval of income we are looking at. This means that within these subgroups there is a lot of heterogeneity in the impacts, heterogeneity that is *not* captured by the CATEs, but it is captured by the utility theory analysis through the stochastic dominance because it uses the *entire* distribution over the subgroup instead of using one single statistic.

5.2 Robustness Checks

A related paper is the study by Bitler et al. [2010] where they further extend their analysis from their previous work[37] to disentangle heterogeneity within- and across-subgroups of the population in a randomized control trial. One of their conclusions is that when only using

[36]There is no need to construct any confidence interval because these graphs were obtained by integrating out all the parameter uncertainty.

[37]Bitler et al. [2006].

covariates with information originated at the time of the program, it is not possible to distinguish heterogeneity of impacts across subgroups. However, when using covariates with information pre-program and contemporaneous covariates the opposite occurs. Our results are consistent with this in the sense that we find substantial heterogeneity across subgroups when using the two types of covariates but not so when we eliminate the use of pre-treatment variables. We evaluated this using the total income outcome and found that Jobs First would lead to higher expected utilities than not using the program over almost the entire population.

If we disregard the censoring problem of the data and only use the normal likelihood instead of the Tobit likelihood, the results are qualitatively identical as the ones shown in the tables for the case of total income. Not so for the case of average earnings where the censored part amounts to 27% of the total sample[38].

As mentioned before, there are other covariates that could be used to evaluate the performance of the program such as a race variable. This would however be the subject of potential conflicts due to basing a decision on the expansion or restriction of the program to a different sample on this type of covariate. We analyze Jobs First when the outcome is total income by substituting the number of kids variable with a race variable that splits the sample into Hispanics or other race, black, and white individuals[39]. In this case four subgroups would not have been picked to be exposed to the treatment by the social planner. This contrasts with the results from the sample CATEs in which only one of those four subgroups had negative impacts. QTEs are relatively constant in four subgroups and only one subgroup exhibits a negatively sloped QTE curve.

Since there are not hyperpriors on the parameters of the prior for the subgroup precisions τ_{μ_j}, we ran a number of different specifications to measure the sensitivity to the hyperparameters α and β. Following our observations in Section 4.1, we ran the entire model with each of the combinations of parameter values shown in Table 7 and Table 8 and the rest of parameter values as in the main specification. Greater values for α and β would dominate the data in the smallest subgroups. None of our results on SOSD changes for any of those

[38]We only present the results for the Tobit model as it handles the censoring problem, the results without the augmentation data step are available upon request.

[39]These results are available upon request.

alternative specifications except for subgroups 9 and 19 for the case of average earnings and only for subgroup 10 for the case of total income when the ratio α/β is relatively large. This might be the result of an exacerbated ratio in the variance for the conditional posterior distribution for τ_{μ_j} since the denominator, when $\epsilon \to 0$, becomes considerably smaller with respect to the numerator.

6 Conclusions

Heterogeneity of treatment effects is a relevant issue when extending or reducing the target population of a welfare program. There has been a literature exploring non-parametric methods to measure this heterogeneity and in a parallel way, a rather small literature has focused on mapping outcome distributions into expected utility theory criteria to make decisions. This paper combines those two literatures to develop a method that finds predictive outcome distributions conditional on covariates for treated and controls and maps those into second order stochastic dominance criteria. This gives a ranking of treatment versus control for each subgroup of the sample without having to specify the utility functional form other than minimal requirements. This can be used to assess whether to extend or restrict the program to a larger or smaller sample to maximize welfare.

We use a Bayesian approach since the size of the subsamples can be very small and classical approaches tend to not be reliable in these cases. Another issue is that the effect of the program in each subgroup is potentially correlated with other subgroups' effects since there might be covariates that make them similar and we did not capture that when dividing into the subsamples despite the random assignment to treatment. Our hierarchical Bayesian model allows for correlation across subgroups.

We find that one fifth of the number of subgroups in which we divided our sample does not maximize expected utility when exposed to the treatment, whereas almost twice as many show negative impacts using simple mean differences. This is because our method sheds light on the benefits from being assigned to the program not just from a point-estimation perspective, but by using the entire distribution of predictive outcomes in order to make a decision. We

can also calculate quantile treatment effects by subgroup showing that as documented in previous studies, there is almost zero impact at low income levels, then positive impacts around the median and then negative impacts for the upper quantiles. This within-variation in part explains the discrepancy between the non-parametric methods and our approach. Even though we have information on race, we do not believe it could successfully be used as a profiling covariate when extending or restricting a program to a different sample.

We let for future research the endogeneity of the number of subgroups and how to split the sample. A close approach to this, although not entirely solving the endogeneity problem, is to use a latent class model or mixture of normal densities to determine the probability of each individual belonging to each subgroup given an exogenous number of subgroups. Then we could compare across different models, each defined as a different number of subgroups, using Bayes factors.

Finally, learning about the differences in impacts across subsamples in a decision making framework can guide policymakers towards welfare improving modifications to the design of welfare programs.

Appendix

Gibbs sampler

To get the posterior distribution of one of the subgroup means, we condition on all the other variables and the parameters of the model, where μ_{-j} is the collection of the μ_k with $k \neq j$, then

$$\begin{aligned}
p(\mu_j|A) \propto & \left(\prod_{i=1}^{n_j} \exp\left(-\tfrac{1}{2}\tau_{\mu_j}(Y_{ij}-\mu_j)^2\right)\right) \times \exp\left(-\tfrac{1}{2}h_1(\mu_j-\mu)^2\right) \\
= & \exp\left(-\tfrac{1}{2}\tau_{\mu_j}\sum_{i=1}^{n_j}(Y_{ij}-\mu_j)^2 - \tfrac{1}{2}h_1(\mu_j-\mu)^2\right) \\
= & \exp\left(-\tfrac{1}{2}\tau_{\mu_j}\left(\sum_{i=1}^{n_j} y_{ij}^2 + n_j\mu_j^2 - 2\mu_j\sum_{i=1}^{n_j} y_{ij}\right) - \tfrac{1}{2}h_1(\mu_j^2+\mu^2-2\mu_j\mu)\right) \\
\propto & \exp\left(\left(-\tfrac{1}{2}\tau_{\mu_j}n_j - \tfrac{1}{2}h_1\right)\mu_j^2 - \tfrac{1}{2}\left(-2\tau_{\mu j}\sum_{i=1}^{n_j} Y_{ij} - 2h_1\mu\right)\mu_j\right) \\
\propto & \exp\left(-\tfrac{1}{2}(\tau_{\mu_j}n_j+h_1)\left(\mu_j - \tfrac{\tau_{\mu_j}\sum_{i=1}^{n_j} Y_{ij}+h_1\mu}{\tau_{\mu_j}n_j+h_1}\right)^2\right), \text{ for } j=1,\ldots,J
\end{aligned}$$

so that each mean subgroup has a normal posterior distribution,

$$\mu_j|A \sim N\left(\frac{\tau_{\mu_j}\sum_{i=1}^{n_j} Y_{ij} + h_1\mu}{\tau_{\mu j}n_j + h_1}, \tau_{\mu_j}n_j + h_1\right), \text{ for } j = 1, \ldots, J$$

or equivalently,

$$\mu_j|A \sim N\left(\frac{\sum_{i=1}^{n_j} Y_{ij} + h_1\mu/\tau_{\mu_j}}{n_j + h_1/\tau_{\mu_j}}, \tau_{\mu_j}n_j + h_1\right), \text{ for } j = 1, \ldots, J.$$

We can see that the posterior mean is a weighted average of the observations within the group and μ. The precision is a weighted sum of the precision of the income for subgroup j and the precision of μ_j.

