I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11863

Distributional Impact Analysis: Toolkit and Illustrations of Impacts beyond the Average Treatment Effect

Guadalupe Bedoya
Luca Bittarello
Jonathan Davis
Nikolas Mittag

SEPTEMBER 2018

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11863

Distributional Impact Analysis: Toolkit and Illustrations of Impacts beyond the Average Treatment Effect

Guadalupe Bedoya
World Bank

Luca Bittarello
Northwestern University

Jonathan Davis
University of Chicago

Nikolas Mittag
CERGE-EI and IZA

SEPTEMBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Distributional Impact Analysis: Toolkit and Illustrations of Impacts beyond the Average Treatment Effect*

Program evaluations often focus on average treatment effects. However, average treatment effects miss important aspects of policy evaluation, such as the impact on inequality and whether treatment harms some individuals. A growing literature develops methods to evaluate such issues by examining the distributional impacts of programs and policies. This toolkit reviews methods to do so, focusing on their application to randomized control trials. The paper emphasizes two strands of the literature: estimation of impacts on outcome distributions and estimation of the distribution of treatment impacts. The article then discusses extensions to conditional treatment effect heterogeneity, that is, to analyses of how treatment impacts vary with observed characteristics. The paper offers advice on inference, testing, and power calculations, which are important when implementing distributional analyses in practice. Finally, the paper illustrates select methods using data from two randomized evaluations.

JEL Classification: C18, C21, C54, C93, D39

Keywords: policy evaluation, distributional impact analysis, heterogeneous treatment effects, impacts on outcome distributions, distribution of treatment effects, random control trials

Corresponding author:
Nikolas Mittag
CERGE-EI
Politických vězňů 7
Prague 1, 110 00
Czech Republic
E-mail: nikolasmittag@posteo.de

* We benefited from the generous advice of Joel L. Horowitz and Andreas Menzel. We thank Arianna Legovini, David Evans, and Caio Piza for making this toolkit possible. We are grateful to Moussa P. Blimpo and participants at the DIME seminar and the Course on Distributional Impact Analysis held at the World Bank. We also benefited from the work and data from "The Impact of High School Financial Education: Experimental Evidence from Brazil" and "Parental Human Capital and Effective School Management: Evidence from the Gambia" and from conversations with members of both teams. We thank Kristoffer Bjärkefur for assistance with programming. This work was generously supported by the Africa Poverty and Social Impact Analysis (PSIA) and the Impact Evaluation to Development Impact (i2i) Trust Funds. This toolkit was written under the guidance of Sergio Firpo and Stephane Bonhomme.

Contents

1. Introduction

Traditional methods to evaluate the impacts of social programs and the vast majority of applied econometric policy evaluations focus on the analysis of means (Carneiro, Hansen and Heckman, 2002; Angrist and Pischke, 2009). However, there is also a large and growing literature on methods to evaluate the effects of programs and policies beyond their mean impact. While less frequently applied, these methods can provide information that is valuable or even necessary in the assessment of the consequences of policies and their desirability. The purpose of this toolkit is to provide an overview of the questions such methods can address and the core approaches that have been developed to answer them, including discussions of the assumptions they require, and practical issues in their implementation.

Mean impacts are a natural first summary statistic to describe the effect of a policy. The mean impact of a policy or intervention tells us by how much the outcome would increase or decrease on average when every member of a particular population is exposed to the policy or intervention. Thereby, they provide the central piece in any cost-benefit analysis. However, a decision-maker usually requires information on the effects of a policy beyond its mean impact. For example, mean impacts allow us to calculate the total gain from a program or policy, but do not allow us to say anything about the distribution of the gain or how the outcome distribution is affected by the program beyond changes in its mean. A positive average program effect tells us that a program can generate social surplus, but it may not be sufficient to allow us to judge whether the program is desirable or not if any weight is placed on distributional concerns, such as whether inequality is affected by the program, whether some people are harmed by the policy or whether a particular demographic group benefits.

Even a purely welfare-maximizing social planner with no normative concerns for particular demographic groups, inequality or not harming anyone will often need information on program impacts beyond their average. For example, judging a program by its mean impact assumes that the welfare consequences of the distributional aspects of programs are either unimportant or are offset by transfers. As Heckman, Smith and Clements (1997) argue, this assumption is strong. Many outcomes, such as educational attainment and health status, cannot feasibly be redistributed themselves. In order to redistribute the welfare gains derived from such outcomes, one needs to know the relation between the outcomes and individual utility, which can usually at best be approximated. In practice, transfers may be costly and implementing the optimal transfer scheme requires some knowledge of the distribution of gains and losses, i.e., an evaluation that goes beyond the mean impact. Finally, some interventions may work well for particular subgroups of the target population, such as those living in urban areas, while there may be better options for rural populations. Knowing which groups benefit more or less can help improve the targeting of policies and programs and thereby help allocate limited resources more effectively.

The common theme of these issues is that they cannot be addressed by mean impacts alone. Finding answers requires thinking about the impact of a program or policy as a collection of distributional parameters rather than a single scalar parameter such as the mean. Hence, we refer to these types of analyses as “distributional impact analysis” (DIA). DIA concerns the study of the distributional consequences of interventions due to participants’ heterogeneous responses or participation decisions. In particular, DIA investigates features beyond the gross total gain of a program by studying where the gains/losses of a program – if any – were produced, and who wins or loses as a result of program participation. The goal of this toolkit is to provide guidance for researchers who want to use RCTs to answer questions for which the mean is insufficient as an answer. We focus on RCTs because they allow us to simplify the exposition of the methods through the use of randomization as a statistical solution to the selection problems that are central to impact evaluation. Subject to addressing these selection problems, the methods we discuss are applicable to non-experimental analyses as well.[1]

In this toolkit, we study two related approaches to distributional impact analysis. The first approach examines a policy's *impact on the outcome distributions*. It concerns differences between (statistics of) the distribution of outcomes with the policy and the distribution of outcomes without it, such as the impact of a policy on the variance or a specific quantile of the outcome distribution. The second approach is to examine the *distribution of treatment impacts*. This approach answers questions such as what fraction of the population is harmed by the policy or what the bottom quartile of the impact of a policy is. It requires (statistics of) the distribution of the policy gains or losses, i.e., the distribution of differences between the outcomes of a given individual with the policy and without it. What distinguishes these two approaches is that the first focuses on how the intervention affects the distribution of the outcome in the population (e.g., how would the bottom quintile of financial literacy test scores change if we were to provide training on financial literacy for the entire population), but disregards how any particular individual is affected by the program. The goal of the second approach is to analyze the distribution of these individual treatment effects (e.g., how individual gains from financial literacy training vary in the population). Due to this additional ambition, the second approach requires stronger assumptions for identification and estimation.

To simplify the exposition, we focus on questions concerning unconditional distributional impact parameters first. That is, we start by discussing questions regarding treatment effect heterogeneity in the entire population and only introduce additional covariates to make the underlying assumptions hold. In practice, one may also be interested in how treatment impacts differ between observable sub-populations such as males and females, or how they vary with continuous covariates such as age or income. Therefore, we then review ways to extend the methods of both approaches to also allow for heterogeneity within observed subpopulations or along continuous covariates. This allows us to answer questions such as whether men or women are more likely to benefit from a program or whether program impacts are increasing in the baseline outcome. Such conditional analyses usually require additional assumptions and (often substantially) larger samples. If this makes them infeasible, conditional means are usually simple to estimate and can still be informative about heterogeneity.

The benefits of DIA outlined above beg the question why we do not see more of these methods applied in practice. Recent studies that estimate distributional impacts have examined earnings and employment (e.g., Abadie, Angrist and Imbens, 2002; Lechner, 1999), safety net programs (e.g., Black et al., 2003; Bitler, Gelbach and Hoynes, 2006), social experiments (e.g., Djebbari and Smith, 2008; Firpo 2007; Heckman, Smith and Clements, 1997) and education (e.g., Carneiro, Hansen and Heckman, 2003; Cunha, Heckman and Schennach, 2010). However, the literature is still nascent compared to the importance of the topic and the tools available. There are often good reasons to focus primarily on mean impact and the motives to not look beyond it depend on the application. For example, DIA usually requires

[1] See, among others, the overviews in Chernozhukov and Hansen (2013), Heckman and Vytlacil (2007) and Abbring and Heckman (2007).

larger sample sizes, and many of the methods below are justified asymptotically and are not necessarily unbiased in finite samples, which can be problematic in applications with small samples such as most RCTs. Some methods rely on additional assumptions that are reasonable in some applications, but not in others. However, part of the hesitation to apply these methods also seems to stem from inertia and the fact that they are relatively new or have only recently become computationally feasible. Inertia may be an obstacle if researchers or their audiences are more used to and hence comfortable interpreting mean impacts and the conditions under which they are valid. To mitigate these issues, we review the practical considerations of statistical inference and power calculations for DIA. Finally, we illustrate select DIA methods and what can be learned from them by re-analyzing the impacts of a financial literacy program in Brazil and a school management program in The Gambia.

We focus on the types of questions introduced in the previous paragraphs, but there are many other parameters of interest, and questions that go beyond mean impacts. The methods and questions we discuss are not meant to be exhaustive, but are a selection to illustrate how they can complement analyses of means in assessing the value or desirability of a policy or program. We intend to provide a set of baseline methods, to discuss common problems and how to detect them, and to shed light on practical issues such as required sample sizes and estimation of standard errors in (potentially dependent) samples. Programs to implement the methods are available online.[2]

This paper is organized as follows. Part 2 introduces notation and distinguishes the types of DIA questions we examine: We first introduce unconditional analyses of impacts on outcome distributions and the distribution of treatment effects. We then point out how these two approaches extend to questions of conditional heterogeneity. Next, we discuss key methods for each approach. Part 3 considers questions on the impact on outcome distributions. Part 4 considers questions on the distribution of treatment effects. Part 5 considers conditional analyses to answer questions of how treatment impacts differ with observables. Part 6 considers practical considerations for RCTs related to statistical inference and power calculations. Part 7 presents our applications.

2. Questions of Interest and Definitions

In this part, we introduce notation and discuss the similarities and differences between the two main approaches, *impact on the outcome distributions* and *distribution of treatment impacts.* We then discuss questions that require each approach to be extended to conditional analyses. Unless stated otherwise, we consider an RCT with full compliance with treatment assignment to illustrate each method without concerns about selection issues.[3]

1. Definitions and Notation

Suppose we have a sample of N observations, indexed by i. The sample is randomized into a treatment group (which received the policy of interest) and a control group (which did not). The indicator variable R denotes treatment assignment: $R_i = 0$ if observation i belongs to the control group and $R_i = 1$ otherwise. We focus on binary policies to simplify the exposi-

[2] The codes can be found at https://github.com/worldbank/DIA-toolkit

[3] Technically, this requires treatment status to be independent of counterfactual outcomes; full compliance is neither necessary nor sufficient for independence.

tion, although most methods in this toolkit extend to more complex interventions. The indicator variable T denotes treatment participation: $T_i = 0$ if observation i did not receive the treatment and $T_i = 1$ otherwise. Full compliance with treatment assignment implies that everyone in the treatment group participates and no one in the control group participates, so $T = R$.

We assume that outcomes are continuous.[4] We write Y_0 for potential outcomes without treatment, with cumulative distribution function F_0 and τ-th quantile $q_0(\tau)$. Potential outcomes under treatment are Y_1, with CDF F_1 and τ-th quantile $q_1(\tau)$.[5] Each of these potential outcomes is defined for all individuals, regardless of their treatment status, which allows us to precisely define counterfactuals such as $F_0(\, y \mid T = 1\,)$, the distribution of outcomes for the treated if they had not received treatment. We observe outcome Y, which is Y_0 if $T = 0$ and Y_1 otherwise. Formally, $Y = (1 - T)\times Y_0 + T\times Y_1$.

Following the practice of most RCTs, we assume the population of interest is the subpopulation of individuals who apply for the program under evaluation. For this reason, estimates from RCTs are often considered to be impacts "on the treated" and we follow this practice. This terminology only depends on what population one considers the available sample to represent, and not on which individuals from this sample participate or the methods used. In practice, there is usually a second selection step, which is compliance with treatment assignment.[6] We consider methods to adjust for this second selection step in Sections 3.3 and 3.4. These methods can be extended to the methods in Sections 4 and 5 or to correct for selective application to participate in the RCT. For mean impacts, researchers often settle for intent-to-treat parameters instead, to avoid further assumptions. Methodologically, this approach extends to many DIA parameters. However, when analyzing heterogeneity, intent-to-treat parameters confound heterogeneity in take-up[7] with heterogeneity in treatment effects. It will usually be more informative to analyze these two sources of heterogeneity separately. Compliance with treatment assignment is observed, so heterogeneity in compliance can be analyzed using the standard methods used to study program take-up. In this toolkit, we focus on heterogeneity in treatment effects, i.e., parameters of treatment on the treated.

2. Two Unconditional Approaches

In terms of *impact on the outcome distributions*, consider questions such as:

- Does microfinance boost average incomes?
  To answer this question, we would estimate the average treatment effect, $\mathbb{E}(Y_1) - \mathbb{E}(Y_0)$, or the average treatment effect on the treated, $\mathbb{E}(Y_1|T = 1) - \mathbb{E}(Y_0|T = 1)$.
- Does hospital regulation raise minimum levels of patient safety?

[4] While these methods can be applied using continuous, discrete or binary outcomes, they add little in the binary case since the distribution has only two points of support and is completely characterized by its mean.

[5] Formally, $q_D(\tau|\cdot) \equiv \arg\inf_x\{\mathrm{P}(Y_D \leq x|\cdot) \geq \tau\}$ for $\tau \in (0,1)$.

[6] Consequently, "on the treated" remains slightly ambiguous, as it could also be defined based on actual treatment receipt within the given sample or population, i.e., in terms of T.

[7] Note that this is take-up given participation in the RCT. For predictions about policy impacts in the population, we need the unconditional take-up model unless the RCT sample is representative of the entire population.

To answer this question, we would estimate the treatment effect on the minimum, $\min Y_1 - \min Y_0$, or on a quantile in the left tail, such as $q_1(0.1) - q_0(0.1)$.

- Does education reform decrease dispersion of student's test scores?
  To answer this question, we could estimate the treatment effect on a measure of inequality, such as the variance, $\text{var}(Y_1) - \text{var}(Y_0)$, or the Gini index.

The answer to the first question is based on the mean, a particular feature of the outcome distribution that is the focus of most conventional approaches to evaluate policies. However, randomization identifies the full distribution of treated and untreated outcomes, F_0 and F_1. R is independent of Y_0 and Y_1, so that $F_0(y_0|R = 0) = F_0(y_0)$ and $F_1(y_1|R = 1) = F_1(y_1)$. Consequently, under full compliance, we can estimate the distribution of treated outcomes, F_1, from the treatment group and the distribution of untreated outcomes, F_0 from the control group. Thus, we can compare different features of the treatment and control distributions to answer the remaining questions. We can measure the impact of a policy on the median, minimum or low quantiles of the distribution of realized outcomes by taking the difference in the median, minimum or bottom decile of the outcome between the treatment and control group. By the same token, we can measure the impact of an intervention on a particular inequality measure, such as the standard deviation or the Gini coefficient, as in the third question.

Importantly, these methods are *not* informative about how a program's impact varies across individuals, because they ignore who within the population belongs in different segments of the outcome distribution. We may want to study *the distribution of treatment effects* if we are concerned about how policy impacts vary across individuals. Analyses of the distribution of treatment effects, discussed in Part 4, can answer questions like:

- What proportion of students benefit from an educational reform?
  For this question, we would compute $\mathrm{P}(Y_1 > Y_0) = \mathrm{P}(\Delta > 0)$.
- Are the improvements in average patient outcomes from health facility inspections driven by a few people who benefit considerably? Formally: is there significant skewness in effects?
  For these questions, we would compute $\mathbb{E}\{[(Y_1 - Y_0 - \mu)/\sigma]^3\}$, where $\mu = \mathbb{E}(Y_1 - Y_0)$ and $\sigma^2 = \mathbb{E}[(Y_1 - Y_0 - \mu\ )^2]$.
- What is the median impact of a microfinance program? More generally, what are the quantiles of the impact distribution, like the minimum or maximum program impact?
  For this question, we would compute the relevant quantile of treatment effects.

Unlike evaluating policy impacts on outcome distributions, studying the distribution of impacts requires additional, sometimes strong, assumptions about how individuals would fare in a counterfactual treatment state. Even a perfect RCT cannot yield this information since the same person cannot be in both the treatment and control group at the same time.

To further illustrate the difference between the two approaches, consider the following fictional example.[8] Researchers select a class of five students to receive finance lessons. Attendance is mandatory and compliance is perfect. At the end of the program, researchers administer tests to measure financial literacy. Table 1 presents the students' test scores. It also shows what their grades would have been if they had not received the lessons. These

[8] Part 7 presents a real RCT with a similar setup.

counterfactual outcomes are of course unobservable in real life, but we nonetheless show them to clarify a few concepts.

In standard impact analysis, we would compute the average treatment effect (ATE): the difference in means between the two potential outcomes. In our example, it is 1.6. Now consider the treatment effect on the median: the difference in the medians between the distributions of potential outcomes. In our example, it is 2. Why is it larger than the ATE? Because the treatment increased lower scores more than those at the top. If the lessons had instead increased all scores by the same amount, the two effects would have agreed. Note that the difference in medians is also higher than the median of individual effects, 1, due to individual mobility across the distribution. For example, student E would have earned the lowest grade without lessons, whereas E achieved the highest score under treatment.

A little formalism is useful for clarifying this point. Suppose that the potential outcomes (Y_{0i}, Y_{1i}) of observation i correspond to quantiles (τ_{0i}, τ_{1i}) of Y_0 and Y_1, respectively. Then the impact of the policy on individual i is given by:

$$\Delta_i \equiv Y_{1i} - Y_{0i} = q_1(\tau_{1i}) - q_0(\tau_{0i}).$$

This individual impact can never be estimated, since even an ideal RCT will only provide an estimate of either $q_1(\tau_{1i})$ or $q_0(\tau_{0i})$. We can rewrite the impact on individual i as:

$$\Delta_i = \underbrace{q_1(\tau_{1i}) - q_0(\tau_{1i})}_{\text{quantile treatment effect at } \tau_{1i}} + \underbrace{q_0(\tau_{1i}) - q_0(\tau_{0i})}_{\text{mobility effect}}.$$

Note that $q_0(\tau_{0i})$ is a counterfactual outcome for individual i corresponding to the τ_{0i}-quantile of F_0. The first difference is a *quantile treatment effect* from the first approach, which is observable. The second difference is a *mobility effect*, which captures the change in outcomes due to the movement of individuals to different quantiles within the same distribution.

This equation clarifies why quantile effects are only equal to individual treatment effects if everyone keeps the same rank in F_0 and F_1. That is, the mobility effect is zero if $\tau_{0i} = \tau_{1i}$ for all individuals, which is called rank invariance. For example, the two black dots in Figure 1, show the potential outcomes for a particular person in the treated (Treatment) and untreated states (Control). This person is hurt by the treatment despite the fact that most of the population benefits, because her relative rank in the distribution is much lower in the treatment group than in the control group. While RCTs identify the quantile treatment effect, unfortunately, the mobility effect can *never* be recovered directly from the data without additional assumptions. This is again because we can never observe the same person in both the treatment and control group at the same time. Even if we observe the same person in the

Table 1: Mock Dataset

Student	Outcome absent treatment	Outcome under treatment	Difference in potential outcomes
A	1	2	1
B	2	4	2
C	3	4	1
D	4	4	0
E	1	5	4

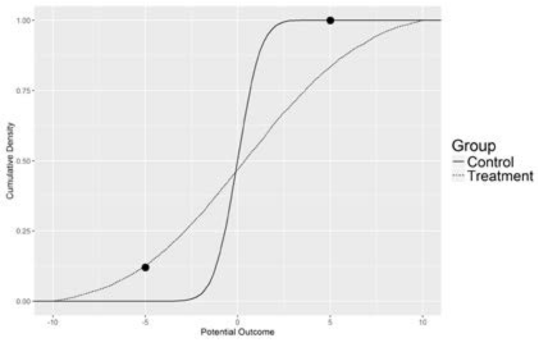


Figure 1: Potential Outcomes and an Individual-Specific Treatment Effect

treatment and in the control group over time, we must still make an assumption about how untreated outcomes vary over time and how they depend on prior treatment status. The key difference between studying impact distributions in Part 4 and impacts *on* distributions in Part 3 is modeling how an individual's relative performance changes when they are in the treatment group instead of the control group, i.e., modeling the mobility effect in the above decomposition. Once we have a credible model of the relationship between τ_{1i} and τ_{0i} and the right data, we can identify the entire distribution of impacts!

3. From Unconditional to Conditional Analysis

The discussions so far concern treatment effect heterogeneity in the entire population. We are often also interested in how treatment effects vary with observable characteristics, such as gender, income or baseline outcomes. As a few concrete examples, consider questions like:

- Are men or women more likely to benefit from a program?
- Are the returns to a financial literacy training program higher for males or for females?
- Are schools with higher baseline test scores more likely to benefit from school management training?

Part 5 provides tools to analyze such questions by discussing methods to study heterogeneity across and, to a lesser extent, within subpopulations defined by observable characteristics. It is important to note that these questions are about how treatment effects correlate with observable characteristics, not necessarily the causal impact of these observable characteristics on the treatment effects. To be sure, the average causal impact of treatment is still identified among subgroups, so long as treatment is randomly assigned within the subgroup. However, the subgroup's characteristics, such as being poor or female, are not randomly assigned. Therefore, if we find that a program's impacts are greater among the poor, we cannot conclude that the treatment effects are great *because* the people are poor. For example, an omitted variable, like neighborhood of residence, may drive the observed correlation.

Nonetheless, answering such questions can provide useful information to policymakers. For example, understanding how treatment effects vary with observable characteristics suggests how to best target a program to maximize its impact (e.g., Manski, 2004) and contributes to the understanding of how particular subgroups of interest, like women or the poorest families, respond to the program. It may also shed light on underlying mechanisms (e.g., Pitt, Rosenzweig and Hassan, 2012) and better predict a program's effects in a population with different characteristics from the original experimental population (e.g., O'Muircheartaigh and Hedges, 2014).

3. Impact on Outcome Distributions

1. Introduction

Policymakers often worry about aggregate treatment effects. For example, they may wish to raise average income, decrease inequality or fight poverty. For such purposes, changes in the level and the shape of the marginal outcome distribution matter more than individual responses to the intervention. This part provides tools to quantify these effects.

There are different approaches to this task. We may analyze simple statistics, such as the mean, an inequality index or a poverty line. This strategy has the advantages of parsimony and familiarity. For a more detailed picture, we may consider shifts in the CDF or its quantiles. We discuss quantiles for the sake of concreteness, but the methods in this part extend to other quantities under minimal conditions and with minimal adjustments.

Quantile treatment effects are the difference between the quantiles of potential outcomes. In graphical terms, they measure the horizontal distance between outcome distributions (Firpo, 2007). We formally define quantile treatment effects on the treated (QTT) as:

$$\Delta_{\mathrm{QTT}}(\tau) \equiv q_1(\tau|T=1) - q_0(\tau|T=1),$$

where $q_D(\tau|T=1)$ is the τ-th quantile of potential outcomes Y_D for the treated.[9] We only observe $q_1(\tau|T=1)$ in the data. The remainder of this part surveys three approaches to recover the counterfactual quantile $q_0(\tau|T=1)$ from untreated observations, which in turn will allow us to estimate $\Delta_{\mathrm{QTT}}(\tau)$.[10]

We focus on unconditional treatment effects. These effects are particularly relevant for policy evaluation, because they capture changes in the dispersion of outcomes both between and within subgroups of the population (Firpo, Fortin and Lemieux, 2009). Since the pioneering work of Koenker and Bassett (1978), a related literature has explored conditional quantile regression (CQR). CQR estimates quantile effects within subgroups under restrictive assumptions. Part 5 discusses conditional effects. Note that unlike average effects, unconditional quantile effects are not weighted averages of subgroup effects.

This part continues as follows. Section 3.2 considers RCTs without endogenous selection. This perfectly randomized setup ensures that the distributions of potential outcomes do not depend on treatment status, so that randomization identifies QTTs without further assumptions. However, many applications deviate from this benchmark. For example, participants might not comply with treatment assignment. The literature offers a wealth of strategies to account for sample selection, as it does for average treatment effects. Section 3.3 surveys inverse probability weighting, which exploits the assumption of selection on observables.

[9] In analogy to average treatment effects, we can also define the quantile treatment effect (QTE): $q_1(\tau) - q_0(\tau)$. See Firpo and Pinto (2016) for additional discussion.

[10] In practice, it is unfeasible to analyze all infinite points of a continuous distribution. Therefore, we limit our endeavor to particular quantiles of interest. If outcomes are discrete, in principle we could measure treatment effects at each value of the support.

Section 3.4 presents an instrumental-variable approach. Section 3.5 concludes with remarks on the interpretation of quantile effects.

2. Randomized Control Trials Without Endogenous Selection

This section considers randomized trials without selection problems, i.e., we assume:

i. Potential outcomes, (Y_0, Y_1), are jointly independent of treatment receipt, T.

