Non-existent outcomes in research on inequality: A causal approach*

Ian Lundberg† Soonhong Cho‡

May 26, 2025

Keywords: partial identification, causal inference, social stratification, inequality, demography

Abstract

Scholars of social stratification often study exposures that shape life outcomes. But some outcomes (such as wage) only exist for some people (such as those who are employed). We show how a common practice—dropping cases with non-existent outcomes—can obscure causal effects when a treatment affects both outcome existence and outcome values. The effects of both beneficial and harmful treatments can be underestimated. Drawing on existing approaches for principal stratification, we show how to study (1) the average effect on whether an outcome exists and (2) the average effect on the outcome among the latent subgroup whose outcome would exist in either treatment condition. To extend our approach to the selection-on-observables settings common in applied research, we develop a framework involving regression and simulation to enable principal stratification estimates that adjust for measured confounders. We illustrate through an empirical example about the effects of parenthood on labor market outcomes.

*An R package in development is available at ilundberg.github.io/pstratreg. For helpful discussions and feedback relevant to this project, we thank Brandon Stewart, Jennie Brand, and members of the Inequality Data Science Lab at UCLA. This research benefited from feedback in presentations at the Cornell Center for the Study of Inequality, the UCLA Department of Sociology, and the American Causal Inference Conference. The authors benefited from facilities and resources provided by the California Center for Population Research at UCLA (CCPR), which receives core support (P2C-HD041022) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

†UCLA Department of Sociology and California Center for Population Research, ianlundberg.org, ianlundberg@ucla.edu.

‡UCLA Department of Political Science, soonhong-cho.github.io, tnsehdtm@gmail.com.

1 Introduction

Many social and economic outcomes exist only for some people. Only employed individuals have wages, only married people can report marital satisfaction, and only those with children can transmit socioeconomic advantages to their descendants. Scholars studying these outcomes often restrict their analysis to those for whom the outcomes exist: the employed, the married, and the parents. Yet treatments that shape the outcome values often also affect whether those outcomes exist at all.

We show how the standard practice of restricting analysis to those with observed outcomes can obscure causal effects, and we provide tools to resolve this problem. We focus on settings where a binary treatment shapes both (1) whether an outcome exists and (2) the value the outcome would take if it were to exist. Drawing on ideas for principal stratification developed for randomized experiments (Frangakis and Rubin, 2002), we define two quantities that researchers might want to study: the average causal effect on outcome existence and the average causal effect on outcome value among the latent subgroup who would have an outcome regardless of treatment condition. To adapt these methods to the selection-on-observables research designs common in quantitative sociology, we develop a new framework using regression to adjust for measured confounders and simulation to carry out principal stratification. We illustrate our approach with an application to the causal effect of motherhood on women's employment and hourly wages.

2 Conceptual Framework

Causal exposures often cause outcomes to come into existence or to cease to exist. This section illustrates the misleading conclusions that can arise when researchers focus solely on those whose outcomes exist.

Figure 1A illustrates four hypothetical people who could receive treatment (job training). Two would have lower wages without job training and higher wages with job training. The other two would be non-employed without job training but would become employed with low wages with job training. A researcher who naively compared the average observed wage under treatment and control, however, would conclude that there was no effect: the observed mean wage among those employed is the same under treatment and control. Thus, a treatment that benefits everyone—either by improving wages or inducing employment—appears (misleadingly) to have zero average

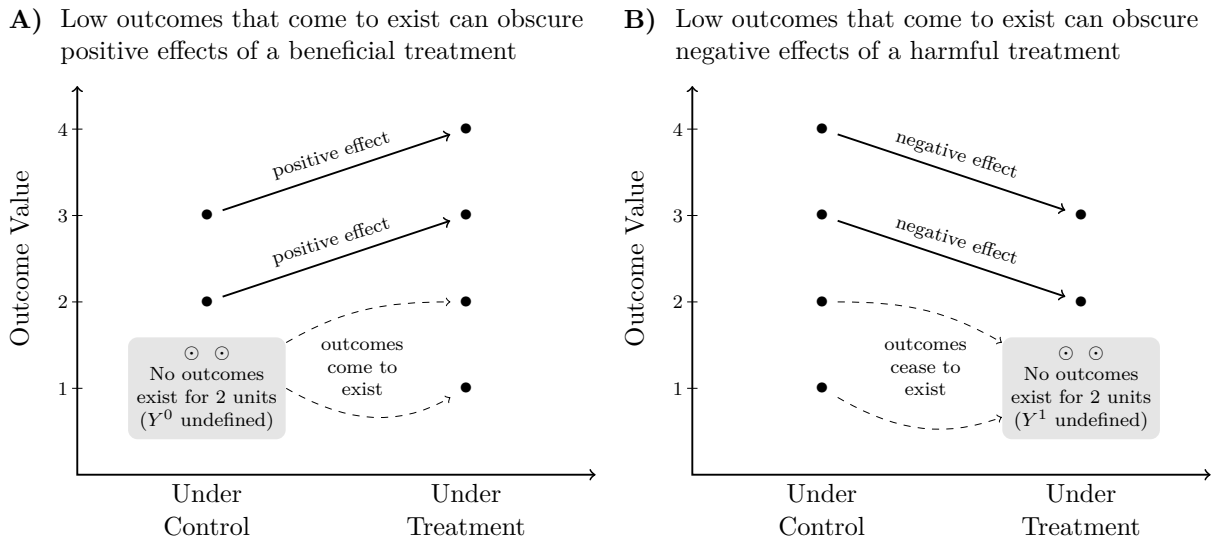


Fig. 1. Non-existent outcomes can obscure beneficial or harmful effects. In Panel A, a treatment with positive effects also causes low outcomes to come to exist. A job training program lifts wages for two workers while helping two lower-paid workers to find jobs. In Panel B, a treatment with negative effects also causes low outcomes to cease to exist. Motherhood causes two women's wages to decline and causes two other women to leave paid employment entirely. In both panels, the mean value of the observed outcomes takes the same value (2.5) in both the control and the treatment conditions.

effect.

Figure 1B illustrates four hypothetical employees facing potential firm downsizing. Two would have higher wages under no change but would experience wage declines under downsizing, while the other two would have lower wages under no change but would lose their jobs if the firm downsized. The naive comparison of average observed wages would again be misleading. A treatment (downsizing) that harms all four workers would appear to have zero average causal effect.

The hypothetical examples in Figure 1 are carefully designed to illustrate the point, but correspond to processes that are widespread in social stratification. As in Panel A, treatments that improve outcomes often bring new people into the population whose outcome values are low. In labor markets, job training programs raise wages for employed participants while helping previously unemployed people secure positions, typically at lower wages (Card et al., 2018; Schochet et al., 2008). Community investment programs in disadvantaged neighborhoods can raise wages for some residents while helping others avoid incarceration through new employment opportunities (Western, 2006), and these otherwise-incarcerated people might join the pool of employed people at low

wages. In educational contexts, tutoring programs boost achievement among enrolled students and prevent at-risk students from dropping out (Nickow et al., 2024), while college completion improves earnings and increases employment rates (Brand, 2023; Brand and Xie, 2010). In each case, the entry of these individuals with lower outcomes into the population of study can obscure the positive effects of the treatment.

As in Panel B, the negative impact of harmful treatments can also be obscured when these treatments cause low outcomes to disappear. In labor markets, automation may depress wages and also eliminate some of the lowest-wage positions held by lower-skilled workers (Acemoglu and Restrepo, 2019; Autor et al., 2003). Employment rates that exclude the incarcerated may underestimate the negative effects of economic shocks on labor market outcomes if those same shocks cause the most vulnerable to engage in criminal activity and become incarcerated (Western, 2002). In families, financial strain might harm marital quality while pushing the most fragile marriages toward dissolution; average marital quality may seem unchanged simply because the worst-off marriages cease to exist (Conger et al., 1990). In urban sociology, local gentrification could weaken social ties among residents, while also displacing those with the most tenuous connections so that they are no longer observed in the neighborhood at all (Desmond, 2016; Hwang and Sampson, 2014).

These patterns—beneficial treatments creating low outcomes and harmful treatments eliminating them—illustrate the problem of non-existent outcomes in research on inequality. Analyses focusing solely on existing outcomes may understate both positive and negative treatment effects. Addressing this challenge requires methodological tools designed to study the existence or non-existence of outcomes as part of the causal process. In the next section, we formalize this problem using the potential outcomes framework (Neyman, 1923; Rubin, 1974) and provide a concrete example examining parenthood's effects on employment and wages.

2.1 Potential outcomes formalize the selection problem

We illustrate with a concrete example motivated by studies of how parenthood affects wages differently by gender. The uneven impact of parenthood on wages is a main source of gender pay inequality: fathers generally do not see a wage penalty or even earn more (Killewald, 2013), while mothers often face significant wage decreases (Budig and England, 2001; England et al., 2016;

		Employment			Hourly wage		
		if a mother − if not = effect			if a mother − if not = effect		
Maya	is a mother	[1]	1	0	[$30]	$40	-$10
Nancy	is not a mother	1	[1]	0	$30	[$40]	-$10
Mia	is a mother	[0]	1	-1	[??]	$20	??
Nia	is not a mother	0	[1]	-1	??	[$20]	??

