(MACHINE) LEARNING WHAT GOVERNMENTS VALUE

DANIEL BJÖRKEGREN, JOSHUA BLUMENSTOCK, AND SAMSUN KNIGHT

Social transfer programs are one of the primary tools used to assist poor families in developing countries. These programs rely critically on targeting criteria that are used to determine eligibility.

This paper develops a method to better understand and align the objectives of a program with the targeting criteria used to implement that program. Our method leverages recent advances in machine learning and causal inference that make it possible to predict the *marginal effects* of a program on a recipient. We show how a given population allocation can be decomposed into three distinct dimensions: (i) differential marginal effects, (ii) traditional welfare weights, which assign higher priority to specific subsets of the population; and (iii) multiple human development objectives, which a policymaker may wish to balance. Taken together, this decomposition makes it possible to infer the policy's implied preferences over households and outcomes. We apply this approach to Mexico's PROGRESA program, one of the world's largest anti-poverty subsidies, to elucidate existing targeting priorities. We find that existing allocations are consistent with the government valuing an additional day of school attendance at 108.07 pesos (of foregone household consumption), and valuing a reduction in illness at 101.98 pesos per day. Allocations imply welfare weights that place 28.3% more value on the median household if indigenous, 7.5% more value for each additional elder household member, and less value on educated and richer households. Alternate eligibility criteria could have marginally improved average health and schooling outcomes at a small cost to average consumption outcomes, or vice versa.

KEYWORDS: targeting, welfare, heterogeneous treatment effects, PROGRESA

Revision January 8, 2020. Preliminary and incomplete.

1. Summary

Targeting is critical to social transfer programs. Governments typically target benefits to households that are observably disadvantaged in some way, such as being poor or disabled. But the rationale behind such criteria is often unclear. Do governments prioritize these households because they are expected to benefit more? Or because these households are intrinsically more valued? The distinction has deep implications for understanding and designing optimal policies [Nichols and Zeckhauser, 1982, Coate and Morris, 1995]. In particular, all members of society may agree to prioritize the households that will benefit most from the program, but may disagree on how much welfare weight to apply to each type of household.

This paper provides a method to infer policymaker preferences from observed transfer allocations. This method makes it possible to separately estimate implied welfare weights (i.e., which households the government prioritizes) and weights on different outcomes (how the government weighs different impacts), from differential marginal effects (which households benefit the most from the transfer). Ex post, we can pose counterfactual welfare weights and valuations of outcomes to produce different allocations, and quantify the welfare impacts of these adjustments. Our approach relies on recent innovations in machine learning that make it possible to estimate the heterogenous treatment effects under random assignment KÃŒnzel et al. [2019], Wager and Athey [2018]. We demonstrate how these advances can be used to better understand and articulate the allocation of social programs.

We consider programs where eligibility is determined based on a score, which encodes the policymaker's ranking between any two households. This ranking implies a system of inequalities, which we use to estimate the value that the government places on different welfare outcomes (estimated using modern methods for heterogenous effects) and different households (based on observed characteristics) by using ordinal support vector machine regression, or "preference-learning" Herbrich et al. [1999], Chu and Keerthi [2007]. We develop a utility model to formalize this approach in section 2 below.

Intuitively, if a government prefers to allocate to a household that would see a low impact but which has some status (e.g., a disability), over a household for which it would have a larger impact on outcomes, that suggests the government places higher welfare weight on households with that status. Or, if a government prefers to allocate a benefit to a household for which it would have a health impact over a household for which it would have a high consumption impact, that implies that it highly values health. Our method can also be used if an observer had only binary information on eligibility, though in that setting it will be less informative.

We show how this approach can be applied to data using the example of PROGRESA, one of the world's largest (and best-studied) anti-poverty programs.[1] We begin by estimating the heterogenous treatment effects of the program. Consistent with prior work using linear models, we find considerable heterogeneity [Djebbari and Smith, 2008]. The joint and marginal distributions of estimated treatment effects over households is shown in Figure **??**. We then use preference-learning techniques based on implicit ranking inequalities to disentangle welfare weights from planner preferences, and find that PROGRESA tends to downweight impacts among highly educated households. Figure **??** shows the estimated heterogeneous treatment effects across three outcomes: children's days of missed school, children's days of sickness, and per-person household consumption.[2]

We further find that the Mexican government's initial allocation rule implies a value of 108.07 pesos for each day of child school attendance and 101.98 pesos for each child sick day. We find that the government would have placed 28.3% more value on the median household if indigenous, 7.5% more value for each additional elder household member, and less value on educated and richer households. The government later changed their allocation rule; estimation using this new rule places more welfare weight on richer households, as well as higher priority on indigenous status of households (35.8% higher value on the median household), and similar priority for household size (12.3% higher value for each additional household member).

Finally, we evaluate the counterfactual allocations that would result from alternate welfare weights. If the government valued only consumption and schooling impacts, the rule would increase the priority of low income and low education households. But if the government valued only schooling, it would have instead decreased the priority of these two groups. If the government cared only about lower-income households, the rule would de-prioritize households with more children. We also assess a technocratic ranking that weighs impacts according to external cost benefit estimates; this would have resulted in very marginally lower average consumption in 1999 and more significantly lower levels of average missed school days and sick days. Finally, we assess the impact of these alternative allocations on consumption, sick days, and missed days of school, and predict that other alternative allocations would have resulted in higher average consumption with no cost in terms of more average sick days or more average missed days of school.

[1]PROGRESA is attractive because it is very well documented in prior work. However, its use in this framework requires several simplifications. In particular, PROGRESA conditioned payouts on certain actions, but we treat the program as an unconditional transfer. We also rule out the possibility that the policy may have differential spillover benefits on different households.
Either could be accommodated in a richer model.

[2]As PROGRESA surveys measured "sick days" for children between 0 and 5 years of age, and "missed days of school" for children between 6 and 16 years of age, each of these outcomes are measured in our data in terms of sicks days per young child in the house and missed school days per school age child in the house, respectively. As consumption was measured for all household members, this outcome is measured in terms of average per-person consumption among household members.

This approach makes it possible to invert the discussion about government programs. Rather than debate the means of the policy (who is eligible, how large are the benefits), this framework makes it possible to debate the ends (how much do we value health, education, or consumption? By how much should we prioritize poor families over middle class families?). The framework naturally applies to the debate about universal basic income versus targeting transfers towards particular households [Hanna and Olken, 2018b].

This paper contributes to literature on optimal targeting and taxation Nichols and Zeckhauser [1982], Barr [2012], Fleurbaey and Maniquet [2018], and especially work focused on targeting in developing countries [Alatas et al., 2012, Hanna and Olken, 2018a]. It builds on prior work that infers policymaker preferences from their actions [Timmins, 2003], extending this approach by capitalizing on recent innovations in machine learning of predicted individual treatment effects KÃŒnzel et al. [2019], Wager and Athey [2018]. Our empirical results also engage with research on the effects [Behrman and Todd, 1999, Gertler, 2004, John Hoddinott, 2004, Djebbari and Smith, 2008] and allocation of of cash transfer programs, particularly PROGRESA [Skoufias et al., 2001, Coady, 2006]. We build on this work by showing how effects and allocations can be combined to audit policymaker priorities, and improve the design of future policies.