This is a strong assumption, which is only likely to hold under idealized conditions. The probability of assignment to treatment must be equal for all participants.[11] Moreover, the probability of treatment take-up must be constant. This condition requires full compliance with treatment assignment ($T = R$) or random noncompliance. We must also observe outcomes for all participants with the same probability. Should there be nonparticipation or nonresponse, they must be independent of treatment status. For simplicity, we refer to RCTs which satisfy assumption (i) as RCTs with perfect compliance or RCTs without endogenous selection.

Assumption (i) ensures that the counterfactual distribution of potential outcomes Y_0 of the treated group, $F_0(y|T = 1)$, is the same as that of the untreated group, $F_0(y|T = 0)$. Therefore, a consistent estimator of the QTT is simply the difference in quantiles between observed outcomes of treated and untreated units:

$$\hat{\Delta}_{\text{QTT}}(\tau) = \hat{q}_1(\tau|T = 1) - \hat{q}_0(\tau|T = 0),$$

where $\hat{q}_D(\tau|T)$ is any consistent estimator of the τ-th quantile of outcomes Y_D for group T, such as the empirical quantile. This estimator is straightforward to extend to other objects, such as the variance or the Gini coefficient: it suffices to take the difference between the statistics for treated and untreated observations. Note that we may also estimate $F_1(y|T = 1)$ from the treatment group and $F_0(y|T = 1)$ from the control group under assumption (i).[12]

It is also possible to recover the QTT from a quantile regression, in the same way that linear regression yields the average treatment effect. To do so, run a quantile regression of observed outcomes, Y, on a constant and a treatment indicator, T. The slope gives the QTT, whereas the intercept gives the quantile $\hat{q}_0(\tau|T = 0)$.

Assumption (i) is strong and fails in many applications. For example, the probability of assignment to treatment may depend on participants' attributes, such as region of residence. The equivalence between the distributions of potential outcomes then breaks down, leading to composition bias. For the same reason, endogenous selection into treatment is a concern when participants do not comply with treatment assignment. Similar to average effects, we may either settle for intent-to-treat effects or account for selection and estimate effects on the treated under alternative frameworks, as the next sections show.

[11] If the probability of assignment to treatment differs across individuals by design, reweighting the sample yields consistent estimators. See Section 3.3.

[12] With random non-compliance, one should include those with $R = 1$ & $T = 0$ in the control group and those with $R = 0$ & $T = 1$ in the treatment group to estimate these distributions.

3. Applications with Selection on Observables: Inverse Probability Weighting

In the ideal RCT of Section 3.2, treated and untreated observations have the same distribution of potential outcomes. Many applications deviate from this benchmark, though. This section considers the weaker assumption of selection on observables.[13] Selection on observables is an assumption of conditional independence between potential outcomes and treatment status,[14] which relaxes our previous assumption of unconditional independence. As the name suggests, we postulate that treatment take-up only depends on observed variables. Therefore, treated and untreated participants differ only in observed characteristics (other than treatment status). Correcting imbalances in covariates should restore the equivalence between outcome distributions and allow us to estimate treatment effects.

This intuition motivates two classes of estimators: inverse probability weighting (IPW) and matching. IPW consists of reweighting the sample to balance covariates across treated and untreated observations, which should then have the same distribution of potential outcomes.[15] Hirano, Imbens and Ridder (2003) propose IPW for average treatment effects, which Firpo (2007) extends to quantile effects. Donald and Hsu (2014) consider CDFs, whereas Firpo and Pinto (2016) discuss inequality indexes.[16] Matching consists of pairing observations according to covariates and differencing outcomes to estimate treatment effects. See Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) for surveys. Matching does not readily extend from average to distributional effects, because it relies on the law of iterated expectations (Frölich, 2007), which fails for quantiles and other nonlinear statistics.

The consistency of IPW estimation of distributional effects relies on the same assumptions and formulas as average effects. It requires:[17]

i. **Selection on observables:** potential outcomes, (Y_0, Y_1), are jointly independent of treatment status, T, given observed covariates, X.
ii. **Common support:** $0 < \mathrm{P}(T = t|X = x) < 1$ for all x and t.

Assumption (i) is the key identification assumption. Assumption (ii) is crucial, albeit standard. Matching requires that covariates take the same value range in the treated and untreated groups. Imbens (2015) presents methods to assess and enforce common support in the data.

We implement the IPW estimator in two steps. First, we compute the weight function:

$$\omega_{\mathsf{IPW}}(T, X) = \frac{T}{\mathrm{P}(T = 1)} + \frac{1 - T}{\mathrm{P}(T = 1)} \times \frac{\mathrm{P}(T = 1|X)}{1 - \mathrm{P}(T = 1|X)}.$$

Here, $\mathrm{P}(T = 1)$ is the unconditional probability of treatment take-up, which we can estimate with the share of the treated in the sample. The conditional probability $\mathrm{P}(T = 1|X)$ is the propensity score, a building block of many estimators of average effects. See Appendix 2 for a discussion of estimation.

[13] We abstract henceforth from nonparticipation and nonresponse.

[14] This assumption is also known as unconfoundedness, ignorability and conditional independence.

[15] DiNardo, Fortin and Lemieux (1996) use reweighting to decompose changes in the density of wages in the U.S. in an early and influential paper. However, they do not investigate the properties of their estimator.

[16] Cattaneo (2010) considers multivalued treatments.

[17] Additional regularity conditions are required. See Firpo (2007, p. 263).

To build intuition for $\omega_{\text{IPW}}(T,X)$, remember that we wish to estimate the quantile treatment effect on the treated. We can still estimate $q_1(\tau|T=1)$ from the treated. Thus, the function $\omega_{\text{IPW}}(T,X)$ gives them equal weight, $1/\mathrm{P}(T=1)$. To recover $q_0(\tau|T=1)$ from the untreated, however, their distribution of potential outcomes must be comparable to that of the treated. Hence, we reweight this subsample. By the assumption of selection on observables, we only need to balance the distribution of X. For that reason, $\omega_{\text{IPW}}(T,X)$ is increasing in the propensity score: we give more weight to untreated observations which resemble the treated and less weight to observations with characteristics that are uncommon among the treated.

After constructing $\widehat{\omega}_{\text{IPW}}(T,X)$,[18] we proceed as above. The QTT estimator is the difference between the quantiles $\hat{q}_1(\tau|T=1)$ and $\hat{q}_0(\tau|T=0)$ of the reweighted sample, which we compute separately for treated and untreated observations. As before, this approach extends to other statistics. It also accommodates alternative treatment effects, such as effects on the entire population: it suffices to adjust the weight function (Firpo and Pinto, 2015).

It is also possible to estimate quantile effects with quantile regressions. One should run a regression of outcomes on an intercept and a treatment indicator on the reweighted sample. Covariates should only enter the model through the weights $\widehat{\omega}_{\text{IPW}}(T,X)$. If they were included as control variables, we would estimate a conditional quantile effect (see Part 5).

The assumption of selection on observables is strong and has received extensive discussion in the program evaluation literature. The covariate set must be rich enough that the unexplained component of treatment take-up is independent of potential outcomes. This approach is popular in observational studies, and may also be useful in many RCTs, since designers often collect extensive data on participants' backgrounds. Imbens (2015) suggests placebo tests to assess plausibility. The assumption becomes testable if a valid instrument is available: see Donald, Hsu and Lieli (2014) for details.

4. Applications with Selection on Unobservables: Instrumental Variables

The previous section leveraged the assumption of selection on observables to address selection bias. This section explores an alternative strategy: instrumental variables (IV).

IV estimation of average effects has a long tradition: see Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) for surveys. Imbens and Rubin (1997) and Abadie (2002) extend it to effects on distributions. Important early contributions for quantile effects are Abadie, Angrist and Imbens (2002) and Chernozhukov and Hansen (2004, 2005). Frölich and Melly (2013a,b) propose estimators of unconditional distributional effects.

As Angrist and Imbens (1994) argue, the IV framework only identifies treatment effects on compliers – the subpopulation whose treatment status depends on the value of the instrument. Further assumptions are necessary for identification of effects on the treated. Frölich and Melly (2013a) exploit one-sided noncompliance: for some value of the instrument, participants never take the intervention. For intuition, consider a clinical trial of a vaccine. Our candidate instrument is randomization. Participants in the treatment group might refuse the vaccination. On the other hand, the control group has no access to it. Noncompliance is one-sided in that it only affects the treatment group.[19] This property implies that all the

[18] If the data include sampling weights, multiply $\widehat{\omega}_{\text{IPW}}(T,X)$ by the weights (Ridgeway et al., 2015).

[19] See the analysis of Head Start by Kline and Walters (2015) for a counter-example. Their paper highlights the importance of verifying the assumption of one-sided non-compliance in practice.

treated are compliers, which allows us to identify effects on the treated from effects on compliers.

This section presents the framework of Frölich and Melly (2013a). We assume the existence of a binary instrument Z, such that $Z = 0$ implies $T = 0$. One example is randomization.[20] Formally, we assume:[21]

i. **Independent instrument:** Y_0 is independent of Z given covariates, X, for all x such that $\mathrm{P}(T = 1|X = x) > 0$.
ii. **One-sided noncompliance:** $\mathrm{P}(T = 0|Z = 0) = 1$.
iii. **Support condition:** $\mathrm{P}(Z = 0|X = x) > 0$ for all x such that $\mathrm{P}(T = 1|X = x) > 0$.

If the assumption of one-sided noncompliance fails, this procedure yields consistent estimates of effects on treated compliers. It is also possible to obtain local treatment effects. See Frölich and Melly (2013b) for details.

Condition (i) ensures that the instrument is valid. We need full independence, which is stronger than the assumption of uncorrelatedness of the linear IV model. In an RCT, it means that Y_0 does not depend on treatment assignment, which is reasonable. Condition (iii) is analogous to the assumption of common support in the previous subsection. We implement the IV estimator in two steps. First we compute the weight function:

$$\omega_{\mathrm{IV}}(T, X, Z) = \frac{T}{\mathrm{P}(T = 1)} + \frac{1 - T}{\mathrm{P}(T = 1)} \times \frac{\mathrm{P}(Z = 0|X) - Z}{\mathrm{P}(Z = 0|X)}.$$

Note the similarity to $\omega_{\mathrm{IPW}}(T, X)$. Appendix 2 discusses the estimation of the conditional probability $\mathrm{P}(Z = 0|X)$.

To build intuition for $\omega_{\mathrm{IV}}(T, X, Z)$, suppose that the instrument is randomization, i.e., $Z = R$. The distribution of potential outcomes is initially balanced across the treatment and control groups, due to random assignment. Noncompliance distorts the distribution for the treated, which now consists of compliers. The untreated include noncompliers from the treatment group, as well as counterfactual compliers and noncompliers from the control group. To recover $q_0(\tau|T = 1)$, we would ideally restrict the untreated to counterfactual compliers, but we do not know who they are. Note however that the distribution of Y_0 for noncompliers is the same in the treatment and control groups, because of randomization. Therefore, giving negative weights to the outcomes of noncompliers from the treatment group makes them “cancel out” counterfactual noncompliers in the control group, leaving us with the distribution of Y_0 for counterfactual compliers!

Accounting for covariates accommodates applications in which the instrument is only conditionally exogenous, such as RCTs with stratified randomization and observational data. We might also want to purge indirect effects of an intervention to focus on particular mechanisms. Although controlling for covariates undoes the equivalence between the treated and compliers, the identification result of Frölich and Melly (2013a) holds nonetheless.

After constructing $\omega_{\mathrm{IV}}(T, X, Z)$,[22] we proceed as above. The QTT estimator is the difference between the quantiles $\hat{q}_1(\tau|T = 1)$ and $\hat{q}_0(\tau|T = 0)$ of the reweighted sample, which

[20] The estimator extends to multivalued and continuous IVs: see Section 3.1 in Frölich and Melly (2013a).

[21] Additional regularity conditions are required. See Frölich and Melly (2013a, p. 391). In particular, we assume that there is no endogenous attrition. The authors propose adjustments for attrition in their Section 4.

[22] If the data include sampling weights, multiply $\hat{\omega}_{\mathrm{IPW}}(T, X)$ by them (Ridgeway et al., 2015).

we compute separately for treated and untreated observations.[23] This approach extends to other statistics, such as the variance.

5. Interpreting Quantile Effects

Quantile effects are often misinterpreted, which can result in unwarranted conclusions for policy. This section discusses and illustrates two common pitfalls: implicit assumptions of rank invariance and extrapolating treatment effects to different populations.

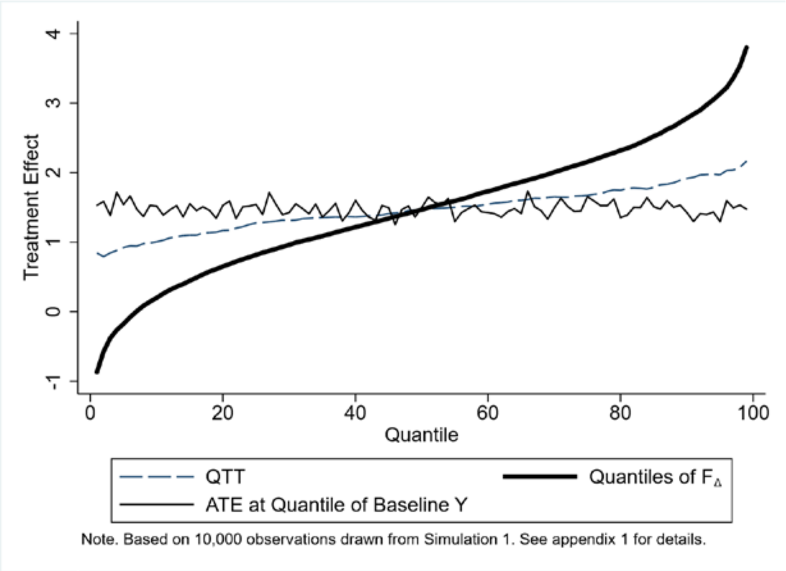
It is easy to conflate quantile effects (i.e., changes in the distribution of outcomes) and individual effects (i.e., the distribution of changes in outcomes). One might naively reason: "The median outcome in the control group is 50. The difference in medians is 10. John is in the control group and his outcome is 50. Therefore, his outcome would be 60 under treatment." Recall that the quantile effect is the difference in quantiles between treated and untreated participants. Thus, we implicitly assumed that John's outcome would be equal to the median of the treated if he underwent treatment himself – in other words, *rank invariance* (cf. Part 2). Rank invariance is a strong assumption, which is implausible when treatment effects differ across observably identical participants.

Assumptions of rank invariance are often subtle. For example, researchers might argue that all participants benefitted from the intervention if all quantiles effects are strictly positive. They would have invoked rank invariance across treatment status. One might also equate quantile effects and changes with respect to baseline outcomes instead of Y_0. This interpretation requires rank invariance both across treatment states and over time.

For illustration, consider Figure 2. It shows estimates of treatment effects, based on data from Simulation 1 (see Appendix 1 for details). We set $S = 2$, so that the first period is the baseline period. Individual effects are independent of baseline outcomes; as a consequence, the average treatment effect at each quantile of baseline outcomes is close to the overall average effect, as can be seen from the line labeled "ATE at Quantile of Baseline Y". Yet, quantile effects are increasing. The QTT (dashed line) clearly differs from both average effects at baseline values (thin line) and quantiles of the distribution of individual effects F_Δ (thick line). This discrepancy underlines the notion that implicitly assuming rank invariance and misinterpreting quantile effect may lead to incorrect conclusions. Note, moreover, that all quantile effects are positive, even though some participants are worse off after treatment.

A different concern is the comparison of treatment effects across (sub-)populations. This part has surveyed the estimation of changes in the distribution of potential outcomes. As Section 4.2.2 argued, RCTs identify these marginal distributions, which allows us to compute treatment effects under minimal assumptions. However, these effects depend on the unidentified joint distribution of (Y_0, Y_1). As a consequence, it is difficult to extrapolate or compare results between different (sub-)populations, unless we can account for differences in both the distributions of Y_0 and the relation between Y_0 and Y_1. For example, suppose that we observe differences in QTTs across genders. By itself, this finding does not tell us whether these discrepancies arise from differences in responses to the intervention between men and women or from differences in outcomes in the absence of treatment. In other words, they can stem from gender differences in the unknown distribution of treatment effects, in the distribution of Y_0 or both. We can estimate the marginal distribution of outcomes, which

[23] As before, it is possible to estimate the QTT from a weighted quantile regression. The slope gives the QTT.



Note. Based on 10,000 observations drawn from Simulation 1. See appendix 1 for details.

Figure 2: QTTs and Common Misinterpretations

allows us to compare treatment effects on quantiles that correspond to the same value of Y_0. See the discussion of Figures 6 and 7 in Bitler, Hoynes and Domina (2014) for an example and strategies to compare quantile effects across groups.

4. Distribution of Treatment Impacts

1. Introduction

1. Definitions and Outline

Part 3 discusses methods to estimate the impact of a policy or program on (functions of) the marginal distribution of outcomes. In this part, we are interested in the distribution of individual specific treatment effects, F_Δ. The distribution of treatment effects is required to answer questions such as:

- What is the variance of treatment effects?
- What was the median impact of the program?
- What proportion of the population was hurt by the program?

As described in Part 2, these questions cannot be answered by the methods in Part 3, because, rather than effects on distributions, they concern effects on individuals. The central difficulty of studying the distribution of individual treatment effects is that they require a counterfactual outcome for every individual. Suppose person i is in the treatment group, so we observe Y_1. In order to estimate the impact of treatment on person i, we need to predict Y_{0i}, say by $\hat{Y}_0$. Then, the individual treatment effect is the difference between the observed outcome in the treatment state and the counterfactual control group outcome *for person* i: $\hat{\Delta} = Y_1 - \hat{Y}_0$. Similarly, for person j in the control group, the researcher needs to predict Y_{1j}, the outcome of person j if j were treated, in order to construct $\hat{\Delta} = \hat{Y}_1 - Y_0$.

Importantly, the individual estimates $\hat{\Delta}_i$ are only of interest to identify the distribution of treatment effects in order to answer questions such as those raised above. The estimated effect on individual i, $\hat{\Delta}$, is of less interest both because it is noisily estimated and because this individual has already been treated, so this treatment effect provides little information about future implementations of the policy. In Part 5, we discuss methods to analyze heterogeneity, which is informative about the types of people who benefit from a program.

In contrast to studying average treatment effects or differences in marginal distributions as in Part 3, individual counterfactuals cannot be identified by randomization alone. Randomly selecting treatment and control groups identifies F_1 and F_0 but not how the outcomes of a single individual vary across the treatment or control states. In general, the marginal distributions and quantile treatment effects from Part 3 inform us of changes in the frequency of outcomes and inequality, but they only provide limited information about idiosyncratic responses to treatment (Bitler, Gelbach and Hoynes, 2014).[24] The distribution of impacts only equals the difference in marginal potential outcome distributions (or quantile effects) when individuals maintain exactly the same rank in both the treatment and control outcome distributions. This *rank preservation* or *invariance* condition implies that observations with the same rank in the treatment and control outcome distributions are appropriate counterfactuals for each other. When the rank invariance condition is satisfied, estimating the distribution of treatment effects only requires the methods discussed in Part 3. Rank invariance is a strong assumption. If it does not hold, the parameters from the previous part are still identified, but their interpretation can be difficult. The interpretation of the distribution of treatment effects is always clear, but the distribution is no longer identified by randomization if rank invariance fails. As a result, we must either make additional, sometimes strong, assumptions that imply individual counterfactuals to point identify the distribution of treatment effects, or we must settle for partial identification, where only a range of parameters is identified.

The empirical strategy depends on the validity of assumptions that we need to assess for the case at hand, so we first provide some background and then discuss more general estimation principles before we discuss a specific model that applies to many evaluations. In particular, we first illustrate the identification problem using the variance of treatment effects as an example. In Section 4.2, we discuss partial identification and how additional assumptions can narrow the bounds from this approach. Section 4.3 discusses point identification. We first illustrate the required assumptions and provide an overview of common approaches to justify them. We then introduce methods to estimate features (moments) of the distribution or, under more stringent assumptions, the entire distribution. In practice, we need to adapt the methods to justify the assumptions and the estimation method to our application and data availability, so we close by giving an overview of a specific panel data model. The model is general enough to cover many common settings and can serve as a blueprint which is adaptable to other situations.

2. *An Example: Variance of Treatment Effects*

The vast majority of impact evaluations focus on average treatment effects: $\mathbb{E}(\Delta) = \mathbb{E}(Y_1 - Y_0)$. This is the first moment of the treatment effect distribution, which provides a measure of location of the distribution, i.e., how much individuals benefit on average. If

[24] As Bitler, Gelbach and Hoynes (2014) point out, quantile effects contain additional information about the distribution of individual effects. If one or more quantile effects are positive, at least one participant benefited from the intervention. The converse is also true. Note that Makarov bounds allow us to quantify the shares of winners and losers under minimal assumptions: see Section 4.3.

treatment effects are constant, it completely describes the effects of the program. However, when there is heterogeneity, a natural extension is to study higher order moments of the distribution. For example, the variance of treatment effects is the second (centered) moment of this distribution and provides a measure of the dispersion of the treatment effects,[25] i.e., how much they vary across individuals. In the fictional RCT in Part 2, the variance of treatment effects summarizes how the impact of financial literacy training varies across students. The variance of treatment effects is a useful measure of the importance of heterogeneity. For instance, if the square root of the variance is close to the average treatment effect, some individuals are likely to be harmed by the program. The variance of individual treatment effects is:

$$\mathrm{var}(\Delta) = \mathrm{var}(Y_1) + \mathrm{var}(Y_0) - 2\,\mathrm{cov}(Y_0, Y_1).$$

The variances of Y_1 and Y_0 are features of their marginal distributions and can be estimated using the methods discussed in Part 3. However, $\mathrm{cov}(Y_0, Y_1)$ requires the researcher to know how Y_1 relates to Y_0. Unfortunately, we can never observe the same person in both the treatment and control states simultaneously. As a result, the data do not identify $\mathrm{cov}(Y_0, Y_1)$ or, by extension, $\mathrm{var}(\Delta)$, without additional assumptions.

Consider the mock data in Table 1, but more realistically, suppose we do not know which outcomes under treatment are paired with outcomes in the absence of treatment. As usual, we can calculate the average treatment effect by taking the difference between the average outcome in the treatment and control groups,[26] which yields 1.6. Similarly, for the variance of treatment effects, we get:

$$\widehat{\mathrm{var}}(Y_1 - Y_0) = \widehat{\mathrm{var}}(Y_1) + \widehat{\mathrm{var}}(Y_0) - 2\,\mathrm{cov}(Y_0, Y_1) = 1.36 + 0.94 - 2\,\mathrm{cov}(Y_0, Y_1).$$

Thus, the variance of treatment effects is not identified from the data alone. We face an analogous identification problem when estimating the entire distribution of individual treatment effects or other features of this distribution: while features of the marginal distributions F_0 and F_1 can be calculated from the data, parameters that describe the relationship between Y_0 and Y_1 are necessary for point identification of the distribution of treatment effects. We discuss two ways to proceed below. First, we can settle for bounds instead of point estimates as discussed in Section 4.2. Bounds are often simple to obtain, but inference can be difficult. They also tend to be too wide to be informative and narrowing them requires additional assumptions. Second, we can estimate the (parameters of the) relationship between counterfactual outcomes, as we discuss in Section 4.3. This requires further assumptions and modeling, such as amending the RCT with a model of participation choice.

2. Bounding the Distribution of Treatment Effects

While the data cannot identify higher order moments of the treatment effects distribution without additional assumptions, they can identify a range of values that must include the true value. Bounds do not tell us anything regarding where in the range the parameter is

[25] Here, the variance of the treatment effect means the variance of individual treatment effects in the population. This is distinct from the variance of the estimate of the average treatment effect due to sampling that is used for inference on average effects.

[26] To keep the example simple, we abstract from sampling variation throughout.

likely to lie, but they rule out parameter values outside this range, because the data are inconsistent with them. For example, if we can find the largest and smallest possible value of $\mathrm{cov}(Y_0, Y_1)$, we can plug these values into the formula above to obtain bounds on the variance of individual treatment effects, i.e., the largest and smallest possible values it can take. We first continue the example of the variance and then extend this idea to the distribution of treatment effects and discuss its advantages and problems.