Now, let's calculate the posterior distribution of μ. From the joint distribution, conditioning on the other variables we have that

$$\begin{aligned}
p(\mu|A) &\propto \left(\Pi_j \exp\left(-\frac{1}{2}h_1(\mu_j - \mu)^2\right)\right) \times \exp\left(-\frac{1}{2}h_2\mu^2\right) \\
&= \exp\left(-\frac{1}{2}h_1\left(\sum_{j=1}^{J}\mu_j^2 - 2\mu\sum_{j=1}^{J}\mu_j + J\mu^2\right) - \frac{1}{2}h_2\mu^2\right) \\
&\propto \exp\left(-\frac{1}{2}h_1\left(-2\mu\sum_{j=1}^{J}\mu_j + J\mu^2\right) - \frac{1}{2}h_2\mu^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\left(Jh_1 + h_2\right)\mu^2 - 2\mu h_1\sum_{j=1}^{J}\mu_j\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(Jh_1 + h_2\right)\left(\mu - \frac{h_1\sum_{j=1}^{J}\mu_j}{Jh_1 + h_2}\right)^2\right)
\end{aligned}$$

so the hyperparameters are

$$\mu|A \sim N\left(\frac{h_1\sum_{j=1}^{J}\mu_j}{Jh_1 + h_2}, Jh_1 + h_2\right),$$

or equivalently,

$$\mu|A \sim N\left(\frac{\sum_{j=1}^{J}\mu_j}{J + h_2/h_1}, Jh_1 + h_2\right),$$

the posterior mean is a weighted average of the subgroup means and the posterior precision is a weighted sum of the precisions of the subgroup means and the precision of μ. Also note that these hyperparameters do not depend explicitly on the value of the observations.

The posterior distribution of τ_{μ_j}.

$$\begin{aligned} p(\tau_{\mu_j}|A) &\propto \tau_{\mu_j}^{n_j/2} \exp\left(-\frac{1}{2}\tau_{\mu j}\sum_{i\in j}(Y_{ij}-\mu_j)^2\right)\tau_{\mu_j}^{\alpha-1}\exp(-\beta\tau_{\mu_j}) \\ &= \tau_{\mu_j}^{\alpha+n_j/2-1}\exp\left(-\left(\beta+\frac{1}{2}\sum_{i\in j}(Y_{ij}-\mu_j)^2\right)\tau_{\mu_j}\right), \text{ for } j=1,\ldots,J. \end{aligned}$$

This expression is the kernel of a gamma distribution, the hyperparameters are

$$\tau_\mu|A \sim G\left(\alpha+n_j/2, \beta+\frac{1}{2}\sum_{i\in j}(Y_{ij}-\mu_j)^2\right).$$

Observe that these hyperparameters do not depend on h_1 or h_2 in the conditional posterior distribution, but in the number of total observations and the overall sum of the squares of the deviations from the subgroup means.

Remember that we gave a prior on h_1, the precision of the group means μ_j,

$$p(h_1) \propto h_1^{\gamma-1}\exp(-\gamma h_1).$$

If A represents all the other parameters, the full conditional distribution for h_1 is

$$\begin{aligned} p(h_1|A) &\propto h_1^{J/2}\exp\left(-\frac{1}{2}h_1\sum_{j=1}^{J}(\mu_j-\mu)^2\right)\times h_1^{\gamma-1}\exp(-\gamma h_1) \\ &= h_1^{\gamma+J/2-1}\exp\left(-h_1\left(\gamma+\frac{1}{2}\sum_{j=1}^{J}(\mu_j-\mu)^2\right)\right) \end{aligned}$$

so $h_1|A \sim G(\gamma+J/2, \gamma+\frac{1}{2}\sum_{j=1}^{J}(\mu_j-\mu)^2)$. In the actual estimation, $\gamma = 0.1$. For h_2 we assume exactly the same prior as for h_1 but its posterior distribution is $h_2|A \sim G(\gamma+1/2, \gamma+\frac{1}{2}\mu^2)$ and $\gamma = 0.1$. Finally, ν was given the prior $\nu \sim N(0, h_3)$. Its posterior distribution is $\nu|A \sim N(\frac{\mu}{1+h_3/h_2}, h_3+h_2)$. Note that if h_2 did not have a posterior distribution then h_3/h_2 would be a constant. The initial parameters for the model are specified in Table 5.

Tobit model

In the case of a mass point at $Y = a$ the strategy to be adopted is a Tobit model (Dehejia [2005]). To simplify the notation we will write everything as if $a = 0$. With this we transform

the likelihood to be the one of a censored normal distribution. With this new likelihood there is a mass point on 0 and then a truncated normal distribution for the rest of the observations. This likelihood has the form

$$p(Y_{ij}|A) = \prod_{i:Y_{ij}=0} (1 - \Phi(\mu_j \tau_\mu^{1/2})) \times \prod_{i:Y_{ij}>0} \tau_\mu^{1/2} \exp\left(-\frac{1}{2}\tau_\mu (Y_{ij} - \mu_j)^2\right).$$

Where A represents all the parameters of the model and Φ is the normal cumulative distribution function. The priors are exactly the same as before. Following the idea of Chib [1992] for Bayesian estimation of a model with censored observations, we can transform the likelihood into a product of normals as before. To do so, let's assume that there exist (hypothetical) negative observations of Y_{ij} when we observe $Y_{ij} = 0$. We can do that by simulating negative observations from truncated normal distributions with parameters μ_j and τ_μ. Now the mass point over 0 has been substituted by a collection of hypothetical negative observations that contribute to the likelihood as normal densities, leading to the same expressions at the beginning of this section.

Assume that the collection of observations for Y is $\{\tilde{Y}_{ij}\}$. But we observe the censored sample[40]

$$Y_{ij} = \begin{cases} \tilde{Y}_{ij} \text{ if } \tilde{Y}_{ij} > 0, \\ 0 \text{ if } \tilde{Y}_{ij} \leq 0. \end{cases}$$

There is a known, set valued function C_{ij} so that

$$C_{ij} = \begin{cases} \tilde{Y}_{ij} \text{ if } \tilde{Y}_{ij} > 0, \\ (-\infty, 0] \text{ if } \tilde{Y}_{ij} \leq 0. \end{cases}$$

Then the joint distribution of observables and unobservables is

$$\begin{aligned} p(Y_{ij}, \mu, \{\mu_j\}_{j=1,\ldots,J}, \{\tau_{\mu_j}\}_{j=1,\ldots,J}|A) = \quad & p(Y_{ij}|\mu, \mu_j, \tau_{\mu_j}, A) p(C|Y_{ij}, A) \times \\ & \times p(\mu|A) p(\{\mu_j\}_{j=1,\ldots,J}|A) p(\{\tau_{\mu_j}\}_{j=1,\ldots,J}|A) \end{aligned}$$

where

$$p(C|Y_{ij}, A) = \prod_{i,j} p(C_{ij}|Y_{ij}, A) = \prod_{i,j} 1_{C_{ij}}(Y_{ij}).$$

[40] As in Geweke [2005].

So that the joint distribution is proportional to

$$\prod_{i,j} 1_{C_{ij}}(Y_{ij}) \times \left(\prod_{i,j} \tau_{\mu j}^{1/2} \exp\left(-\frac{1}{2} \tau_{\mu_j} (Y_{ij} - \mu_j)^2 \right) \right) \times$$
$$\left(\prod_{j=1}^{J} h_1^{1/2} \exp\left(-\frac{1}{2} h_1 (\mu_j - \mu)^2 \right) \right) \times$$
$$h_2^{1/2} \exp\left(-\frac{1}{2} h_2 \mu^2 \right) \times$$
$$\prod_{j=1}^{J} \tau_{\mu_j}^{\alpha - 1} \exp(-\beta \tau_{\mu_j}).$$

This expression is almost identical to the one presented at the beginning of this section. The censoring data is not a problem in the model because the posterior distributions will have the same kernels as before. But now the posterior kernels include those Y_{ij}'s that are negative too. Then the posterior distributions will be slightly different.