Finally, our efforts relate broadly to recent work on *fairness in machine learning* (Dwork et al. [2012], Barocas et al. [2018]). Within this subfield, several papers have studied the social welfare implications of algorithmic decisions, and how social welfare concerns relate to different notions of fairness Ensign et al. [2017], Hu and Chen [2018], Mouzannar et al. [2018], Liu et al. [2018]. Most directly related, Noriega et al. Noriega et al. [2018] discuss how different constraints to targeting can impact efficiency and fairness. Our approach is distinct, however, in that we show how using machine learning tools can be used to better characterize and audit the implied priorities of a program, as revealed in the program's observed allocation. We hope that by providing increased visibility into these revealed preferences, future policies can be better aligned with stated preferences and explicit policy objectives.

2. Model

We show how to back out the notion of welfare implied by the combination of the decisions encoded in policy, and the impacts of those policies on different households.

We consider a planner deciding how to allocate treatment among N entities, which may be individuals or households. For convenience, we refer to entities as households. The planner selects a treatment status $y_i \in \{0, 1\}$ for each household i. Household i has characteristics $\mathbf{x}_i$. The planner earns total welfare according to:

$$S = \sum_i \mu(\mathbf{x}_i) \cdot u(y_i, \mathbf{x}_i)$$

where $\mu(\mathbf{x}_i)$ represents the welfare weight of a person with characteristics $\mathbf{x}_i$, and $u(\cdot,\cdot)$ represents the policymaker's evaluation of the household's utility. This evaluation of utility may be a linear combination of multiple outcome measures (such as consumption, health, and education):

$$u(y_i, \mathbf{x}_i) = C + \sum_j \lambda_j g_j(y_i, \mathbf{x}_i)$$

where $g_j(\cdot,\cdot)$ represents outcome j and λ_j represents its relative value, or 'impact weight', and C is a constant representing the base value of providing the program, regardless of its impact.

Imagine we have an experimental design that has recovered the (potentially heterogeneous) effect of treatment on each outcome j as a function of covariates $\mathbf{x}_i$:

$$\Delta g_j(\mathbf{x}_i) \coloneqq g_j(1, \mathbf{x}_i) - g_j(0, \mathbf{x}_i)$$

The impact on social welfare is then:

$$\Delta S = \sum_i \mu(\mathbf{x}_i) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i))$$

The policymaker assigns each household a score $z(\mathbf{x})$, representing the priority order in which they would receive program benefits.

We denominate welfare weights in units of the least preferred household, as the problem is invariant to multiplicative scaling ($\mu(\underline{\mathbf{x}}) = 1$, for $\underline{\mathbf{x}}$ such that $z(\underline{\mathbf{x}}) \leq z(\mathbf{x}) \forall \mathbf{x}$). In our application, we denominate λ_j weights in units of consumption, so $\lambda_{cons.} = 1$.

3. Intuition

To demonstrate the intuition behind our method, we consider a simple example in Figure 1. Consider the case of a single outcome and one dimension of heterogeneity, x, which corresponds with consumption. A policymaker allocates a program by ordering households by the function $z(x)$, prioritizing poor households. As shown in Figure 1, the same allocation rule could imply higher welfare weights on the poor, higher welfare weights on the rich, or equal welfare weights, depending on how treatment effects vary with x.

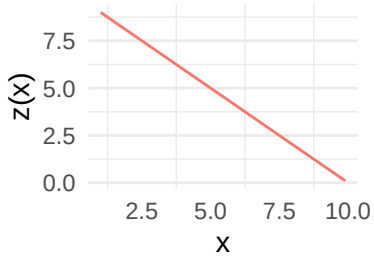
The next section demonstrates how to empirically recover welfare and impact weights from data in when there are multiple dimensions of heterogeneity and multiple outcomes of interest.

4. Estimation

If we observe the planner allocate treatment, what can we infer about a policymaker's preferences ($\mu(\mathbf{x}_i)$, C, $\boldsymbol{\lambda}$)?

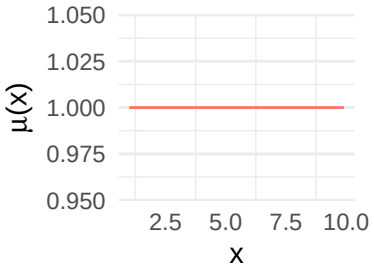
FIGURE 1. Intuitive Example

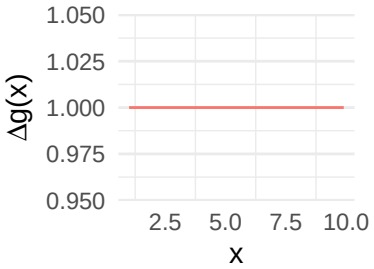
An allocation rule that prefers the poor (low x)...



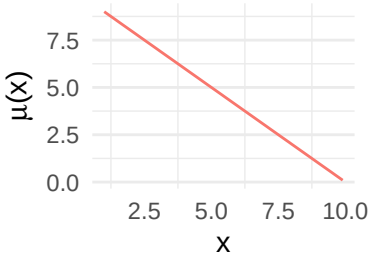
Could result from

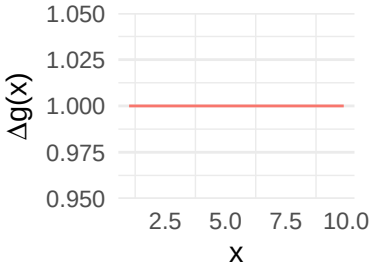
Equal welfare weights on households if treatment effects decline linearly in x



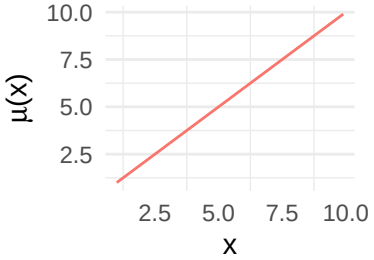


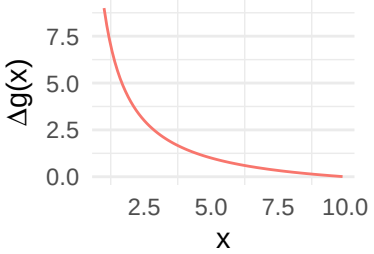
Higher welfare weight on the poor if treatment effects are constant





Higher welfare weight on the rich if treatment effects are much higher for the poor





1. **What we can infer from cardinal treatment scores.** First, consider a policymaker who ranks households according to the score $\tilde{z}(\mathbf{x}_i)$, which is a cardinal ranking, so that

$$\tilde{z}(\mathbf{x}_i) = \mu(\mathbf{x}_i) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i))$$

We assume that $\mu(\mathbf{x}_i)$ takes a linear functional form,

$$\mu(\mathbf{x}_i) = \beta \mathbf{x}_i + \varepsilon_i$$

where β is an vector of coefficients with the same dimension as $\mathbf{x}$. Then we can infer the welfare weights with

$$E\left(\tilde{z}(\mathbf{x}_i)\right) = E\left(\beta \mathbf{x}_i \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i))\right)$$

so long as the degrees of freedom $(dim(\beta) + dim(\lambda))$ are less than or equal to the number of observations and $\varepsilon_i \perp \Delta g_j(\mathbf{x}_i) \forall i, j$, which is to say, any estimation error of the welfare weights is distributed independently of the treatment effects. This assumes that there is no omitted variable $\tilde{\mathbf{x}}_i$ that drives both policymaker preferences over households and treatment effects.