Fig. 2. Conceptual illustration: Non-existent outcomes can obscure the motherhood wage penalty. For a hypothetical set of four women, the figure depicts potential employment and wage outcomes that each individual would realize as a mother and as a non-mother. Boxes denote factual outcomes that would be observed in data. In this illustration, motherhood reduces employment by 50 percentage points on average. The motherhood wage penalty is -$10 for Maya and Nancy, but it is an undefined quantity for Mia and Nia because their wage as a mother does not exist; they would not be employed in that condition. Despite beginning with a perfectly matched set of two mothers and two non-mothers, a researcher who dropped the non-employed (Mia) would induce selection bias such that there would incorrectly appear to be zero motherhood wage penalty on average.

Gough and Noonan, 2013; Staff and Mortimer, 2012; Waldfogel, 1997). It is common in this literature to either condition on employment by focusing the analysis on employed individuals with wages, or to impute wages for the non-employed (often with simply zero or with predicted values from assumed models on missingness). Both approaches involve steps that might make researchers and readers uneasy. Conditioning on employment creates selection bias because employment is itself affected by the treatment, making employed and non-employed groups non-comparable. Wage imputation requires strong assumptions about outcomes that fundamentally do not exist: someone who refuses to tell you their wage has an unobserved wage, but someone who is not employed truly has no wage at all. These standard approaches obscure important aspects of inequality and produce unwarranted confidence in estimated effects on wage outcomes.

We illustrate the problem using four hypothetical women: Maya, Nancy, Mia, and Nia (Figure 2). For analytical clarity, we assume we know their potential wages under both motherhood and non-motherhood conditions—information unavailable in real data but useful for understanding the underlying logic.

Maya and Nancy are two women whose potential hourly wages are identical. Motherhood has no effect on their employment: they would remain employed regardless of motherhood. However, motherhood has a causal effect on their hourly wages: Each would earn $30 as a mother but $40 if not a mother. The only difference between Maya and Nancy is their realized treatment status:

Maya is a mother while Nancy is not. Consequently, a researcher would only observe outcomes in their respective treatment conditions. If the researcher knew that Maya and Nancy had identical potential outcomes, the researcher could correctly estimate their $10 motherhood wage penalty by matching them and subtracting Nancy's wage as a non-mother from Maya's wage as a mother.

Mia and Nia present another scenario. Neither would work for pay if they became mothers, but both would work at $20 per hour if they remained childless. For Mia and Nia, motherhood would have a large negative effect on employment. The effect of motherhood on hourly wage, however, is an undefined quantity: the difference between a non-existent wage as a non-employed mother and a $20 per hour wage as an employed non-mother. For Mia and Nia, employment status, rather than wage, constitutes the relevant labor market outcome.

Because the motherhood wage penalty is undefined for Mia and Nia, the average treatment effect is undefined for the population. The average effect of motherhood on hourly wages may not be the right causal question in this population. Instead, two causal questions emerge. (1) What is the average effect of motherhood on employment? In this population, motherhood reduces employment by 50% because it has an effect of -1 in half of the population (Mia and Nia) and has an effect of 0 in the other half (Maya and Nancy). (2) Among those who would be employed regardless, what is the average effect of motherhood on hourly wage? In the subpopulation of Maya and Nancy, motherhood reduces the hourly wage by $10 on average.

2.2 How standard practice can mislead

Standard practice produce misleading conclusions in this setting. The boxed values in Figure 2 show the outcomes observable to researchers. When Mia is a mother, motherhood prevents her employment, resulting in no observable wage. A researcher following standard practice would exclude her from the analysis, resulting in a dataset with only one employed mother (Maya, $30/hour) and two employed non-mothers (Nancy, $40/hour; Nia, $20/hour). While these women have different outcomes, we assume for this illustration that all four are identical on observed pre-treatment variables. If conducting matching, for instance, there would be no way to know whether to match Maya to Nancy or to Nia.

Comparing the average wage of employed mothers ($30) with employed non-mothers ($\frac{\$40+\$20}{2}$ = $30) would erroneously suggest no motherhood wage penalty. This conclusion is particularly

troubling given the true causal structure: motherhood harms all four women's labor market outcomes, either preventing employment (Mia and Nia) or reducing wages (Maya and Nancy). By excluding non-employed women, researchers inadvertently condition on a post-treatment variable (employment status), inducing post-treatment bias (Montgomery et al., 2018).

2.3 Heckman selection models: A solution that does not work in our setting

Our example is similar to the classic selection problem in labor market studies: people can report wages only if they are employed (Blau and Kahn, 2017; Gronau, 1974; Winship and Mare, 1992). A popular solution is the Heckman selection model (Heckman, 1979), which begins by defining an outcome that exists for all units. For a non-employed mother, we might define her outcome Y to equal the wage she would be paid if counterfactually employed. Under this perspective, outcomes exist for all units but are observed only for the employed. We do not adopt this solution for two reasons. First, we find it philosophically difficult to define potential wages for those who are not employed: doing so changes the outcome of interest from realized wages to potential wages that would be realized under employment. Our approach instead takes the realized wage as the outcome.

The second reason concerns causal identification: the selection model approach requires an instrument that affects employment but is independent of the potential wages that would be realized if employed. With such an instrument and additional assumptions, it becomes possible to reweight the cases with observed outcomes to draw inference about all cases. But we would argue that such an instrument rarely exists for the kinds of questions that arise in social stratification. For example, it is difficult to imagine a variable that strongly predicts employment but is independent of the wage that would be realized if employed, thus satisfying what is sometimes called the exclusion restriction. Such a variable is hard to imagine because many causes of employment also shape wages. We instead prefer a solution that focuses on realized wages and allows the full causal process shaping employment to also shape wages.

2.4 Notation

Before presenting our preferred approach, we define notation that will be used throughout the paper. Let S^a and Y^a be potential outcome existence and outcome value under treatment value a. In the example above, outcome existence S corresponds to employment and the outcome value Y

corresponds to wage. For units with $S^a = 0$ the outcome Y^a does not exist. The treatment values a indicate parenthood or non-parenthood in our example. Denote $\mu(a, \vec{x}) = \mathbf{E}(Y \mid S = 1, A = a, \vec{X} = \vec{x})$ the mean observed outcome among those whose outcome exists with treatment value a and confounders $\vec{x}$. Let $\mu^a(\vec{x}) = \mathbf{E}(Y^a \mid S^0 = S^1 = 1, \vec{X} = \vec{x})$ be the mean potential outcome under treatment value a among those with confounder vector $\vec{x}$ whose outcome would exist under either treatment. Denote $\tau(\vec{X}) = \mu^1(\vec{X}) - \mu^0(\vec{X})$ the difference in these mean functions, corresponding to a conditional average causal effect on outcome values. We use subscripts "Lower" and "Upper" to refer to lower and upper bounds, e.g. $\mu^0_{\text{Lower}} \leq \mu^0(\vec{X}) \leq \mu^0_{\text{Upper}}$. Let $\pi(a, \vec{x}) = \mathrm{P}(S = 1 \mid A = a, \vec{X} = \vec{x})$ be the probability of outcome existence given treatment value a and confounder vector $\vec{x}$. Let $\pi_{S^0=S^1=1}(\vec{x}) = \mathrm{P}(S^0 = S^1 = 1 \mid \vec{X} = \vec{x})$ be the probability of having an outcome regardless of treatment, conditional on confounder vector $\vec{x}$, and $\pi_{S^0=S^1=1|S=1}(a, \vec{x}) = \mathrm{P}(S^0 = S^1 = 1 \mid S = 1, A = a, \vec{X} = \vec{x})$ be the probability of outcome existence under either treatment, conditional on observed outcome existence under treatment value a.

2.5 Principal stratification provides better causal estimands when some outcomes do not exist

The challenge exemplified by Mia and Nia—where treatment affects whether an outcome exists at all—has been well-studied in biostatistics and epidemiology for medical trials where some participants do not survive to the end of the trial ("truncation by death"), possibly due to the intervention itself (Cheng and Small, 2006; Ding et al., 2011; Zhang and Rubin, 2003). The analytical structure remains analogous; just as death renders health outcomes non-existent, non-employment likewise renders wages non-existent. Economists have applied similar methods to job-training programs where training influences both employment probability and wages, creating non-random selection into the observed wage sample (Frumento et al., 2012; Lee, 2009; Zhang et al., 2009). Political scientists have used principal stratification to study racial disparities in police use of force, where interactions with police only exist among those stopped by police, with race potentially affecting both the stopping decision and subsequent force (Knox et al., 2020).[1] Our approach builds upon these applications but makes a distinct contribution by using parametric models to adjust for measured confounders.

[1] More political science examples can be found in Slough (2023).

The approach recognizes that units belong to latent groups termed *principal strata*, defined by their potential survival outcomes S^1 under treatment and S^0 under control. A unit "survives" if it has a defined outcome Y at the end of the study. In our example, survival corresponds to employment. The four women in our example fall into two principal strata: Maya and Nancy would be employed regardless of motherhood ($S^1 = 1, S^0 = 1$), while Mia and Nia would be employed only as non-mothers ($S^1 = 0, S^0 = 1$). With binary treatment and employment status, four principal strata exist: employed regardless of motherhood, employed only as non-mothers, employed only as mothers, and non-employed regardless of motherhood. Only the first two appear in Figure 2.