The Gibbs sampler for this likelihood is like the one from the previous section except that now there is a data augmentation step. The Gibbs sampler from the previous section is now nested in the data augmentation step. In the first iteration, negative values are drawn from a truncated normal distribution to replace the observations of zero income. Then the Gibbs sampler from the previous section is carried out exactly as in the normal model. In order to get predictive distributions we need to slightly modify the algorithm from the normal model:

1. Get a random draw from the posterior distributions $p(\mu_j|A)$ and $p(\tau_{\mu_j}|A)$.

2. Get a random draw from

$$p(Y_{ij}|A) = \prod_{i:Y_{ij}=0} (1 - \Phi(\mu_j \tau_{\mu_j}^{1/2})) \times \prod_{i:Y_{ij}>0} \tau_{\mu_j}^{1/2} \exp\left(-\frac{1}{2} \tau_{\mu j} (Y_{ij} - \mu_j)^2 \right).$$

The last step can be done for treated and controls and for each $j = 1, \ldots, J$ this way:

1. get a random number u from a uniform density on $[0, 1]$.

2. if $u < 1 - \Phi(\mu_j \tau_{\mu_j}^{1/2})$ then $Y_{ij} = 0$. Otherwise, get a random draw from a truncated normal on $[0, \infty)$ with the appropriate parameters.

Tables

Table 1: Subgroups definitions

group	age > 30	Prev. employed	0 or 1 kids	2 kids	3+ kids	high school degree
1			X			
2				X		
3					X	
4	X		X			
5	X			X		
6	X				X	
7		X	X			
8		X		X		
9		X			X	
10	X	X	X			
11	X	X		X		
12	X	X			X	
13			X			X
14				X		X
15					X	X
16	X		X			X
17	X			X		X
18	X				X	X
19		X	X			X
20		X		X		X
21		X			X	X
22	X	X	X			X
23	X	X		X		X
24	X	X			X	X

Note: Subgroups are exogenously defined by covariates used in similar studies.

Table 2: Summary statistics

	$T = 0$		$T = 1$	
	mean	std. dev.	mean	std. dev.
age > 30	0.491	0.500	0.495	0.500
prev. employed	0.598	0.491	0.564	0.496
0 - 1 kids	0.505	0.500	0.497	0.500
2 kids	0.278	0.448	0.266	0.442
3+ kids	0.217	0.412	0.236	0.425
high school diploma	0.498	0.500	0.462	0.499
female	0.960	0.195	0.968	0.177
black	0.383	0.486	0.381	0.486
white	0.364	0.481	0.377	0.485
hispanic or other	0.234	0.424	0.223	0.417
avg total income per quarter ($)	2,449	1,390	2,758	1,572
avg earnings per quarter ($)	1,120	1,583	1,173	1,511

Note: Summary statistics for the entire sample by treated and controls. Each of the covariates is coded as a dummy variable equal to 1 if the described characteristic is true for the individual, except for the last two rows which are averages over the first 7 quarters since inception to the program.

Table 3: Sample and posterior means. Average earnings.

group	n	T = 1 sample mean		n	T = 0 sample mean		sample CATE		t stat.	posterior mean CATE	
1	132	441	(865)	95	414	(936)	27	(120)	0.2	16	(52)
2	87	520	(846)	77	333	(624)	186	(117)	1.6	139	(134)
3	70	457	(626)	52	556	(1,190)	-99	(166)	-0.6	254	(141)
4	106	460	(855)	113	301	(662)	159	(103)	1.5	33	(64)
5	94	826	(1,225)	82	475	(929)	351	(166)	2.1	317	(190)
6	110	687	(1,169)	102	382	(1,196)	305	(163)	1.9	113	(133)
7	223	1,069	(1,261)	228	1,176	(1,361)	-107	(124)	-0.9	-92	(139)
8	67	1,325	(1,229)	79	1,440	(1,768)	-115	(257)	-0.4	34	(271)
9	50	1,145	(1,393)	44	1,251	(1,604)	-106	(309)	-0.3	-84	(360)
10	108	1,734	(2,429)	106	1,940	(2,394)	-206	(330)	-0.6	-225	(386)
11	92	1,776	(1,931)	82	1,338	(1,644)	438	(274)	1.6	561	(305)
12	76	1,586	(1,652)	73	1,438	(2,119)	147	(311)	0.5	596	(379)
13	94	837	(1,182)	73	533	(856)	303	(164)	1.8	311	(220)
14	35	828	(1,255)	50	602	(1,091)	226	(256)	0.9	314	(326)
15	24	457	(1,095)	26	511	(796)	-54	(269)	-0.2	-134	(243)
16	81	883	(1,227)	103	422	(971)	462	(162)	2.9	387	(211)
17	60	688	(1,202)	66	551	(1,078)	137	(203)	0.7	80	(157)
18	93	905	(1,282)	68	578	(1,267)	327	(204)	1.6	522	(192)
19	249	1,569	(1,486)	307	1,689	(1,629)	-120	(134)	-0.9	-93	(144)
20	77	1,979	(1,763)	77	1,828	(1,551)	150	(268)	0.6	118	(275)
21	33	1,535	(1,270)	41	1,551	(1,586)	-17	(340)	0.0	220	(397)
22	131	1,821	(1,636)	113	1,575	(1,696)	246	(214)	1.2	356	(249)
23	90	1,632	(1,791)	115	1,780	(1,812)	-148	(254)	-0.6	-171	(281)
24	78	2,088	(1,751)	83	1,874	(1,944)	214	(292)	0.7	187	(327)

Note: Sample standard errors in parentheses. Total sample size for $T = 1$ is 2,227 and for $T = 0$ is 2,236. Contrasts between the sample CATE estimates and the CATEs were obtained by taking the difference $\mu_j^{treated} - \mu_j^{control}$ using the Markov chains from the Gibbs sampler. Tobit likelihood specification. Sample standard errors in parentheses. See main text for details.

Table 4: Sample and posterior means. Average total income.

group	n	T = 1		n	T = 0		sample		t stat.	posterior mean	
		sample mean			sample mean		CATE			CATE	
1	132	1,852	(1,072)	95	1,658	(980)	194	(142)	1.4	183	(137)
2	87	2,324	(1,103)	77	2,086	(857)	238	(157)	1.5	243	(155)
3	70	2,832	(1,322)	52	2,906	(1,120)	–74	(229)	–0.3	–43	(218)
4	106	1,725	(1,054)	113	1,526	(900)	199	(134)	1.5	203	(130)
5	94	2,596	(1,235)	82	2,035	(1,019)	561	(175)	3.2	557	(170)
6	110	2,948	(1,493)	102	2,561	(1,270)	387	(193)	2.0	385	(185)
7	223	2,466	(1,212)	228	2,332	(1,127)	134	(110)	1.2	143	(109)
8	67	3,225	(1,170)	79	2,871	(1,486)	354	(225)	1.6	373	(210)
9	50	3,609	(1,360)	44	3,229	(1,193)	380	(268)	1.4	396	(265)
10	108	2,837	(2,254)	106	2,819	(2,119)	18	(300)	0.1	67	(288)
11	92	3,204	(1,808)	82	2,707	(1,373)	497	(246)	2.0	490	(240)
12	76	3,846	(1,928)	73	3,064	(2,002)	782	(323)	2.4	769	(291)
13	94	2,216	(1,240)	73	1,883	(854)	333	(175)	1.9	338	(169)
14	35	2,528	(1,317)	50	2,149	(1,065)	379	(260)	1.5	399	(256)
15	24	2,722	(1,143)	26	2,622	(1,016)	100	(308)	0.3	135	(299)
16	81	2,084	(1,261)	103	1,670	(956)	414	(163)	2.5	429	(170)
17	60	2,386	(1,236)	66	2,051	(971)	336	(198)	1.7	344	(206)
18	93	3,291	(1,469)	68	2,763	(1,205)	528	(218)	2.4	527	(209)
19	249	2,784	(1,371)	307	2,664	(1,326)	120	(115)	1.0	126	(111)
20	77	3,611	(1,459)	77	2,987	(1,026)	624	(203)	3.1	599	(195)
21	33	4,016	(1,354)	41	3,257	(1,127)	759	(288)	2.6	677	(285)
22	131	2,983	(1,477)	113	2,373	(1,488)	610	(191)	3.2	584	(192)
23	90	3,034	(1,901)	115	2,941	(1,360)	92	(228)	0.4	98	(222)
24	78	3,988	(1,762)	83	3,240	(1,566)	748	(262)	2.9	719	(260)

Note: Sample standard errors in parentheses. Total sample size for $T = 1$ is 2,227 and for $T = 0$ is 2,236. Contrasts between the sample CATE estimates and the CATEs were obtained by taking the difference $\mu_j^{treated} - \mu_j^{control}$ using the Markov chains from the Gibbs sampler. Tobit likelihood specification. Sample standard errors in parentheses. See main text for details.