Crucially, the extent to which policymakers prioritize households intrinsically, without any consideration for the relative marginal benefits of treatment, is captured by the comparative size of the C term against the λ_j weights.

Estimation Method. Linear regression can be used to estimate the best-fit parameters for β and λ.

2. **What we can infer from ordinal treatment scores or treatment status.**

If we observe ordinal treatment scores. Now consider the case where we don't observe the underlying cardinal score, but some monotone transformation $z(\mathbf{x}_i) = f(\tilde{z}(\mathbf{x}_i))$. This transformation preserves the priority order of who would receive treatment, but no longer describes the intensity of preferences.

For each pair of households i and i', with $z_i > z_{i'}$, we must have:

$$\mu(\mathbf{x}_i) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i)) \geq \mu(\mathbf{x}_{i'}) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_{i'}))$$

As above, we assume that $\mu(\mathbf{x}_i)$ takes a linear functional form,

$$\mu(\mathbf{x}_i) = \beta \mathbf{x}_i + \varepsilon_i$$

We then combine the above two expressions to formulate the following inequalities:

$$E\left[h_{mi}\left(\beta \mathbf{x}_i \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i)) - \beta \mathbf{x}_{i'} \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_{i'}))\right)\right] \geq 0$$

which holds for any instruments $\mathbf{h}_m$ such that $h_{mi} \perp \varepsilon_i \forall i$. So long as any error in $\mu(.)$ is independently distributed from the household-level variable h_{mi}, this formula should hold in expectation. We include as instruments $\mathbf{x}_i$, $\Delta g_j(\mathbf{x}_i)$ for each j, and the constant 1.[3]

These x_{ik} conditions then serve as the basis for a squared loss function $Loss(\beta, \lambda, \mathbf{x}_i)$, where, writing $\beta \mathbf{x}_i \cdot \left(C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i)\right)$ as $\delta_i(\beta)$,

(1)

$$Loss(\beta, \lambda, \mathbf{x}_i) = \sum_{\mathbf{x}_{i'}: z(\mathbf{x}_{i'}) < z(\mathbf{x}_i)} \left[1_{\delta_i(\beta) < \delta_{i'}(\beta)} + \sum_j 1_{\Delta g_j(\mathbf{x}_i)\delta_i(\beta) < \Delta g_j(\mathbf{x}_{i'})\delta_{i'}(\beta)} + \sum_k *1_{x_{ik}\delta_i(\beta) < x_{i'k}\delta_{i'}(\beta)} \right]$$

We then sum this loss over all $\mathbf{x_i}$ to compute total aggregate loss, $\sum_{\mathbf{x_i}} Loss(\beta, \lambda, \mathbf{x}_i)$. In the estimation procedure we also add a regularization parameter, determined using 3-fold cross-validation, to control for overfitting, as is standard in preference-learning method applications.

If we observe only allocation status. Now imagine we observe a policymaker selecting only a final allocation $\mathbf{y}$. Then, if the policymaker treated household i but not household i', it must be that was preferable to reversing the allocation:

$$\mu(\mathbf{x}_i) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_i)) \geq \mu(\mathbf{x}_{i'}) \cdot (C + \sum_j \lambda_j \Delta g_j(\mathbf{x}_{i'}))$$

This inequality condition is simply an ordinal ranking where $z(\mathbf{x}_i)$ is a binary indicator. We are therefore able to estimate $\hat{\beta}$ and $\hat{\lambda}$ exactly as described above, using the same conditions to formulate the same loss functions.

Estimation Method. In the case where there is only one outcome of interest j, then the above loss function in (1) reduces to:

(2)

$$Loss(\beta, x_i) = \sum_{\mathbf{x}_{i'}: z(\mathbf{x}_{i'}) < z(\mathbf{x}_i)} \left[1_{\delta_i(\beta) < \delta_{i'}(\beta)} + 1_{\Delta g(\mathbf{x}_i)\delta_i(\beta) < \Delta g(\mathbf{x}_{i'})\delta_{i'}(\beta)} + \sum_k *1_{x_{ik}\delta_i(\beta) < x_{i'k}\delta_{i'}(\beta)} \right]$$

Minimizing the sum of this loss function over all observations with respect to β then allows us to back out the estimated parameters of $\mu(\mathbf{x}_i)$, $\hat{\beta}$. If there exists $\hat{\beta}$ such that loss is zero for all observations, we can recover the set of all parameter values that produce zero loss for the dataset of $\{\mathbf{x}_i, \Delta g_j(\mathbf{x}\prime), z(\mathbf{x}_i)\}$. If there are no such parameter values of $\hat{\beta}$ that produce zero loss for the dataset, we will estimate the unique vector $\hat{\beta}$ that minimizes the loss over the observed data.[4]

[3] In the case that $\mathbf{h}_{mi} < 0$, the inequality is flipped, and the subsequent loss is calculated accordingly.

[4] It is also possible to have a range of parameter values that produce minimum loss if the objective function is flat over the local region around the minimum. In our application, this only occurs when $z(.)$ is very coarse (such as a 0-1 binary treatment indicator) and treatment effects are estimated without heterogeneity, leaving the algorithm little data to discriminate between different impact weights λ.

In the case of $j > 1$ outcomes of interest, we must infer λ_j as well. We therefore jointly estimate $\hat{\beta}$ and $\hat{\lambda}$ as the parameter values that minimize the full loss function as described in equation (1), as derived directly from the utility inequality conditions. [5]

Implicitly, the values of $\hat{\beta}$ are being identified off of observations with relatively similar treatment effects across outcomes but distinct rankings, as the difference between their ranks must be driven by $\mu(\mathbf{x}_i)$; remaining variation in the ranking then identifies $\hat{\lambda}$.

5. Empirical Example

We demonstrate our method on the Mexican PROGRESA conditional cash transfer (CCT) program.

1. **Context.** PROGRESA began in the late 1990s and served as the inspiration and model for a number of similar conditional cash transfer programs across Latin America, including a large welfare program in Mexico, Oportunidades, that has served millions of poor families in the years since. PROGRESA, first implemented by the Mexican federal government in 1997, was specifically designed to improve poor families' investment in the human capital of their children by incentivizing both health investments in pregnant women and very young children (0-5 years in age) and incentivizing school attendance for families with children enrolled in grades 3-9. Bi-monthly cash grants (equal to roughly 20% of pre-survey monthly consumption) were offered to family mothers conditional on regular doctor's visits and/or regular school attendance, depending on eligibility.[6]

1.1. *Targeting.* PROGRESA targeted poor communities on the basis of a 'village marginality index' (VMI), and targeted poor households within these communities on the basis of a 'household poverty score' (HPS). The VMI was based on a series of village-level variables, including the proportion of households living in poverty, population density, and health and education infrastructure. The HPS was based on a household-level proxy means test: surveyors collected data on easily observable characteristics (such as housing materials, family structure, etc.) on all households in eligible communities through a census, and for a small sample, also collected in-depth information on per-capita consumption. The coefficients from a regression of these observable characteristics on per-capita consumption for the in-depth sample then served as the weights for constructing the HPS from these more easily-collected data.