The first causal estimand of interest is the average causal effect on the existence of the outcome (i.e., survival), which in our example is the average difference in employment that would be realized as a mother (S^1) and employment that would be realized as a non-mother (S^0).

$$\mathbf{E}\left(S^1 - S^0\right) \tag{1}$$

When the effect on survival is large, many units have an outcome under only one treatment condition. For these units, the average causal effect on the outcome Y is undefined since either Y^0 or Y^1 does not exist. The insight of principal stratification is that a meaningful contrast $Y^1 - Y^0$ can only be well-defined for the principal stratum whose outcome exists regardless of treatment ($S^1 = S^0 = 1$).[2] Principal stratification targets the average causal effect on the outcome among the latent subgroup who would have an outcome regardless of treatment.[3]

$$\tau = \mathbf{E}\Bigg(\underbrace{Y^1 - Y^0}_{\substack{\text{Effect of Treatment} \\ \text{on Outcome Value}}} \quad | \quad \underbrace{S^1 = S^0 = 1}_{\substack{\text{Among Those Whose Outcome} \\ \text{Exists Regardless of Treatment}}} \Bigg) \tag{2}$$

In our example, this is the average effect on wage Y among those who would be employed regardless of parenthood.

[2]Principal stratification extends beyond addressing non-existent outcomes, providing a framework for handling various post-treatment complications. The method's most prominent application appears in studies of imperfect compliance, where instrumental variables estimate effects for the "complier" principal stratum—those units whose treatment status would align with their assignment under both treatment and control conditions (Angrist et al., 1996).

[3]This local causal estimand is termed the "Survivor Average Causal Effect" (SACE) in biostatistics.

Estimating τ is difficult for two reasons. First, as with all causal estimands, at most one of Y^1 or Y^0 (potential outcomes) is observed for any unit. Second, only one of S^1 or S^0 (potential outcome existence) is observed. In concrete terms, for an employed non-mother like Nia, we observe $S^0_{\text{Nia}} = 1$ but cannot observe her counterfactual employment status as a mother $S^1_{\text{Nia}} = 0$. Data therefore cannot tell us if Nia belongs to our target subgroup (those employed regardless of motherhood) or to the subgroup that creates problems for the analysis (employed only if a non-mother). Because the estimand involves counterfactuals for both Y and S, identification requires assumptions beyond those typically used for causal effects.

3 Assumptions for principal stratification

Principal stratum causal effects like Eq. 2 can be point-identified only under very strong assumptions, but can be set-identified within bounds under weaker, more credible assumptions (Knox et al., 2020; Miratrix et al., 2018). The stronger our assumptions, the tighter the bounds we are able to construct. This section introduces three assumptions one might make: conditional exchangeability, monotonicity, and mean dominance.

3.1 Assumption: Conditional exchangeability

We first assume conditional exchangeability (no unmeasured confounding): potential outcome existence is independent of treatment given measured confounders $\vec{X}$.

$$\{S^0, S^1\} \perp\!\!\!\perp A \mid \vec{X} \quad \text{for } a \in \{0,1\} \qquad \text{(conditional exchangeability for outcome existence)} \tag{3}$$

This assumption enables identification of the average causal effect on outcome existence by conditioning on $\vec{X}$, by an identification result standard in observational causal inference.

$$\mathbf{E}(S^1 - S^0) = \overbrace{\mathbf{E}_{\vec{X}}}^{\substack{\text{Population}\\ \text{Mean}}} \overbrace{\left(\underbrace{\pi(1,\vec{X})}_{\substack{\text{Among Treated}\\ \text{Given } \vec{X}}} - \underbrace{\pi(0,\vec{X})}_{\substack{\text{Among Untreated}\\ \text{Given } \vec{X}}} \right)}^{\substack{\text{Difference in Outcome}\\ \text{Existence Probabilities}}} \tag{4}$$

Because Y^a is only defined when $S^a = 1$, conditional exchangeability for outcomes is more

complex. At each treatment value $a = 0$ and $a = 1$, we assume that the potential outcome value Y^a is independent of treatment among those whose outcome would exist under that treatment condition ($S^a = 1$).

$$Y^a \perp\!\!\!\perp A \mid S^a = 1, \vec{X} \quad \text{for } a \in \{0, 1\} \qquad \text{(conditional exchangeability for outcome value)} \tag{5}$$

The assumptions of conditional exchangeability are analogous to their forms used in standard causal inference designs that rely on selection on observables. Section 6.1 returns to these assumptions to discuss their plausibility in our concrete example using a Directed Acyclic Graph.

3.2 Assumption: Monotonicity

Monotonicity assumes treatment has a one-directional effect on outcome existence, taking one of two forms: positive or negative monotonicity.[4] Under positive monotonicity, treatment may cause an outcome to exist but never causes it to cease existing:

$$S_i^1 \geq S_i^0 \quad \text{for all } i \qquad \text{(positive monotonicity)} \tag{6}$$

For example, job training may increase employment but never decrease it: anyone employed without training would also be employed with training. This aids identification because when we observe an untreated unit with an observed outcome (employed without training, $A_i = 0, S_i^0 = 1$), positive monotonicity implies this unit would have an outcome under either treatment condition ($S_i^0 = S_i^1 = 1$).

Under negative monotonicity, treatment may cause an outcome to cease existing but never to come into existence:

$$S_i^1 \leq S_i^0 \quad \text{for all } i \qquad \text{(negative monotonicity)} \tag{7}$$

In our motherhood example, this means motherhood may reduce employment but never causes non-employed women to enter paid employment. When we observe a treated unit whose outcome exists (employed mother, $A_i = 1, S_i^1 = 1$), negative monotonicity implies she would also be employed if

[4] Monotonicity assumption has been widely applied in causal inference under selection (Lee, 2009) and instrumental variables approaches (Imbens and Angrist, 1994).

she counterfactually had no child ($S_i^0 = 1$), placing her in our target subgroup.

3.3 Assumption: Mean dominance

Mean dominance assumes a certain ordering of potential outcomes across principal strata. For example, women employed regardless of motherhood may have higher potential wages as non-motherhood, on average, than those employed only as non-mothers. Formally:

$$\mathbf{E}(Y^0 \mid S^0 = S^1 = 1, \vec{X} = \vec{x}) \geq \mathbf{E}(Y^0 \mid S^0 = 1, S^1 = 0, \vec{X} = \vec{x}) \quad \forall \vec{x} \qquad \text{(mean dominance)} \quad (8)$$

Mean dominance can be positive or negative, for Y^0 or Y^1, depending on the setting. Here we focus on positive mean dominance for Y^0. This holds in our four-woman example, where always-employed women (Maya and Nancy, \$40) have higher potential non-mother hourly wages than those employed only as non-mothers (Mia and Nia, \$20).

Mean dominance would be plausible if women employed regardless of motherhood have characteristics associated with higher wages compared to those who would exit upon motherhood: better human capital (e.g. education, skills, work experience), stronger labor market attachment (e.g., career ambition, occupational choice), or more favorable job characteristics (e.g., family-friendly policies, flexible hours). The assumption could fail if low-wage women tend to continue working due to financial constraints, or if high-wage women exit employment upon motherhood—either because their jobs are more family-unfriendly or because they have higher-earning spouses providing financial flexibility. Researchers who make the assumption of mean dominance should present an argument for why it may hold in their particular setting.

4 Bounding effects on Y: Nonparametric set identification

Conditional exchangeability, monotonicity, and mean dominance enable nonparametric set identification of average causal effects among those who would have an outcome regardless of treatment. We first discuss a particularly tractable setting: bounding the average treatment effect on the treated under negative monotonicity. Then we discuss more general bounding results for average treatment effects. This section focuses on nonparametric identification, and the following section

introduces parametric estimation.

4.1 Average treatment effect on the treated under conditional exchangeability and negative monotonicity

We first consider the average treatment effect on treated units who would have an outcome regardless of treatment (ATT). In our example, this is the effect of motherhood on wages among mothers who would be employed regardless of motherhood.

$$\tau^{\text{ATT}} = \mathbf{E}\Bigg(\overbrace{Y^1 - Y^0}^{\substack{\text{Effect of Treatment} \\ \text{on Outcome Value} \\ \text{(e.g., wage)}}} \;\Big|\; \overbrace{A = 1}^{\substack{\text{Among} \\ \text{Treated Units} \\ \text{(e.g., parents)}}} , \quad \overbrace{S^1 = S^0 = 1}^{\substack{\text{Whose Outcome Would Exist} \\ \text{Regardless of Treatment} \\ \text{(e.g., employed regardless)}}} \Bigg) \tag{9}$$

$$= \mathbf{E}_{\vec{X}} \left(\mu^1(\vec{X}) \mid A = 1 \right) - \mathbf{E}_{\vec{X}} \left(\mu^0(\vec{X}) \mid A = 1 \right) \tag{10}$$

where the second line decomposes τ_{ATT} into two components using the law of iterated expectation and linearity of expectation.