Recall that $\mathrm{cov}(Y_0, Y_1) = \rho_{01}\sigma_1\sigma_0$, where ρ_{01} is the correlation between Y_1 and Y_0, σ_1 is the standard deviation of Y_1 and σ_0 is the standard deviation of Y_0. We can estimate σ_1 and σ_0 from the treatment and control group. The only remaining unknown is the correlation coefficient, ρ_{01}, which must lie between -1 and 1 by definition. Therefore, without additional assumptions,

$$-\sigma_1\sigma_0 \leq \mathrm{cov}(Y_0, Y_1) \leq \sigma_1\sigma_0.$$

Substituting these bounds for $\mathrm{cov}(Y_0, Y_1)$ in the formula for $\mathrm{var}(\mathit{\Delta})$ yields bounds for $\mathrm{var}(\mathit{\Delta})$. Using the data in Table 1, the bounds are:

$$1.36 + 0.94 - 2\sqrt{1.36\times 0.94} \leq \widehat{\mathrm{var}}(Y_1 - Y_0) \leq 1.36 + 0.94 + 2\sqrt{1.36\times 0.94},$$
$$0.04 \leq \widehat{\mathrm{var}}(Y_1 - Y_0) \leq 4.56.$$

What can we learn from these bounds? At the upper bound, the average treatment effect is only 1.3 standard deviations from zero, so the data do not seem to rule out negative treatment effects. Since the lower bound is greater than zero, there has to be at least some treatment effect heterogeneity. In this way, bounds can also be used to test relevant hypotheses. For example, the classical standard approach to impact evaluation assumes that treatment effects are constant, so $\mathrm{var}(\mathit{\Delta}) = 0$. If the bounds do not include zero, as above, the data are not consistent with constant treatment effects (Heckman, Smith and Clements, 1997).[27] In Simulation 1, where treatment effects are normally distributed with unit mean and variance, the bounds on the standard deviation of treatment effects tell us that it must be between 0.78 and 3.13. Therefore, the bounds rule out a constant treatment effect, but include estimates over three times higher than the true standard deviation.

The bounds can be tightened using additional assumptions. The results in Heckman, Smith and Clements (1997) suggest assuming that potential outcomes are positively correlated ($\rho_{01} \geq 0$) may be reasonable in some cases. If we are willing to assume that people who do well in the absence of treatment also do well with treatment, the bounds become:

$$0 \leq \mathrm{cov}(Y_0, Y_1) \leq \sigma_1\sigma_0.$$

Substituting these bounds into the formula for $\widehat{\mathrm{var}}(Y_1 - Y_0)$ yields the narrower bounds:

$$1.36 + 0.94 - 2\sqrt{1.36\times 0.94} \leq \widehat{\mathrm{var}}(Y_1 - Y_0) \leq 1.36 + 0.94$$
$$0.04 \leq \widehat{\mathrm{var}}(Y_1 - Y_0) \leq 2.3.$$

[27] To conduct a formal hypothesis test, the researcher needs to calculate standard errors for the bounds. We return to this problem at the end of this section.

The data in Table 1 actually imply $\rho_{01} = 0.21$. Therefore, the true variance of the treatment effect is 1.84, which lies within these bounds.

How does the example extend to learning about the distribution of treatment effects without making assumptions beyond what was required in Part 3? Just as in the case of the variance above, one can still use the marginal outcome distributions to calculate bounds on the entire distribution of treatment effects. Assume we have estimates of the marginal potential outcome distributions $F_1(\cdot\,|X)$ and $F_0(\cdot\,|X)$ from the treatment and control groups. Then the distribution of treatment effects, $F_\Delta(d) = P(\Delta \leq d)$, can be bounded at any point d using the following Makarov Bounds (Makarov, 1982; Firpo and Ridder, 2008):[28]

$$\sup_t \max\{F_1(t) - F_0(t-d), 0\} \leq F_\Delta(d) \leq \inf_t \min\{F_1(t) - F_0(t-d) + 1\,, 1\}.$$

Bounds may also be informative about specific questions of interest. For example, marginal potential outcome distributions are sometimes sufficient to find whether anyone was hurt by the program or the share of treatment effects that are negative, i.e., $F_\Delta(0)$. The Makarov bounds indicate the range of values for the joint distribution that are consistent with the observed marginal distributions. We break down the pieces of the lower bound as an illustration. The maximum function in the lower bound just imposes the logical restriction that a CDF cannot be negative. If we ignore the maximum for now, the bound simplifies to:

$$\sup_t F_1(t) - F_0(t),$$

which is just the largest vertical difference between the treatment and control group CDFs.

As with the lower bound, the minimum in the upper bound imposes the restriction that the CDF can be no greater than 1. Ignoring the minimum, the upper bound becomes:

$$\inf_t F_1(t) - F_0(t) + 1,$$

which is determined by the point where the treatment group looks best in comparison to the control group.

The first panel of Figure 3 illustrates Makarov Bounds for the two data generating processes described in simulation 2. In both cases, the average treatment effect is 1, but the standard deviation of treatment effects is 0 in the first case and 5 in the second case. Despite the fact that treatment effects are constant in the first case, the bounds do not rule out sizeable heterogeneity. On the other hand, the bounds in the second case clearly rule out constant treatment effects.

The second panel of Figure 3 plots the densities of outcomes generated from the second data generating process. The treated outcome density has a slightly higher mean than the untreated density, but is much more dispersed. Consequently, a non-zero share of the treated density's mass falls below the minimum value of the untreated potential outcome density. This indicates that *someone* was hurt by the program in this hypothetical example. We cannot say who was hurt by the program, but we know that the proportion of people hurt by the program is at least $F_1[\min(Y_0)]$. In this simulation, this bound indicates that at least 26% (and at most 59%) of individuals were hurt by the treatment. In fact, based on the simulated data, 40% of individuals were hurt by the treatment.

[28] Refer to Firpo and Ridder (2008) for a formula that yields narrower bounds by averaging across bounds from conditional distributions. Note that we have simplified the notation by assuming continuous CDFs.

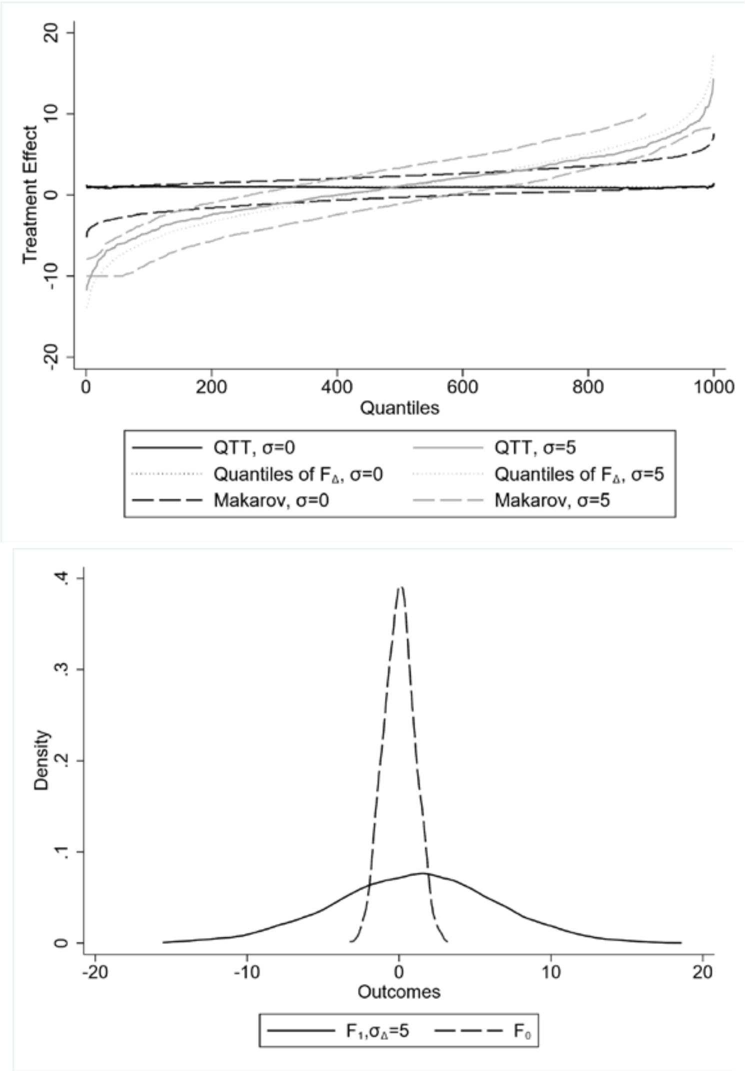


Figure 3: Simulation 3, DGP and Makarov Bounds with $\sigma_{\Delta} = 5$ and $\sigma_{\Delta} = 0.5$

In these examples, we have ignored sampling variation, i.e., that the marginal distributions are estimates rather than the true population distributions. Sampling variation exacerbates the problem that bounds are often wide (Heckman, Smith and Clements 1997), since estimation error makes the possible range of parameters even wider than the point estimates of the upper and lower bounds. Moreover, obtaining the relevant standard errors is often difficult (see Subsection 6.2.1). Nonetheless, bounds are usually easy to calculate and clearly demonstrate what the data imply about the distribution. Thereby, they can provide a useful informal assessment of what can, and (more often) what cannot, be learned from the data alone. For example, if the point estimates of the variance bounds include zero, the data are not informative about heterogeneity of treatment effects without further assumptions.

3. Point Identification of Features of the Distribution of Treatment Effects

When we are willing to make additional assumptions, we can make more progress in point identifying the distribution of treatment effects. We first continue the example of the variance of treatment effects and illustrate what kind of assumptions are required for point identification and what previous studies have done to justify them. The conditions under which these assumptions are plausible crucially depend on the substantive problem and the available data. We use cross-sectional notation to keep the exposition simple. In practice, one can easily adapt the methods to richer models, such as panel data models, by including individual fixed effects or lagged values in X_i. To allow researchers to choose and adapt the methods to their cases, we discuss two general estimation frameworks in the next subsection. We conclude with an example of a specific panel data model that can easily be adapted to many common empirical settings.

Continuing the example of the variance of treatment effects, we can point identify $\operatorname{var}(\Delta)$ if we are willing to assume a particular value of the correlation between the treated and the untreated outcome. For example, if potential outcomes are uncorrelated $\operatorname{cov}(Y_0, Y_1) = 0$. In our mock example, this implies: $\widehat{\operatorname{var}}(Y_1 - Y_0) = 1.36 + 0.94 = 2.3$. As we have seen above, the true covariance in our example is positive, so assuming that it is zero leads to an overestimate of the variance of treatment effects.

More generally, we need to be able to identify the dependence between the treated and untreated outcomes to identify the distribution of treatment effects. So far, we have primarily dealt with extreme cases (such as no or perfect dependence), in which the researcher explicitly chooses values for the parameters that determine the dependence of potential outcomes. While additional assumptions are always required, a much better case for them can be made in practice for two reasons. First, the researcher may have a model that includes the relevant dependence parameters and may be able to estimate them from this model under weaker conditions. For example, $\operatorname{cov}(Y_0, Y_1)$ is a crucial parameter of models of individual choice such as the (generalized) Roy model, which can be estimated under conditions outlined in Heckman and Honoré (1990) and extended in Abbring and Heckman (2007). Thus, rather than drawing a convenient value from thin air as we do here for ease of exposition, we may be able to estimate the required parameters under plausible assumptions by amending the RCT with a model of individual choice. Second, the assumptions are usually only required to hold conditional on observables, i.e., after controlling for X, as Subsection 4.3.2 describes.

1. *Identification and Key Assumptions*

Identification of the distribution of treatment effects comes from restrictions of the dependence between the part of the untreated outcome that is not explained by the covariates and the size of the treatment effect. In this subsection, we discuss why these assumptions are necessary, how they solve the identification problem and why covariates are important even in an RCT. Throughout this part, we assume perfect compliance, so that one can use R and T interchangeably. We define parameters and estimators in terms of T here, as the distribution of individual "intent-to-treat" parameters is unlikely to be interesting and is not clearly defined. If perfect compliance fails, one needs to adapt the re-weighting or IV methods from part 3. For simplicity, we assume that the effect of covariates is linear and additively separable:

$$\begin{aligned} Y_0 &= X\beta + \varepsilon, \\ Y_1 &= Y_0 + \Delta = X\beta + \Delta + \varepsilon, \end{aligned}$$

where Δ is an individual specific impact of the treatment. This can conveniently be written in one equation as:

$$Y = X\beta + \Delta T + \varepsilon.$$

This model is simple to extend: e.g., including non-linear functions of X is straightforward. It is also more general than the resemblance to a standard regression equation suggests. In impact evaluation, we are interested in Δ and not in the effect of X on Y_0. Thus, one can think of β as the linear projection coefficient, so that X and ε are uncorrelated by construction (but not necessarily independent). If Δ depends on X, the control group still identifies β, so that the treatment group identifies the (not necessarily causal) relation of the treatment effect to observables. Thus, as long as randomization works, we can purge the observable part of the model by partial regression, i.e., by regressing Y on X (and XT if treatment effects depend on observables) and working with residuals from this regression as if X does not matter.[29]

In practice, including X in the model and estimating it in one step may be more convenient than partial regression. However, partial regression provides a useful thought device, as it leaves only the unobservable part of the model, which consists of $\Delta + \varepsilon$ for the treated and ε_i only for the control group: $\Delta T + \varepsilon$. This is a benefit of randomization that helps to identify the distribution of treatment effects. To see that it is not sufficient, consider the analogy to identifying mean effects, where one would compute the mean of ε from the control group, the mean of $\Delta + \varepsilon$ from the treatment group, and obtain the mean of Δ as their difference. Extending this to distributions, the treatment group identifies the distribution of the sum of Δ and ε and the control group identifies the distribution of ε. However, one cannot back out the distribution of Δ from these two distributions: contrary to means, the difference between two distributions is not the distribution of the differences.

To make progress, additional restrictions on the dependence of ε, the part of the untreated outcome that is not related to the covariates X, and the treatment effect are required. We saw above that, if we are willing to assume that individuals' potential outcomes are uncorrelated across treatment states, then the variance of treatment effects can be point identified. Technically, we have achieved identification by imposing a moment restriction. All three terms in the formula for the variance of treatment effects above are second moments of the data and the covariance is the only moment that depends on the joint distribution. Restricting it to zero leaves only terms that we can estimate, since the other two terms, the variances of Y_0 and Y_1, only depend on the marginal distributions. Similarly, in the conditional case, we can identify the variance of individual treatment effects if $\operatorname{cov}(\Delta, \varepsilon) = 0$, since:

$$\operatorname{var}(\Delta) = \operatorname{var}(\Delta + \varepsilon) - \operatorname{var}(\varepsilon) - 2\operatorname{cov}(\Delta, \varepsilon) = \operatorname{var}(\Delta + \varepsilon) - \operatorname{var}(\varepsilon).$$

We can estimate the first term using the treatment group and the second from the control group. This idea generalizes. If we are willing to assume that all third moments that are not determined by the marginal distributions are zero, for example, the third moment of the distribution of treatment effects is identified. We provide more detail on moment estimation in Subsection 4.4.1. The limiting case of this idea is to assume that *all* moments of the joint distribution only depend on the moments of the marginal distribution. This implies that treatment effects and the unexplained part of the untreated outcome are independent condi-

[29] Note that if randomization has been compromised, e.g., by noncompliance, the same has to be done for T.

tional on X. Then the entire distribution of treatment effects can often be estimated using the method of deconvolution, as discussed in Subsection 4.4.2.

2. Justifying Identification Assumptions

The discussion above shows that identification of (features of) the distribution of treatment effects requires independence assumptions. How can we justify such assumptions? They are satisfied if treatment effect heterogeneity is unrelated to any unobservable aspects of individual i, which underscores the importance of covariates. Unlike methods relying on randomization, conditioning on covariates is typically required for identification here. The covariates control for variation in Y_0 that is potentially correlated with impact heterogeneity. Consequently, the required assumptions to estimate (moments of) the distribution of treatment effects may be plausible when the data include a rich set of individual characteristics. This is because the conditional independence assumption allows Δ to depend on observable characteristics of person i but not on any unobservable characteristics or variables that have been excluded from the model. Such a restriction seems more plausible the better the model of untreated outcomes is. As a simple example, when the model contains no covariates, $\varepsilon = Y_0$, so assuming Δ and ε are uncorrelated amounts to the (usually unrealistic) assumption that levels ($Y_0 = \varepsilon$) and gains (Δ) are not related. However, RCTs in economics often collect detailed information including baseline values. Including baseline outcomes changes this restriction to assuming that gains from the program are unrelated to deviations of the outcome from its expected path. While the independence assumption usually seems unrealistic in cross-sectional applications, it often seems more plausible in panel data settings.

Consequently, a key component of making the conditional independence assumption credible is to attempt to control for all variables that are related to both the size of the treatment effect and the untreated outcome. Subject to the usual caveats (see, e.g., Angrist and Pischke, 2009), this suggests controlling for many characteristics in a flexible way and assessing the robustness of results to changes in the conditioning variables. However, just as in a standard regression, regardless of data availability and modeling, whether the assumption can be justified or not depends on the application at hand.

If we find the assumption that Δ is conditionally independent of ε too strong, we may turn to several tools from the literature to enhance credibility. Aakvik, Heckman and Vytlacil (2005) show that this assumption can be weakened by taking a random effects factor approach. Suppose the data are generated by:

$$Y = X\beta + \Delta T + \theta + \varepsilon,$$

where θ is a vector of all unobserved variables which are potentially correlated with Δ but are independent of X and T. If we can control for θ, the distribution of Δ can be recovered using deconvolution. Aakvik, Heckman and Vytlacil (2005) assume that θ is an individual specific random effect with a particular distribution and estimate the model using maximum likelihood. If panel data are available, we can model θ using a less restrictive fixed effects approach, as in the example in Subsection 4.4.4 below.

If we have access to multiple related measures of important omitted confounders (or proxies), we can estimate the distribution of θ using a *measurement system* (Carneiro, Hansen and Heckman, 2003). To be concrete, suppose θ represents ability. If we observe three test scores, we may be willing to assume:

$$\begin{aligned} M_1 &= \theta + u_1, \\ M_2 &= \alpha_2\theta + u_2, \\ M_3 &= \alpha_3\theta + u_3 \end{aligned}$$

where θ, u_1, u_2 and u_3 are all mutually independent and identically distributed across individuals and $\mathbb{E}(u_{1i}) = \mathbb{E}(u_{2i}) = \mathbb{E}(u_{3i}) = 0$. Then, by Kotlarski (1967), $F_\theta(\cdot)$ is nonparameterically identified, which implies $F_\Delta(\cdot)$ can be recovered using deconvolution. The supplemental appendix of Arcidiacono et al. (2011) provides more detailed instructions on how to use Kotlarski's Theorem to identify $F_\Delta(\cdot)$. The recent literature on dynamic factor models (e.g., Cunha, Heckman and Schennach, 2010) combines this approach with the advantages of panel data.

4. Estimation Methods

This section presents two approaches to identifying and estimating features of the distribution of treatment effects. In Subsection 4.4.1, we extend our discussion about estimating the variance of treatment effects to higher order moments of the distribution. This first approach relaxes some of the assumptions required to estimate the full distribution of treatment effects. It is therefore preferable for questions that can be answered by moments of the distribution of treatment effects. For example, what is the variance of treatment effects? In Subsection 4.4.2, we show how to use deconvolution to estimate the entire distribution of treatment effects, $F_\Delta(\cdot)$, when our model sufficiently captures the variation in the untreated outcome that is related to treatment effect heterogeneity, so that Δ_i and ε_i are independent. All of the policy questions raised at the beginning of this section can be answered by computing features of this distribution. Moreover, it is simple to calculate the variance of the estimated distribution to assess the variability of treatment effects or the fraction of individuals hurt by the training, $\hat{F}_\Delta(0)$. We present simulation results to illustrate these two methods in Subsection 4.4.3.

We continue to use our basic model from the previous section to keep the discussion simple. However, caution is warranted when applying the methods in practice, as the required conditional independence assumption is strong. This assumption likely requires a richer underlying model and may thereby limit how useful these methods are in practice. Subsection 4.4.4 presents additional details of these methods in a specific panel data context where the conditional independence assumption may be plausible.

1. Identifying and Estimating Moments

As the example of the variance shows, moments of the distribution of treatment effects are identified under weaker conditions and often provide sufficient information to answer policy questions.[30] The first two moments of a distribution imply its variance and the first three moments imply its skewness. Skewness is informative about how lopsided the distribution is. For example, a high and negative skewness indicates that there is an important number of individuals whose treatment effect is not far above the mean impact and a few individuals with treatment effects far below the mean impact. With many vaccines and medications,

[30] As a reminder, the k-th moment of the distribution of treatment effects is defined as $\mathbb{E}(\Delta^{\mathrm{k}})$.

one may be worried about very lopsided distributions where most people benefit modestly, but a few people are severely harmed. Examining whether the skewness is large and negative provides evidence on this issue.

While moments are usually easy to estimate by their sample analog, $\mathit{\Delta}$ is never observed, so we cannot estimate its moment directly. However, we can calculate the residuals from a partial regression of Y on X for the treatment and control group. Recall from above that if covariates are linearly related to Y_0 and Y_1, the residual from a partial regression of Y on X corresponds to $\mathit{\Delta} + \varepsilon$ for treatment group observations and ε for control group observations.

We can calculate the variance of treatment effects from the first two moments of the treatment effects distribution, which we can estimate as moments of our partial residuals:

$$\mathrm{var}(\mathit{\Delta}) = \mathbb{E}(\mathit{\Delta}^2) - \mathbb{E}(\mathit{\Delta})^2 = \mathbb{E}[(\mathit{\Delta} + \varepsilon)^2] - \mathbb{E}(\varepsilon^2) - \mathbb{E}(\mathit{\Delta})^2.$$

The first two terms are the second moments of treated and untreated partial residuals and $\mathbb{E}(\mathit{\Delta})$ is just the average treatment effect. Thus, each of these expectations can be estimated using partial residuals. This is possible because we have assumed that $\mathit{\Delta}$ and ε are uncorrelated, so $\mathbb{E}(\mathit{\Delta}\varepsilon) = 0$. This is just a conditional version of the assumption that $\mathrm{cov}(Y_0, Y_1) = 0$ that we used to identify the variance of treatment effects in our example above.

The skewness of the distribution of treatment effects is given by:

$$\mathrm{skewness}(\mathit{\Delta}) = \frac{\mathbb{E}(\mathit{\Delta}^3\ ) - 3\mathbb{E}(\mathit{\Delta})\,\mathrm{var}(\mathit{\Delta}) - \mathbb{E}(\mathit{\Delta})^3}{\mathrm{var}(\mathit{\Delta})^{3/2}}.$$

Once we know the variance of treatment effects, the only unknown term in this equation is the third moment, $\mathbb{E}(\mathit{\Delta}^3)$. To estimate it from our partial regression residuals, we need to extend our assumption that terms which do not depend solely on the marginal distributions of potential outcomes are zero to $\mathbb{E}(\mathit{\Delta}\varepsilon^2)$ and $\mathbb{E}(\mathit{\Delta}^2\varepsilon)$. This is a restriction on how the variance of the error term is related to the treatment effect and vice versa. Under this assumption, the third moment of the distribution of treatment impacts is given by:

$$\mathbb{E}(\mathit{\Delta}^3) = \mathbb{E}[(\mathit{\Delta} + \varepsilon)^3] - 3\mathbb{E}(\mathit{\Delta})\mathbb{E}(\varepsilon^2) - \mathbb{E}(\varepsilon^3),$$

Thus, we can estimate the skewness of the distribution of treatment effects from moments of our partial residuals. Appendix 3 shows how to use the binomial formula to solve for higher order moments.

In theory, we can use this method to estimate *all* moments (or $K = \infty$). This requires *all* moments that are not determined by the marginal distributions to be zero. Of course, estimating infinitely many moments is impractical. Thus, one may prefer to estimate the first K moments and select one of the (usually many) distributions they are consistent with. For example, in Section 7.2, we estimate the entire distribution of treatment effects by assuming they are normally distributed and by plugging in mean and variance estimates. Wu and Perloff (2006) estimate the first four moments and recommend selecting the distribution according to the Principle of Maximum Entropy. This distribution is unique and is "maximally noncommittal with regard to missing information" (Jaynes, 1957; Wu and Perloff, 2006). See their paper for details.

2. Deconvolution

Deconvolution methods estimate the distribution of treatment effects by removing the variation due to the error term from treatment group observations so the only remaining variation is due to the treatment effect. Disentangling the variation of the treatment effect from that of the error term requires a strong conditional independence assumption. More precisely, conditional independence requires that the random component of Y conditional on (X, R), $\mathit{\Delta} R + \varepsilon$, is the sum of two independent random variables for the treatment group. The distribution of the sum of two random variables is called the convolution of these random variables. Deconvolution methods undo this convolution by removing the distribution of one of the random variables from the sum. In our simple model, deconvolution requires $\mathit{\Delta}$ to be statistically independent of ε given X.[31]

A testable implication of these assumptions is that:

$$\operatorname{var}(Y_1|X) = \operatorname{var}(\mathit{\Delta}) + \operatorname{var}(\varepsilon) \geq \operatorname{var}(\varepsilon) = \operatorname{var}(Y_0|X).$$

This can be tested using the marginal outcome distributions. If $\operatorname{var}(Y_1)$ is not greater than $\operatorname{var}(Y_0)$, we should be wary of using this approach. Note that $\operatorname{var}(Y_1)$ is greater than $\operatorname{var}(Y_0)$ in all of the simulations in Subsection 4.4.3 below, even though the underlying assumptions are violated in some cases.