[5] We use only the indicator in the λ loss function in order to avoid endogeneity in penalization weights within the loss function. For more details, see the Appendix.

[6] For a more detailed treatment of PROGRESA and its background, see Emmanuel Skoufias [2008], John Hoddinott [2004], and Simone Boyce [2003].

1.2. *Experimental Design.* During the early years of its implementation, PROGRESA administrators used a randomized experimental design as part of its staggered rollout across communities: approximately 10% (506) of the 5,000 eligible communities were selected to be part of the evaluation, with 320 assigned to the treatment group and 185 to the control group. Behrman and Todd [1999] show that the randomization across communities was successful in that treatment and control communities were statistically indistinguishable across a wide array of observable covariates. Treatment communities were initiated into the PROGRESA program in the summer 1998 while control groups were not initiated into the program until 2000.

1.3. *Data.* Households in both groups were surveyed over five rounds. First, households were surveyed as part of the census that served as the basis of the HPS, and then they were surveyed again in May 1998 prior to treatment-group program initiation, and then follow-up surveys were conducted four more times at approximate six month intervals thereafter. These surveys asked household demographic and socioeconomic characteristics, as well as questions about health care utilization and educational attendance. We focus on the follow-up surveys between treatment, October 1998 and November 1999. Summary statistics for the matched data sample of households present in both periods are presented in Table 1. These data contain information on approximately 15,000 households over the entire experiment period.

1.4. *Outcomes.* We focus on three outcomes of interest that we focus on in this application are changes in per-capita monthly consumption, changes in health status of children, and changes in educational attendance of children. In the terminology of the above theory section, these outcomes are our $\Delta g_j(x_i)$, where $j = 3$. All of these are plausibly target outcomes that a social planner or government might prioritize when designing a welfare program, although only the latter two outcomes were explicitly prioritized by the Mexican federal government as part of the official goals of the program. Previous studies have estimated significant treatment impacts of PROGRESA for all three of these outcomes using the same data sample that we explore here (John Hoddinott [2004], Emmanuel Skoufias [2008], Simone Boyce [2003], Djebbari and Smith [2008]).

2. **Estimation.** To estimate the potentially heterogeneous impacts of PROGRESA on our set of outcome variables, we follow Djebbari and Smith [2008] and estimate a regression equation that allows the effect of treatment to vary by household characteristics. (This step can also be done using a more sophisticated model such as Wager and Athey [2018].) We allow treatment to vary by education level of household head, indigenous status of houshold head, gender of household head, a binary indicator for household head working in the agricultural sector, the age of the household head, number of children, number of school-age children,

TABLE 1. Descriptive Statistics

	October 1998 mean	November 1999 mean
Monthly average per capita consumption (pesos)	234.508	178.185
Assigned to treatment group	0.606	0.606
Household poverty score (1997)	695.700	695.700
Village marginality index (1997)	0.470	0.470
Household size	5.75	5.75
... Number of children less than 2 years old	0.692	0.702
... Number of children 3-5 years old	0.577	0.565
... Number of children 6-10 years old	0.948	0.928
... Number of boys 11-14 years old	0.356	0.350
... Number of girls 11-14 years old	0.338	0.332
... Number of boys 15-19 years old	0.318	0.316
... Number of girls 15-19 years old	0.310	0.308
... Number of men 20-34 years old	0.492	0.500
... Number of women 20-34 years old	0.5497	0.555
... Number of men 35-54 years old	0.444	0.445
... Number of women 35-54 years old	0.438	0.439
... Number of men at least 55 years old	0.253	0.254
... Number of women at least 55 years old	0.251	0.253
Head of household:		
... Is male	0.902	0.902
... Is an agricultural worker	0.596	0.600
... Education (in years)	2.703	2.704
... Is indigenous	0.386	0.386
... Age	45.47	45.50
Number of days a child is sick	1.310	0.857
Number of days a child misses school	0.567	0.249
N	14949	14949

and total household size. For clarity we focus on only outcomes in 1999.[7] Formally, we define $D_i \in \{0, 1\}$ as a dummy variable for treatment status of household i, $Y_{i,1999}$ as the outcome variable as measured in 1999, $X_{i,1998}$ as the vector of covariates as measured in 1998, and estimate the following regression:

$$\Delta Y_{i,1999} = \beta_0 + \beta_X X_{i,1998} + (\beta_D + \beta_{DX} X_{i,1998}) D_i + \varepsilon_i \tag{3}$$

This model allows post period (1999) outcomes to differ systematically according to household covariates, and additionally allows the treatment effect of PROGRESA to differ

[7]We also depart from Djebbari and Smith [2008] in not including poverty scores and village marginality index or their respective interactions in the list of covariates, to avoid potention correlated errors from using these rankings in both the TE estimates and in the preference-learning method.

across households according to their respective covariate profiles, as captured in the β_D and β_{DX} variables.

We construct our variables for treatment effects from the predicted values from our estimated formula (3), as

$$\Delta\hat{g}(\mathbf{x}_i) = (\hat{\beta}_D + \hat{\beta}_{DX} X_i)$$

We use these estimated treatment effects $\Delta\hat{g}(\mathbf{x}_i)$, and search for welfare weight parameters $\hat{\beta}$, impact weight parameters $\hat{\lambda}$, and constant C that minimize the loss function (Equation 2). We allow household priority weights $\mu(\mathbf{x}_i)$ to vary over education levels of the household heads, the size of households, the indigenous status of the household head, and the level of non-food consumption in October 1998 (a proxy for economic status). The final estimates of these vectors then serve as our estimates of the implied welfare weighting across covariates and impact weighting across outcomes, as inferred from the ranking of social priority and the distribution of program impacts across households.

3. **Estimates.**

3.1. *Treatment Effects.* To begin with, we present the results from the estimation of heterogeneous treatment effects from equation (2). We estimate considerable heterogeneity across households for all three outcomes, and find modest treatment effects on average for all three.

On average, PROGRESA is estimated to have increased household monthly consumption by 12.21 pesos, to have reduced the number of sick days per child by 0.064, and reduced the number of school days missed per child by 0.10.[8]