The first component is the outcome under treatment for treated units whose outcome would exist regardless of treatment. Under negative monotonicity ($S_i^1 \leq S_i^0$), any treated unit with an outcome ($S_i^1 = 1$) would also have an outcome in the absence of treatment ($S_i^0 = 1$). In our concrete example, Maya, an employed mother, would (by monotonicity) also be employed as a non-mother. Thus, the first component is point-identified as the mean among treated units with outcomes.

$$\mathbf{E}_{\vec{X}} \left(\mu^1(\vec{X}) \mid A = 1 \right) = \mathbf{E}(Y^1 \mid S^0 = S^1 = 1, A = 1) \qquad \text{by iterated expectation} \tag{11}$$

$$= \mathbf{E}(Y^1 \mid S = 1, A = 1) \qquad \text{by negative monotonicity} \tag{12}$$

$$= \mathbf{E}(Y \mid S = 1, A = 1) \qquad \text{by consistency} \tag{13}$$

The second component of τ^{ATT} is more challenging. Under conditional exchangeability, counterfactual outcomes (without motherhood) can be identified from comparable non-mothers (condi-

tional on $\vec{X}$ and $S^0 = S^1 = 1$).

$$
\begin{aligned}
&\mathbf{E}_{\vec{X}}\left(\mu^0(\vec{X}) \mid A=1\right) \\
&= \mathbf{E}_{\vec{X}}\left(\mathbf{E}(Y^0 \mid S^0 = S^1 = 1, A=1, \vec{X})\right) && \text{by iterated expectation} && (14)\\
&= \mathbf{E}_{\vec{X}}\left(\mathbf{E}(Y^0 \mid S^0 = S^1 = 1, A=0, \vec{X})\right) && \text{by conditional exchangeability} && (15)\\
&= \mathbf{E}_{\vec{X}}\Bigg(\mathbf{E}(Y \mid \underbrace{S^0 = S^1 = 1}_{\substack{\text{Difficulty:}\\ \text{Latent Subgroup}}}, A=0, \vec{X})\Bigg) && \text{by consistency} && (16)
\end{aligned}
$$

However, the conditioning on $S^0 = S^1 = 1$ involves unobservable potential outcomes: for non-mothers, we cannot observe S^1. In our concrete example, we know Nancy ($S^1_{\text{Nancy}} = 1$) belongs to our target stratum while Nia ($S^1_{\text{Nia}} = 0$) does not, but this is unobservable in real data because S^1 is counterfactual for them. The challenge is that the employed non-mothers are a mixture of two principal strata: one stratum employed regardless (Nancy) and another stratum who would only be employed as non-mothers (Nia).

Under negative monotonicity, we can point-identify the proportion of employed non-mothers who are in the target stratum, within any subgroup taking a particular confounder vector value $\vec{X}$.

$$
\begin{aligned}
\overbrace{\pi_{S^0=S^1=1|S=1}(A=0,\vec{X})}^{\substack{\text{Proportion Employed Regardless}\\ \text{Among Employed Non-Mothers}}} &= \mathrm{P}(S^1 = 1 \mid \vec{X}, A=0, S^0=1) && && (17)\\
&= \frac{\mathrm{P}(S^1=1, S^0=1 \mid \vec{X}, A=0)}{\mathrm{P}(S^0=1 \mid \vec{X}, A=0)} && \text{by def. of conditional prob.} && (18)\\
&= \frac{\mathrm{P}(S^1=1 \mid \vec{X}, A=0)}{\mathrm{P}(S^0=1 \mid \vec{X}, A=0)} && \text{by negative monotonicity} && (19)\\
&= \frac{\mathrm{P}(S^1=1 \mid \vec{X}, A=1)}{\mathrm{P}(S^0=1 \mid \vec{X}, A=0)} && \text{by conditional exchangeability} && (20)\\
&= \frac{\mathrm{P}(S=1 \mid \vec{X}, A=1)}{\mathrm{P}(S=1 \mid \vec{X}, A=0)} && \text{by consistency} && (21)\\
&= \frac{\pi(1,\vec{X})}{\pi(0,\vec{X})} && && (22)
\end{aligned}
$$

Intuitively, under negative monotonicity, this equation says that the proportion who would have an outcome under $A = 1$, among those whose outcome exists under $A = 0$, equals the ratio of outcome existence under $A = 1$ to outcome existence under $A = 0$.

The second step is to use this proportion to bound the target estimand. Among employed non-mothers with covariate $\vec{X}$, a proportion $\pi(\vec{X})$ are in the stratum of interest and a proportion $1-\pi(\vec{X})$ are not. We create lower (upper) bounds on $\mu^0(\vec{x})$ for all $\vec{x}$ by considering the extreme cases where the $\pi(\vec{X})$ fraction of units in the always-employed stratum are from the lowest- (highest-)valued portions of the observed distribution:

$$\mu^0_{\text{Lower}}(\vec{X}) = \mathbf{E}\Bigg(Y \mid S=1, A=0, \vec{X}, \underbrace{F_{Y|S=1,A=0,\vec{X}}(Y) < \pi_{S^1=1|S=1,A=0}(\vec{X})}_{\substack{\text{averaging over the employed-regardless} \\ \text{assumed to be the lower} \\ \text{portion of the distribution}}}\Bigg) \tag{23}$$

$$\mu^0_{\text{Upper}}(\vec{X}) = \mathbf{E}\Bigg(Y \mid S=1, A=0, \vec{X}, \underbrace{F_{Y|S=1,A=0,\vec{X}}(Y) > 1-\pi_{S^1=1|S=1,A=0}(\vec{X})}_{\substack{\text{averaging over the employed-regardless} \\ \text{assumed to be the upper} \\ \text{portion of the distribution}}}\Bigg) \tag{24}$$

Finally, we bound τ_{ATT} by taking the difference between (1) the sample mean of the treated units' outcomes and (2) the bounds on the estimates for their outcome in the absence of treatment.

$$\mathbf{E}\Big(Y - \mu^0_{\text{Upper}}(\vec{X}) \mid S=1, A=1\Big) \leq \tau^{\text{ATT}} \leq \mathbf{E}\Big(Y - \mu^0_{\text{Lower}}(\vec{X}) \mid S=1, A=1\Big) \tag{25}$$

Our four-women example illustrates these bounds. The mean outcome of employed mothers (Maya) is \$30. The employed non-mothers Nia (\$20) and Nancy (\$40) are a mixture: one is employed-regardless and one would leave employment if she became a mother. Without knowing which non-mother belongs to which group, we construct bounds by considering both possibilities: if Nancy is the always-employed non-mother, the effect is (Maya - Nancy) = (\$30 - \$40) = \$10; if Nia is the always-employed non-mother, the effect is (Maya - Nia) = (\$30 - \$20) = +\$10. Our assumptions and evidence thus bound the average effect between −\$10 and +\$10.

Bounding the ATT under negative monotonicity is straightforward because treated units with outcomes $\{i : S_i = 1, A_i = 1\}$ coincide exactly with the target population $\{i : S_i^0 = S_i^1 = 1, A_i = 1\}$. This simplifies two steps. First, the expected outcome under treatment equals the observed mean of Y among these units. Second, the conditional bounds $\mu^0_{\text{Lower}}(\vec{X})$ and $\mu^1_{\text{Upper}}(\vec{X})$ could be aggregated over the observed distribution of $\vec{X}$: the distribution $\vec{X} \mid S=1, A=1$ among treated units with

existing outcomes (Eq. 25).

4.2 Average treatment effect under negative monotonicity

For the average treatment effect over all units who have outcomes regardless, two additional challenges arise: (1) estimating potential outcomes under treatment for control units, and (2) aggregating over the conditional distribution $\vec{X} \mid S^0 = S^1 = 1$.

Under conditional exchangeability and negative monotonicity, the conditional mean outcome under treatment for those whose outcome exists regardless is point-identified.

$$
\begin{aligned}
&\mathbf{E}(Y^1 \mid S^0 = S^1 = 1, \vec{X}) && && (26)\\
&= \mathbf{E}(Y^1 \mid S^0 = S^1 = 1, A = 1, \vec{X}) && \text{by conditional exchangeability} && (27)\\
&= \mathbf{E}(Y^| S^1 = 1, A = 1, \vec{X}) && \text{by negative monotonicity} && (28)\\
&= \mathbf{E}(Y \mid S = 1, A = 1, \vec{X}) && \text{by consistency} && (29)\\
&= \mu^1(\vec{X}) && && (30)
\end{aligned}
$$

Similarly, the conditional proportion whose outcome exists regardless given confounders $\vec{X}$ is identified by the conditional proportion of treated units who have an outcome.

$$
\begin{aligned}
\pi_{S^0=S^1=1}(\vec{X}) &= \mathrm{P}(S^0 = S^1 = 1 \mid \vec{X}) && \text{defining abbreviated notation} && (31)\\
&= \mathrm{P}(S^1 = 1 \mid \vec{X}) && \text{by negative monotonicity} && (32)\\
&= \mathrm{P}(S^1 = 1 \mid A = 1, \vec{X}) && \text{by conditional exchangeability} && (33)\\
&= \mathrm{P}(S = 1 \mid A = 1, \vec{X}) && \text{by consistency} && (34)
\end{aligned}
$$

These results bounds τ_{ATE} by the weighted average of conditional estimates, with weights proportional to stratum membership probabilities.