Deconvolution is often implemented via estimated characteristic functions (e.g., Bonhomme and Robin, 2010). Here, we present a simpler algorithm by Mallows (2007), which Arellano and Bonhomme (2012) find works well in practice. It estimates $F_{\mathit{\Delta}}(\cdot)$ by randomly matching treatment group residuals with control group residuals for the full sample many times. To build intuition for this algorithm, consider again the partial residuals from above. For members of the control group, they contain just the error term, ε. For someone in the treatment group, the partial residual is the sum of the treatment effect and the error term, $\mathit{\Delta} + \varepsilon$. Since $\mathit{\Delta}$ and ε are independent by assumption, we can create pseudo-draws of $\mathit{\Delta}$ to approximate its distribution using:

$$\hat{\mathit{\Delta}} = \widehat{\mathit{\Delta} + \varepsilon} - \hat{\varepsilon},$$

where $\hat{\varepsilon}$ is a control-group partial residual and $\widehat{\mathit{\Delta} + \varepsilon}$ is a treatment-group partial residual. Mallows' Algorithm "shrinks" estimates of $\hat{\mathit{\Delta}}$ by matching treatment- and control-group partial residuals under the assumption that large values of treatment group partial residuals are associated with larger than usual error terms and small treatment group partial residuals are associated with smaller than usual error terms. The algorithm then draws a new pseudo-sample of treatment residuals by randomly matching its shrinkage estimate of $\hat{\mathit{\Delta}}$ with residuals and estimating a new shrinkage distribution.

Mallows' Algorithm is described in more detail for the conventional linear model below:

$$Y = X\beta + \mathit{\Delta} T + \varepsilon\,.$$

[31] The model also requires additive separability and restrictions on the dependence of X and $\mathit{\Delta}$ or ε as discussed above. However, these assumptions are specific to the simple model that we use for clarity here. They can potentially be relaxed, e.g., when many periods of pre- and post-treatment observations are available, one can use flexible panel data models, such as in Jacobson, LaLonde and Sullivan (1993).

To better distinguish treatment and control partial residuals, we will follow Arellano and Bonhomme's (2012) notation: $A = \Delta + \varepsilon$, $B = \Delta$ and $C = \varepsilon$. Note that $A = B + C$. To be sure, A and C are just vectors of partial residual estimates, whereas B is a vector of simulated draws from the (unknown) distribution of treatment effects.

To prepare the data for Mallows' Algorithm, use estimates from the conventional linear model above to form A and C using treatment and control group values of $Y - X\hat{\beta}$, respectively.[32] Then, run the algorithm as follows.

1. Set $B_0 = \mathrm{sort}(A - C)$.
2. Let $\tilde{B}_0$ be a random permutation of B_0.
3. Let $\tilde{A}$ denote A sorted according to the order of $\tilde{B}_0 + C$.
4. Set $B_k = \mathrm{sort}\left(\tilde{A} - C\right)$. Repeat steps 2 through 4 many times.[33]

Each B_k is a pseudo-sample drawn from $F_\Delta(\cdot)$. The empirical distribution of the full set of B_ks is therefore an estimate of the distribution of $F_\Delta(\cdot)$. An example program is in the supplemental material available online.

3. Simulation Results

Table 2 shows the results of Simulation 3. We present results using 6 different combinations of parameters which determine the correlation between treatment effects and omitted variables and the independence of treatment effects from potential outcomes. The moment and deconvolution estimation methods always yield the same average effect, but the differences are particularly pronounced in the respective estimates of the standard deviation of treatment effects. As we would expect, both the deconvolution and moment methods perform relatively better when their assumptions are satisfied, which is the case for deconvolution in column one only and for the moment-based approach in column two only. In our simulation setup, the moment approach yields standard deviation estimates, which are closer to the truth when an omitted variable is negatively correlated with treatment effects, regardless of whether the covariance satisfies the moment or deconvolution assumptions. In contrast, the deconvolution standard deviation estimates are a better approximation of the truth when the omitted variable is positively correlated with treatment effects.

[32] The full C vector or complete replications of the vector should be used in the algorithm since it is mean zero by construction. A and C must be the same size. If C is larger than A, replace A with m bootstrap draws with replication where m is the length of C. If A is larger than C, either replace A with m bootstrap draws with replication or with $2m$ draws and replace C with $[C' \quad C']'$.

[33] Arellano and Bonhomme (2012) repeat this procedure 2000 times and discard the first 500 iterations. In practice, the largest number of iterations that computing and time constraints allow is desirable.

Table 2: Results from Simulation 3

Correlation with Omitted Variable	None		Negative		Positive	
Y_0 independent of Δ	Yes	No	Yes	No	Yes	No
A. Summary Statistics of Actual Effects						
Mean	1.05	1.06	− 0.38	− 0.36	1.52	1.54
Standard deviation	0.70	1.22	1.63	1.92	0.86	1.32
Skewness	0.06	0.06	0.02	− 0.01	0.12	0.12
B. Summary Statistics from Moments						
Mean	1.14	1.16	− 0.34	− 0.32	1.64	1.65
Standard deviation	1.40	1.38	1.63	1.62	1.63	1.62
Skewness	0.05	0.05	–0.01	− 0.02	0.04	0.03
C. Summary Statistics from Deconvolution						
Mean	1.14	1.16	− 0.34	− 0.32	1.64	1.65
Standard deviation	0.69	0.72	1.09	1.10	1.09	1.10
Skewness	0.03	0.03	− 0.03	− 0.04	0.03	0.03

Notes: Estimates based on Simulation 3, with σ_{01} equal to 0.5 or 0 and c equal to − 1, 0, or 1. When σ_{01} is 0.5, treatment effects are independent of potential outcomes, so the deconvolution assumptions are satisfied. When σ_{01} is 0, the moments assumptions are satisfied. When c is 0.5, there is no omitted variable. When c is 1 or − 1, there is an omitted variable which is positively or negatively correlated with both treatment outcomes and treatment effects, respectively.

4. A Specific Application with Panel Data

In practice, we likely need to extend the simple model from above and adapt the methods to a specific setting. To provide some guidance for this, we next discuss the panel data setting from Arellano and Bonhomme (2012), which can easily be adapted if the assumptions above do not seem plausible or the available data set has a different structure. The conditional independence assumption required for deconvolution is more likely to be satisfied when the researcher can control for a rich set of individual characteristics. Panel data are a special case that also allows to flexibly control for unobserved individual heterogeneity.

As discussed previously, the central difficulty in studying the distribution of treatment impacts is that a counterfactual outcome is needed for every individual in the sample. In a panel, the same individual can often be observed in the treated and control states at different times. This allows one to model individual heterogeneity quite flexibly, but instead requires restrictions on how the treatment effect and the residual are allowed to vary over time, as we show below. Arellano and Bonhomme (2012) show how to estimate higher order moments of the treatment effect distribution and apply deconvolution methods to estimate the full distribution of treatment effects when panel data are available.

Consider the panel version of the linear model we work with throughout this section:

$$Y_{is} = \alpha_i + X_{is}\beta + \Delta_i T_{is} + v_{is}, \qquad s = 1, \ldots, S.$$

Call this model (1). Here, T_{is} is an indicator for whether person i has ever been treated by period s. Outcomes are shifted by the treatment effect Δ_i in all periods after an individual is treated, so this model implies that the treatment effect is constant over the time horizon of the panel.[34] This may be a strong assumption, since treatment effects may not begin immediately and/or may fade over time. For example, when analyzing the impact of a school management intervention on students' test scores, this amounts to assuming the intervention had constant impacts over the entire period of data collection (i.e., three years in the application in Section 7.2). With very long panels, this can be relaxed by modeling decay parametrically. A more general model is:

$$y_{is} = \alpha_i + X'_{is}\beta + \delta_s + \Delta_i T_{is} + \rho_i(s - s_i^*)T_{is} + v_{is}, s = 1, \dots, S.$$

where s_i^* is the period in which individual i received treatment. Here, the treatment effect is assumed to evolve linearly over time according to the (non-stochastic) linear equation $\Delta_{i,s-s^*} = \Delta_i + \rho_i(s - s_i^*)$.

Arellano and Bonhomme (2012) show that the variance and distribution of Δ_i in model (1) are identified under the following two assumptions:

i. **Conditional independence:** Δ_i is statistically independent of v_{is}.
ii. **Error dynamics:** The model requires restriction of the dependence of errors over time, so that we can estimate the time series process of v_i. See Arellano and Bonhomme (2012) for further detail. In their application, they assume that the errors follow an autoregressive or moving-average process with independent and identically distributed innovations in each period.

Note that the conditional independence assumption is now conditional on an individual fixed effect and the time varying controls, X_{is}. The only remaining source of potentially problematic residual variation is time-varying unobservables. While this is still a strong independence assumption, it is likely more plausible than in the cross-sectional setting. The second assumption requires that the v_{is} are not too correlated across periods. This is required to ensure that the treatment effect can be disentangled from the persistent component of error terms. Importantly, the relationship between Δ_i and α_i is left unrestricted.[35]

With panel data, the identification of the distribution of individual treatment effects relies primarily on how outcomes of treated individuals vary between treatment and control states.[36] While this allows us to estimate particular Δ_i, we are primarily interested in the distribution of treatment effects, as these individual estimates will be quite noisy in most applications and most policy relevant questions relate to future realizations of treatment effects. Therefore, estimating the distribution of treatment effects using these models requires the data to have several features. Here, we will focus on model (1) under the assumptions of Arellano and Bonhomme (2012). In particular, we assume someone is "treated" if they have ever received the treatment. In order to identify this model, individuals must be observed at least three times, including at least one period before and another after the individual receives the treatment. Suppose individual i is observed three times and receives the treatment in the third period. Then, her observed outcomes are given by:

$$Y_{i,s} = \alpha_i + \delta_s + X'_{is}\beta + v_{is},$$

[34] Alternatively, we could define T_{is} as an indicator for being treated in period s. In this case, treatment is assumed to affect the individual only in treated periods.

[35] In non-experimental settings, the relationship between Δ_i, α_i, and T_{is} is unrestricted.

[36] Control group variation aids estimation of common parameters that vary over time.

$$Y_{i,s+1} = \alpha_i + \delta_{s+1} + X'_{i,s+1}\beta + v_{i,s+1},$$
$$Y_{i,s+2} = \alpha_i + \Delta_i + \delta_{s+2} + X'_{i,s+2}\beta + v_{i,s+2}.$$

Consider the variation in the data needed to identify each of the above parameters. First, observe that α_i and Δ_i are individual-specific, so that only individual i's observations are informative about their values. If not for δ_s and β, the parameters that are common across individuals, α_i and β_i could be estimated separately for each individual. For this reason, Δ_i is only identified for the subpopulation of individuals who are observed in both the treatment and control states. In contrast, untreated observations are informative about the common parameters, δ_s and β.

If, instead, individual i were treated in the second and third periods, her observed outcomes would be:

$$Y_{i,s} = \alpha_i + \delta_s + X'_{is}\beta + v_{is},$$
$$Y_{i,s+1} = \alpha_i + \Delta_i + \delta_{s+1} + X'_{i,s+1}\beta + v_{i,s+1},$$
$$Y_{i,s+2} = \alpha_i + \Delta_i + \delta_{s+2} + X'_{i,s+2}\beta + v_{i,s+2}.$$

In principle, the distribution of Δ_i is identified if each individual is only treated in one period or, even in a cross-section, if the errors v_{is} are independent and identically distributed. This assumption implies that the residual from any individual j at any time period s, v_{js}, could be used as a counterfactual for the residual of individual i at time s. If one only uses a cross-section, this reduces to the assumption that residuals from another individual at the same time, v_{js}, are valid counterfactuals for v_{is}. Exploiting panel data to use only within individual variation, as we do here, relaxes this assumption to require only residuals from the same person, $v_{is'}$, to be valid counterfactuals for v_{is} for $s' \neq s$.

We use only individual i's observations to estimate Δ_i, so consistency is in the number of treated and untreated periods, not the number of individuals. This is illustrated in Figure 4 below which is generated from Simulation 1 with $S = 4$ (left panel) and $S = 16$ (right panel). In particular, notice that the distribution of regression estimates of the treatment effects is quite different from the true distribution and the deconvolution estimates when $S = 4$. The estimates are much more similar when $S = 16$.[37] In general, estimates of Δ_i will be noisy when individuals are only treated in one, or even a few, periods. Moreover, estimates of

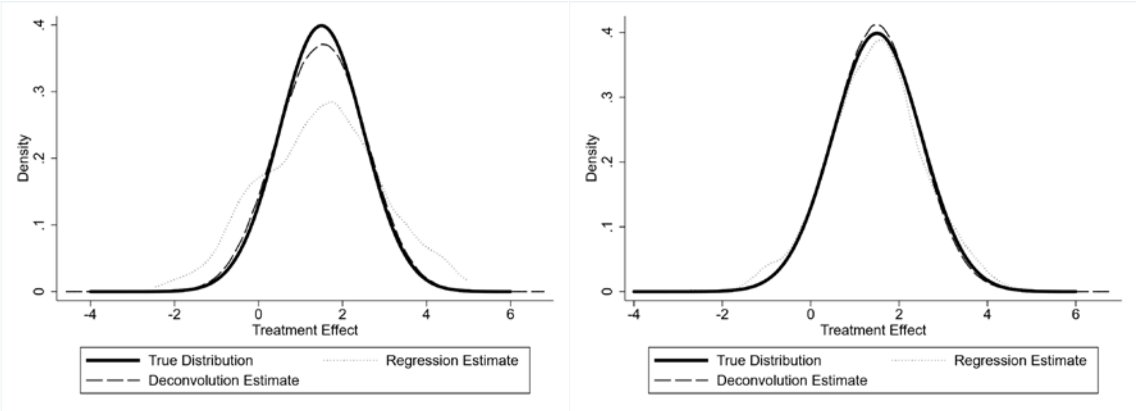


Figure 4: Estimates of Distribution of Treatment Effects (Left: $S = 4$; Right: $S = 16$)

[37] The required number of time periods depends on the SEs of Δ_i, so there is no general rule. However, whether the variation due to the estimation error is minimal can easily be examined in any given application.

treatment impacts on a particular individual are not of particular policy interest, since the individual has already received the treatment in question. The distribution of treatment effects is potentially more informative about how individuals are likely to be affected by future applications of a program. Estimates of individual treatment effects may be of interest for analyzing how Δ_i varies by individual characteristics, which we discuss in more detail in Part 5. For this purpose, the individual estimates can be regressed on covariates, but standard errors should be adjusted to account for the fact that the outcomes are themselves estimated.

A less ambitious goal than estimating Δ_i for each treated individual is to estimate the distribution of treatment effects, $F_\Delta(\cdot)$. When S, the number of periods in the panel, is large, the distribution of estimated individual treatment impacts, $\hat{F}_{\hat{\Delta}}(\cdot)$ is likely to be an accurate approximation of $F_\Delta(\cdot)$. In such cases, one may only need to calculate or plot the distribution of the estimated treatment effects. Unfortunately, S is typically relatively small in RCTs. When S is small, each Δ_i is estimated using few observations, so $\hat{F}_{\hat{\Delta}}(\cdot)$ is an inflated version of $F_\Delta(\cdot)$. Therefore, one needs to correct for this estimation error using deconvolution as specified in step 3 below.

When the data requirements and assumptions discussed above are satisfied, this model can be estimated using the following procedure:

1. **Setup:** Set up the data so that there is one observation for each unique individual-time period combination.
2. **Estimate regression:** Regress the outcome on individual fixed effects, an indicator for having been randomly assigned to treatment by round s interacted with individual fixed effects and covariates with common effects across individuals.
3. **Estimate feature of interest:** If interested in the variance of treatment effects, use Arellano and Bonhomme (2012), eq. (47). If interested in the full distribution of treatment effects, estimate $F_\Delta(\cdot)$ by applying Mallows' algorithm using treatment and control period values of $y_{is} - X_{is}'\hat{\beta} - \hat{\alpha}_i$.

Applications of these methods are shown using data from an actual RCT conducted in The Gambia in Section 7.2 and using simulated data in Section 7.3. Example code appears in the supplementary materials available online.

5. Distributional Impacts and Conditional Analyses

1. Introduction

Thus far, we have considered unconditional heterogeneity, i.e., how treatment impacts vary in the population overall. We may also be interested in questions such as whether and how the impact of a program on inequality differs by gender or whether program impacts are increasing or decreasing with baseline income. These questions concern (features of) the joint distribution of treatment impacts and other variables. Thus, as discussed in Part 2, they are not answered directly by the methods above and call for *conditional analysis*. In this part, we discuss how to study conditional treatment effect heterogeneity.

We first describe subgroup analysis, i.e., how the methods above can be applied to subpopulations, such as by gender. Conditional versions of both questions concerning impacts on outcome distributions, such as whether the effect on inequality varies by gender, and questions on the distribution of outcomes, such as whether the fraction of individuals who benefit from the program varies by gender, may be of interest. Section 5.2 shows that for discrete covariates, such as gender, the problem is tractable, but often requires larger sample sizes than those typically available. For continuous variables, such extensions would require estimating the joint distribution of two continuous variables, such as treatment effects and baseline income – which requires even stronger assumptions than the one discussed above and exceptionally large data. However, estimating features of this joint distribution, such as conditional means, may still answer important questions. For example, how would the average treatment effect change if the 1985 age distribution were replaced with the current distribution, holding all else constant (Rothe, 2012)? Therefore, researchers often estimate conditional average treatment effects. We discuss their estimation and interpretation in Section 5.3.

2. Subgroup Analysis

We start with subgroup analysis, i.e., applying the methods discussed above separately to two or more groups of interest. While straightforward in theory, this approach requires common support and large sample sizes. In particular, we can implement exactly the same methods used to study the full sample on each of the different groups of interest. This conditional approach can be applied to a wide range of methods, including average treatment effects, quantile treatment effects and distributions of treatment effects. For example, the QTTs for men and women are given by:

$$\begin{gathered}\Delta_{\text{QTT}}(\tau|\text{gender}=\text{male})=q_1(\tau|T=1,\text{male})-q_0(\tau|T=1,\text{male}),\\ \Delta_{\text{QTT}}(\tau|\text{gender}=\text{female})=q_1(\tau|T=1,\text{female})-q_0(\tau|T=1,\text{female}).\end{gathered}$$

If compliance is not perfect, so that independence of T and (Y_0, Y_1) fails, it is important to take this into account as discussed in Part 3. Just as QTTs are interpretable as the difference between the treatment and control marginal distributions at a particular quantile, the subgroup QTT for, say, males, is the difference in the marginal distributions of outcome *among males* at a particular quantile.[38] These subgroup distributional impacts are of interest because, as in the full sample, means may mask substantial within-group treatment heterogeneity. Note that if the distribution of the outcome variable is very different across groups, quantile treatment effects for each group may look different even if the effects are identical at the same values of the outcome variable. This is due to the fact that the quantiles of the distribution of untreated outcomes, $q_0(\tau)$, differ between the groups: see the discussions in Section 3.5 and around Figure 6 of Bitler, Hoynes and Domina (2014).

In principle, the above approach of estimating parameters separately by subgroup works whenever the subgroup has both treatment and control observations, i.e., *common sup-*

[38] Note that subgroup QTTs differ from the estimates of conditional quantile regression (CQR). CQR models each quantile as a parametric function of covariates. The CQR coefficient on the interaction of the subpopulation indicator and T would indicate how the τ^{th} quantile of y differs between the treated in the subpopulation and the *overall* population (rather than the untreated subpopulation). It assumes that treatment effects are constant across subgroups. Note that conditional quantile effects do not average to the unconditional effect, so that CQR provides little information about treatment effects for the population of participants.

port.[39] However, subgroup analysis requires even larger sample sizes, because separate analyses are being conducted across each subgroup of interest. Thus, it may often be infeasible, particularly when subgroups are defined by several characteristics or continuous measures. Taking baseline income as an example, it would be quite surprising if every unique income level in the study was reported by at least two families, let alone one treatment and one control family. Even at income levels where there are multiple observations, there generally will not be enough observations to yield a precise enough estimate of the mean, not to mention the entire conditional distribution. Adjusting for multiple hypotheses tests as we discuss in Part 6 further increases the required sample size. Of course, we can partially overcome this issue by collapsing continuous variables into a small number of categories, like using income deciles, instead of income directly. This approach gains feasibility at the cost of potentially missing some heterogeneity, for example, within observations in the same income decile. These assumptions may allow learning about heterogeneity, but they can also mask it if they are not chosen well, as we discuss further in Section 5.3.

3. Conditional Average Treatment Effects

In this section, we focus on conditional average treatment effects. In practice, most researchers settle for showing that average effects change with covariates. This is generally not because learning about how the full distribution changes with covariates is not of interest, but because the subgroup sample sizes are too small to estimate separate distributions within each subgroup. However, we may still be able to estimate the average effect of interest using only the relevant subsample. For example, if we are interested in how average impacts vary across men and women, we can simply estimate Conditional Average Treatment Effects (CATEs)[40] separately using the subsamples of men and women:

$$\mathbb{E}(\Delta_i|\text{gender} = \text{male}) = \mathbb{E}(Y|T = 1, \text{male}) - \mathbb{E}(Y|T = 0, \text{male}),$$
$$\mathbb{E}(\Delta_i|\text{gender} = \text{female}) = \mathbb{E}(Y|T = 1, \text{female}) - \mathbb{E}(Y|T = 0, \text{female}).$$

As above, this approach requires every subgroup of interest to have a sufficient number of observations in both treatment and control groups. If perfect compliance fails, T may be endogenous and bias CATEs, just as it biases ATEs. Consequently, reweighting or IV methods are required to recover effects on the treated, as in Part 3. One can also still use randomization instead of treatment status to obtain intent-to-treat parameters. However, for their interpretation it is important to take into account that both take-up and treatment effects may vary between groups. Subpopulation means share most of the advantages and shortcomings of average treatment effects discussed in Part 0, with the added benefit that comparing means across subpopulations can shed some light on heterogeneity. Djebbari and Smith (2008) call heterogeneity explained by differences in subgroup means “systematic heterogeneity” and heterogeneity remaining after controlling for these differences “idiosyncratic heterogeneity”. Subgroup means are therefore informative about systematic heterogeneity, but not idiosyncratic heterogeneity. In some cases, this can lead to incorrect conclusions about

[39] Selecting the relevant subgroups is a critical step in conditional analysis. It is recommended to define the groups of interest at the design stage and include them as part of the data collection. See Section 6.4. Despite extensive planning, however, researchers may not anticipate the most relevant subgroups.

[40] We follow common terminology in calling this parameter CATE rather than “on the treated”, as the discussion in Part 2 suggests.

the nature and extent of treatment effect heterogeneity. For example, Bitler, Gelbach and Hoynes (2014) re-analyze a welfare experiment to investigate the extent to which allowing for heterogeneity in CATEs can explain the heterogeneity in quantile treatment effects in Bitler, Gelbach and Hoynes (2006). Using a simulation exercise, they conclude heterogeneity across subgroups is unable to explain the observed treatment effect heterogeneity.

As pointed out above, estimating how average treatment effects vary with covariates non-parametrically is often infeasible when there are many subgroups or when treatment effects vary with a continuous variable. These problems can be mitigated by making parametric assumptions on the treatment effect heterogeneity. We often have prior beliefs that suggest a certain functional form for treatment effect heterogeneity. For example, we may be willing to assume treatment effects vary linearly or quadratically with income. At the risk of misspecification, these assumptions increase power and allow for identification of effects without full common support. Parametric models of treatment effects can be implemented by interacting covariates with treatment status:

$$Y = X\beta + \tilde{X}T\gamma + \delta T + \varepsilon.$$

As above, T is unlikely to be exogenous when perfect compliance fails, so re-weighting or IV methods are necessary to obtain consistent estimates. The methods from sections 3.3 and 3.4 are directly applicable, since conditional means are features of the marginal outcome distributions. One may be tempted to continue to use randomization status R instead of treatment status T in order to estimate an intent-to-treat parameter. However, as pointed out above, ITT parameters conflate heterogeneity in take-up with heterogeneity in treatment effects, which makes them hard to interpret. It will usually be preferable to separately estimate take-up given participation in the RCT and treatment effects for the treated separately.

We use $\tilde{X}$ instead of X here to indicate that all covariates do not need to be interacted with treatment. For example, $\tilde{X}$ might include only gender or income if the researcher is primarily concerned with that dimension of treatment heterogeneity. The coefficients on the interactions, γ, describe how the average treatment effect varies with $\tilde{X}$. To be sure, this model imposes several parametric assumptions. Most importantly, potential outcomes, Y_0 and Y_1, are assumed to be linearly related to the covariates, since $Y_0 = X\beta + \varepsilon$ and $Y_1 = X\beta + \tilde{X}\gamma + \varepsilon$. When the true data generating process is nonlinear, the estimates can be interpreted as a linear approximation of the conditional expectation function and provide a useful first impression of the existence and direction of heterogeneity. Also, covariates included in X but not $\tilde{X}$ are assumed to have the same impact on treatment and control outcomes.

While linear models can be made quite flexible by including powers of covariates and interaction terms, they may still struggle to detect complex relationships. A common exercise in regression analysis is to add covariates to the model to see how a coefficient of interest changes. To explore whether observed subgroup impacts are driven by an omitted variable, we can add additional covariates to $\tilde{X}$. Recent work proposes using machine learning algorithms as a data-driven approach to "build" the model (Imai and Ratkovik, 2013; Athey and Imbens, 2015; Wager and Athey, 2015) and include only the most important dimensions of treatment heterogeneity.