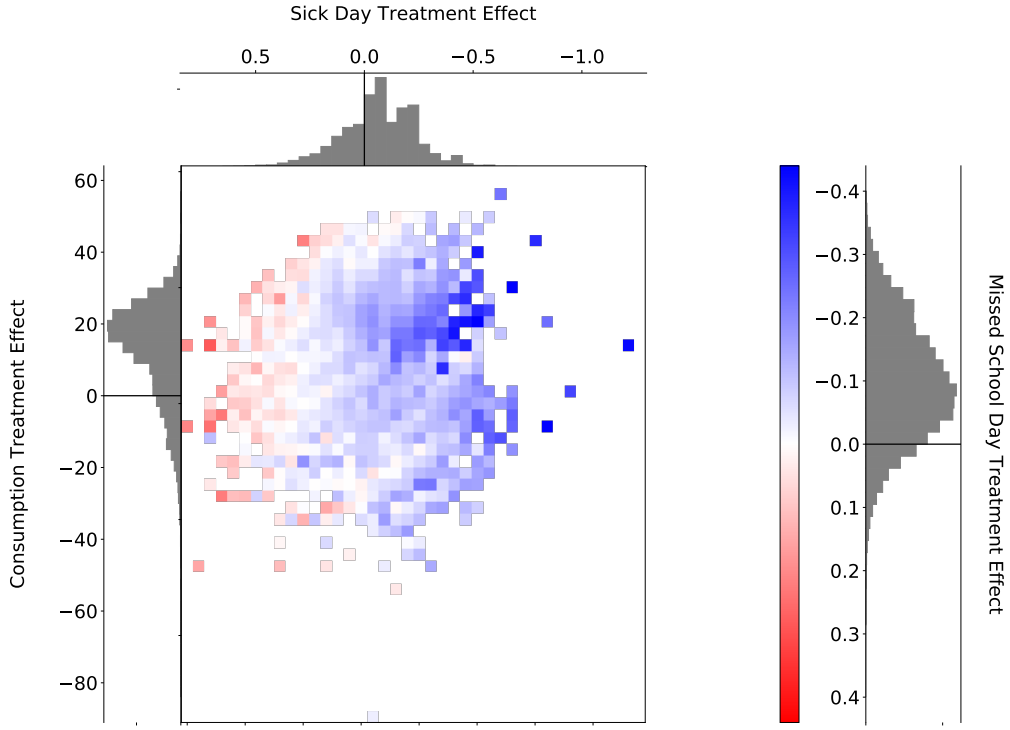
We next report heterogeneous treatment effects. The overall distributions of treatment effects by outcome are presented in Figure 2, and coefficient estimates are presented in Table 2, with standard errors in parentheses. Similar to Djebbari and Smith [2008], we find that consumption treatment impacts are higher for households with indigenous status and male heads of households, and lower for larger households.

3.2. *Welfare Weight and Impact Weight Estimates.* We then combine these heterogeneous treatment effect estimates $\Delta\hat{g}_j(\mathbf{x}_i)$ with the household poverty scores as our priority ranking over households, $z(.)$, and estimate welfare weights and impact weights using the preference-learning method. While final eligibility for PROGRESA was determined by a combination of village marginality index and poverty scores, within each village, poverty scores were the basis for determining a household's treatment versus non-treatment status, and as such, we use them as our preferred ranking.

Estimates are presented in column 1 of Table 3.

[8]Both "number of days sick" and "number of school days missed" refer to the single month before the survey was conducted. In the case of the Nov. 1999 survey, this reflects the number of days in October for either outcome.

FIGURE 2. Distribution of Estimated Treatment Effects



Joint and marginal distributions of estimated treatment effects of PROGRESA conditional cash transfer on schooling, health, and consumption. Schooling treatment effects are measured over the number of missed school days per school-age child in a given household. Health treatment effects are measured over the number of sick days per young (0-5 years old) child in a given household. Consumption treatment effects are measured over per-person consumption in pesos in a given household. Marginal distributions for consumption and health treatment effects are shown over the y and x axes, respectively, and are binned together in the center figure. Average schooling treatment effects in each consumption-health-treatment-effect bin is shown by the fill color of the bin, according to the index of the legend on the right. The marginal distribution of schooling treatment effects is shown in parallel to this legend. Note that missed school days and sick days are inferred to be "bads", according to our estimated weights, and so higher negative values for these treatment effects are associated with higher social utility.

TABLE 2. Treatment Effect Coefficient Estimates

	Consumption (Monthly avg per-person, pesos)	Health (sick days per child)	Schooling (missed school days per child)
Treatment	5.0746	-0.5809	-0.1985
	(16.812)	(0.423)	(0.192)
Treatment X Num Young Children (< 5 y.o.)	1.0894	-0.0787	-0.0467
	(2.702)	(0.065)	(0.028)
Treatment X Num Children (6-16 y.o.)	5.4304	-0.0917	0.0018
	(2.577)	(0.06)	(0.028)
Treatment X Num Elders (> 55 y.o.)	-9.7668	-0.0234	-0.0011
	(4.742)	(0.128)	(0.052)
Treatment X Total Household Size	-3.1224	0.0877	0.0272
	(1.934)	(0.05)	(0.021)
Treatment X Head Education	-0.6038	-0.0014	-0.0024
	(1.059)	(0.024)	(0.011)
Treatment X Indigenous Status	4.2948	0.1319	0.001
	(5.048)	(0.121)	(0.054)
Treatment X Male Head of Household	30.9864	0.2982	0.0192
	(9.002)	(0.262)	(0.103)
Treatment X Head of HH Agricultural Worker	-7.1956	-0.1568	-0.0595
	(5.364)	(0.129)	(0.058)
Treatment X Age of Household Head	-0.0353	0.0011	0.0027
	(0.254)	(0.007)	(0.003)
Treatment X Household Income in 1997	-0.7145	0.0052	-0.0185
	(1.44)	(0.035)	(0.016)

We find that the government would have placed 8.0% higher value on the median household if they were of indigenous status, 16.9% higher value for each additional household member, 0.3% lower value for each additional peso of non-food per person average consumption, and 0.6% lower value for each additional year of household head education.

We estimate that the government's allocations are consistent with valuing each child sick day as 101.98 pesos of consumption, and each missed school days as 108.07 pesos of consumption. Across impacts, 70% of the utility weight is derived through intrinsic valuation of the household, and 30% through valuation of changes in outcomes: for the median household, 14% of utility weight is derived through consumption, as compared to 9% through schooling impacts and 6% through health impacts.

These valuations can be compared against other estimates of the value of health and education:

If we assume a mapping of sick days to disability adjusted life years (DALYs) based simply on the number of days lost to sickness, then the government's allocation implies a valuation of $101.98 * 365 = 37222.7$ pesos or $3722.27 per DALY. This valuation is roughly thirty-seven times larger than standard recommendations for cost-effective health interventions [Laxminarayan et al., 2006]. It is considerably higher than the revealed-preference inferences of valuations per DALY of $23.68 that Kremer et al. [2011] infer for Kenyan households, based on how far they are willing to walk for clean water.

The government's allocation implies a value of $108.07 * 180 = 19452.6$ pesos or \$1945.26 per missed year of school. This is also about an order of magnitude larger than a back-of-the-envelope calculation of the value of each school year: based on a review of multiple studies, Psacharopoulos and Patrinos [2018] suggest a 9% average return to a year of schooling. If we proxy for income by consumption levels, 9% of average consumption in 1999 is approximately equal to 16 pesos. Assuming a lifetime of 40 years of work, with later years discounted at a rate of 3%, this corresponds to lifetime present-discounted earnings of 424.91 pesos per missed year of school.

4. **Counterfactuals.** With these estimates in hand, we compare the estimates from the government allocation to counterfactual allocations.

4.1. *Alternate welfare weights.* In columns 2-3 of Table 3, we assess the scores and outcomes that result when using the empirically estimated impact weights but alternate welfare weights. We present the implied distribution of the $z'(.)$ ranking, as described by a linear regression of the implied $z'(.)$ over the four covariate dimensions inspected in the model. We report the coefficients of this linear regression over the constructed $z'(.)$ with α_k, for each covariate k.