$$
\underbrace{\mathbf{E}_{\vec{X}}}_{\substack{\text{Mean}\\ \text{over}\\ \vec{X}}}\left(\underbrace{\tau_{\text{Lower}}(\vec{X})}_{\substack{\text{Lower bound}\\ \text{on average}\\ \text{effect given } \vec{X}}} \underbrace{\frac{\pi_{S^0=S^1=1}(\vec{X})}{\mathrm{E}(\pi_{S^0=S^1=1}(\vec{X}))}}_{\substack{\text{Weighted by the rate}\\ \text{of having an outcome}\\ \text{regardless of treatment}\\ \text{given } \vec{X}}} \right) \leq \tau_{\text{ATE}} \leq \mathbf{E}\left(\underbrace{\tau_{\text{Upper}}(\vec{X})}_{\substack{\text{Upper bound}\\ \text{on average}\\ \text{effect given } \vec{X}}} \underbrace{\frac{\pi_{S^0=S^1=1}(\vec{X})}{\mathrm{E}(\pi_{S^0=S^1=1}(\vec{X}))}}_{\substack{\text{Weighted by the rate}\\ \text{of having an outcome}\\ \text{regardless of treatment}\\ \text{given } \vec{X}}} \right) \tag{35}
$$

4.3 Average treatment effect without assuming monotonicity

Without monotonicity, set identification becomes more challenging because even the weight $\pi_{S^0=S^1=1}(\vec{X})$ is only set-identified:

$$\pi_{\text{Lower},S^0=S^1=1}(\vec{x}) \quad \leq \quad \pi_{S^0=S^1=1}(\vec{x}) \quad \leq \quad \pi_{\text{Upper},S^0=S^1=1}(\vec{x}), \tag{36}$$

with formulas for these bounds provided in Appendix A.1. A general procedure is to first create bounds on the conditional average treatment effect and the proportion whose outcomes exist regardless, both conditional on $\vec{X}$:

$$\tau_{\text{Lower}}(\vec{X}) \quad \leq \quad \mathbf{E}(Y^1 - Y^0 \mid S^0 = S^1 = 1, \vec{X}) \quad \leq \quad \tau_{\text{Upper}}(\vec{X}). \tag{37}$$

Aggregating these conditional bounds over $\vec{X}$ is challenging because the weight for each covariate value $\vec{x}$ depends on the proportion of units who would have an outcome regardless in that subgroup—a quantity that is only set-identified. We propose a sequential procedure that strategically allocates weights: lower bounds emerge by weighting observations with smaller conditional effects, while upper bounds prioritize observations with larger conditional effects.

For the lower bound: (1) Sort observations by $\tau_{\text{Lower}}(\vec{x}_i)$ from smallest to largest; (2) Initialize all stratum probabilities to $\tilde{\pi}(\vec{x}_i) = \pi_{\text{Lower}}(\vec{x}_i)$ and calculate the (tentative) initial weighted estimate $\tilde{\tau}_{\text{Lower}} = \frac{1}{\sum_i \tilde{\pi}(\vec{x}_i)} \sum_i \tilde{\pi}(\vec{x}_i)\tau_{\text{Lower}}(\vec{x}_i)$; (3) Starting with the smallest effect observation, sequentially increase each probability to its maximum, $\pi_{\text{Upper}}(\vec{x}_i)$, and re-calculate $\tilde{\tau}_{\text{Lower}}$ after each update; (4) Continue this process for observations with $\tau_{\text{Lower}}(\vec{x}_i)$ less than the current estimate until increasing weights would no longer reduce $\tilde{\tau}_{\text{Lower}}$. This yields a valid lower bound $\tau_{\text{Lower}} = \tilde{\tau}_{\text{Lower}}$.

For the upper bound, we carry out the same procedure in reverse, starting with observations having the largest $\tau_{\text{Upper}}(\vec{x}_i)$ values and sequentially increasing their weights to maximize the estimate. While valid, we anticipate that its usefulness may be rare in practice as it may yield very wide bounds in many settings. Since monotonicity is often credible in social science applications and produces narrower, simpler bounds, we focus primarily on monotonicity-based bounds.

5 Parametric estimation by regression and simulation

In randomized settings or settings with few discrete confounders $\vec{X}$, the identification results can be applied with nonparametric plug-in estimators, as in (Miratrix et al., 2018; Zhang and Rubin, 2003). However, in many observational studies, confounders $\vec{X}$ may be continuous or high-dimensional. It may be unusual to observe multiple units taking a single value of the confounders, so that nonparametric estimation is simply infeasible. This section presents our main methodological contribution: an estimation strategy that relies on parametric regression and simulation (Figure 3).

5.1 Parametrically model the existence of outcomes

We model the probability of outcome existence using logistic regression with treatment-specific coefficients:

$$\pi(a, \vec{x}) = \text{logit}^{-1}\left(\alpha_a + \vec{x}'\vec{\beta}_a\right) \tag{38}$$

5.2 Parametrically model outcome values given existence

For outcome Y among those for whom this outcome exists, we assume conditional normality with treatment-specific mean and variance functions. We estimate the variance model by a generalized linear model assuming a Gamma distribution on squared residuals (Western and Bloome, 2009).

$$Y \mid S = 1, A = a, \vec{X} = \vec{x} \sim \mathcal{N}\Big(\mu(a, \vec{x}), \sigma^2(a, \vec{x})\Big) \tag{39}$$

$$\mu(a, \vec{x}) = \eta_a + \vec{X}'\vec{\lambda}_a \tag{40}$$

$$\log\big(\sigma^2(a, \vec{x})\big) = \nu_a + \vec{X}'\vec{\gamma}_a \tag{41}$$

Two important considerations are relevant to this model. First, units with existing outcomes ($S = 1$) represent a mixture of latent principal strata in proportions that differ by treatment group. The simulation step that follows, thus, is essential to yield interpretable causal estimates from the outcome model. Second, while many researchers focus solely on conditional mean functions $\mu(a, \vec{x})$, our estimators require simulating from the full conditional outcome distribution. Hence, in this model the variance term $\sigma^2(a, \vec{x})$ and the assumption of conditional normality are also modeling

1) Define causal estimands.

$\mathbf{E}(S^1 - S^0)$ (effect on outcome existence)

$\mathbf{E}(Y^1 - Y^0 \mid S^0 = S^1 = 1)$ (effect on Y among those with outcomes regardless)

2) Make causal assumptions.

$S^a \perp\!\!\!\perp A \mid \vec{X}$ for all a (conditional exchangeability for survival)

$Y^a \perp\!\!\!\perp A \mid S^a = 1, \vec{X}$ for all a (conditional exchangeability for outcome)

$S_i^1 \geq S_i^0$ for all i (positive monotonicity)

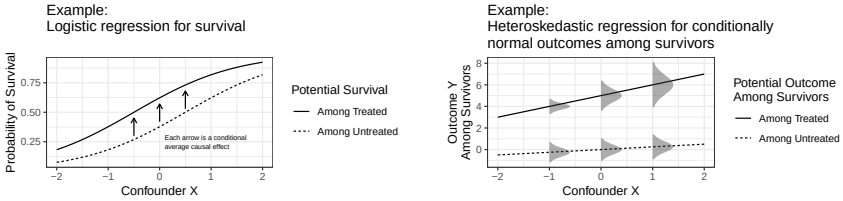
3) Make statistical modeling assumptions.

Model outcome existence:

$S \mid A, \vec{X} \sim \text{Bernoulli}\big(\pi(A, \vec{X})\big)$

Model outcome given its existence:

$Y \mid S = 1, A, \vec{X} \sim \text{Normal}\big(\mu(A, \vec{X}), \sigma^2(A, \vec{X})\big)$



4) Simulate outcomes at each $\vec{X} = \vec{x}_i$. Bound $\mu^a(\vec{x}_i) = \mathbf{E}(Y^a \mid S^0 = S^1 = 1, \vec{X} = \vec{x}_i)$.

Simulated Y of **treated** survivors who observationally match unit i:

By positive monotonicity, $\frac{\pi(0,\vec{x}_i)}{\pi(1,\vec{x}_i)}$ of this density are those who survive regardless

Lower Bound: Mean of Blue Region $\mu^1_{\text{Lower}}(\vec{x}_i)$

Upper Bound: Mean of Blue Region $\mu^1_{\text{Upper}}(\vec{x}_i)$

Simulated Y of **untreated** survivors who observationally match unit i:

By positive monotonicity, 100% of this density are those who survive regardless

Point Identified: Mean of Blue Region $\mu^0(\vec{x}_i)$

5) Difference to bound the conditional average effect on Y among those who survive regardless.

$$\hat{\tau}_{\text{Lower}}(\vec{x}_i) = \hat{\mu}^1_{\text{Lower}}(\vec{x}_i) - \hat{\mu}(0, \vec{x}_i) \qquad \hat{\tau}_{\text{Upper}}(\vec{x}_i) = \hat{\mu}^1_{\text{Upper}}(\vec{x}_i) - \hat{\mu}(0, \vec{x}_i)$$

6) Aggregate over $\vec{X}$ to produce average causal effect estimates.

$$\underbrace{\hat{\mathbf{E}}(S^1 - S^0)}_{\text{effect on survival}} = \underbrace{\tfrac{1}{n}\textstyle\sum_{i=1}^n}_{\text{average over units}} \Big(\underbrace{\hat{\pi}(1, \vec{x}_i)}_{\text{predicted survival under treatment}} - \underbrace{\hat{\pi}(0, \vec{x}_i)}_{\text{predicted survival under control}} \Big)$$

$$\hat{\mathbf{E}}(Y^1 - Y^0 \mid S^0 = S^1 = 1) \in \left[\frac{1}{\sum_i \hat{\pi}(0,\vec{x}_i)} \sum_i \hat{\pi}(0, \vec{x}_i) \hat{\tau}^{\text{Lower}}(\vec{x}_i), \quad \frac{1}{\sum_i \hat{\pi}(0,\vec{x}_i)} \sum_i \hat{\pi}(0, \vec{x}_i) \hat{\tau}^{\text{Upper}}(\vec{x}_i) \right]$$

effect on Y among those whose outcome exists regardless

conditional bounds weighted by conditional rate of outcome existing regardless (under positive monotonicity)

Fig. 3. Principal stratification with regression: Illustrated with positive monotonicity. In the figure, outcome Y_i exists for any unit that survives ($S_i = 1$), such as a wage existing only for employed individuals. Positive monotonicity assumes the intervention (e.g., $A_i = 1$ indicates receipt of job training) never decreases employment probability. Note that this figure illustrates positive monotonicity ($S^1 \geq S^0$) as applies in job training scenarios, whereas our motherhood application uses negative monotonicity ($S^0 \geq S^1$). Steps 4–6 become more complicated without monotonicity, because the conditional probability of surviving regardless is only set-identified.

assumptions of critical importance.