Conditional average treatment effects provide a simple measure of how treatment effects vary with observable characteristics. As with the average treatment effect, however, there is a distribution around the conditional mean. For example, even if all conditional means are positive, some individuals may still be hurt by the program. This idiosyncratic heterogeneity may pose a threat to the external validity of the results. For example, using subgroup effects to inform future targeting of a program may backfire if the impact on the marginal participant is very different from the average treatment effect among the current set of participants, even within each subgroup. Smith (2015) provides the following example. Suppose there is a program in which half of men have an impact of 10 and half have an impact of 4. Similarly, half of women have an impact of 12 and the other half have an impact of 1. If the cost of participating in the program is 5, only the "big responders" will choose to participate and the program will appear more effective for women than men. However, the impact on the marginal male participant (4) is larger than the impact on the marginal female participant (1).

We can attempt to overcome the above issue by estimating a model of participation along with the outcome model. To address the concerns raised in the previous paragraph, we need to understand how treatment effects vary with both the observables and the unobservables, i.e., the error term V of this participation model. The methods in this section can be used to study how treatment effects vary with observed covariates. Heckman and Vytlacil (1999, 2005, 2007) analogously define marginal treatment effects (MTE) as the average treatment effect at a specific value of the unobservable component of the participation equation:

$$\Delta_{\mathrm{MTE}} = \mathbb{E}(Y_1 - Y_0 | V = v, X = x).$$

This is a conditional average treatment effect just as are those described below, but one conditioning variable, V, is not observed. Thus, estimation is more complex. Under the assumptions of Heckman and Vytlacil, it can be estimated as the derivative of the outcome, Y, with respect to the exogenous variation in the propensity to participate. Brinch, Mogstad and Wiswall (forthcoming) and Kowalski (2016) are examples of recent studies which use MTEs to examine treatment heterogeneity. Cornelissen et al. (2016) provide a review and discuss the relation to LATE.

6. Statistical Inference and Power Calculations

1. Introduction

Parts 3 and 4 discuss identification and estimation of treatment effects. However, meaningful policy evaluation must also account for the precision of point estimates. This part surveys some tools for formal hypothesis tests and power calculations in DIA: Section 6.2 discusses statistical inference for DIA. It focuses on repeated testing and functional hypotheses. Section 6.3 surveys tests of a particular distributional hypothesis: whether individual responses to treatment were heterogeneous. Section 6.4 considers the choice of sample size for RCTs. It aims to help researchers design RCTs to address distributional hypotheses.

These three sections draw on approximations to the finite-sample behavior of estimators of treatment effects. Inference and power calculations would ideally use the exact distribution of the estimator in question for each sample size. However, finite-sample results are only available for special cases.[41] Hence, we must rely on asymptotic approximations, for which there are two main strategies. Traditional asymptotic theory exploits the limiting distribution of estimators as the sample size goes to infinity, which is normal in most cases. This approach has two drawbacks. First, the resulting approximation might be poor. Second, the asymptotic distribution might be impractical. For example, the asymptotic variance of the IPW estimator of QTTs (Section 3.3) depends on a conditional expectation (Firpo, 2007), which is difficult to estimate due to the curse of dimensionality. Therefore, we recommend simulation-based inference instead. The procedure is straightforward: estimate the effect of interest on many different samples; pool the estimates across replications to construct a sample of estimators; then use the distribution of the estimator from this sample for inference.

2. Statistical Inference

When researchers conduct DIA, they often wish to address multiple hypotheses or hypotheses about multiple parameters. However, repeated testing increases the probability of false positives, distorting significance and power levels. Therefore, critical values need careful adjustment according to the question at hand.[42]

This section surveys three categories of hypotheses, which often occur in DIA:

- *Single hypotheses about a finite number of points*, which lend themselves to standard pointwise inference.
- *Multiple hypotheses about a finite number of points*, which also pertain to pointwise inference, although critical values need adjustment to correct significance levels.
- *Single hypotheses about a function*, which require uniform inference.

The next subsections discuss each category in turn, including examples of relevant hypotheses and algorithms for valid inference.

1. Pointwise Inference

This subsection focuses on pointwise inference: testing a single hypothesis about a finite number of points. For instance, consider the following questions:

- Is the average treatment effect significantly different from zero?
- Is the distribution of treatment effects symmetric?
- Is the change in the first quartile of outcomes larger than the third quartile?

The first hypothesis concerns a single statistic of the distribution of treatment effects. The second also concerns a single statistic of this distribution if we reformulate it in terms of skewness.[43] The third concerns the treatment effect on two points of the outcome distribution. Thus, they all require pointwise inference.

[41] For instance, finite-sample theory offers exact formulas for the distribution of quantile estimators at any sample size if the data are independent and identically distributed. See Koenker (2005) and Chernozhukov, Hansen and Jansson (2009).

[42] The design of the RCT should also account for multiple testing. For example, optimal sample sizes are larger. See Section 6.4.

[43] Note that no skewness is a necessary condition for symmetry, but it is not sufficient.

Following the discussion in the introduction of this part, we recommend basing inference on the bootstrap. Instead of hypothesis tests, we consider the equivalent problem of constructing confidence intervals. We accept the null hypothesis if it falls entirely within the confidence interval and reject it otherwise.

Suppose that the sample consists of N observations. To construct a confidence interval at the 95% level,[44] Horowitz (2001) recommends the percentile method:

1. Sample N observations at random with replacement from the data. You now have a bootstrap sample.
2. Estimate the treatment effect of interest on the bootstrap sample and store it.
3. Repeat steps 1 and 2 B times. You now have a sample of estimates of size B.
4. Compute the quantiles 0.025 and 0.975 of the sample of estimators, $z_{0.025}$ and $z_{0.975}$. You now have bootstrap critical values. Your confidence interval is: $[z_{0.025}, z_{0.975}]$.

For different significance levels, adjust step 4. Alternative resampling procedures in step 1 are possible. For example, we can sample clusters instead of individual observations to account for within-cluster dependence.[45] We can also use the bootstrap to obtain different statistics. For example, compute the standard deviation of the sample of estimators in step 4 to estimate standard errors.

How many bootstrap repetitions are necessary? Computation is often trivial, so that a conservatively large B comes at almost no cost. Based on Andrews and Buchinsky (2000), Cameron and Trivedi (2005) suggest at least 348 repetitions for confidence intervals at level 0.05 and 685 for level 0.01. In applied work, however, researchers tend to run at least a thousand repetitions and often as many as ten thousand.

To achieve correct inference with the bootstrap, step 2 should account for the interaction between different sources of estimation noise. Consider, for example, the IPW estimator of treatment effects (see Section 3.3). We first compute weights for each observation, before taking the difference between weighted quantiles (or averages, variances, etc.). One should not overlook estimation error from the first step, which depends on the weight estimator. In the case of parametric models, it is sufficient to re-estimate the weights at each iteration in step 2 to adjust percentile intervals.[46] Nonparametric or semiparametric estimators are more challenging, because their rates of convergence depend on tuning parameters (e.g., the kernel bandwidth). Most data-dependent procedures pick optimal tuning parameters, in the sense that they minimize mean squared error. However, optimal convergence rates often generate asymptotic bias, which invalidates inference (whether it draws on asymptotic theory or the bootstrap). Then it is necessary to either correct for bias or to adjust the tuning parameter. Moreover, these procedures themselves introduce additional estimation noise. It is unclear, however, whether one should correct for it. In a study of kernel density estimators, for example, Hall and Kang (2001) argue against re-estimating the bandwidth at each bootstrap round.

[44] Analogous algorithms yield one-sided or asymmetric tests. See Horowitz (2001).

[45] The usual caveats about cluster-robust estimation apply. We refer the reader to the excellent survey of Cameron and Miller (2015) for details, in particular Subsection VI.C.3 (“Bootstrap with caution”).

[46] For most estimators, analytical corrections for asymptotic confidence intervals are possible. See Newey and McFadden (2001) for a general discussion of inference for two-step estimators.

Note also that the bootstrap requires continuity conditions.[47] The estimators of changes in the outcome distribution in Part 3 generally satisfy these assumptions. The exception are extremal effects: very low and very high quantiles, for which the scarcity of data affects convergence rates. Zhang (2016) considers inference for such extremal quantile effects. Moment estimators for the distribution of treatment effects (Subsection 4.4.1) also satisfy the bootstrap regularity conditions, as long as this distribution is not degenerate. On the other hand, the validity of resampling for the deconvolution estimator of Subsections 4.4.2 and 4.4.4 is an open question,[48] because the asymptotic properties of Mallows' Algorithm are unknown. As the literature stands, the bootstrap provides some insight into the precision of this estimator, but it does not allow formal hypothesis tests.

The theory of pointwise inference focuses on point estimates. However, it is also possible to test hypotheses about partially identified parameters. For example, Section 4.2 discussed the estimation of bounds on features of the distribution of treatment effects. As Imbens and Manski (2004) and Stoye (2009) note, inference raises additional questions in this case. They propose a strategy to construct confidence intervals that cover the true, unknown parameter within the bounds with the correct probability. The limits of these intervals depend on the asymptotic distribution of the bound estimators. This task is often difficult because many bounds involve extrema, such as the sample maxima and minima in the Makarov bounds in Section 4.2. These discontinuities violate the smoothness assumptions of standard methods of inference. Fan and Park (2010) develop a subsampling strategy to perform inference on Makarov bounds, which is valid under weak conditions.

2. *Multiple Hypothesis Testing (MHT)*

Policy evaluation often entails multiple hypothesis tests. For example, consider the questions:

- Do average treatment effects vary across subgroups?
- What is the average treatment effect for various outcomes?
- Which quantile treatment effects are positive?

Repeated testing distorts significance levels, necessitating adjustments to critical values for correct inference. To illustrate this point, suppose that researchers worry about heterogeneity across ethnicities. They estimate average treatment effects for five ethnic groups and test each estimate against the zero null hypothesis at the 5-percent level. Thus, the probability of a false positive is five percent for each test. Across all tests, however, it is larger. If the tests are independent, it rises to $1 – 0.95^5 = 0.23$!

Adjusting critical values requires us to extend the concept of test size (i.e., probability of making a Type I error) to multiple hypotheses. Several extensions exist and researchers should base their choice of rejection rule on substantive considerations.[49] One approach

[47] It is often possible to adjust the simulation algorithm to obtain valid inference under weaker assumptions. For instance, Otsu and Rai (forthcoming) develop a valid weighted bootstrap approach for matching estimators. Different resampling methods, such as subsampling, are often valid under weaker conditions than the bootstrap, although their theoretical properties are not as advantageous. See Horowitz (2001).

[48] The nonparametric bootstrap is valid for the kernel deconvolution estimator of the distribution of treatment effects (Bissantz et al., 2007).

[49] Researchers should ideally define their preferred error rate at the design stage.

considers the Familywise Error Rate (FWER). The FWER is the probability of falsely rejecting at least one true null hypothesis. This criterion is stringent: it preserves significance levels at a heavy cost of test power. Some authors consider the *k*-FWER instead, which generalizes the FWER to *k* false rejections. Also popular is the False Discovery Rate (FDR): the expected proportion of false positives across all hypothesis tests. As Romano and Wolf (2010) point out, however, control of the FDR does not allow us to make precise statements about the realized probability of false discoveries, which might remain quite high.

Romano and Wolf (2010) discuss these different concepts of error rates in more detail. They propose a general method to construct simultaneous confidence regions, using resampling and an iterative step-down algorithm to preserve power.[50] The resulting intervals achieve balance, in the sense that each marginal interval covers the true parameter with the same probability. Here we present a special case of their algorithm, which controls the FWER.

We wish to test K hypothesis about parameters $\{\theta_k\}_{k=1}^{K}$ at level α. The sample size is n. We assume that there are estimators $\hat{\theta}_k$ for each parameter θ_k, such that $\tau_n \theta_k$ converges to a nondegenerate distribution, where τ_n is a nonnegative sequence. For example, θ_k might be the average treatment effect or a quantile effect, in which case $\tau_n = \sqrt{n}$. We will proceed in a series of single-step tests. At each iteration, we remove rejected hypotheses from consideration. The single-step algorithm is:

1. Estimate $\hat{\theta}_k$ for each k.
2. Sample n observations at random with replacement from the data. You now have a bootstrap sample.
3. Estimate $\hat{\theta}_{kb}^*$ for each k on the bootstrap sample. Use the same estimator of step 1. Compute and store: $\tau_n \left|\hat{\theta}_{kb}^* - \hat{\theta}_k\right|$.
4. Repeat steps 2 and 3 B times.
5. Compute the empirical CDF $\hat{H}$ of $\tau_n \left|\hat{\theta}_{kb}^* - \hat{\theta}_k\right|$ across bootstrap replications. For each parameter k and bootstrap replication b, you will have $\hat{H}\left(\tau_n \left|\hat{\theta}_{kb}^* - \hat{\theta}_k\right|\right)$.
6. For each bootstrap replication b, compute $z_b^* = \max_k \hat{H}\left(\tau_n \left|\hat{\theta}_{kb}^* - \hat{\theta}_k\right|\right)$.
7. Compute the α-th quantile of z_b^*, $c(\alpha)$. Note that it lies in the unit interval.
8. For each parameter k, compute the $c(\alpha)$-th quantile of $\tau_n \left|\hat{\theta}_{kb}^* - \hat{\theta}_k\right|$ across bootstrap replicates. This quantile is the critical value for the k-th hypothesis test.

To start the step-down algorithm, run the single-step algorithm. If you do not reject any hypothesis, stop. If you do reject, go back to step 6. When you compute the maximum across parameters, ignore previously rejected hypotheses. Step 7 will yield a different percentile for use as a critical value in step 8. Continue in this fashion until you stop rejecting.

3. Uniform Inference

This subsection discusses uniform inference: testing a single hypothesis about a continuous object, such as a function. Uniform inference allows us to address such questions as:

- Are all quantile treatment effects positive?

[50] Lee and Shaikh (2014) and List, Shaikh and Xu (2016) illustrate the algorithm with experimental data.

- Is the distribution of treatment effects normal?

The distinction between uniform inference and corrections for multiple testing can be subtle. Uniform inference is concerned with *single* hypotheses about *single continuous* objects, such as the quantile process or the distribution function. Testing whether there was any effect on the outcome distribution pertains to uniform inference, because we would test the equality of two distribution functions. On the other hand, testing which quantile effects are significant is a problem of multiple testing, because we would perform a sequence of pointwise significance tests for each estimate.

Statisticians and econometricians have developed various tests of popular distributional hypotheses, which are often based on scalar statistics. For example, consider the hypothesis of positive quantile effects. It is equivalent to first-order stochastic dominance, for which we have the Kolmogorov-Smirnov statistic (the largest absolute difference between the outcome distributions of treated and untreated participants). Chernozhukov, Fernández-Val and Melly (2013) exploit this approach to construct bootstrap uniform confidence bands, whose width depends on the critical values of the Kolmogorov-Smirnov statistic.[51] Uniform confidence bands cover the entire function of interest with the correct probability, in the same way that confidence intervals cover point estimates. They have three advantages over scalar test statistics: they are easy to display on a graph; they allow researchers to test hypotheses for which no known statistics are available; and they allow readers to test hypotheses which the authors of a paper may not have considered.[52]

3. Tests of Heterogeneous Treatment Effects

DIA builds on the premise that treatment responses vary across individuals. How can we test this hypothesis?

As far as heterogeneity relates to discrete covariates, significant differences in CATEs indicate heterogeneity in individual responses. Such heterogeneity across subgroups is often interesting in itself. Tests are standard, but researchers should adjust critical values for MHT, which often takes a heavy toll on power. Crump et al. (2008) generalize this idea to accommodate discrete and continuous covariates in both parametric and non-parametric estimators of conditional means. They propose a scalar statistic, thereby avoiding MHT.

The tests consider heterogeneity as it relates to observables and will thereby fail to detect heterogeneity that is unrelated to observables. To test for *any* heterogeneity, we can use the fact that the distributions of Y_0 and Y_1 only differ in means under the null. To test whether treatment effects are constant, therefore, we may test whether the variance of outcomes is equal for the treated and untreated – i.e., test whether $\operatorname{var}(Y_0) = \operatorname{var}(Y_1)$. Significant differences in the variance allow us to reject the null hypothesis. We can implement such a pointwise test with the bootstrap algorithm of Subsection 6.2.1. However, this test may have low or no power if the covariance of Y_0 and Δ is negative: $\operatorname{var}(Y_1) = \operatorname{var}(Y_0) + \operatorname{var}(\Delta) + 2\operatorname{cov}(\Delta, Y_0)$, so the negative covariance would offset the variance from treatment effect heterogeneity. The analytical example in Section 3.5 illustrates this point. The test may also falsely reject in the presence of heteroskedasticity, i.e., if the variances of the error terms differ between treatment and control. Heteroskedasticity may be a concern if random-

[51] See algorithms 2 (p. 2221) and 3 (p. 2222) in Chernozhukov, Fernández-Val and Melly (2013).

[52] For some estimators, the literature provides asymptotic theory to construct confidence bands as well. For quantile treatment effects, see Chernozhukov, Fernández-Val and Melly (2013) and Koenker and Xiao (2002).

ization was compromised. Conducting the test on residuals may address this concern and increase power, but it may also lead to false rejections if the model is misspecified.

The idea in the previous paragraph generalizes to higher moments. Joint tests of multiple moments may achieve greater power. One can also use a test of equality of distributions, such as the Kolmogorov-Smirnov test, to assess whether the distributions of deviations from the mean differ between the treated and untreated.[53] Similar to the variance, one may be worried that the distributions differ for other reasons if randomization was compromised.

Further evidence against the null hypothesis of constant treatment effects may come from quantile effects or bounds. If treatment effects were constant, all quantile effects would be equal to the average treatment effect. Implementing this approach requires either corrections for multiple testing (Subsection 6.2.2) or uniform inference (Subsection 6.2.3). One may also be able to reject the null if the variance bounds do not include zero, as Section 4.2 discusses.

4. Power Calculations

This section considers the choice of sample size for DIA in RCTs. We first adapt the analytic framework of List, Sadoff and Wagner (2011) to DIA. The complexities of DIA often render an analytic approach undesirable or unfeasible, so we close with some guidance on power calculations via simulation.

1. Analytic Approach

An important choice in experimental design is the sample size. Due to cost considerations, we seek the smallest sample for a desired level of precision.

As a first step, we must define precision. Our criterion will be the significance level and the power of a particular pointwise test against the zero null hypothesis (ZNH).[54] We wish to determine the optimal sample size, n^*, and the optimal probability of assignment to treatment, p^*, such that a test of the ZNH will reject a given minimum effect size at significance level α with power $1-\beta$. These parameters reflect the experimenter's tolerance of statistical error. Here, n^* includes both treated and untreated participants. To derive the power of a test, we need the asymptotic distribution of the estimator of interest. For concreteness, we focus on quantile treatment effects under the independence assumption of Section 3.2.[55] Similar to ATTs, QTTs are asymptotically normal. Therefore, we only need to specify their asymptotic variance. Power calculations for QTTs and ATTs only differ in the exact form of this variance.

We proceed in four steps. For a quantile τ of interest:

[53] Note that this test involves estimated parameters, which the bootstrap should account for. See Subsection 6.2.1.

[54] The ZNH states that the treatment had no effect. It is the main hypothesis of interest in most RCTs. Other null hypotheses are possible with the necessary modifications. As a reminder, the significance level is the probability of a type I error in a two-sided test: rejecting the null hypothesis when it is true (a false positive). The power is the probability of rejecting the null hypothesis when the alternative is true.

[55] We assume that the data are IID. For a discussion of clustered data and stratified designs, see Section 4 of List, Sadoff and Wagner (2011), and especially McConnell and Vera-Hernández (2015).

1. Specify the quantile $q_0(\tau)$ and the density $f_0[q_0(\tau)]$ of potential outcomes Y_0. Note that this quantile is the same under the null and the alternative hypotheses.
2. Specify the quantile $q_1(\tau)$ and the density $f_1[q_1(\tau)]$ of potential outcomes Y_1 under the alternative hypothesis. Note that $q_1(\tau) = q_0(\tau)$ and $f_0[q_0(\tau)] = f_1[q_1(\tau)]$ under the null hypothesis of no treatment effect.
3. Define a minimum detectable effect size $q_1(\tau) - q_0(\tau)$, a significance level α and a test power $1 - \beta$. The RCT will detect treatment effects equal to or greater than $q_1(\tau) - q_0(\tau)$ with error rates α and β.
4. Under the alternative hypothesis, the asymptotic variance of the estimator $\hat{q}_T(\tau)$ is:

$$\sigma_T^2(\tau) = \frac{\tau(1-\tau)}{f_T[q_T(\tau)]^2}.$$

The optimal randomization rate, $p^*(\tau)$, and the sample size, $n^*(\tau)$, are:

$$p^*(\tau) = \frac{\sigma_1(\tau)}{\sigma_0(\tau) + \sigma_1(\tau)},$$
$$n^*(\tau) = \left[\frac{z_{\alpha/2} + z_\beta}{q_1(\tau) - q_0(\tau)}\right]^2 \left[\frac{\sigma_0^2(\tau)}{1 - p^*(\tau)} + \frac{\sigma_1^2(\tau)}{p^*(\tau)}\right],$$

where $z_{\alpha/2}$ and z_β are the quantiles of the standard normal distribution. Note that sample sizes are larger for quantiles in low-density regions, such as the tails of the outcome distribution, because the scarcity of observations reduces accuracy.

These same steps yield optimal sample sizes for other quantities. The formulas for n^* and p^* in step 4 go through as long as the estimator of interest is asymptotically normal.[56] Adjust steps 1 and 2 according to the object of interest and its asymptotic variance σ_T^2. For the average treatment effect, for example, one should specify the mean and the variance of potential outcomes instead of the quantile and the density, since $\sigma_T^2 = \text{var}(Y_T)$.

The procedure above considers a particular quantile of interest. In most applications, however, researchers compute quantile effects on a grid. Three difficulties arise. The first consists of formulating distinct hypothesis tests for each quantile. It is often simpler to think in terms of the entire distribution. For example, a location model assumes that the distribution shifts by a constant. Then all quantile effects are equal to that constant and $\sigma_0^2(\tau) = \sigma_1^2(\tau)$, which reduces the number of free parameters by two thirds. Secondly, the sample size $n^*(\tau)$ and the randomization rate $p^*(\tau)$ differ for each quantile.[57] One should choose the largest sample size across τ to ensure correct test size and power at all points. Note, though, that the formulas above are invalid for extremal effects. It is advisable to avoid estimation of extremal effects – beyond the 5th and 95th percentiles, say. Lastly, grid estimation implies multiple hypothesis testing, invalidating the formulas above. There are two solutions to this problem. It is possible to adjust critical values for repeated testing (cf. Section 6.2). However, this approach involves complex algorithms and many unknown quantities, which limits its usefulness. Simulation offers a more feasible route, which we survey in the next

[56] It is also possible to use finite-sample distributions instead of the asymptotic normal approximation when they are available. See Koenker (2005) and Chernozhukov, Hansen and Jansson (2009) for quantile treatment effects.

[57] For a location model, $p^*(\tau) = 0.5$ for all τ.

subsection. The closed-form formulas provide a first guess of sample size to start the simulation algorithm.

If researchers intend to conduct subgroup analysis, the choice of sample size and randomization rate should also take MHT into account. Otherwise, significance levels will be higher and/or test power will be lower than planned. Note that the number of observations might dramatically increase with the number of tests. Splitting samples might help keep the RCT financially viable. This approach splits the sample into two subsamples. In one subsample, an automated procedure selects a small set of relevant covariates. In the second subsample, we estimate the resulting model and test the significance of each variable, limiting data requirements. For further details, see Wasserman and Roeder (2009), Fithian, Sun and Taylor (2017) and Fafchamps and Labonne (2016) and the references therein.

2. Simulation Approach

Simulation is an alternative strategy to determine the required sample size. Simulation for DIA presents no particular challenge, so we only sketch a brief overview here. See McConnell and Vera-Hernández (2015) for details. They also provide an algorithm and sample code.

Similar to the previous subsection, the selection criteria are the significance and power of a hypothesis test against the ZNH. Instead of computing these error probabilities with asymptotic methods, however, we use simulation. To be precise, we simulate pseudo-samples from the model of interest under the null and the alternative hypothesis. For each pseudo-sample, we perform the relevant hypothesis test. Then we adjust the number of observations until empirical rejection rates satisfy the desired level of accuracy.

Simulation offers a number of advantages over asymptotic theory. Firstly, it accommodates estimators with unknown or excessively complex asymptotic distribution (e.g., the IPW estimator of QTTs), as well as difficult inference problems (e.g., uniform inference or multiple hypotheses). Secondly, it reflects the finite-sample behavior of estimators which might converge slowly to their limiting distribution. Thirdly, it is straightforward to incorporate complex designs, such as panel data.

To conduct a simulation exercise, it is necessary to choose the number of pseudo-samples. Unless the computational burden is high, it should be large: five thousand or ten thousand are good figures. One also needs to specify the data generating process. Choosing distributions of covariates and the error term is often challenging: they should reproduce the conditions of the RCT, including dependence patterns in the error term (such as clusters or serial correlation in panel data). Previous studies might provide some guidance. If an existing dataset contains the covariates of interest (or some of them), it is possible to draw pseudo-samples from it to minimize distributional assumptions. Sensitivity analyses are always advisable.