When welfare weights are set equal across households (column 2), the resulting score puts much more weight on the measures of household size and household age composition relative to the other variables. The effective priority ranking no longer positively correlates with indigenous status as well.

When welfare weights are set to order households by income (column 3), the resulting score sets positive weight on the number of elder household members and negative weight on the number of children, as well as a larger negative weight on education.

4.2. *Technocratic impact weights.* In columns 4-5 of Table 3, we keep the original welfare weights but use assumed technocratic impact weights. We assume 50 pesos per DALY and roughly 16 pesos per missed school day, and an assumption that the consumption increase is permanent, so that the present-value discounted weight on each peso is roughly 70.

The $z'(.)$ ranking implied by these weights covaries more tightly with poverty and indigenous status, and covaries inversely with household size, but otherwise appears broadly similar.[9] By contrast, changing the household covariate welfare weights to an equal weighting across households so that $\mu(\mathbf{x}_i) \equiv 1$ leads to more weight on the number of elders in the household.

[9]Simone Boyce [2003] find that PROGRESA reduces childhood illness rates by roughly 25%; as the average sick day impact of PROGRESA is 0.032, we make the very strong assumptions that a 0.032 reduction in average sick days thereby corresponds to a reduction in illness rates of 25%, and that a 25% reduction in illness rates corresponds to an increase in average DALY of 0.25. We thereby deduce that a reduction of 1.0 in average sick days would correspond to an increase in average DALY of 6.9, which on the very low-end would cost roughly \$10 for an intervention, leading to a value of \$69. We round down to account for the potential overestimated value of reduced illness rates in terms of DALY.

4.3. *Focus on different outcomes.* In columns 6-8 of Table 3, we present alternative specifications that reset impact weights to be entirely determined by one of the three outcomes inspected. When the impact weights are determined entirely by schooling effects, $z'(.)$ covaries positively with income, education, and indigenous status; when the impact weights are determined entirely by consumption outcomes, $z'(.)$ covaries negatively with education and income and positively with indigenous status; when impact weights are determined by health outcomes, $z'(.)$ covaries negatively with all three.

4.4. *An alternative government scoring rule.* In column 9 of Table 3, we present the weight estimates of an alternative specification used by the government. After 2003, the Mexican government began using a different priority ranking that increased the priority of older, childless households [Skoufias et al., 1999, 2001]. In this 'densified' ranking, Village Marginality Index (VMI) mattered more and poverty scores less relative to when the program evaluation was begun; originally, only 50% of households within eligible communities were eligible for the program, but after the change, roughly 80% of households were eligible. We approximate the change in ranking from this reordering by setting $z(.)$ to the VMI instead of poverty scores.

This rule actually places more welfare weight on richer households (for the median household, 0.9% higher value for each additional peso of non-food per person average consumption), as well as higher priority on indigenous status of households (35.8% higher value on the median household), and similar priority for household size (12.3% higher value for each additional household member).

The impact weights on health outcomes are about 20% higher, estimated at 125.41 pesos for each fewer sick day per child, while the impact weights on schooling outcomes are 40% lower, estimated at 60.85 pesos for each fewer missed school day per child. Overall, however, the biggest change is the relative increase of the constant term, leading to a shrinking weight of all outcomes in the total welfare calculation: for the median household, now 85% of total utility summation derives through the constant term / intrinsic valuation, 3% through schooling impacts, 4% through health impacts, and 7.8% through consumption impacts.

In this way, we compare how the implicit ranking from different specifications of welfare weights and impact weights changes as we change the weights. We also identify the group that would have received treatment status under these alternative allocations and present counterfactual average outcomes in 1999 among all households under an alternative treatment assignment regime, by constructing predicted values if the top 60% of the constructed priority ranking had received treatment.

The $z'(.)$ distributions inferred from the alternative specifications are broadly similar except in the differences implied by their weights on covariates: under the udpated ranking, the $z'(.)$ covaries with indigenous status almost three times as much, and covaries positively with higher-income households.

Table 3. Allocation Rules

	Actual	**Counterfactual**							
	Household Poverty Score	Empirical Impact Weights		Technocratic Impact Weights		Only Value Consumption Impacts	Only Value Health Impacts	Only Value Education Impacts	2003 Household Poverty Score
		Equal Welfare Weights	Prefer Poorest	Empirical Welfare Weights	Equal Welfare Weights	Empirical Welfare Weights	Empirical Welfare Weights	Empirical Welfare Weights	
Welfare weights $\mu(\mathbf{x}_i)$									
β_{HHSize}	3.815	0	0	3.815	0	3.815	3.815	3.815	4.293
β_{Educ}	-0.156	0	0	-0.156	0	-0.156	-0.156	-0.156	-0.565
$\beta_{1997Inc.}$	-0.236	0	-1	-0.236	0	-0.236	-0.236	-0.236	0.3
$\beta_{Indigenous}$	9.866	0	0	9.866	0	9.866	9.866	9.866	12.438
$\beta_{NumElders}$	2.61	0	0	2.61	0	2.61	2.61	2.61	3.653
$\beta_{NumChildren}$	4.564	0	0	4.564	0	4.564	4.564	4.564	2.813
$\beta_{constant}$	(.)	1	(.)	(.)	1	(.)	(.)	(.)	(.)
Impact weights $\boldsymbol{\lambda}$									
$\lambda_{Consumption}$ (pesos)	1.0	1.0	1.0	1.0	1.0	1	0.0	0.0	1.0
λ_{Health} (pesos per child sick day)	-101.98	-101.98	-101.98	-0.725	-0.725	0	-101.98	0.0	-125.407
$\lambda_{Schooling}$ (pesos per missed school day)	-108.069	-108.069	-108.069	-0.307	-0.307	0	0.0	-108.069	-60.85
C (pesos regardless of impact)	77.165	77.165	77.165	77.165	77.165	0	0	0	167.233
Score Function (Linear regression coeffs.) $z(\mathbf{x}_i)$									
α_{HHSize}	-1.41	-5.008	4.36	1.394	-1.333	-0.847	-2.237	-1.716	-0.142
α_{Educ}	-0.037	0.22	-0.13	-0.291	-0.154	-0.228	-0.04	0.398	-0.457
$\alpha_{1997Inc.}$	-0.03	-0.003	-5.115	-0.203	-0.285	-0.077	-0.154	0.547	0.163
$\alpha_{Indigenous}$	2.421	-2.579	1.775	7.591	2.657	2.234	-2.126	1.604	7.137
$\alpha_{NumElders}$	-0.687	-4.468	3.588	-0.998	-6.906	-2.794	0.015	-0.839	0.734
$\alpha_{NumChildren}$	4.273	4.973	-4.311	4.47	2.139	2.067	2.419	2.155	5.115
Pred. outcomes under alt. treatment assignment									
Monthly consumption, 1999 (pesos)	181.40865	181.18703	177.27299	181.31144	183.73169	183.32466	178.54919	179.96725	180.40554
Monthly sick days, 1999 (per child)	1.09544	1.04921	1.12106	1.1213	1.09799	1.10781	1.03562	1.07744	1.11261
Monthly missed school days, 1999 (per school-age child)	0.29713	0.27346	0.32039	0.30798	0.29417	0.29898	0.28164	0.26928	0.30614
N	14300	14300	14300	14300	14300	14300	14300	14300	14300