Table 5: Initial parameter values

Parameter	Value
α	0.1
β	0.1
h_3	$1E-12$
γ	0.1

Table 6: Posterior means. Average earnings.

subgroup	μ_T	s.d.	τ_T	s.d.	μ_U	s.d.	τ_U	s.d.
1	19	49	5.09E-07	9.44E-08	3	20	3.75E-07	8.53E-08
2	163	125	6.32E-07	1.32E-07	24	51	9.71E-07	2.35E-07
3	294	104	1.49E-06	3.11E-07	41	94	2.74E-07	7.84E-08
4	33	64	5.17E-07	1.13E-07	0	0	4.93E-07	1.20E-07
5	323	187	3.05E-07	6.25E-08	5	28	3.51E-07	9.07E-08
6	113	133	3.00E-07	5.86E-08	0	0	1.85E-07	4.55E-08
7	969	96	5.17E-07	5.27E-08	1062	102	4.44E-07	4.44E-08
8	1282	152	6.30E-07	1.10E-07	1248	225	2.66E-07	4.51E-08
9	975	232	3.94E-07	8.52E-08	1060	263	3.22E-07	7.53E-08
10	1277	261	1.23E-07	1.88E-08	1502	285	1.27E-07	2.05E-08
11	1582	206	2.27E-07	3.67E-08	1021	230	2.49E-07	4.78E-08
12	1471	211	3.30E-07	5.71E-08	875	315	1.42E-07	2.87E-08
13	489	168	3.99E-07	7.02E-08	178	139	6.49E-07	1.47E-07
14	438	273	3.36E-07	1.06E-07	125	163	3.63E-07	1.05E-07
15	58	139	2.51E-07	1.21E-07	192	201	7.10E-07	2.69E-07
16	388	211	3.00E-07	6.78E-08	0	6	2.83E-07	7.16E-08
17	95	148	2.53E-07	7.09E-08	15	50	2.72E-07	7.70E-08
18	524	191	3.43E-07	6.33E-08	2	17	1.64E-07	4.69E-08
19	1476	101	3.96E-07	3.87E-08	1569	104	3.12E-07	2.71E-08
20	1851	200	3.02E-07	5.10E-08	1732	184	3.75E-07	6.32E-08
21	1420	232	5.55E-07	1.39E-07	1200	314	2.64E-07	6.79E-08
22	1667	161	3.04E-07	4.32E-08	1311	197	2.52E-07	3.92E-08
23	1444	214	2.52E-07	4.21E-08	1615	182	2.56E-07	3.59E-08
24	1833	228	2.51E-07	4.50E-08	1645	228	2.17E-07	3.67E-08

	$T=1$	μ	s.d.	h_1	s.d.	h_2	s.d.
		806	168	1.92E-06	6.51E-07	2.16E-06	4.04E-06

	$T=0$	μ	s.d.	h_1	s.d.	h_2	s.d.
		428	239	1.14E-06	3.70E-07	1.55E-02	2.55E-01

Note: Columns 2-5 show the posterior means for the parameters indicated for the controls by subgroup. Columns 6-9 show the posterior means for parameters from the treated sample. At bottom of the table the rest of the posterior mean parameters for controls and treated. Re-censoring was applied to eliminate negative mean draws after using the Tobit model.

Table 7: Initial parameter values. Robustness checks. Average earnings.

α	β	Subgroups*
0.1	0.1	7, 9, 10, 19, 23
0.2	0.2	7, 10, 19, 23
1	0.2	7, 9, 10, 23
2	2	7, 10, 23
5	2	7, 9, 10, 15, 23
10	2	7, 9, 10, 15, 23

Note: *These subgroups exhibit SOSD from being exposed to the control over being exposed to the treatment. The subgroups listed are those for which the differences between the G functions are strictly positive over the interval $[0, 5000]$. Subgroups 3 and 9 would otherwise be in the table above for every specification if we allowed for very small negative values at the very low quarterly average earnings.

Table 8: Initial parameter values. Robustness checks. Average total income.

α	β	Subgroups*
0.1	0.1	3, 10, 23
0.2	0.2	3, 7, 10, 23
1	0.2	3, 23
2	2	3, 23
5	2	3, 7, 10, 23
10	2	3, 23

Note: *These subgroups exhibit SOSD from being exposed to the control over being exposed to the treatment. The subgroups listed are those for which the differences between the G functions are strictly positive over the interval $[0, 5000]$.

Table 9: Posterior means. Average total income.

subgroup	μ_T	s.d.	τ_T	s.d.	μ_U	s.d.	τ_U	s.d.
1	1,874	95	8.69E-07	1.10E-07	1,691	100	1.04E-06	1.53E-07
2	2,345	118	8.21E-07	1.27E-07	2,102	96	1.36E-06	2.08E-07
3	2,833	153	5.73E-07	9.63E-08	2,876	150	7.92E-07	1.64E-07
4	1,759	98	9.05E-07	1.29E-07	1,556	87	1.23E-06	1.69E-07
5	2,609	126	6.57E-07	9.82E-08	2,053	114	9.68E-07	1.50E-07
6	2,942	143	4.49E-07	5.84E-08	2,557	118	6.22E-07	8.99E-08
7	2,474	80	6.82E-07	6.51E-08	2,331	75	7.90E-07	7.29E-08
8	3,202	138	7.35E-07	1.23E-07	2,830	165	4.57E-07	7.31E-08
9	3,540	190	5.39E-07	1.16E-07	3,144	180	7.02E-07	1.50E-07
10	2,841	211	1.97E-07	2.73E-08	2,774	191	2.24E-07	3.14E-08
11	3,176	181	3.10E-07	4.53E-08	2,686	150	5.33E-07	8.35E-08
12	3,733	218	2.70E-07	4.58E-08	2,955	209	2.51E-07	4.28E-08
13	2,250	128	6.47E-07	9.76E-08	1,911	106	1.37E-06	2.38E-07
14	2,575	220	5.81E-07	1.38E-07	2,176	148	8.89E-07	1.80E-07
15	2,740	238	7.73E-07	2.24E-07	2,607	203	9.70E-07	2.71E-07
16	2,123	144	6.31E-07	9.84E-08	1,694	92	1.09E-06	1.59E-07
17	2,422	161	6.54E-07	1.20E-07	2,078	122	1.06E-06	1.84E-07
18	3,265	154	4.60E-07	6.73E-08	2,738	143	6.88E-07	1.19E-07
19	2,788	85	5.33E-07	5.04E-08	2,662	75	5.70E-07	4.66E-08
20	3,560	161	4.71E-07	7.58E-08	2,961	113	9.48E-07	1.51E-07
21	3,855	230	5.41E-07	1.38E-07	3,172	173	7.86E-07	1.74E-07
22	2,975	132	4.59E-07	5.41E-08	2,391	133	4.53E-07	6.01E-08
23	3,015	182	2.77E-07	4.04E-08	2,918	125	5.43E-07	7.38E-08
24	3,885	206	3.22E-07	5.18E-08	3,163	164	4.04E-07	6.50E-08