5.3 Simulate to estimate principal stratification estimands

The final step uses the estimated model to simulate principal stratification estimands. The first estimand of interest—the average causal effect on outcome existence—can be point estimated by predicting $\hat{\pi}(a, \vec{x}_i)$ from the model for outcome existence under each treatment value a, plugging these into Eq. 4, and estimating the population mean with the sample mean.

$$\hat{\mathbf{E}}\left(S^1 - S^0\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\pi}(1, \vec{x}_i) - \hat{\pi}(0, \vec{x}_i)\right) \tag{42}$$

Turning attention to outcomes, the simulation procedure varies by estimand and the assumptions maintained. For concreteness, here we consider ATT among those whose outcome would exist regardless, under negative monotonicity.

$$\tau^{\text{ATT}} = \mathbf{E}(Y^1 - Y^0 \mid A = 1, S^0 = S^1 = 1) \tag{43}$$

Using notation from Section 2.4, we estimate this quantity as:

$$\hat{\tau}^{\text{ATT}} = \frac{1}{\sum_i \mathbb{I}\{A_i = 1\}} \sum_{i:A_i=1}\left(Y_i - \hat{\mu}^0(\vec{x}_i)\right) \tag{44}$$

where $\hat{\mu}^0(\vec{x}_i)$ is the predicted mean potential outcome under control for unit i, and $\mathbb{I}\{A_i = a\}$ is an indicator variable taking the value 1 when $A_i = a$ and 0 otherwise. As shown earlier, the outcome under control is set-identified given the conditional distribution of Y. Using our parametric outcome model, we estimate the bounds by first simulating R draws from the conditional outcome distribution for each unit i,

$$\tilde{Y}_{ir} \overset{\text{iid}}{\sim} \mathcal{N}\left(\hat{\mu}(0, \vec{x}_i), \hat{\sigma}^2(0, \vec{x}_i)\right), \quad r = 1, \ldots, R \tag{45}$$

The nonparametric bounds retain specific quantiles of the conditional distribution. Our simulation-

based estimators take the same quantities from the simulated distribution.

$$\hat{\mu}^0_{\text{Lower}}(\vec{x}_i) = \frac{1}{R\pi_{S^1=1|S=1,A=0}(\vec{x}_i)} \sum_{r=1}^{R} \tilde{Y}_{ir} \mathbb{I}\left\{\tilde{Y}_{ir} < \text{Quantile}(\tilde{Y}_i, \pi_{S^1=1|S=1,A=0}(\vec{x}_i))\right\} \tag{46}$$

$$\hat{\mu}^0_{\text{Upper}}(\vec{x}_i) = \frac{1}{R\pi_{S^1=1|S=1,A=0}(\vec{x}_i)} \sum_{r=1}^{R} \tilde{Y}_{ir} \mathbb{I}\left\{\tilde{Y}_{ir} > \text{Quantile}(\tilde{Y}_i, 1 - \pi_{S^1=1|S=1,A=0}(\vec{x}_i))\right\} \tag{47}$$

Finally, we construct overall bound estimates by averaging the conditional bounds across the sample.

$$\hat{\tau}^{\text{ATT}}_{\text{Lower}} = \frac{1}{\sum_i \mathbb{I}\{A_i = 1\}} \sum_{i:A_i=1} \left(Y_i - \hat{\mu}^0_{\text{Upper}}(\vec{x}_i)\right) \tag{48}$$

$$\hat{\tau}^{\text{ATT}}_{\text{Upper}} = \frac{1}{\sum_i \mathbb{I}\{A_i = 1\}} \sum_{i:A_i=1} \left(Y_i - \hat{\mu}^0_{\text{Lower}}(\vec{x}_i)\right) \tag{49}$$

While formulas vary by estimand and assumptions, the general procedure remains the same: use the nonparametric formulas but plug in simulated conditional distributions in place of true conditional distributions and sample means in place of population means.

5.4 Inference: Construct confidence intervals by the bootstrap

We quantify statistical uncertainty using the nonparametric bootstrap to capture sampling variation across all model components. For each of B bootstrap samples ($b = 1, \ldots, B$), we re-estimate all models and calculate bounds $\hat{\tau}_{\text{Lower},b}$ and $\hat{\tau}_{\text{Upper},b}$ (either for ATE or ATT). Our $(1-\alpha)$-level confidence interval takes the empirical quantiles of order $(\alpha/2)$ and $(1-\alpha/2)$ of the lower and upper bounds, respectively. This intersection of two one-sided confidence intervals provides coverage for the identified set—an appropriate approach for our partial identification setting where parameters of interest are "bounded" (Imbens and Manski, 2004; Zhao et al., 2019).

5.5 Generalization to weighted samples

Social surveys often employ sampling weights to account for unequal selection probabilities. The procedure above generalizes to these settings by substituting weighted sample means for unweighted sample means whenever estimating population quantities. Likewise, weights can be incorporated when estimating the parametric models for outcome existence and outcome values.

6 Empirical illustration: The effect of parenthood on hourly wage

We illustrate with an empirical analysis of the causal effect of parenthood on hourly wages for men and women. A long line of research has documented that becoming a parent is associated with wage losses and employment declines for women (Budig and England, 2001; England et al., 2016; Gough and Noonan, 2013; Staff and Mortimer, 2012; Waldfogel, 1997). However, recent evidence suggests that the motherhood wage penalty may be disappearing over time as the effect size moves closer to zero (Buchmann and McDaniel, 2016; Pal and Waldfogel, 2016), though other scholars find that it is stable (Jee et al., 2019). Our empirical illustration defines the wage effect among those employed regardless of motherhood and shows that evidence is consistent with a range of possible estimates—including estimates near zero and estimates far from zero—depending on the assumptions one is willing to make about who comprises this latent set of people.

We analyze data from the 1997 National Longitudinal Survey of Youth Cohort. We construct a sample of 1,985 mothers and 1,837 fathers observed in the year immediately after giving birth and a comparison group of 20,543 (25,902) person-year observations on women (men) who have not yet given birth. Our confounders include pre-parenthood characteristics: race, age, education, marital status, job tenure, work experience, and employment status and hourly wage in the previous year (Fig 4).

We study two causal estimands: (1) the average effect of parenthood on employment among parents, and (2) the average effect of parenthood on wage among parents who would be employed regardless of parenthood (ATT as in Eq. 10).

6.1 Nonparametric causal assumptions and parametric statistical estimators

Our first causal assumption is conditional exchangeability: potential employment statuses (S^a) and potential wages (Y^a) are independent of parenthood (A) conditional on measured confounders $\vec{X}$. After adjusting for pre-parenthood characteristics, parenthood can be considered as-if randomly assigned with respect to potential outcomes. While untestable, this assumption is made more plausible by our rich set of pre-treatment covariates. Fig 5 presents a causal Directed Acyclic Graph (DAG) where $\vec{X}$ represents a sufficient adjustment set. The validity of this assumption relies on the richness of the available covariates and the researchers' domain knowledge about