TABLE 4. Weight Estimates under Coarse Data Conditions

	$z(.)$:	Poverty Score	Binary
	$\Delta g_j(\mathbf{x_i})$:	Heterog.	Heterog.
Welfare weights			
	β_{HHSize}	3.815	1.979
	β_{Educ}	-0.156	-0.074
	$\beta_{1997Inc.}$	-0.236	-0.052
	$\beta_{Indigenous}$	9.866	5.308
	$\beta_{NumElders}$	2.61	2.014
	$\beta_{NumChildren}$	4.564	2.5120
Impact weights			
	$\lambda_{Consumption}$	1.0	1
	λ_{Health}	-101.98	-88.126
	$\lambda_{Schooling}$	-108.069	-117.127
	C	77.165	68.87
	N	14949	14949

5. **Performance in Settings with Coarser Data.**

5.1. *If we observe final allocations, not a score.* In many cases, researchers may only have access to a binary $z(.)$ that connotes treatment eligibility. We demonstrate how our method can be applied by estimating our main specification with a binary indicator of above- or below-median poverty score. Welfare weight estimates from this lower-information $z(.)$ are presented in column 2 of Table 4, where column 1 present the standard specification results. With the exception of the impact weight on health, the relative sizes of the estimated weights are strikingly similar between these two columns, demonstrating the method's viability in situations where precise ranking data is limited or unavailable.

6. CONCLUSION

While economists reason about primitives of utility and welfare weights, policy discussions commonly revolve instead around the mechanics of implementation. This paper demonstrates how heterogeneous treatment effect estimates can be used to bridge between these two conceptions. We imagine this framework could be used in several ways. First, it could be used to characterise existing allocations, to provide an indication of the revealed preferences of the policymakers. This, in turn, provides an ex-post auditing mechanism that can help hold policymakers accountable for past transfers – and in particular, to evaluate whether the implemented allocation reflects the stated goals of the policy. Perhaps most importantly, this approach can be used to amend existing policies and guide future allocations. In particular, it can demonstrate how different priorities over welfare outcomes and population subgroups would produce different allocations, and quantifies the welfare impacts of these adjustments.

References

Vivi Alatas, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4):1206–1240, June 2012. ISSN 0002-8282. doi: 10.1257/aer.102.4.1206. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.4.1206.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018.

Nicholas Barr. *Economics of the welfare state*. Oxford university press, 2012.

Jere R. Behrman and Petra E. Todd. Randomness in the experimental samples of PROGRESA (education, health, and nutrition program). *International Food Policy Research Institute, Washington, DC*, 1999.

Wei Chu and S. Sathiya Keerthi. Support Vector Ordinal Regression. *Neural Computation*, 19:792–815, 2007. doi: 10.1162/neco.2007.19.3.792.

David P. Coady. The Welfare Returns to Finer Targeting: The Case of The Progresa Program in Mexico. *International Tax and Public Finance*, 13(2-3):217–239, May 2006. ISSN 0927-5940, 1573-6970. doi: 10.1007/s10797-006-4824-2. URL https://link.springer.com/article/10.1007/s10797-006-4824-2.

Stephen Coate and Stephen Morris. On the Form of Transfers to Special Interests. *Journal of Political Economy*, 103(6):1210–1235, December 1995. ISSN 0022-3808. doi: 10.1086/601449. URL https://www.journals.uchicago.edu/doi/abs/10.1086/601449.

Habiba Djebbari and Jeffrey Smith. Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145(1):64–80, July 2008. ISSN 0304-4076. doi: 10.1016/j.jeconom.2008.05.012. URL http://www.sciencedirect.com/science/article/pii/S0304407608000493.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Vincenzo Di Maro Emmanuel Skoufias. Conditional Cash Transfers, Adult Work Incentives, and Poverty. *Journal of Development Studies*, 44(7):935–960, 2008.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]*, June 2017. URL http://arxiv.org/abs/1706.09847. arXiv: 1706.09847.

Marc Fleurbaey and Francois Maniquet. Optimal income taxation theory and principles of fairness. *Journal of Economic Literature*, 56(3):1029–79, 2018.

Paul Gertler. Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. *The American Economic Review*, 94(2): 336–341, 2004. ISSN 0002-8282. URL https://www.jstor.org/stable/3592906.

Rema Hanna and Benjamin A. Olken. Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives*,

32(4):201–226, November 2018a. ISSN 0895-3309. doi: 10.1257/jep.32.4.201. URL https://www.aeaweb.org/articles?id=10.1257/jep.32.4.201.

Rema Hanna and Benjamin A. Olken. Who should receive anti-poverty programs? Universal basic incomes vs. targeted transfers in developing countries. *Journal of Economic Perspectives*, 2018b.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support Vector Learning for Ordinal Regression. In *In International Conference on Artificial Neural Networks*, pages 97–102, 1999.

Lily Hu and Yiling Chen. Welfare and Distributional Impacts of Fair Classification. *arXiv:1807.01134 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1807.01134. arXiv: 1807.01134.

Emmanuel Skoufias John Hoddinott. The Impact of PROGRESA on Food Consumption. *Economic Development and Cultural Change*, 53(1):37–61, 2004.

Michael Kremer, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions. *The Quarterly Journal of Economics*, 126(1):145–205, February 2011. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjq010. URL http://qje.oxfordjournals.org/content/126/1/145.

SÃ¶ren R. KÃŒnzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1804597116. URL https://www.pnas.org/content/116/10/4156.

Ramanan Laxminarayan, Jeffrey Chow, and Sonbol A. Shahid-Salles. *Intervention Cost-Effectiveness: Overview of Main Messages*. The International Bank for Reconstruction and Development / The World Bank, 2006. URL https://www.ncbi.nlm.nih.gov/books/NBK11784/.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3156–3164, Stockholm, Sweden, 2018.

Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From Fair Decision Making to Social Equality. *arXiv:1812.02952 [cs, stat]*, December 2018. URL http://arxiv.org/abs/1812.02952. arXiv: 1812.02952.

Albert L. Nichols and Richard J. Zeckhauser. Targeting Transfers through Restrictions on Recipients. *The American Economic Review*, 72(2):372–377, 1982. ISSN 0002-8282. URL http://www.jstor.org/stable/1802361.

Alejandro Noriega, Bernardo Garcia-Bulle, Luis Tejerina, and Alex Pentland. Algorithmic Fairness and Efficiency in Targeting Social Welfare Programs at Scale. *Bloomberg Data for Good Exchange Conference*, 2018.

George Psacharopoulos and Harry Anthony Patrinos. Returns to investment in education. 2018.

Paul Gertler Simone Boyce. An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico. volume 85. Royal Economic Society, 2003.

Emmanuel Skoufias, Benjamin Davis, and Jere R. Behrman. An evaluation of the selection of beneficiary households in the education, health, and nutrition program (PROGRESA) of Mexico. *International Food Policy Research Institute, Washington, DC*, 1999.

Emmanuel Skoufias, Benjamin Davis, and Sergio de la Vega. Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA. *World Development*, 29 (10):1769–1784, October 2001. ISSN 0305-750X. doi: 10.1016/S0305-750X(01)00060-2. URL http://www.sciencedirect.com/science/article/pii/S0305750X01000602.