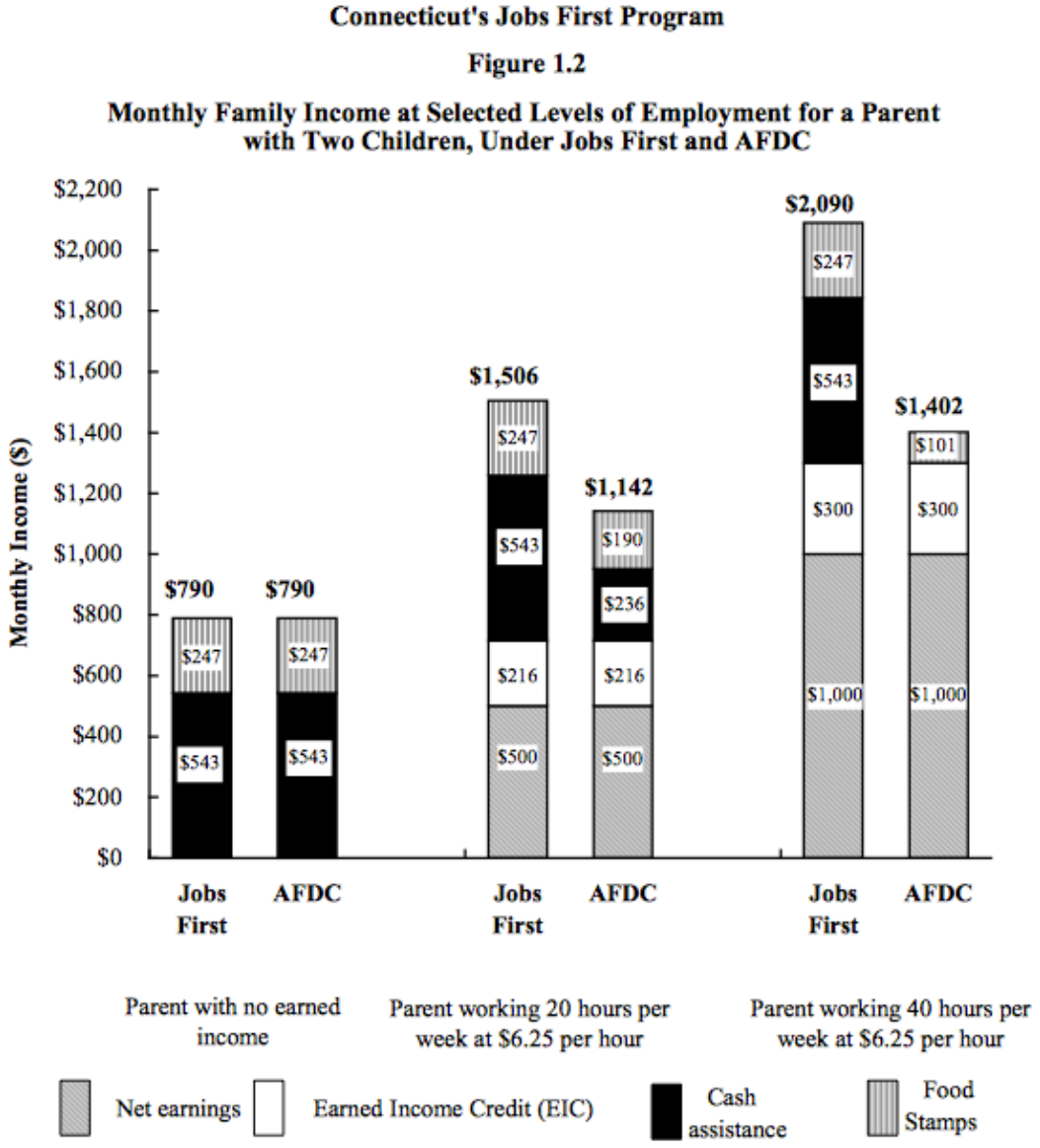
	$T=1$	μ	s.d.	h_1	s.d.	h_2	s.d.
		2,870	134	2.71E-06	8.71E-07	1.52E-07	1.93E-07

	$T=0$	μ	s.d.	h_1	s.d.	h_2	s.d.
		2,500	115	3.92E-06	1.19E-06	2.06E-07	2.70E-07

Note: Columns 2-5 show the posterior means for the parameters indicated for the controls by subgroup. Columns 6-9 show the posterior means for parameters from the treated sample. At bottom of the table the rest of the posterior mean parameters for controls and treated. Re-censoring was applied to eliminate negative mean draws after using the Tobit model.