7. Applications

In this part, we present applications of the methods discussed above using data from two RCTs. Section 7.1 revisits the study of financial education by Bruhn et al. (2016). In Section 7.2, we study the impact of a school-development program on students' test scores, building on Blimpo, Evans and Lahire (2016). The goal of this part is threefold: (1) to demonstrate what we can learn from DIA beyond standard mean analyses; (2) to provide examples of how to choose appropriate methods and how to assess their assumptions and address common concerns; and (3) to illustrate briefly the implementation of the methods and their output.[58] The programs corresponding to these analyses can be found in the supplemental materials available online.

1. Financial Education RCT in Brazil

This section revisits the financial education program evaluated by Bruhn et al. (2016). Comprehensive lessons in basic finance and responsible intertemporal choices were integrated into regular classroom curricula of randomly selected high schools. The program increased average financial proficiency by a quarter of a standard deviation, which is large in comparison with similar programs (Bruhn et al., 2016).

However, policymakers may wish for more information than average effects. For instance, they may take particular interest in low values of financial proficiency, which might put consumers at risk of pyramid schemes and other predatory tactics. They may want to fine tune the program if they observe no change in the frequency of bottom scores. In a similar vein, governments may worry about dispersion in financial proficiency, which could affect the rewards from financial inclusion. They may also ask whether effects differed across subgroups, which could help them target the intervention and understand why it works (or not). If poverty is associated with low financial proficiency, for example, we might advocate additional financial education for children from poor households. To explore these questions, we estimate several DIA parameters, mostly related to changes in the outcome distribution (Part 3) and conditional analysis (Part 5).

1. Data

Our sample includes 892 schools and 18,276 students in six states in Brazil.[59] We focus on the short-term impact between the baseline survey (August 2010) and the first follow-up (December 2010).[60] We observe three-quarters of the original baseline sample of around 25,000 students at the first follow-up. The authors report that the groups are overall balanced across treatment and control groups at each round of data collection.[61]

We focus on financial proficiency as the outcome of interest. Each survey included a multiple-choice test, from which Bruhn et al. (2016) construct an index of financial knowledge on a hundred-point scale to track students' progress. We use this score for our analyses. For future reference, the average score in the control group at the first follow-up is 56.05 and the standard deviation is 14.81.

2. Changes in the Distribution of Financial Proficiency

Did the intervention change the frequency of very low test scores? Did it change dispersion in financial proficiency? Did it reshape the proficiency distribution?

[58] We focus on DIA and DIA-specific issues. The original papers discuss solutions to other common problems that affect both DIA and the estimation of average treatment effects.

[59] From an original list of 910 schools, nineteen did not participate for unknown reasons.

[60] The dataset is available from the American Economic Journal: Applied Economics.

[61] As an exception, the authors find a small difference in the gender ratio, which is significant at the 10% level (these tests are unadjusted for multiple hypothesis testing).

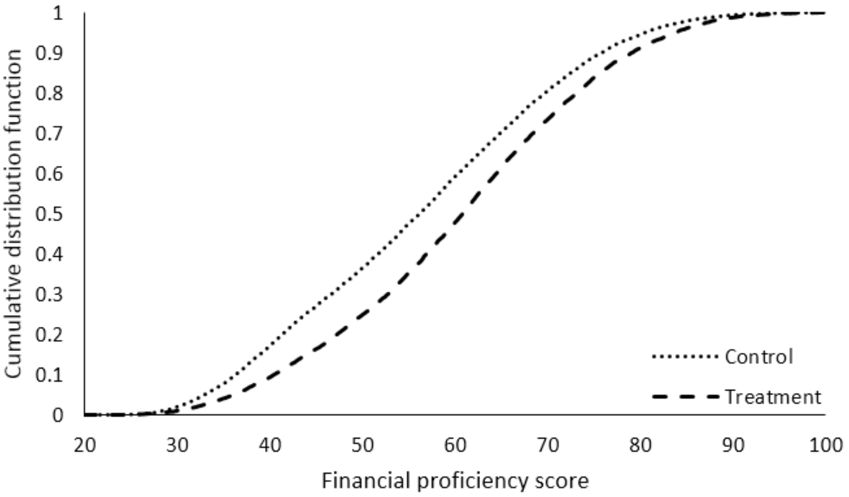


Figure 5: Outcome Distributions for Financial Education Program

To answer these questions, we examine treatment effects on a broad set of parameters: the mean, the standard deviation and the ratio of the 75th to the 25th percentile; every percentile between the 5th and the 95th; and the CDF for every integer score between 33 and 81.[62] We consider a total of 143 parameters. We assume independence between outcomes and treatment status (cf. Section 3.2). The intervention began halfway into the school year, so students could not change schools or classes before the first follow-up, limiting noncompliance with treatment assignment. On the other hand, nonresponse may be a concern: only three-quarters of the baseline sample took the follow-up survey. However, Bruhn et al. (2016) find no significant differences in the means of background variables at each round, which suggests that nonresponse is as good as random.

Figure 5 plots the outcome distribution for each group. Table 3 and Table 4 show estimates of treatment effects.[63] All confidence intervals control the FWER at the 95-percent level using the step-down bootstrap algorithm of Romano and Wolf (2010), see Subsection 6.2.2. Thus, the probability of at least one false rejection across all significance tests is asymptotically smaller than five percent.

We find a large increase in average financial proficiency: 4.3 points (0.29 SD or 7.6%), in line with Bruhn et al. (2016). There was also a reduction in the proportion of very low scores. For example, the share of scores below 40 points is lower by 0.08, a 45.8% reduction from 0.17 in the control group. Outcomes are also less dispersed in the treatment group: the standard deviation is smaller by 0.46 points, a 3.12% reduction of the 14.81 points in the control group, and the ratio of the 75/25 quartiles is reduced by 7.5% from 1.53 in the control group. All of these effects are statistically significant. In summary, the intervention not only increased average scores, but also decreased inequality in financial proficiency.

Quantile effects provide further insight into these results. As Figure 6 shows, the intervention shifted the entire CDF to the right. Recall that quantile effects are the horizontal distance between these curves; hence, they are all positive (and statistically significant). Moreover, the change in lower quantiles is relatively larger. The first decile increased by 4.2 points or 11.5%. The first quartile is greater by 6.2 points (14.1%), whereas the difference in the ninth decile is 3.2 points (4.27%). A Kolmogorov-Smirnov test rejects equality of quantile

[62] The 5th percentile of control-group outcomes is 32.95. The 95th is 80.46.

[63] Figure 12 in Appendix 4 shows changes in the distribution function (the vertical distance between the CDFs in Figure 5). These effects mirror quantile effects.

Table 3: Treatment Effects Estimates for Financial Education Program

Statistic	**Control group value**	**Effect estimate**		**Standard error**	**Simultaneous conf. region (95%)**	
		Value	Percent			
Mean	56.050	4.266	7.611%	0.571	2.671	5.861
Standard deviation	14.808	−0.462	−3.119%	0.201	−0.853	−0.071
75/25 perc. ratio	1.530	−0.119	−7.748%	0.020	−0.172	−0.066
10th percentile	36.302	4.177	11.505%	0.582	2.552	5.802
25th percentile	43.867	6.189	14.109%	0.843	3.933	8.446
50th percentile	56.157	4.642	8.266%	0.700	2.717	6.568
75th percentile	67.138	3.537	5.268%	0.575	1.889	5.185
90th percentile	75.812	3.236	4.269%	0.572	1.536	4.937
CDF at 40 points	0.172	−0.079	−45.742%	0.011	−0.109	−0.049
CDF at 50 points	0.365	−0.116	−31.806%	0.016	−0.160	−0.073
CDF at 60 points	0.593	−0.114	−19.236%	0.016	−0.159	−0.070
CDF at 70 points	0.806	−0.071	−8.864%	0.012	−0.106	−0.038
CDF at 80 points	0.945	−0.032	−3.373%	0.006	−0.049	−0.015

Notes: Standard errors and confidence region based on the bootstrap (five thousand replications) and clustered at the school level. The confidence region controls the FWER (probability of at least one false rejection across tests), following Romano and Wolf (2010).

effects at the 5% level. This pattern reveals that the reduction in outcome inequality occurred despite increases of the upper percentiles (which are desirable in an educational program). It was a consequence of proportionally larger gains in the left tail.

How would we estimate these treatment effects under the assumption of selection on observables? For illustration purposes, suppose that a small imbalance in the gender ratio and baseline scores generated suspicion of selection bias. Following Section 3.3, we can use inverse probability weighting to rebalance groups. We compute the weights with a logistic regression of treatment status on a quadratic polynomial of baseline scores, an indicator for missing baseline score and an indicator for gender. Table 9 in Appendix 4 show our results. Although the rebalanced estimates are smaller than their unweighted counterparts, the differences are not statistically significant, which suggests that endogenous selection did not affect our main results. Bruhn et al. (2016) arrived at a similar conclusion, although their

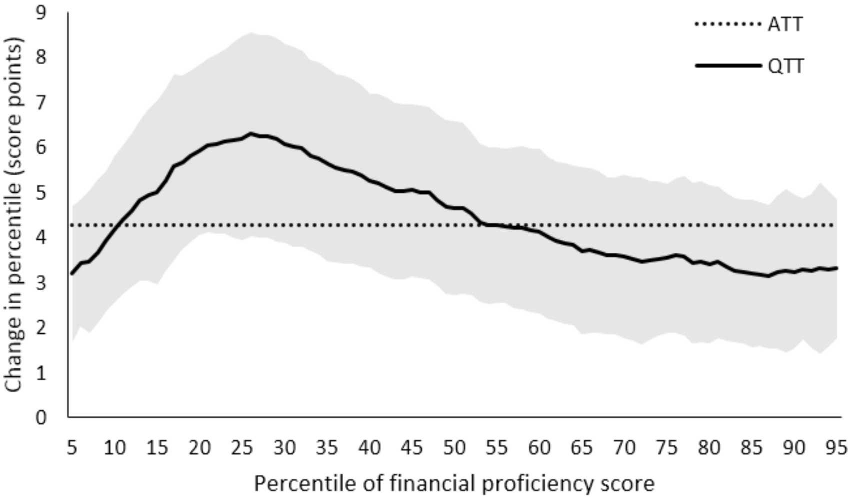


Figure 6: Quantile Effects Estimates for Financial Education Program

methodology differs.[64]

3. Distribution of Individual Treatment Effects

Quantile analysis helps us address a broad range of policy questions. Nonetheless, the possibility of mobility effects constrains our interpretation of quantile effects, in that the change in quantile τ may differ from the individual treatment effect at quantile τ. Consider Figure 6. It is tempting to believe that low-scoring students benefited more than average, since the change in bottom percentiles is relatively larger. As Part 2 and Section 3.5 argue, however, this reasoning hinges on an implicit assumption of rank invariance: the same students are in the bottom and the top of the distribution in both treated and untreated states.

Despite this limitation, our estimates in Table 3 and Figure 6 have implications for individual effects. Dispersion decreased, so responses to treatment must have been heterogeneous. Moreover, positive average and quantile effects imply that at least some participants benefited from the intervention. However, the methods of Part 3 do not allow us to quantify these qualitative results.

Estimating features of the distribution of individual effects requires strong assumptions (cf. Section 4.3), which are hard to justify in this application. As an alternative, Section 4.2 shows that we can construct bounds under minimal conditions. As a first exercise, we estimate the Makarov bounds on the share of positive effects, $\mathrm{P}(Y_1 > Y_0)$. Recall that the bounds comprise the range of values which are consistent with the observed marginal distributions. We find that the probability of positive individual treatment effects lies between 12% and 100%.[65] These bounds confirm that at least some participants benefited from the intervention; moreover, the data are compatible with no one being hurt, because the bounds include 100% positive effects. Next, we bound the standard deviation of individual effects. Assuming that the correlation between potential outcomes is positive, we obtain that it lies between 0.46 and 20.62 points. The lower limit is positive,[66] which rules out constant treatment effects. At 0.46 points, the lower bound represents nearly 11% of the ATT, suggesting that the dispersion in treatment effects is non-negligible. To summarize, these bounds suggest the existence of heterogeneous individual treatment effects under minimal assumptions on the data.

4. Conditional Analysis

The previous subsections showed that students' responses to the financial education program were heterogeneous. This section investigates the relation between treatment impacts and pupils' background characteristics. For that purpose, we compute conditional average effects (CATEs) and conditional QTTs for different subgroups of participants, following Section 5.2.[67]

[64] Bruhn et al. (2016) compute average treatment effects by linear regression. They include a quadratic polynomial of baseline scores, an indicator for missing baseline score and an indicator for gender as controls. The resulting estimate is not statistically different from their baseline regression, which does not include controls.

[65] Note that we do not perform inference to assess the precision of these bounds.

[66] The 95% confidence interval is [0.07, 20.89] (Imbens and Manski, 2004). Under no assumptions on the correlation between potential outcomes, the upper bound is 29.15 and the confidence region is [0.07, 29.54].

[67] We compute treatment effects as the difference in the relevant statistic between the treatment and control subsamples of each subgroup (cf. Section 5.2), as in Subsection 7.1.2.

Table 4: Conditional Average Treatment Effects for Financial Education Program

Subgroup		**Sample size**	**Effect estimate**		**Standard error**	**Simultaneous conf. region (95%)**	
			Value	Percent			
Baseline score above median	No	7960	4.120	8.417%	0.476	2.891	5.349
	Yes	7960	4.069	6.345%	0.588	2.508	5.631
Student has repeated grade	No	10949	4.472	7.634%	0.607	2.866	6.079
	Yes	4437	4.653	9.112%	0.665	2.950	6.355
Student is female	No	6941	4.977	8.588%	0.673	3.227	6.728
	Yes	8720	3.806	6.855%	0.609	2.209	5.403
Student works	No	10898	4.188	7.311%	0.613	2.589	5.788
	Yes	5612	4.722	8.671%	0.702	2.919	6.525
Student earns income	No	5713	3.585	6.163%	0.624	1.943	5.228
	Yes	10812	4.682	8.877%	0.628	3.061	6.304
Family is on welfare (Bolsa Família)	No	10216	5.020	8.301%	0.658	3.294	6.746
	Yes	5334	3.386	5.833%	0.678	1.671	5.101

Notes: The first four specifications include 848 schools. The last two include 851 schools. Standard errors and confidence region based on the bootstrap (five thousand replications) and clustered at the school level. The confidence region controls the FWER (probability of at least one false rejection across tests), following Romano and Wolf (2010).

Based on the available data, we consider six background variables: (1) baseline proficiency score (above or below median); (2) grade repetition; (3) gender; (4) family welfare status (Bolsa Família); (5) working status; and (6) income-earning status (i.e., whether the student earns any income, including pocket money). Each variable defines two subgroups. [68] Table 1 in Bruhn et al. (2016) reports summary statistics and balance tests by variable.

Table 4 displays our estimates of CATEs. Table 5 reports *t*-statistics for the difference in CATEs. We perform six equality tests in total; hence, we must adjust for multiple testing (MHT). In Subsection 7.1.2, we controlled the FWER (the probability of at least one false rejection). Although this error rate is the most stringent, all effects were significant at the five-percent level. Evidence of conditional heterogeneity is less robust. Therefore, Table 5 shows critical values for the 2-FWER (the probability of at least two false rejections) as well. We use the step-down bootstrap algorithm of Romano and Wolf (2010). (See Subsection 6.2.2.)

Responses to treatment seem to correlate with socioeconomic background. The families of a third of our sample participate in Bolsa Família, a welfare program for low-income households. On average, these students gained significantly less from the intervention (3.39 points, against 5.02 for the remainder). This difference is significant at the ten-percent level if we control the FWER. We only find weak evidence of heterogeneity otherwise. Boys had higher average gains than girls (4.98, against 3.81). The average effect was also larger for students with earned income (4.68, against 3.59). However, these differences are only significant if we control the 2-FWER, i.e., if we are willing to tolerate a five-percent chance of two false rejections in six tests. Other criteria do not yield significant effects, whether we control the FWER or the 2-FWER.

[68] A number of participants declined to answer the relevant questions in the surveys, depressing response rates to 64 %. For rigorous policy evaluation, it would be necessary to assess robustness to missing observations.

Table 5: Critical Values for Test of Equality of Average Effects for Financial Education Program

Subgroup	*t*-stat.	5% confidence level			10% confidence level		
		Point-wise	1-FWER	2-FWER	Point-wise	1-FWER	2-FWER
Baseline score above median	−0.095	1.913	2.615	1.890	1.619	2.325	1.696
Student has repeated grade	0.284	1.914	2.498	1.884	1.609	2.284	1.688
Student is female	−2.219	**1.979**	2.622	**1.955**	**1.660**	2.374	**1.759**
Student works	0.850	1.957	2.695	1.930	1.677	2.427	1.752
Student earns income	2.139	**1.970**	2.597	**1.929**	**1.648**	2.344	**1.711**
Family is on welfare (Bolsa Família)	−2.407	**1.998**	2.645	**1.957**	**1.636**	**2.357**	**1.723**

Notes: Bold values indicate rejection of the null hypothesis. The first four specifications include 848 schools. The last two include 851 schools. Critical values based on bootstrapping *t*-statistics (five thousand replications) and clustered at the school level. Columns 4, 5, 7 and 8 control the *k*-FWER via the step-down algorithm of Romano and Wolf (2010).

Figure 7 plots estimates of five conditional quantile effects for each subgroup.[69] Patterns are similar to Figure 6, with effects peaking around the first quartile. The only clear exception is students whose baseline score was above the median, for which quantile effects are monotonically decreasing. Are differences between subgroups significant? There are two approaches to this question. On the one hand, we could test whether the difference between each quantile effect in each pair of subgroups is significant. However, this approach would require thirty hypothesis tests; adjusting for MHT would take a heavy toll on power. On the other hand, we could test whether all quantile effects are equal in each pair of subgroups. To do so, we could test the maximum absolute difference in quantile effects across quantiles,[70] which only requires six tests. This second procedure reveals that quantile effects only differ between subgroups defined by baseline scores and gender, controlling the FWER at the ten-percent level. These results point to heterogeneity both between and within subgroups, which average effects did not uncover.

Note that these estimates are not informative about causal effects of the characteristics that define the subgroups, because students do not sort into subgroups at random. Note also that we did not take the overlaps between subgroups into account. For example, students might be more likely to work if their families are on welfare. Subgroup analysis might provide additional insights as we break down cells into smaller groups. However, the number of observations per cell would decrease just as the number of hypothesis tests increased, to the detriment of statistical power.

[69] The quantiles are: the 10th, the 25th, the 50th, the 75th and the 90th. Table 10 in Appendix 3 reports point estimates and standard errors.

[70] This procedure tests for significant differences in at least one quantile effect for each pair of subgroups. It cannot detect all quantile effects for which there are significant differences.

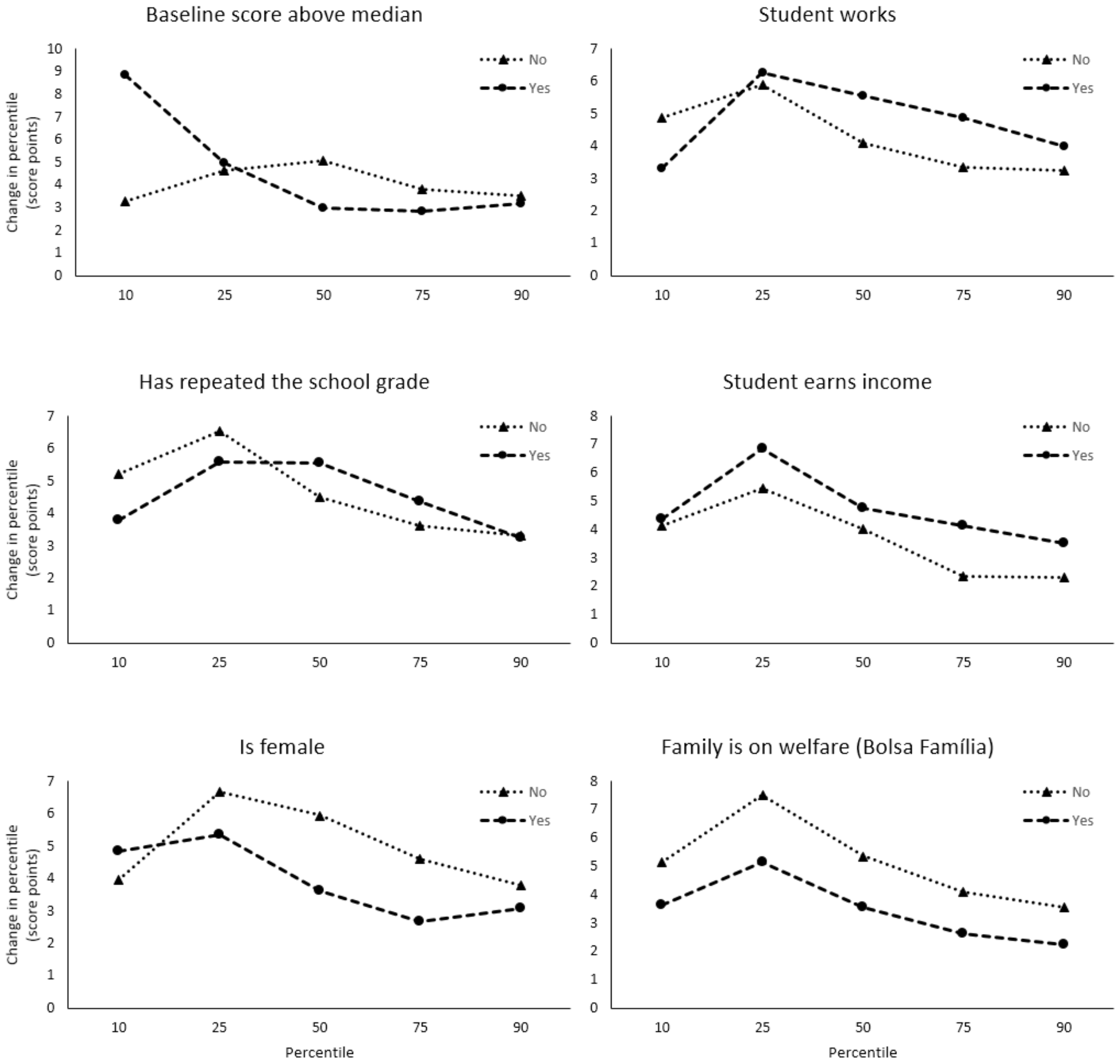


Figure 7: Conditional Quantile Treatment Effects for Financial Education Program

On a methodological note, Table 5 highlights the importance of correcting for multiple tests (even at large sample sizes). Controlling the FWER inflates critical values by 30 to 50 percent. As a consequence, only one difference in average effects remains significant, compared to three if we do not adjust. The choice of error rate is also consequential. Whereas control of the FWER always widens confidence regions, other criteria may not. For example, control of the 2-FWER leads to smaller critical values at the 95-percent level and larger cutoffs at the 90-percent level. To understand this puzzling feature, recall that the probability of two or more false rejections across six pointwise tests at the five-percent level is 3.28%, which is lower than the nominal five-percent level. Therefore, critical values adjust downwards. At the 10-percent level, it is 11.43%. Hence, critical values adjust upwards.

5. Power Calculations

This dataset includes 18,276 observations, which allowed us to obtain precise estimates of a rich parameter set. However, DIA is feasible with much smaller samples. This subsection looks into the choice of sample size for this RCT.

Table 6: Optimal Sample Sizes for Financial Education Program by Quantile Treatment Effect

	Minimum detectable effect							
	+1%	**+5%**	**+10%**	**+15%**	**+1 point**	**+5 pts.**	**+10 pts.**	**+15 pts.**
10th perc.	67920	2828	743	347	8862	985	354	181
20th perc.	72109	3002	789	368	12194	1355	488	249
30th perc.	90478	3767	990	462	19492	2166	780	398
40th perc.	61993	2581	678	317	16371	1819	655	334
50th perc.	46217	1924	506	236	14430	1603	577	294
60th perc.	38287	1594	419	196	13792	1532	552	281
70th perc.	33043	1376	361	169	13753	1528	550	281
80th perc.	28562	1189	312	146	13731	1526	549	280
90th perc.	27957	1164	306	143	15908	1768	636	325

Notes: The table shows the smallest sample size, such that a pointwise two-sided test of the zero null hypothesis detects a given treatment effect with size 5% and power 80%. See Subsection 6.3.1.

We follow the approach of Subsection 6.4.1. We compute the optimal sample size to detect a given minimum effect on each decile. We fix the significance level at 5% and power at 80%. We must also parametrize the asymptotic variance of our estimator of interest. For quantile effects, it depends on the density of outcomes, which we compute from the control group.[71] In an application, researchers might obtain the density function based on a baseline sample, previous studies and parametric assumptions.

Table 6 reports our results for a typical random control trial without any adjustments. Small effects require large samples: detecting an increase of 1% in the 10th percentile requires a total of 67,920 observations. However, moderate samples suffice for modest effects. For instance, to detect increases around 10% in each decile would have required a sample for around 990 in the 30th decile.

Table 6 does not adjust sample sizes for multiple hypothesis tests or clustering, which would raise data requirements. The formula for the optimal sample size in Subsection 6.4.1 depends on the square of the critical values. In Subsection 7.1.2, we controlled the FWER across 143 tests, inflating cutoffs by up to 50%. Such a correction would increase the recommended sample size by a factor of 2.25. Estimating fewer parameters would reduce this adjustment. For instance, suppose that we focus on the mean and the deciles (a total of ten parameters). Then, control of the FWER increases critical values by 30%, which implies multiplying the sample size by a factor of 1.7. As for clustering, the optimal sample size is proportional to the variance of the estimator of interest. In our application, accounting for clustering increases variances by a factor between 4 and 7.