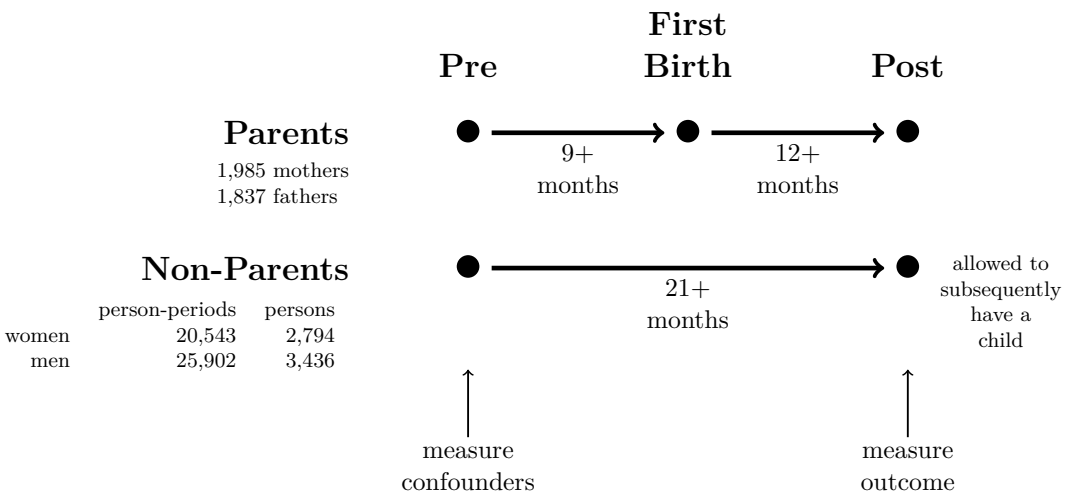


Fig. 4. Data structure for illustration with the NLSY97. We begin by arranging the data so that each observation on a person takes the role of a pre-observation and is paired with the soonest post-observation on that same person at least 1.75 years into the future and no more than 6 years into the future. We label observations as non-parent observations if a first birth occurs in the window and parent observations if a birth does not occur in the window, dropping observations where the pre-period is after the first birth. We keep parent observations only if the pre- and post-observations are each no more than 3 years from the birth (28,135 childless men pairs, 2,029 father pairs, 21,704 childless women pairs, 2,098 mother pairs). All of our cases have valid reports of employment. We drop cases who reported employment but did not report a wage, so that missing wages always correspond to non-employment in our analysis, producing the analytical sample sizes reported above.

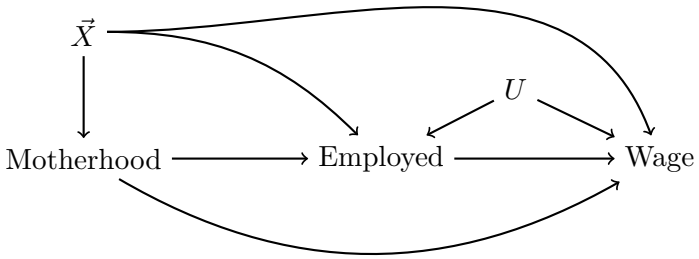


Fig. 5. Nonparametric causal assumptions to select confounders. We assume that all backdoor paths between motherhood and employment and between motherhood and wage are blocked by the confounder set $\vec{X}$, which in our example includes age, education, marital status, full-time employment, job tenure, work experience, and wage and employment each lagged by one year. Our approach is valid even when U exists and is unmeasured.

potential confounders. Importantly, our causal identification remains valid even in the presence of unobserved variables U that may affect both employment and wages, such as access to advantageous job opportunities within one's social network.

We present a series of estimates under different assumptions. For women, we present estimates assuming negative monotonicity: motherhood may cause women to leave paid employment but never causes non-employed women to enter the workforce. This assumption is supported by theory, literature, and our findings (in Section 6.3), though untestable at the individual level.

For both women and men, we consider positive mean dominance: within any covariate subgroup $\vec{x}$, potential wages are higher for always-employed individuals, on average, than for those employed under only one treatment condition. This assumption is plausible if those with stronger labor force attachment have higher wages, though one could argue that financial necessity might keep low-wage workers employed regardless of parenthood.

Our estimators follow those specified in Section 5, using a logistic regression model for employment and a conditionally normal outcome model for log hourly wage, with quantile-based nonparametric bootstrap confidence intervals.

6.2 Estimated effects on employment

We estimate that motherhood reduces employment by 13.9 percentage points on average in our sample (Figure 6). The effect is also heterogeneous across subgroups: motherhood reduces post-

birth employment most strongly among women who before birth either were not employed or were employed for less than $15 per hour, among whom motherhood reduces employment by 15.7 percentage points. Among high-wage women whose pre-birth wages were more than $20 per hour, motherhood reduces employment by only 10.5 percentage points. No such pattern is apparent among men, for whom fatherhood has approximately zero effect on employment in all subgroups visualized.

The pattern of employment effects supports the posited negative monotonicity assumption: motherhood either reduces employment or has no effect, but never increases it. While untestable, this assumption gains credibility from the consistently negative effect estimates in Figure 6A across population subgroups. Mean dominance assumes women whose employment is reduced by motherhood would have lower potential wages as non-mothers than those whose employment is unaffected. This assumption, though also untestable, is supported by our finding that motherhood reduces employment most among women with lower pre-birth wages.

With approximately zero employment effects for men across all subgroups (Figure 6B), monotonicity is less defensible for fathers. Given these null effects, we do not report results that assume monotonicity for fathers.

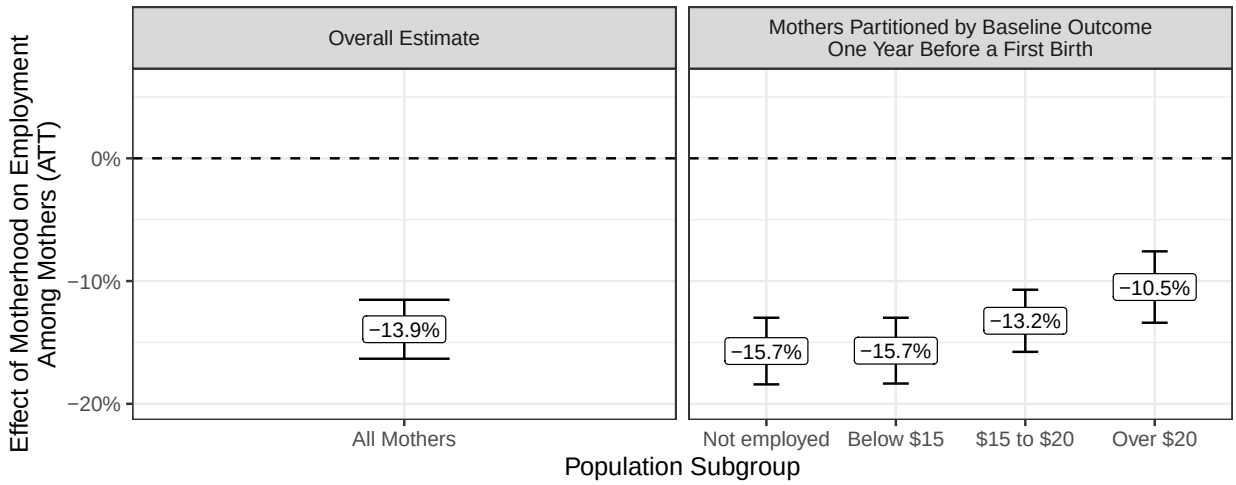
6.3 Estimated effects on wages

We now illustrate bounded estimates on wages, which demonstrate that additional assumptions transparently tighten the estimated bounds (Manski, 1995, 2003). Figure 7 presents the estimated effect of parenthood on hourly wage among always-employed individuals. Under exchangeability alone, bounds are wide: fatherhood's effect on log wages ranges from -0.38 to +0.57, while motherhood's ranges from -0.61 to +0.60. These wide bounds reflect fundamental uncertainty about each person's counterfactual employment status under the treatment value that did not occur.

Adding negative monotonicity narrows bounds for women to [-0.15, 0.14], demonstrating the identifying power of theoretically-justified assumptions. For men, however, monotonicity is questionable. Some men might become employed upon fatherhood to support their family, while others (e.g., those with a high-earning spouse) might step back from employment to care for children. We therefore choose not to report estimates for men that rely on monotonicity.

Imposing only mean dominance—that always-employed men have higher potential wages—

A) Effect of motherhood on employment



B) Effect of fatherhood on employment

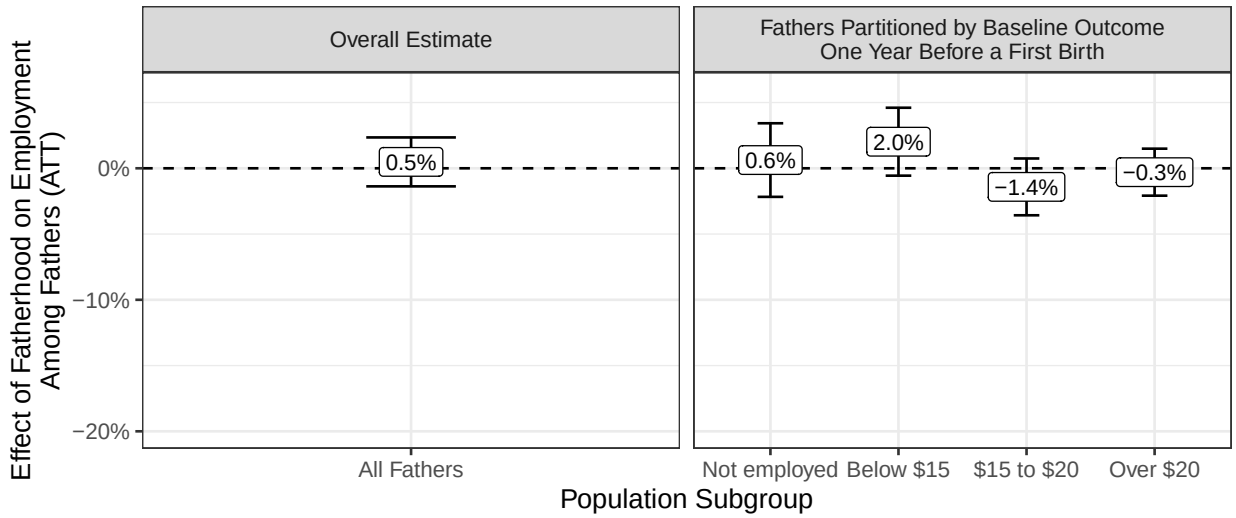


Fig. 6. Effect of parenthood on employment. All estimates correspond to average treatment effects on the treated (effects among parents). The facet on the right groups new parents by their observed employment outcome one year before parenthood.