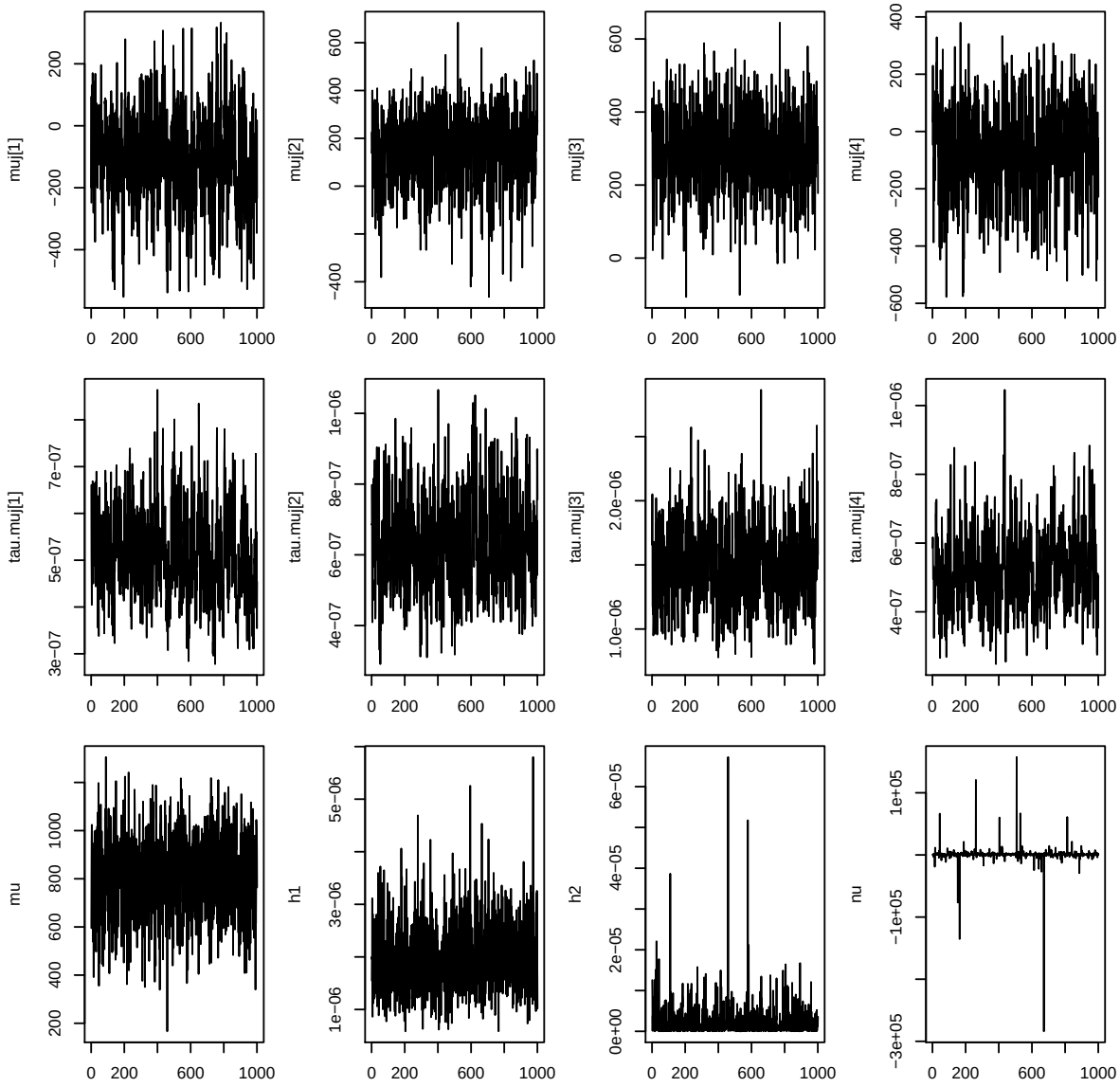
Figures

Figure 1: Connecticut's Jobs First



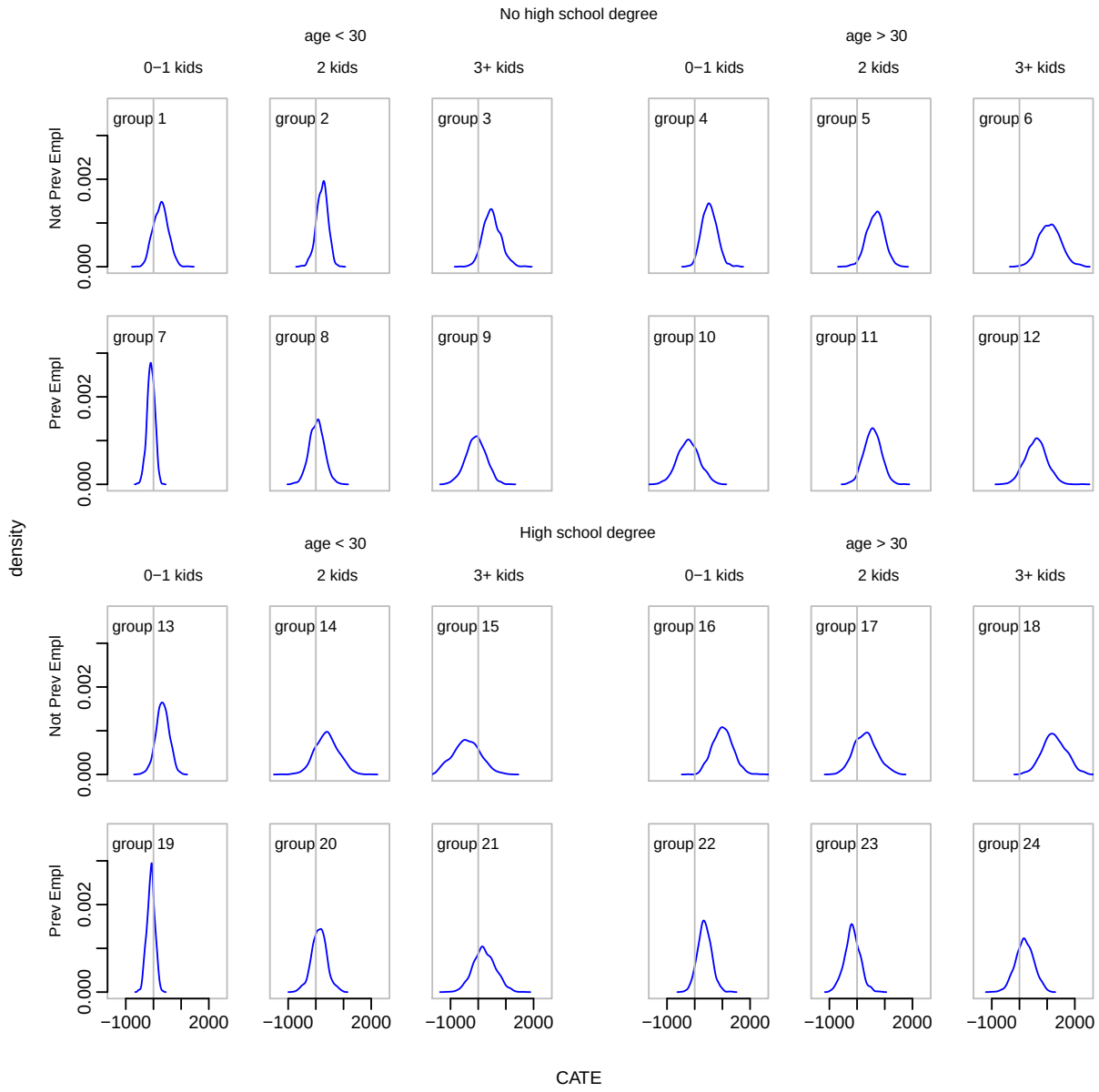
Note: Source: MDRC's Final Report.

Figure 2: Times series of draws from the Gibbs sampler. Average earnings.



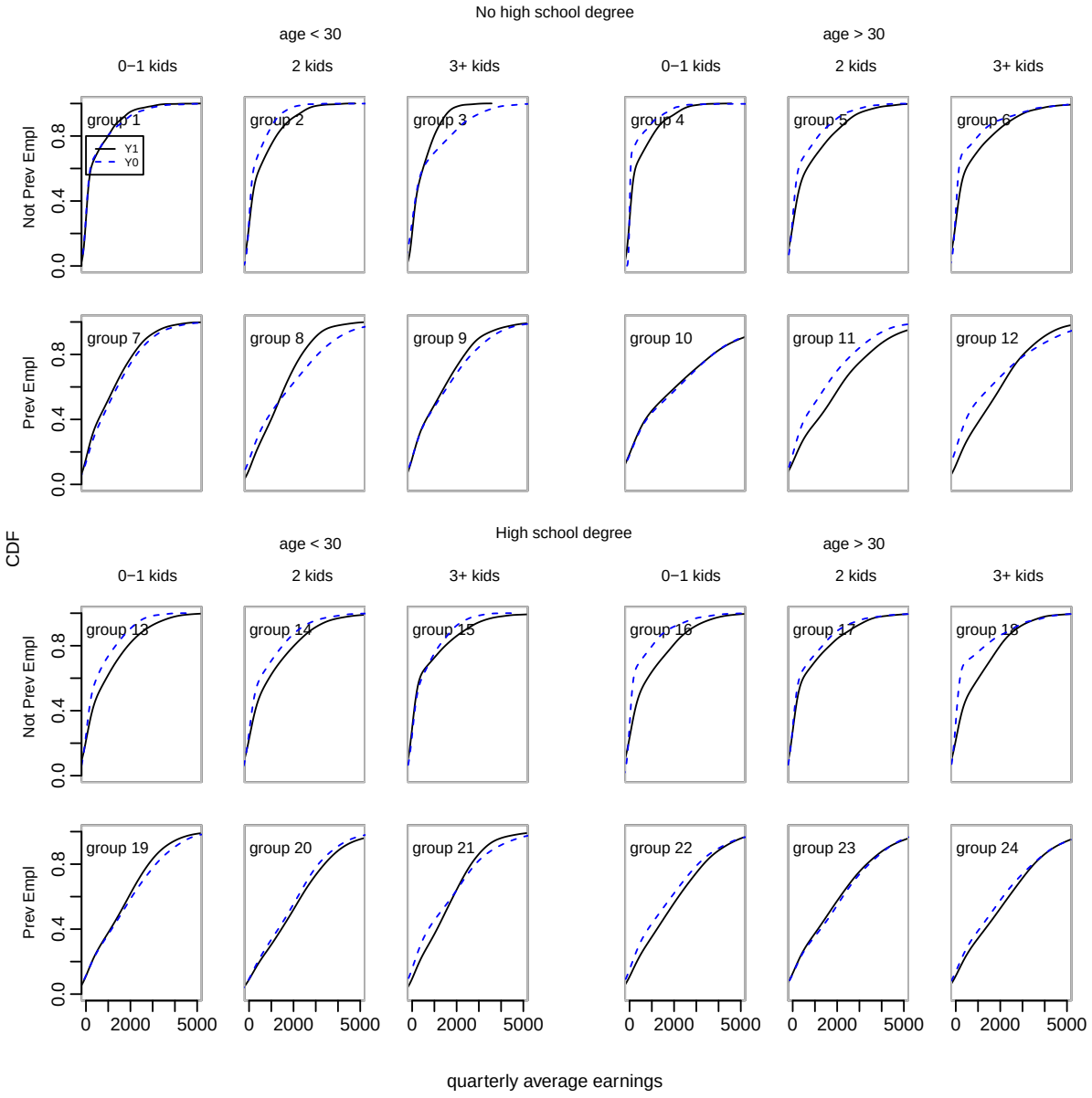
Note: Each graph shows the time series of random draws that form the posterior distributions. These are only the *last* 1,000 draws from the Gibbs sampler for selected parameters from the treated subsample.