Combining these rules of thumb, our estimates of each quantile effect (see Table 3 and Figure 6) and the formula for the optimal sample size (see Subsection 6.4.1), we find that the optimal sample sizes for this RCT would be 12,000 observations if we control the FWER across all 143 parameters and 9,000 observations if we only control the FWER across the mean and the deciles. Note that adjustments for multiple testing and clustering are unknown at the design stage. Simulations may help researchers incorporate these features into their choice of sample size. See Subsection 6.4.2.

[71] We use the Epanechnikov kernel and the Sheather-Jones plug-in bandwidth.

6. Takeaways

These empirical examples highlight several points about the importance of going beyond average treatment effects. First, it is difficult to summarize the impact of an intervention with any single statistic. In particular, average effects often hide significant heterogeneity. Second, we gain important insights from the changes in the shape of the distribution. Quantile analysis and other DIA methods allow us to address a broader range of policy concerns than means alone under minimal assumptions. Third, conditional analysis is a complement and not a substitute for unconditional approaches. Each method reveals a different dimension of heterogeneity. Finally, our analysis illustrates methodological issues in DIA, such as correcting for multiple testing.

2. School Management RCT in The Gambia

In this section, we use the methods in this toolkit, focusing on those in Part 4, to re-analyze the impact of the Whole School Development Program (WSDP). The WSDP was administered as part of an RCT in The Gambia from 2007 to 2011 and evaluated by Blimpo, Evans and Lahire (2016). The program aims to improve school quality by training school leaders and community members in school management techniques. Each of 273 Gambian primary schools was randomly assigned to one of three groups. Ninety schools in the first treatment arm were offered the WSDP, which provided principals, certain teachers and community members with a comprehensive training program in school management. These schools were also given a $500 grant to help cover costs associated with implementing new initiatives based on the training. To disentangle the impact of the WSDP training from the impact of the grant, 94 schools in the second treatment arm ("grant only") received the $500 grant without any additional training. In the third arm, 89 schools served as the control group and received no treatment.

Blimpo, Evans and Lahire (2016) find that the average impact of the WSDP treatment on test scores is approximately zero and statistically insignificant when looking at impacts separately by grade and year. Zero average effects are consistent with a program having no impact or with a program having heterogeneous impacts, benefitting some schools and hurting others. In order to distinguish between these two cases we use the methods discussed in Part 4 to estimate the variance of treatment effects and, relying on strong assumptions, the entire distribution of treatment effects. The results of these exercises are important for policy: If there is treatment effect heterogeneity the program could become effective if better targeted at those with positive treatment effects. This is obviously not possible if the treatment effect is (very close to) zero for everyone.

The WSDP is a good application to demonstrate the methods discussed in Part 4 because high quality panel data were collected for the evaluation. However, we do not find evidence of heterogeneity in treatment effects: Our results are consistent with the WSDP having no impact on any school. To illustrate the methods further, and demonstrate that our finding of no impact is not due to a general problem of these methods, we simulate an analogous dataset, based on the features of the data collected for this RCT, using the Simulation 1 data generating process described in Appendix 1.

1. *Data*

We use the data from Blimpo, Evans and Lahire (2016). Baseline data were collected in 2008 at the start of the program. Follow-up data were collected in 2009, 2010 and 2011. Blimpo, Evans and Lahire (2016) demonstrate that schools' baseline characteristics are balanced across the three arms. Here, we focus on the impact of the WSDP on student test scores. Math and literacy scores were collected for 3rd and 5th graders in 2008 and 2010 and for 4th and 6th graders in 2009 and 2011.

We construct a balanced panel of test scores by taking the average test score by school-grade-year and restricting attention to WSDP and control group observations where 2 grades were observed in each year from 2008 to 2011. This yields a sample of 960 total observations with 8 observations for each of 120 schools, including 61 schools in the WSDP treatment arm and 59 schools in the control group. This differs from Blimpo, Evans and Lahire (2016), who use student level data clustered by school-year for their cross-sectional analysis. We instead focus on a balanced panel of school level observations. As Table 7 shows, this approach yields similar results for the mean impacts, but has the advantage of simplifying the implementation of Mallows' deconvolution algorithm, which we use in Subsection 7.2.2.[72]

2. *Going Beyond the Mean of the Distribution of Treatment Effects*

Blimpo, Evans and Lahire (2016) find the average impact of the WSDP treatment on test scores is approximately zero when looking at impacts separately by grade and year. We find a similar small and statistically insignificant effect, using pooled OLS on the stacked cross-sections. But does this null average effect mask treatment effect heterogeneity? Of course, in order for a null average effect to mask treatment heterogeneity, some schools must benefit while others are hurt by the program. While policymakers usually implement programs because they think they will benefit participants, programs may have unintended negative consequences if, for example, they disrupt effective systems. If this is the case here, what proportion of schools benefitted from the WSDP?

In this subsection, we answer these questions by studying features of the distribution of the WSDP's effects beyond the mean. Unlike average effects and quantile treatment effects, features of the distribution of impacts beyond the mean are not identified by random variation induced by an RCT alone; additional assumptions are always required. The WSDP RCT is a good candidate for satisfying these assumptions because it includes four waves of data, and treated schools are observed both before and after the implementation of the WSDP. This allows us to analyze the data using panel data methods, like those discussed in Subsection 4.4.4. Conditional on school fixed effects, the variance of treatment effects is identified so long as the treatment effect is uncorrelated with all time-varying components of the error term, and these time varying components are not themselves too correlated over time. The full distribution of treatment effects is identified under the stronger assumption that treatment effects are independent of all time varying components of the error term. While these are not directly testable assumptions, they may be plausible. They would be violated, for ex-

[72] Specifically, by focusing on a balanced panel, all schools' residual vectors are of the same length and receive equal weight.

ample, if the school management training is particularly effective in areas with the highest (unobserved) economic growth during the study period.

In order to move beyond the mean, we use the following variant of the panel data specification from Subsection 4.4.4. to account for heterogeneity in the WSDP's impact:

$$Y_{igs} = \varDelta_i \mathrm{WSDP}_{is} + \alpha_i + \beta_g + \gamma_s + \varepsilon_{igs}.$$

Note two features of this specification: First, we control for school fixed effects, α_i. Second, the treatment effects are no longer assumed to be constant across schools. Instead, $\varDelta_i$ is the impact of the WSDP on school i. Importantly, $\varDelta_i$ is only identified for the subpopulation of schools in the WSDP treatment arm, since identification relies on variation in WSDP_{is} within school i. Control group schools contribute to estimating the coefficients β_g and γ_s which are common across all schools.

The first row of Table 7 shows that the estimate of the average impact across schools, 0.01, is quite similar to the estimate assuming treatment effects are constant across schools. This estimate is just $\widehat{\mathbb{E}}(\hat{\varDelta}_i | T = 1)$. Standard errors are given by conventional formulas for method of moments estimators.

The second row shows the estimate of the variance of treatment effects using the formula given in Arellano and Bonhomme's (2012) equation (50). The estimate, 0.005, is about half the size of the mean impact and is also statistically insignificant. The validity of this estimate relies on two assumptions. First, treatment effects are conditionally mean independent of all past, current and future error terms (i.e., strict exogeneity):

$$\mathbb{E}(\varDelta_i | \mathrm{WSDP}_{is}, \alpha_i, \beta_g, \gamma_s, \varepsilon_{ig1}, \dots, \varepsilon_{igS}) = \mathbb{E}(\varDelta_i | \mathrm{WSDP}_{is}, \alpha_i, \beta_g, \gamma_s).$$

Second, the covariance matrix of $\varepsilon_i = (\varepsilon_{ig1}, \dots, \varepsilon_{igS})$ is given by $\varOmega_i = \sigma_{igs}^2 I_S$. The notation σ_{igs}^2 indicates standard errors may be a function of individual covariates. The key component of this latter assumption is that a school's errors are not correlated across periods. This can be relaxed so long as errors are not "too correlated".

As discussed in Part 4, we can identify the entire distribution of treatment effects if we are willing to make even stronger assumptions. In many applications, these assumptions are unjustified. We include this analysis here primarily for the purpose of illustration.

First, if we are willing to assume treatment effects are normally distributed, then the mean and variance estimates are sufficient statistics for the distribution. This distribution is shown with the dotted line in Figure 8. The distribution is tightly concentrated around zero. To be sure, we cannot rule out that the distribution is degenerate at zero, since neither the mean or variance is statistically significant.

Alternatively, we can recover the entire distribution using deconvolution if we are willing

Table 7: Beyond the Mean Impact of WSDP on Numeracy and Literacy Test Scores

	Estimate	Standard Error
$\mathbb{E}(\varDelta_i)$	0.01	0.03
$\mathrm{var}(\varDelta_i)$	0.005	0.01
Makarov Bounds on $\mathrm{P}(\varDelta_i \geq 0)$	[0.05, 0.95]	

Notes: N = 960 with 8 observations on each of 120 schools. Mean and variance of impacts estimates and standard errors calculated using Arellano and Bonhomme's (2012) mean group and robust variance estimators. See Section 4.4 for details.

to assume that treatment effects are conditionally independent of the error terms ε_{igs}. While this assumption is not directly testable, a necessary condition for it being satisfied is that the variance of $\Delta_i W_{igs} + \varepsilon_{igs}$ is greater for treatment observations than for control observations. This condition is satisfied in our sample; the variance of these partial residuals is 0.133 for treatment observations and 0.114 for control observations.

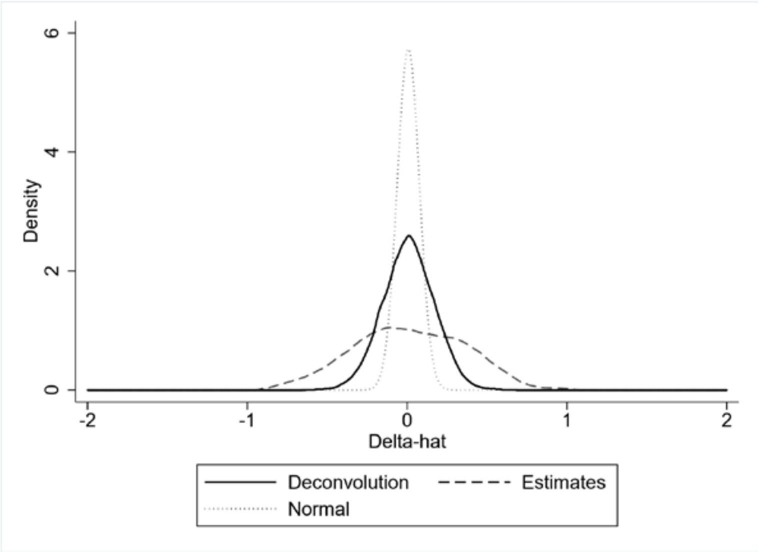


Figure 8: Estimates of Distribution of WSDP Impacts

The estimates of $\hat{\Delta}_i$ are very noisy because they are estimated using only 3 treatment observations and one control observation per treated school. Deconvolution attempts to disentangle the variation due to the treatment effects from the variation attributable to noise. The dashed line in Figure 8 shows the distribution of the unadjusted $\hat{\Delta}_i$ estimates. In contrast, the solid line shows the distribution of treatment effects recovered from applying deconvolution via Mallows' algorithm. As expected, the deconvolution estimate is much less dispersed than the unadjusted distribution. However, it is more dispersed and has heavier tails than the normal approximation to the distribution of treatment effects.

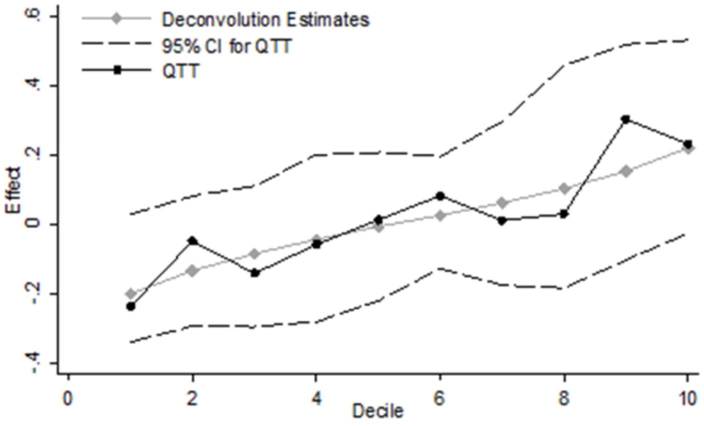


Figure 9: Deciles of Treatment Effects and QTTs

With knowledge of the full distribution of treatment effects, we can calculate any feature of the distribution. For example, we can calculate any quantile of the treatment effect distribution or any moment of the distribution, like the mean, variance, skewness or kurtosis of the distribution. In Figure 9, we present deciles of the deconvolution estimates of the distribution of treatment effects in gray. For comparison, we show quantile treatment effects at each decile in black. We calculate these quantile treatment effects using the empirical deciles of the partial residuals of a regression of average test scores in 2011 on grade level. While the distribution of treatment effects is calculated using the full panel, our panel functional form implies that treatment effects are the same in every year and should therefore be comparable to the quantile treatment effects calculated using only 2011 data. Pointwise confidence intervals for the QTTs are calculated using one thousand iterations of the pointwise bootstrap described in Subsection 6.2.1.

In this case, deciles of the distribution of treatment impacts look quite similar to the quantile treatment effects at each decile. The deciles of the distribution of treatment impacts are increasing by construction, but the QTTs trend upwards at almost exactly the same rate. As discussed in Parts 2 and 4, this is not the case in general. The similarity of these estimates suggests the WSDP did not cause much mobility across the potential outcome distributions. In other words, schools that ranked high in the control group distribution also ranked high in treatment group distribution. While none of the decile QTTs are significantly different from 0 based on the pointwise 95% confidence interval, the effect at the top decile, 0.23, is significant at the 10-percent level.

3. Simulation Results

The analyses in this section so far amend Blimpo, Evans and Lahire's (2016) finding of a zero average effect by providing evidence that this is due to zero or small individual effects rather than offsetting gains and losses. This demonstrates that DIA can provide further guidance for policy even in cases with no average impact. However, the lack of significance impairs the illustration of what else one can learn from DIA and such panel data and how failures of the assumptions can affect the results. To improve the illustration, we replicate our distributional analysis of the WSDP using simulated data mirroring the sample used for the WSDP. Specifically, we draw a simulated panel of 8 observations for each of 120 "schools" from the following data generating process:

$$y_{it} = \alpha_i + \Delta_i T_{it} + 0.5\sqrt{t}\gamma_i + \varepsilon_{it}.$$

With the following distributional assumptions:

$$\alpha_i \sim N(0,1), \begin{pmatrix} \Delta_i \\ \gamma_i \end{pmatrix} \sim N\left(\begin{bmatrix} 1.5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sqrt{0.5} \\ \rho\sqrt{0.5} & 0.5 \end{bmatrix} \right) \text{ and } \varepsilon_{it} \sim N(0, 0.01).$$

We include the $\sqrt{t}\gamma_{\mathrm{i}}$ term to explore the impact of violating the deconvolution assumptions on the estimates. In particular, Δ_i will be correlated with this term unless $\rho = 0$. As in the WSDP sample, 61 schools are randomly assigned to treatment beginning in year 2.

Table 8 presents estimates of the average and variance of treatment effects and Makarov Bounds on the proportion of schools who benefit from the program when ρ is 0, 0.1 and 0.8.[73] Panel A shows results when ρ is 0 so that the deconvolution assumptions are satisfied.

[73] Note that each simulated dataset was drawn with the same random seed so that the only difference in the estimated is due to changing ρ. Mirroring the above analysis, our controls include an indicator for whether the

Table 8: Illustrating Analyses Beyong the Mean Impact using Simulated Data

	True value from DGP	Estimate	Std. error
A. $\text{cov}(\Delta_i, \gamma_i) = 0$			
$\mathbb{E}(\Delta_i)$	1.50	1.57	0.15
$\text{var}(\Delta_i)$	1.00	1.37	0.23
Makarov Bounds on $\text{P}(\Delta_i \geq 0)$	0.93	[0.49, 1.00]	
B. $\text{cov}(\Delta_i, \gamma_i) = 0.1$			
$\mathbb{E}(\Delta_i)$	1.50	1.57	0.16
$\text{var}(\Delta_i)$	1.00	1.41	0.24
Makarov Bounds on $\text{P}(\Delta_i \geq 0)$	0.93	[0.48, 1.00]	
C. $\text{cov}(\Delta_i, \gamma_i) = 0.8$			
$\mathbb{E}(\Delta_i)$	1.50	1.55	0.18
$\text{var}(\Delta_i)$	1.00	1.64	0.30
Makarov Bounds on $\text{P}(\Delta_i \geq 0)$	0.93	[0.46, 1.00]	

Notes: N = 960 with 8 observations on each of 120 schools. Based on simulated data with identical structure to WSDP sample. Average and variance of impacts estimates and standard errors calculated using Arellano and Bonhomme's (2012) mean group and robust variance estimators. See Section 4.4.4 for additional details.

In this case, the estimated mean and variance of treatment effects are both statistically significant, but not significantly different from the true values. The Makarov Bounds for the proportion of schools which were hurt range from 0.49 to 1, which is consistent with roughly half of schools being hurt/benefit by the program to everyone benefitting, but include the true value of 0.93.

Panel B shows results when ρ is 0.1 so that the deconvolution assumptions are violated, but the correlation between the treatment effect and the omitted variable is relatively weak. Since the omitted variable is mean zero, the estimated average effect remains unbiased.

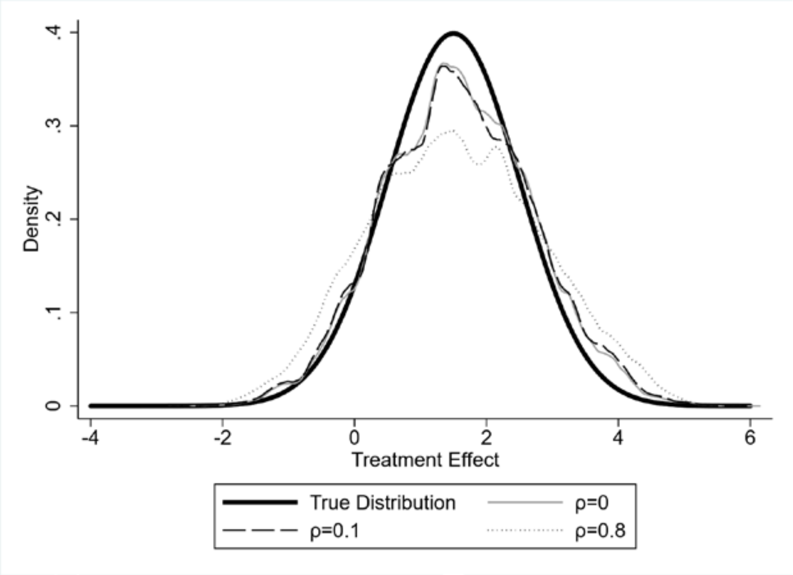


Figure 10: Estimates of Distribution of Simulated Impacts

observation is from an “older” grade (which are randomly selected since this coefficient does not enter the DGP) and year indicators.

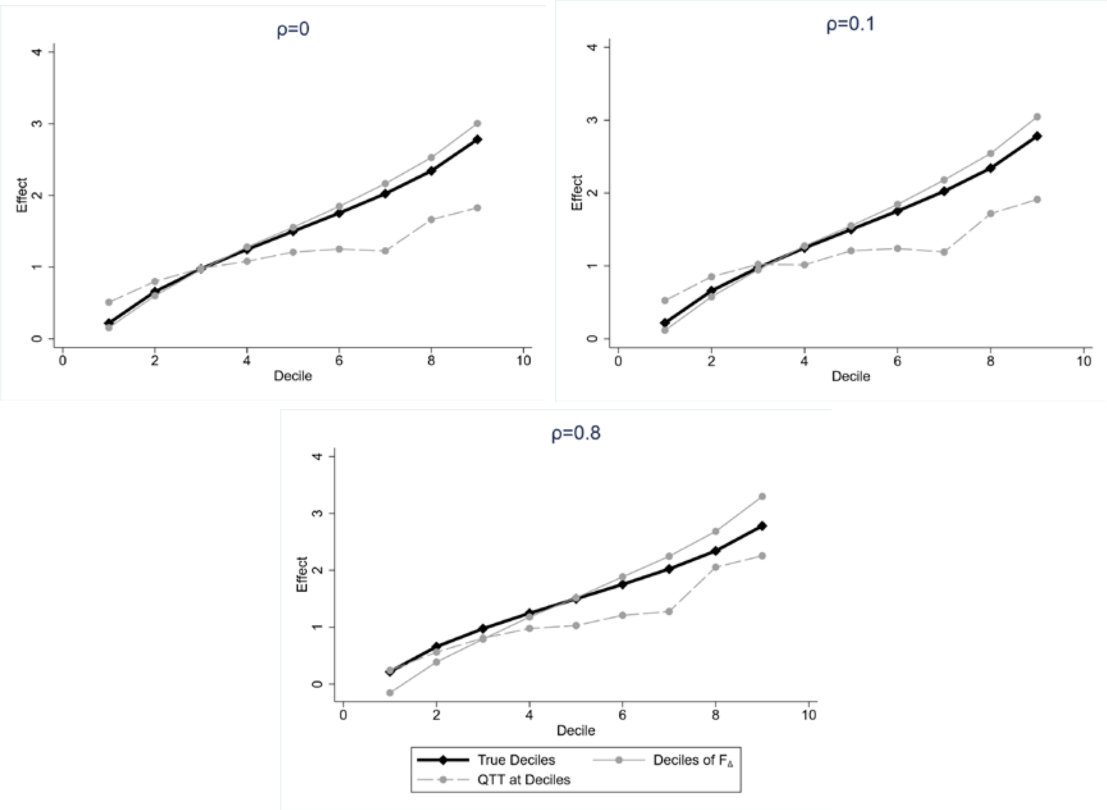


Figure 11: Deciles of Simulated Treatment Effects and QTTs

However, the estimated variance of treatment effects increased slightly and as a result is significantly different from the true variance at the ten-percent level. Panel C shows the case when ρ is 0.8, so that treatment effects are strongly correlated with the omitted variable. In this case, the estimated variance increases further and is significantly different from the true variance at the five-percent level. In both cases, the Makarov Bounds become somewhat wider, but are qualitatively quite similar.

Figure 11 plots the true distribution of treatment effects and the deconvolution estimates for each of the above cases. As Table 8 suggests, when ρ = 0 or 0.1, the estimated distribution is slightly more dispersed but generally similar to the true distribution of treatment effects. When ρ = 0.8, the estimated distribution is even more over dispersed.

The three figures in Figure 11 plot the deciles of the true distribution of treatment effects, the deconvolution estimate of the distribution of treatment effects, and estimated QTTs at each decile when ρ equals 0, 0.1, or 0.8, respectively. Similarly to Figure 10, the estimates are quite similar when ρ is 0 or 0.1. In both cases, the estimated deciles of the distribution of treatment effects are quite similar to the deciles from the true distribution. The correlation between the estimated and true deciles is about 0.92 in both cases. The deciles of QTTs are much flatter and generally smaller than the true deciles of treatment effects above the 4th decile. When ρ is 0.8, so that the deconvolution assumptions are violated, the deciles of the estimated distribution of treatment effects are still quite similar to the true deciles. But in this case, the QTTs at the deciles are about equicorrelated with the true deciles as the deconvolution estimates, whereas the QTTs were less correlated with the true deciles of the treatment effect in the other cases.

4. *Takeaways*

By going beyond the mean and estimating the distribution of treatment effects we were able to test whether the null average effect found in Blimpo, Evans, and Lahire (2015) was masking policy relevant treatment heterogeneity. The zero average effect does not appear to be masking heterogeneity in schools' response to the Whole School Development Program. In fact, the estimated variance implies that the standard deviation is about 70% of the estimated average effect and is also statistically insignificant. This finding of little treatment heterogeneity is corroborated by both the deconvolution estimate of the distribution and estimates of QTTs.

While relevant for policy, such a null result falls short of illustrating what we can learn from the methods we apply. The simulated results demonstrate that the null results are a feature of the particular program being studied rather than a deficiency of the distributional methods. Our estimates of both the variance and distribution of treatment effects are statistically indistinguishable from the truth even when the strong assumptions are violated. However, it is not clear how sensitive the methods are to violations of their assumptions in general, and we do not mean to suggest that the robustness of the estimates in our application generalizes.

References

AAKVIK, A., J.J. HECKMAN AND E.J. VYTLACIL (2005): “Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs”, *Journal of Econometrics* 125(1–2), 15–21.

ABADIE, A. (2002): “Bootstrap tests for distributional treatment effects in instrumental variable models”, *Journal of the American Statistical Association* 97(457), 284–292.

ABADIE, A., J. ANGRIST AND G.W. IMBENS (2002): “Instrumental variables estimation of quantile treatment effects”, *Econometrica* 70(1), 91–117.

ABBRING, J., AND J. HECKMAN (2007): “Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation”, *Handbook of Econometrics* 6, 5145–5303.