Christopher Timmins. Measuring the Dynamic Efficiency Costs of Regulators' Preferences: Municipal Water Utilities in the Arid West. *Econometrica*, 70(2): 603–629, December 2003. ISSN 1468-0262. doi: 10.1111/1468-0262.00297. URL http://onlinelibrary.wiley.com.ezp-prod1.hul.harvard.edu/doi/10.1111/1468-0262.00297/abstra

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1319839. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2017.1319839.

7. Appendix

1. **Remarks about the functional form for $Loss(\beta, \lambda, \mathbf{x_i})$.** We design $Loss(\beta, \lambda, \mathbf{x_i})$ based on binary indicator functions, rather than on the squared loss, because the penalization weights in a squared loss function for $Loss(\beta, \lambda, \mathbf{x_i})$ may otherwise be asymmetric, leading to systematic misestimation of $\hat{\lambda}$.

For ease of explanation, consider the ideal case where we know the true welfare weights $\mu(x_i)$ and where $j = 2$, and so $\lambda_2 = 1 - \lambda_1$. We are therefore only attempting to recover a single parameter, λ_1. Using a squared loss formula for $Loss(\beta, \lambda, \mathbf{x_i})$, we would have

$$Loss(\beta, \lambda, \mathbf{x_i}) = (\mu(\mathbf{x}_i) * (\lambda_1 \Delta g_1(\mathbf{x}_i) + (1 - \lambda_1)\Delta g_2(\mathbf{x}_i)) - \mu(\mathbf{x}_j) * (\lambda_1 \Delta g_1(\mathbf{x}_j) + (1 - \lambda_1)\Delta g_2(\mathbf{x}_j)))^2$$

if $\mu(\mathbf{x}_i) * (\lambda_1 \Delta g_1(\mathbf{x}_i) + (1 - \lambda_1)\Delta g_2(\mathbf{x}_i)) < \mu(\mathbf{x}_j) * (\lambda_1 \Delta g_1(\mathbf{x}_j) + (1 - \lambda_1)\Delta g_2(\mathbf{x}_j))$ and 0 otherwise. ($z(\mathbf{x}_i) > z(\mathbf{x}_j)$, here.) Rearranging terms, this becomes

$$Loss(\beta, \lambda, \mathbf{x_i}) = (\lambda_1 (\mu(x_i)\Delta g_1(\mathbf{x}_i) - \mu(x_j)\Delta g_1(\mathbf{x_j}) + \mu(x_j)\Delta g_2(\mathbf{x}_j) - \mu(x_i)\Delta g_2(\mathbf{x}_i)) + \mu(x_i)\Delta g_2(\mathbf{x}_i) - \mu(x_j)\Delta g_2(\mathbf{x}_j))^2$$

when loss is positive. The penalization weights for λ_1 are therefore functionally being determined by the relative values of $\mu(\mathbf{x}_i)\Delta g_1(\mathbf{x}_i) - \mu(\mathbf{x}_j)\Delta g_1(\mathbf{x}_j)$ and $\mu(\mathbf{x}_j)\Delta g_2(\mathbf{x}_j) - \mu(\mathbf{x}_i)\Delta g_2(\mathbf{x}_i)$. This is easiest to see in the extreme cases where $\lambda_1 = 1$ or $\lambda_1 = 0$. Then

$$Loss(\beta, \lambda, \mathbf{x_i}) = (\mu(\mathbf{x}_i)\Delta g_1(\mathbf{x}_i) - \mu(\mathbf{x}_j)\Delta g_1(\mathbf{x}_j))^2$$

or

$$Loss(\beta, \lambda, \mathbf{x_i}) = (\mu(\mathbf{x}_i)\Delta g_2(\mathbf{x}_i) - \mu(\mathbf{x}_j)\Delta g_2(\mathbf{x_j}))^2$$

respectively.

This is where the problem arises: the relative sizes of $\mu(\mathbf{x_i})\Delta g_1(\mathbf{x_i}) - \mu(\mathbf{x}_j)\Delta g_1(\mathbf{x_j})$ and $\mu(\mathbf{x}_j)\Delta g_2(\mathbf{x}_j) - \mu(\mathbf{x}_i)\Delta g_2(\mathbf{x}_i)$ may be unequal for reasons unrelated to specification of λ_1, and if so, they may bias estimation of λ_1 that uses the above loss function. For example, if $\Delta g_2(.)$ is more tightly positively correlated with $\mu(.)$ for any given x_i than $\Delta g_1(.)$ is, then the difference $\mu(\mathbf{x_i})\Delta g_2(\mathbf{x_i}) - \mu(\mathbf{x}_j)\Delta g_2(\mathbf{x_j})$ may also be systematically larger, which means that values of $\hat{\lambda_1}$ that are 'too small' will be penalized more than values of $\hat{\lambda_1}$ that are 'too large' and the optimizer may prefer a final estimate of $\hat{\lambda}_1 > \lambda_1$, assuming that there are observations with positive loss on either side.[10] Even in the measure-zero case where the optimization algorithm starts on the true λ_1, if there are equal violations of the inequality

[10] E.g., considering $\lambda_1 = 0$, then in cases of positive loss we have $\mu(x_i)\Delta g_2(x_i) < \mu(x_j)\Delta g_2(x_j)$. If $\mu(x_i) = 1$, then $Loss_\lambda = (\mu(x_i)\Delta g_2(x_i) - \mu(x_j)\Delta g_2(x_j))^2 = (\Delta g_2(x_i) - \Delta g_2(x_j))^2$. But if $\mu(x_i) = \Delta g_2(x_i)$, then $Loss_\lambda = \left(\Delta g_2(x_i)^2 - \Delta g_2(x_j)^2\right)^2$. This may be smaller or larger, depending, but in either case, the penalty from loss for misspecification of $\hat{\lambda}$ will be asymmetrically different.

in both directions due to noise, loss function minimization may prefer estimates of $\hat{\lambda_1}$ that are smaller or larger than λ_1 in order to shrink the loss from the direction with the outsized 'penalization weight'. In Monte Carlo simulations, we have confirmed that this effect can severely bias estimated $\hat{\lambda}$ away from true λ when using squared loss at this step.

Therefore, we use a loss function based on indicator functions, in order to completely sidestep this issue of endogenously asymmetric penalization weights:

$$Loss(\lambda, x_i) = 1_{\delta_i(\beta)<\delta_{i'}(\beta)} + \sum_j 1_{\Delta g_j(\mathbf{x}_i)\delta_i(\beta)<\Delta g_j(\mathbf{x}_{i'})\delta_{i'}(\beta)} + \sum_k *1_{x_{ik}\delta_i(\beta)<x_{i'k}\delta_{i'}(\beta)}$$

where $\delta_i(\beta)$ represents $\beta\mathbf{x}_i \cdot \Big(\sum_j \lambda_j \Delta g_j(\mathbf{x}_i)\Big)$.

Monte Carlo simulations confirm that the above loss function accurately recovers estimated $\hat{\lambda}$ that are close to true λ values under a wide variety of specifications and parameter values. In particular, the above loss function functions well even in cases where $\mu(.)$ is much more correlated with one treatment effect than another.