Figure 3: Kernel estimates for the CATEs posterior densities by subgroup. Average earnings.



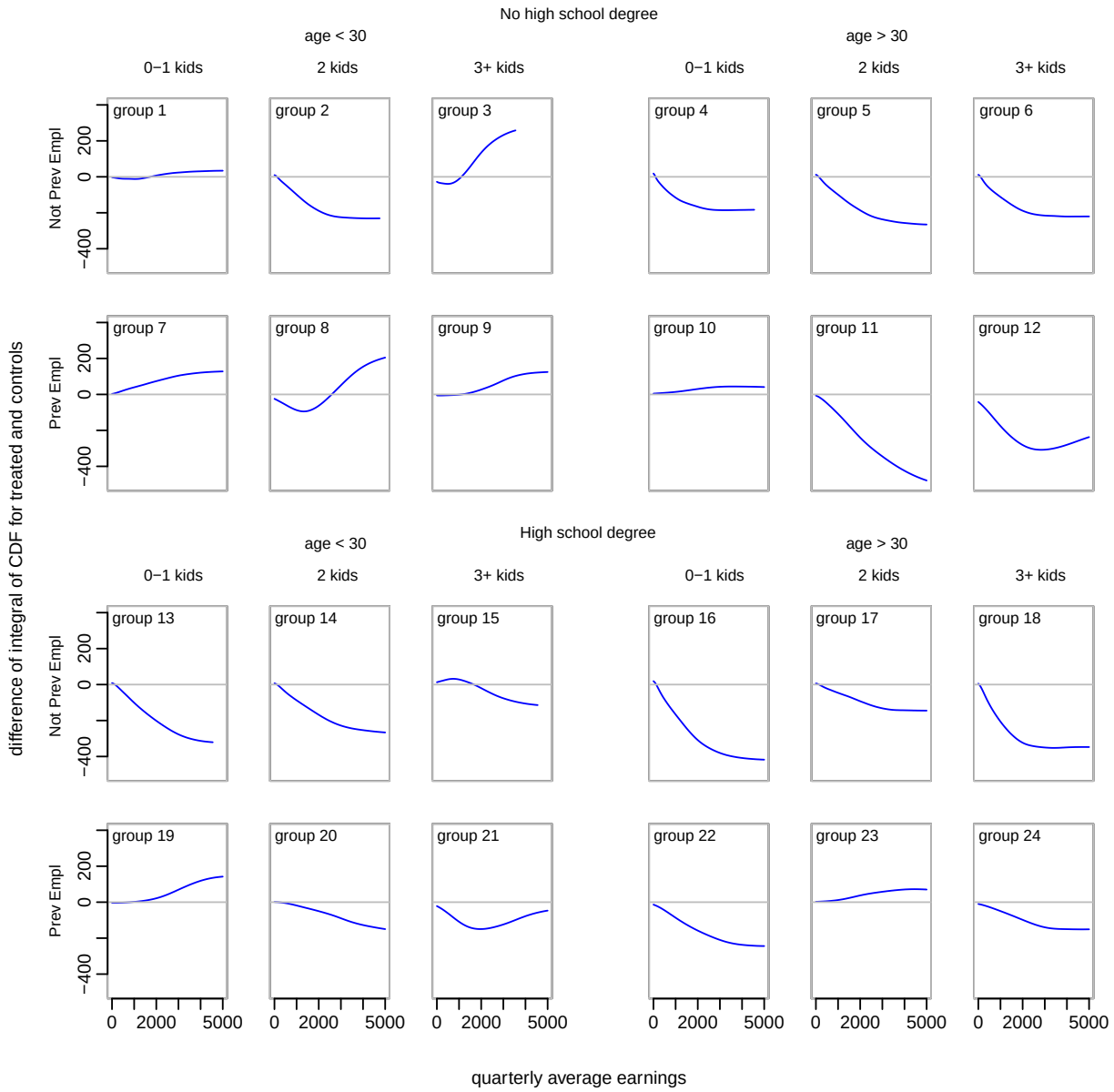
Note: Each graph represents the kernel density estimate for the distribution of the CATE for that subgroup obtained by taking the difference $\mu_j^{treated} - \mu_j^{control}$ using the draws from the Gibbs sampler. See main text for details.

Figure 4: Cumulative distribution functions of predictive income. Average earnings.



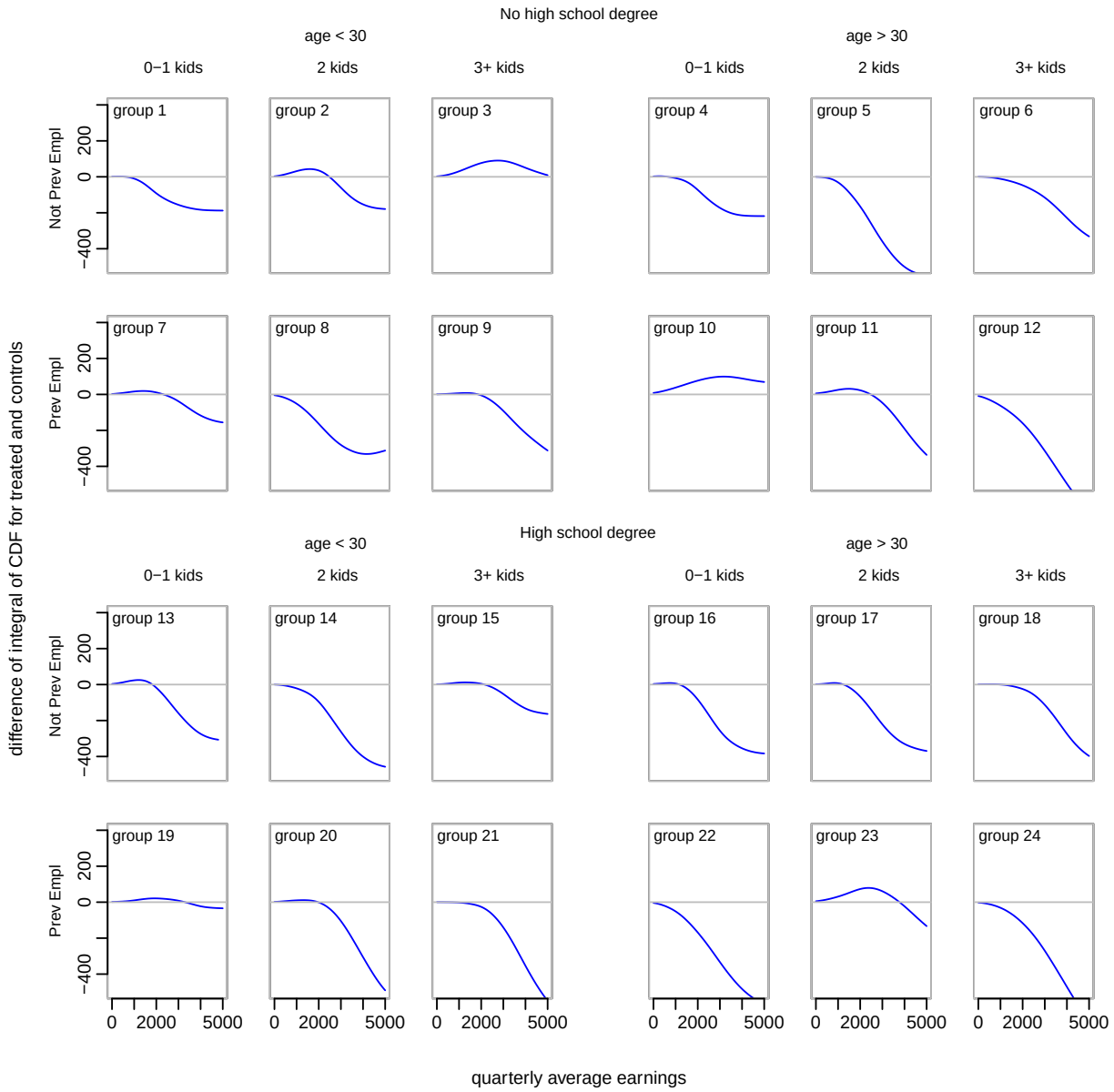
Note: Each graph represents the cumulative distribution functions of income for treated and controls over an interval of quarterly average income that covers the data. First order stochastic dominance of being treated is evident in most subgroups. The horizontal difference of the two lines for a given subgroup and for a given number on the vertical axis is the QTE.