ANDREWS, D.W.K., AND M. BUCHINSKY (2000): “A three-step method for choosing the number of bootstrap repetitions”, *Econometrica* 68(1), 23–51.

ANGRIST, J.D., AND G.W. IMBENS (1994): “Identification and estimation of local average treatment effects”, *Econometrica* 62(2), 467–475.

ANGRIST, J., AND J.S. PISCHKE (2009): *Mostly harmless econometrics: An empiricist's companion*. Princeton (NJ): Princeton University Press.

ARCIDIACONO, P., E. AUCEJO, H. FANG AND K. SPENNER (2011): “Does affirmative action lead to mismatch? A new test and evidence”, *Quantitative Economics* 2(3), 303–333.

ARELLANO, M., AND S. BONHOMME (2012): “Identifying distributional characteristics in random coefficients panel data models”, *Review of Economic Studies* 79(3), 987–1020.

ATHEY, S., AND G.W. IMBENS (2002): “Recursive partitioning for heterogeneous causal effects”, arXiv:1504.01132.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): “Inference on treatment effects after selection among high-dimensional controls”, *Review of Economic Studies* 81(2), 608–650.

BITLER, M., J. GELBACH AND H.W. HOYNES (2006): “What mean impacts miss: Distributional effects of welfare reform experiments”, *American Economic Review* 96(4), 988–1012.

——— (2014): “Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment”, *NBER Working Paper* 20142.

BITLER, M., H.W. HOYNES AND T. DOMINA (2014): “Experimental evidence on distributional effects of Head Start”, *NBER Working Paper* 20434.

BISSANTZ, N., L. DÜMBGEN, H. HOLZMANN AND A. MUNK (2007): "Non-parametric confidence bands in deconvolution density estimation”, *Journal of the Royal Statistical Society B* 69(3), 483–506.

Black, D.A., J.A. Smith, M.C. Berger and B. J. Noel (2003): "Is the threat of reemployment services more effective than the services themselves? Experimental evidence from the UI system", *American Economic Review* 93(3), 1313–1327.

Blimpo, M., D.K. Evans and N. Lahire (2015): "Parental human capital and effective school management: Evidence from The Gambia", *World Bank Policy Research Working Paper* 7238.

Bonhomme, S., and J. Robin (2010): "Generalized nonparameteric deconvolution with an application to earnings dynamics", *Review of Economic Studies* 77(2), 491–533.

Brinch, C., M. Mogstad and M. Wiswall (forthcoming): "Beyond LATE with a discrete instrument", *Journal of Political Economy*.

Bruhn, M., L. de Souza Leão, A. Legovini, R. Marchetti and B. Zia (2016): "The impact of high school financial education: Evidence from a large-scale evaluation in Brazil", *American Economic Journal: Applied Economics* 8(4), 256–295.

Cameron, C., and D.L. Miller (2015): "A practitioner's guide to cluster-robust inference", *Journal of Human Resources* 50(2), 317–372.

Cameron, C., and P. Trivedi (2005): *Microeconometrics: Methods and applications*. New York (NY): Cambridge University Press.

Carneiro, P., K. Hansen and J. Heckman (2002): "Removing the veil of ignorance in assessing the distributional impacts of social policies", *NBER Working Papers* 8840.

——— (2003): "Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice", *International Economic Review* 44(2), 361–422.

Cattaneo, M. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability", *Journal of Econometrics* 155(2), 138–154.

Chernozhukov, V., I. Fernández-Val and A. Galichon (2010): "Quantile and probability curves without crossing", *Econometrica* 78(3), 1093–1125.

Chernozhukov, V., I. Fernández-Val and B. Melly (2013): "Inference on counterfactual distributions", *Econometrica* 81(6), 2205–2268.

Chernozhukov, V., and C. Hansen (2004): "The impact of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis", *Review of Economics and Statistics* 86(3), 735–751.

——— (2005): "An IV model of quantile treatment effects", *Econometrica* 73(1), 245–261.

——— (2013): "Quantile models with endogeneity", *Annual Review of Economics* 5, 57–81.

Chernozhukov, V., C. Hansen and M. Jansson (2009): "Finite sample inference in econometric models via quantile restrictions", *Journal of Econometrics* 152(2), 93–103.

Cornelissen, T., C. Dustmann, A. Raute and U. Schönberg (2016): "From LATE to MTE: Alternative methods for the evaluation of policy interventions", *Labour Economics* 41, 47–60.

Crump, R.K., V.J. Hotz, G.W. Imbens and O.A. Mitnik (2008): "Nonparametric tests for treatment effect heterogeneity", *Review of Economics and Statistics* 90(3), 389–405.

——— (2009): "Dealing with limited overlap in estimation of average treatment effects", *Biometrika* 96(1), 187–199.

Cunha, F., J. Heckman and S. Schennach (2010): "Estimating the technology of cognitive and non-cognitive skill formation", *Econometrica* 78(3), 883–931.

DiNardo, J., N.M. Fortin and T. Lemieux (1996): "Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach", *Econometrica* 64(5), 1001–1044.

Djebbari, H., and J.A. Smith (2008): "Heterogeneous impacts in progresa", *Journal of Econometrics* 145(1), 64–80.

Donald, S.G. and Y.C. Hsu (2014): "Estimation and inference for distribution functions and quantile functions in treatment effect models", *Journal of Econometrics* 178(3), 383–397.

Donald, S.G., Y.C. Hsu and R.P. Lieli (2014): "Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT", *Journal of Business & Economic Statistics* 32(3), 395–415.

Fafchamps, M., and J. Labonne (2016): "Using split samples to improve inference about causal effects", *NBER Working Paper* 21842.

Fan, Y., and S. Park (2010): "Sharp bounds on the distribution of treatment effects and their statistical inference", *Econometric Theory* 26(3), 931–951.

Fithian, W., D. Sun and J. Taylor (2017). "Optimal inference after model selection", arXiv:1410.2597v4.

Firpo, S. (2007): "Efficient semiparametric estimation of quantile treatment effects", *Econometrica*, 75(1), 259–276.

FIRPO, S., N.M. FORTIN AND T. LEMIEUX (2009): “Unconditional quantile regressions”, *Econometrica*, 77(3), 953–973.

FIRPO, S., AND C. PINTO (2015): “Identification and estimation of distributional impacts of interventions using changes in inequality measures”, *Journal of Applied Econometrics* 31(3): 457–486.

FIRPO, S., and G. RIDDER (2008). “Bounds on functionals of the distribution of treatment effects”, *Textos para Discussão* 201. São Paulo, SP: Escola de Economia de São Paulo.

FRÖLICH, M. (2006): “Non-parametric regression for binary dependent variables”, *Econometrics Journal* 9(3), 511–540.

FRÖLICH, M. (2007): “Propensity score matching without conditional independence assumption—with an application to the gender gap in the United Kingdom”, *Econometrics Journal* 10(2), 359–407.

FRÖLICH, M., AND B. MELLY (2013a): “Identification of treatment effects on the treated with one-sided non-compliance”, *Econometric Reviews* 32(3), 384–414.

——— (2013b): “Unconditional quantile treatment effects under endogeneity”, *Journal of Business & Economic Statistics* 31(3), 346–357.

HECKMAN, J.J., AND B. HONORÉ (1990): “The empirical content of the Roy model”, *Econometrica* 58(5), 1121–1149.

HECKMAN, J.J., J. SMITH AND N. CLEMENTS (1997): “Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts”, *Review of Economic Studies* 64(4), 487–535.

HECKMAN, J.J., AND E.J. VYTLACIL (1999): “Local instrumental variables and latent variable models for identifying and bounding treatment effects”, *Proceedings of the National Academy of Sciences*, vol. 96, 4730–4734.

——— (2005): “Structural equations, treatment effects, and econometric policy evaluation”, *Econometrica* 73(3), 669–738.

——— (2007): “Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments”, *Handbook of Econometrics* 6b, 4875–5143.

HIRANO, K., G.W. IMBENS AND G. RIDDER (2003): “Efficient estimation of average treatment effects using the estimated propensity score”, *Econometrica* 71(4), 1161–1189.

HOROWITZ, J.L. (2001): “The bootstrap”, *Handbook of Econometrics* 5, 3159–3228.

HOROWITZ, J.L., AND N.E. SAVIN (2001): “Binary response models: Logits, probits and semiparametrics”, *Journal of Economic Perspectives* 15(4), 43–56.

IMAI, K., AND M. RATKOVIC (2013): “Estimating treatment effect heterogeneity in randomized program evaluation”, *Annals of Applied Statistics* 7(1): 443–470.

IMBENS, G.W. (2015): “Matching methods in practice: Three Examples”, *Journal of Human Resources* 50(2), 373–419.

IMBENS, G.W., AND C.F. MANSKI (2004): “Confidence intervals for partially identified parameters”, *Econometrica* 72(6), 1845–1857.

IMBENS, G.W AND D.B. RUBIN (1997): “Estimating outcome distributions for compliers in instrumental variables models”, *Review of Economic Studies* 64(4), 555–574.

IMBENS, G.W AND J.M. WOOLDRIDGE (2009): “Recent developments in the econometrics of program evaluation”, *Journal of Economic Literature* 47(1), 5–86.

JACOBSON, L.S., R.J. LALONDE AND D.G. SULLIVAN (1993): “Earnings losses of displaced workers”, *American Economic Review* 83(4), 685–709.

JAYNES, E. (1957): “Information theory and statistical mechanics”, *Physics Review* 106(4), 620–630.

KLINE, P., AND C. WALTERS (2015): “Evaluating public programs with close substitutes: The case of Head Start”, *NBER Working Paper* 21658.

KOENKER, R. (2005): *Quantile regression*. Cambridge, UK: Cambridge University Press.

KOENKER, R., AND G. BASSETT (1978): “Regression quantiles”, *Econometrica* 46(1), 33–50.

KOENKER, R., AND Z. XIAO (2002): “Inference on the quantile regression process”, *Econometrica* 70(4), 1583–1612.

KOTLARSKI, I. (1967): “On characterizing the gamma and the normal distribution”, *Pacific Journal of Mathematics* 20(1), 69–76.

KOWALSKI, A.E. (2016): “Doing more when you’re running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments”, *NBER Working Paper Series* 22363.

LECHNER, M. (1999): “Earnings and employment effects of continuous off-the-job training in East Germany after unification”, *Journal of Business and Economic Statistics* 17(1), 74–90.

LEE, S., AND A.M. SHAIKH (2014): “Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment”, *Journal of Applied Econometrics* 29(4), 612–626.

LIST, J.A., A.M. SHAIKH AND Y. XU (2016): “Multiple hypothesis testing in experimental economics”, *NBER Working Paper* 21875.

MAKAROV, G. (1982): “Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed”, *Theory of Probability and its Applications* 26(4), 803–806.

MALLOWS, C. (2007): “Deconvolution by simulation”, in R. Liu, W. Strawderman and C.H. Zhang (eds.), *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, IMS Lecture Notes – Monograph Series, vol. 54, 1–11. New Brunswick, NJ: Rutgers University.

MANSKI, C.F. (2004): “Statistical treatment rules for heterogeneous populations”, *Econometrica* 72(4), 1221–1246.

MCCONNELL, B., AND VERA-HERNÁNDEZ, M. (2015): “Going beyond simple sample size calculations: A practitioner’s guide”, *IFS Working Paper* W15/17.

NEWEY, WHITNEY K., AND DANIEL L. MCFADDEN. (1994): “Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*. Vol. 4, ed. Robert F. Engle and Daniel L. McFadden, Chapter 36, 2111-2245. Amsterdam:Elsevier.

O’MUIRCHEARTAIGH, C., AND L.V. HEDGES (2014): “Generalizing from unrepresentative experiments: a stratified propensity score approach”, *Applied Statistics*, 63(2), 195–210.

PITT, M.M., M.R. ROSENZWEIG AND M.N. HASSAN (2012): “Human capital investment and the gender division of labor in a brawn-based economy”, *American Economic Review* 102(7), 3531–3560.

OTSU, T., AND Y. RAI (forthcoming): “Bootstrap inference of matching estimators for average treatment effects”, *Journal of the American Statistical Association*. DOI: 10.1080/01621459.2016.1231613.

RIDGEWAY, G., S.A. KOVALCHIK, B.A. GRIFFIN AND M.U. KABETO (2015): “Propensity score analysis with survey weighted data”, *Journal of Causal Inference* 3(2), 237–249.

ROMANO, J.P., AND M. WOLF (2010): “Balanced control of generalized error rates”, *Annals of Statistics* 38(1), 598–633.

ROTHE, C. (2012): “Partial distributional policy effects”, *Econometrica* 80(5), 2269–2301.

SCHNENNACH, S. (2013): “Convolution without independence”, *CEMMAP working paper* CWP46/13.

SMITH, J. (2015): “The important role of heterogeneity in social and biological models”, presentation at the RCGD/IHPI Seminar.

STOYE, J. (2009): “More on confidence intervals for partially identified parameters”, *Econometrica* 77(4), 1299-1315.

WAGER, S., AND S. ATHEY (2015): “Estimation and inference of heterogeneous treatment effects using random forests”, arXiv:1510.04342.

WASSERMAN, L., AND K. ROEDER (2009): “High-dimensional variable selection”, *Annals of Statistics* 37(5A), 2178–2201.

WU, X., and J. PERLOFF (2006): “Information-theoretic deconvolution approximation of treatment effect distribution”, unpublished manuscript. College Station, TX: Texas A&M University.

ZHANG, Y. (2016): *Three Essays on Extremal Quantiles*, Ph.D. dissertation. Durham, NC: Duke University. Available at: hdl.handle.net/10161/12160.

Appendix 1. Simulation Details

Throughout the toolkit, we use results from simulation exercises for illustration. In this appendix, we describe the data generating processes used for each of these simulations.

Simulation 1

Simulation 1 is our workhorse simulation, because it can be used to generate cross-sectional data with baseline outcome measures or panel data with an arbitrary number of periods.

In this simulation, individual i’s outcome in period s is:

$$Y_{is} = \alpha_i + \varDelta_i R_{is} + \varepsilon_{is}.$$

Moreover, we assume

$$\begin{pmatrix}\alpha_i\\ \Delta_i\end{pmatrix}\sim N\left(\begin{bmatrix}0\\1.5\end{bmatrix},\begin{bmatrix}0.5 & 0.5\\0.5 & 1.0\end{bmatrix}\right),$$

and $\varepsilon_{is}\sim N(0,1)$. We require exactly half of observations to be randomly selected for treatment beginning in period $S/2$.[74]

Simulation 2

Simulation 2 draws data from two data generating processes with identical average treatment effects, but very different levels of heterogeneity in treatment effects. Specifically, we assume the following data-generating process:

$$Y_i^1 = \Delta_i^1 R_i + \varepsilon_i,$$
$$Y_i^2 = \Delta_i^2 R_i + \varepsilon_i.$$

For each individual, we draw a single ε_i and R_i. We assume $\varepsilon_i\sim N(0,1)$. Furthermore, we require exactly half of observations be selected for treatment by drawing a uniform random variable for each individual and selecting the treatment cutoff using the median value across all draws. As for treatment effects, we assume $\Delta_i^1\sim N(1,0.5^2)$ and $\Delta_i^2\sim N(1,5^2)$.

Simulation 3

In Simulation 3, potential outcomes are:

$$Y_1 = 1 + cX_1 + \varepsilon_1,$$
$$Y_0 = 0.5X_1 + \varepsilon_0,$$

where

$$\begin{pmatrix}\varepsilon_1\\ \varepsilon_0\end{pmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1.0 & \sigma_{12}\\ \sigma_{12} & 0.5\end{bmatrix}\right).$$

Note that Y_1 and Y_0 are independent when $\sigma_{12}=0$, whereas Δ and Y_0 are independent when $\sigma_{12}=0.5$.

In addition, we assume $X_1\sim N(1,1)$. Treatment effects are then:

$$\Delta = Y_1 - Y_0 = 1 + (c-0.5)X_1 + \varepsilon_1 - \varepsilon_0.$$

We do not condition on X_1 in our simulations. Consequently, there is an omitted variable which is positively correlated with treatment effects when $c>0.5$ and negatively correlated with treatment effects when $c<0.5$. There is no omitted variable when $c=0.5$.

Appendix 2. Estimating Conditional Probabilities

[74] If S is odd, we use $\lceil S/2 \rceil$.

To construct the IPW and IV weights, we need conditional probabilities $\mathrm{P}(\cdot\,|X)$. We might know them a priori. If our instrument is treatment assignment and we only want to adjust for stratification, for example, the randomization rules provide the stratum specific treatment probabilities. Otherwise, we need to estimate and predict $\mathrm{P}(\cdot\,|X)$.

There are many strategies to predict $\mathrm{P}(\cdot\,|X)$. If the covariates are discrete, Abadie, Angrist and Imbens (2002) recommend sorting the data into cells and using the proportion of treated observations within each cell. This method is nonparametric and efficient. For continuous covariates, Hirano, Imbens and Ridder (2003) suggest a flexible logistic specification, including polynomials of covariates and interaction terms. The authors give conditions under which the resulting estimator is nonparametric. The appendix of Imbens (2015) presents an algorithm to select the higher-order terms. One can also use a LASSO procedure to select controls: see Athey and Imbens (2015) and Belloni, Chernozhukov and Hansen (2014) for references. Alternative semiparametric or nonparametric strategies are available. Because the variance of binary outcomes is intrinsically bounded, these methods perform well. See Horowiz and Savin (2001) and Frölich (2006) for references.

Both the IV and the IPW estimators are sensitive to observations in the tails of $\mathrm{P}(\cdot\,|X)$. Unfortunately, it is difficult to predict this conditional probability with much accuracy near its boundaries. Crump et al. (2009) suggest trimming the sample and ignoring observations for which $\mathrm{P}(\cdot\,|X)$ is close to zero or one when estimating quantiles. If the rate of trimming decreases with sample size, the resulting estimator remains consistent. The authors suggest using only observations with predicted probability between 0.1 and 0.9 as a rule of thumb.

Appendix 3. Recursively Solving for Higher Order Moments

The Binomial Theorem implies:

$$\mathbb{E}[(\Delta_i + \varepsilon_i)^K] = \mathbb{E}\left[\sum_{k=1}^{K}\binom{K}{k}\Delta_i^{K-k}\varepsilon_i^k\right] = \mathbb{E}(\Delta_i^K) + \sum_{k=1}^{K}\binom{K}{k}\mathbb{E}(\Delta_i^{K-k})\mathbb{E}(\varepsilon_i^k).$$

For any k, $\mathbb{E}[(\Delta_i + \varepsilon_i)^K]$ can be estimated from the treatment group residuals and each $\mathbb{E}(\varepsilon_i^k)$ can be estimated from the control group residuals. Therefore, this is a system of K independent equations and K unknowns that we can solve for $\mathbb{E}(\Delta_i^K)$. This gives us K equations for the first K moments of the treatment effect distribution:

$$\mathbb{E}(\Delta_i^k) = \mathbb{E}[(\Delta_i + \varepsilon_i)^K] - \sum_{k=1}^{K}\binom{K}{k}\mathbb{E}(\Delta_i^{K-k})\mathbb{E}(\varepsilon_i^k).$$

These moments can be estimated recursively using the sample counterparts to each term in the above equation. Specifically,

$$\widehat{\mathbb{E}}[(\Delta_i + \varepsilon_i)^K] = \frac{1}{N_1}\sum_{i=1}^{N_1}(Y_{1i} - X_i\,\hat{\beta})^K,$$

$$\widehat{\mathbb{E}}(\varepsilon_i^k) = \frac{1}{N_0}\sum_{i=1}^{N_0}(Y_{0i} - X_i\hat{\beta})^k.$$

where $\hat{\beta}$ is the OLS coefficient from a regression of Y on X and N_1 and N_0 are the number of treatment and control group observations, respectively.

Appendix 4. Additional Results from Applications

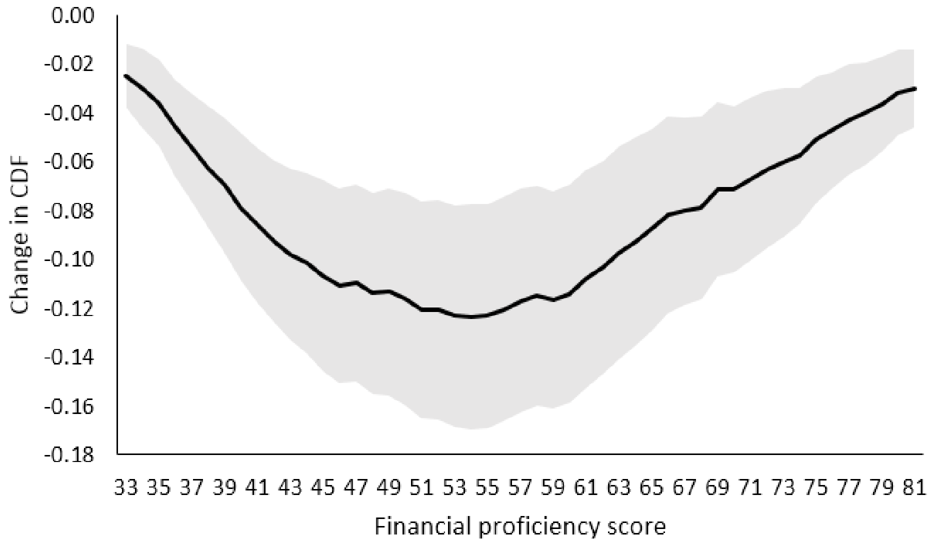


Figure 12: Effects on the CDF for Financial Education Program

Table 9: Reweighted Treatment Effects for Financial Education Program

Statistic	**Control-group value**	**Effect estimate**		**Standard error**	**Simultaneous conf. region (95%)**	
		Value	Percent			
Mean	56.260	3.859	6.886%	0.427	2.636	5.083
Standard deviation	14.768	−0.382	−2.583%	0.187	−0.753	−0.012
75/25 perc. ratio	1.525	−0.106	−6.904%	0.019	−0.157	−0.054
10th percentile	33.047	3.887	10.706%	0.543	2.311	5.462
25th percentile	44.134	5.613	12.795%	0.733	3.564	7.662
50th percentile	56.438	4.043	7.200%	0.512	2.563	5.524
75th percentile	67.326	3.228	4.808%	0.404	2.023	4.434
90th percentile	75.894	3.099	4.088%	0.404	1.720	4.480
CDF at 40 points	0.164	−0.070	−40.897%	0.009	−0.098	−0.043
CDF at 50 points	0.352	−0.103	−28.144%	0.013	−0.140	−0.066
CDF at 60 points	0.581	−0.102	−17.232%	0.012	−0.137	−0.068
CDF at 70 points	0.801	−0.066	−8.188%	0.009	−0.093	−0.040
CDF at 80 points	0.945	−0.031	−3.315%	0.005	−0.045	−0.017

Notes: Effects based on inverse probability weighting (Firpo and Pinto, 2016). Weights based on logistic regression of treatment on gender indicator and quadratic polynomial of baseline scores. Standard errors and confidence region based on the bootstrap (five thousand replications) and clustered at the school level. The confidence region controls the FWER (probability of at least one false rejection across tests), following Romano and Wolf (2010).

Table 10: Conditional Quantile Effects for Financial Education Program

Subgroup		Sample size	Percentile				
			10th	25th	50th	75th	90th
Baseline score above median	No	7960	3.263	4.645	5.060	3.793	3.497
			(0.460)	(0.664)	(0.635)	(0.554)	(0.615)
	Yes	7960	8.875	4.953	2.992	2.853	3.157
			(1.275)	(0.864)	(0.572)	(0.482)	(0.543)
Student has repeated grade	No	10949	5.197	6.522	4.516	3.608	3.306
			(0.716)	(0.867)	(0.670)	(0.642)	(0.691)
	Yes	4437	3.791	5.592	5.568	4.372	3.261
			(0.661)	(0.773)	(0.794)	(0.889)	(0.914)
Student is female	No	6941	3.959	6.657	5.943	4.601	3.803
			(0.662)	(0.863)	(0.938)	(0.861)	(0.875)
	Yes	8720	4.843	5.350	3.626	2.659	3.080
			(0.870)	(0.842)	(0.634)	(0.684)	(0.631)
Student works	No	10898	4.872	5.906	4.088	3.342	3.232
			(0.636)	(0.874)	(0.713)	(0.682)	(0.629)
	Yes	5612	3.313	6.265	5.550	4.865	3.983
			(0.740)	(0.868)	(0.922)	(0.912)	(0.871)
Student earns income	No	5713	4.117	5.467	4.025	2.337	2.331
			(0.724)	(0.990)	(0.800)	(0.733)	(0.747)
	Yes	10812	4.386	6.864	4.773	4.136	3.515
			(0.648)	(0.936)	(0.709)	(0.692)	(0.583)
Family is on welfare (Bolsa Família)	No	10216	5.145	7.522	5.356	4.081	3.548
			(0.704)	(0.985)	(0.757)	(0.751)	(0.679)
	Yes	5334	3.618	5.152	3.559	2.628	2.245
			(0.787)	(0.894)	(0.900)	(0.869)	(0.884)

Notes: The first four specifications include 848 schools. The last two include 851 schools. Bootstrap standard errors in parentheses, based on five thousand replications and clustered at the school level.