Figure 5: Second order stochastic dominance by subgroup. Average earnings.



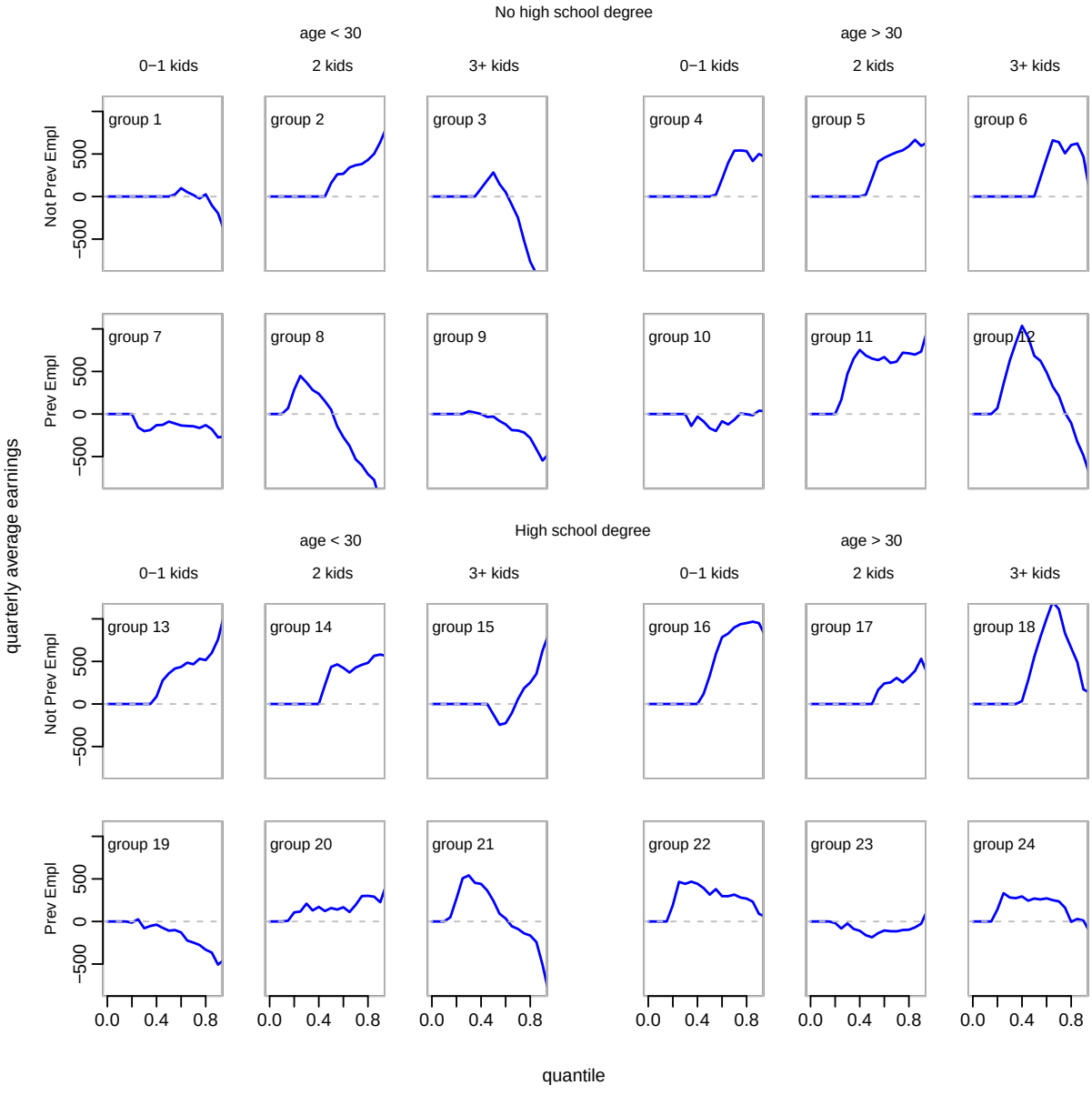
Note: Each graph represents the difference (treatment minus controls) between the graphs of $G_X(a) = \int^a F_X(t)dt$ where F_X is the predictive distribution's CDF for earnings. If the resulting difference is negative then there is SOSD of being treated with respect to being assigned to the control group.

Figure 6: Second order stochastic dominance by subgroup. Average total income.



Note: Each graph represents the difference (treatment minus controls) between the graphs of $G_X(a) = \int^a F_X(t)dt$ where F_X is the predictive distribution's CDF for income. If the resulting difference is negative then there is SOSD of being treated with respect to being assigned to the control group.

Figure 7: Quantile treatment effects by subgroup. Average earnings.



Note: Continuous lines represent the QTEs over the earnings distribution.

References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2012). Estimating conditional average treatment effects. *Working paper*.

Adams-Ciardullo, D., Bloom, D., Hendra, R., Michalopoulos, C., Morris, P., Serivener, S., and Walter, J. (2002). *Jobs First: Final report on Connecticut's welfare reform initiative.* Manpower Demonstration Research Corporation.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167:168–196.

Bitler, M. P., Gelbach, J., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Bitler, M. P., Gelbach, J., and Hoynes, H. W. (2010). Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment. *Working paper*.

Chamberlain, G. (2011). *Bayesian Aspects of Treatment Choice*. The Oxford Handbook of Bayesian Econometrics. New York: Oxford University Press.

Chib, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics, 51, 79-99*, 51:79–99.

Chib, S., Greenberg, E., and Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, 18:321–348.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125:141–173.

Dehejia, R. H. (2008). *When is ATE Enough? Rules of Thumb and Decision Analysis in Evaluating Training Programs*. Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics. New York: Elsevier-Science.

Efron, B. (2011). Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, 104:1015–1028.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, chapter 4. Wiley-Interscience.

Graham, B. and Hirano, K. (2011). Robustness to parametric assumptions in missing data models. *American Economic Review: Papers & Proceedings*, 101:3:538–543.

Green, J., Mas-Colell, A., and Whinston, M. (1995). *Microeconomic Theory*, chapter 6.D. Oxford University Press.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *NBER Working paper 16110*.

Hirano, K. and Porter, J. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Hu, X. (2011). Modeling endogenous treatment effects with heterogeneity: a bayesian nonparametric approach. `http://scholarcommons.usf.edu/etd/3159`. Dissertation.

Hu, X., Munkin, M., and Trivedi, P. (2011). Estimating incentive and selection effects in Medigap insurance. *Working paper*.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86:4–29.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Kasy, M. (2013). Using data to inform policy. *Working paper*.

Kline, P. and Tartari, M. (2013). What distributional impacts mean: Welfare reform experiments and competing margins of adjustment. *Working paper*.

Li, M. and Tobias, J. (2011). Bayesian inference in a correlated random coefficients model: Modeling causal effect heterogeneity with an application to heterogenous returns to schooling. *Journal of Econometrics*, 162:345–361.

Manski, C. (2001). Designing programs for heterogeneous populations: The value of covariate information. *American Economic Review: Papers & Proceedings*, 91:103–106.

Manski, C. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72:1221–1246.

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

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.