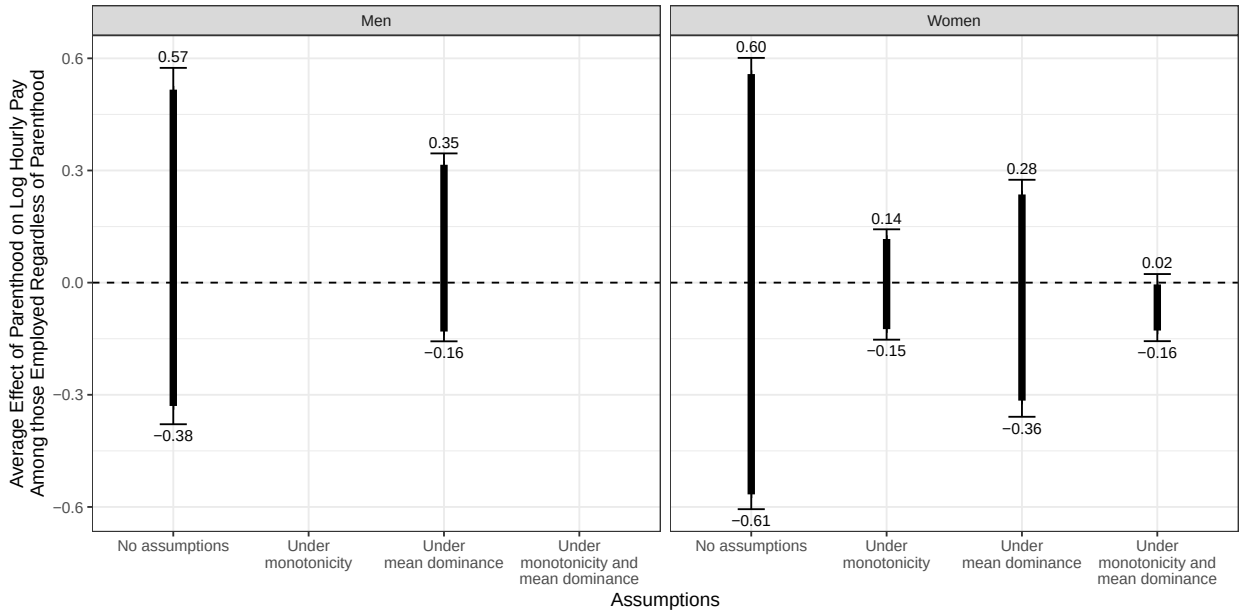


Fig. 7. Effect of parenthood on wage among those employed regardless.

provides different estimates. Mean dominance yields bounds of [-0.16, 0.35] for fathers, suggesting a possible fatherhood premium. For mothers, this assumption alone suggests bounds of [-0.36, 0.28], moderately tighter than the baseline bounds. When combining both monotonicity and mean dominance, we obtain the tightest bounds of [-0.16, 0.02] for the motherhood wage effect. This suggests a predominantly negative effect with a wage penalty of up to 16%.

The sequential incorporation of assumptions demonstrates both the value and limitations of theory-driven causal inference. While successive assumptions based on qualitative understanding narrow our bounds (Coppock and Kaur, 2022), substantial uncertainty (due to outcomes that may not exist) remains even under our strongest assumptions. This contrasts with standard practice: regression excluding non-employed women yields a deceptively precise estimate near zero (estimate = -0.01, CI = [-0.03, 0.02]). This false precision stems from ignoring the uncertainty about which individuals would remain employed under different treatment conditions. Our approach explicitly acknowledges this uncertainty, revealing a wider range of plausible effects. Appendix B formally demonstrates how analyses restricted to units with existing outcomes suffer from selection bias.

Our results contribute to the literature on parenthood wage effects by showing how selection into employment misleads inferences about wage penalties of motherhood. The bounds under various assumptions suggest standard estimates need reconsideration in light of selective employ-

ment. Moreover, the stronger employment effects among lower-wage women suggest that standard analyses may understate motherhood's total impact by focusing solely on wages conditional on employment.

7 Discussion

Research in social stratification regularly faces a difficult problem: treatments that shape the values of outcomes often also determine whether those outcomes exist at all. We showed how the standard practice of restricting analysis to observed outcomes obscures causal effects and produces misleading inferences about inequality. For researchers facing this problem, there are two primary recommendations. First, before studying the values of outcomes researchers should estimate and report effects of the treatment on outcome existence. When a treatment has a large effect on employment, for example, analyses that jump to wage as the outcome may miss an essential part of the story. To estimate effects on the existence of the outcome requires no new tools beyond standard causal inference methods, and could yield valuable new insights. Second, when researchers move on to study the value of selectively-existing outcomes they should explicitly focus on the latent subgroup who would have an outcome under either treatment condition. When producing results, these researchers should make transparent assumptions such as monotonicity and mean dominance, and they should report interval estimates that correctly incorporate their fundamental uncertainty about which units are in the latent stratum of interest: those whose outcome would exist under either treatment condition.

The primary contribution of this paper is showing how bounding strategies that originated in biostatistics can be adapted for use in studies of social stratification. To do so, our technical contribution is an approach involving parametric regression, simulation, and aggregation of simulated conditional estimates to marginal summaries. The parametric approach we develop enables researchers to use concepts originally designed for randomized experiments and apply them in settings where it is necessary to adjust for a large set of measured confounders.

A second contribution of this paper is to demonstrate through an empirical example that standard practice can produce misleadingly tight confidence intervals centered near zero when true effects may be much larger in magnitude. Our reexamination of the motherhood wage penalty

shows that while conventional estimates produce a motherhood effect near zero with a tight confidence interval, our strongest assumptions yield bounds of [-0.16, 0.02] that suggest the possibility of a sizeable negative effect. Our bounds reveal that the motherhood wage penalty could be substantially more negative than previous research suggests—an uncertainty masked by conventional approaches that restrict analyses to employed women.

Most broadly, our framework provides a template for studying social stratification when outcome existence is itself selective. We believe there exist many processes in social stratification in which inputs that shape the values of an outcome also determine its existence. Rather than treating non-existent outcomes as missing data to be handled through deletion or imputation, their explicit incorporation can deepen understanding of inequality.

References

Acemoglu, D. and Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of economic perspectives*, 33(2):3–30.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865.

Brand, J. E. (2023). *Overcoming the odds: The benefits of completing college for unlikely graduates*. Russell Sage Foundation.

Brand, J. E. and Xie, Y. (2010). Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75(2):273–302.

Buchmann, C. and McDaniel, A. (2016). Motherhood and the wages of women in professional occupations. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(4):128–150.

Budig, M. J. and England, P. (2001). The wage penalty for motherhood. *American sociological review*, pages 204–225.

Card, D., Kluve, J., and Weber, A. (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(5):815–836.

Conger, R. D., Elder Jr, G. H., Lorenz, F. O., Conger, K. J., Simons, R. L., Whitbeck, L. B., Huck, S., and Melby, J. N. (1990). Linking economic hardship to marital quality and instability. *Journal of Marriage and the Family*, pages 643–656.

Coppock, A. and Kaur, D. (2022). Qualitative imputation of missing potential outcomes. *American Journal of Political Science*, 66(3):681–695.

Desmond, M. (2016). *Evicted: Poverty and Profit in the American City*. Crown Publishing Group.

Ding, P., Geng, Z., Yan, W., and Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106(496):1578–1591.

England, P., Bearak, J., Budig, M. J., and Hodges, M. J. (2016). Do highly paid, highly skilled women experience the largest motherhood penalty? *American sociological review*, 81(6):1161–1189.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498):450–466.

Gough, M. and Noonan, M. (2013). A review of the motherhood wage penalty in the united states. *Sociology Compass*, 7(4):328–342.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of political Economy*, 82(6):1119–1143.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Hwang, J. and Sampson, R. J. (2014). Divergent pathways of gentrification: Racial inequality and the social order of renewal in chicago neighborhoods. *American sociological review*, 79(4):726–751.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Jee, E., Misra, J., and Murray-Close, M. (2019). Motherhood penalties in the us, 1986–2014. *Journal of Marriage and Family*, 81(2):434–449.

Killewald, A. (2013). A reconsideration of the fatherhood premium: Marriage, coresidence, biology, and fathers' wages. *American Sociological Review*, 78(1):96–116.

Knox, D., Lowe, W., and Mummolo, J. (2020). Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76:1071–1102.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Miratrix, L., Furey, J., Feller, A., Grindal, T., and Page, L. C. (2018). Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness*, 11(1):133–162.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles (with discussion). section 9 (translated). *Statistical Science*, 5(4):465—472.

Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for prek–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1):74–107.

Pal, I. and Waldfogel, J. (2016). The family gap in pay: New evidence for 1967 to 2013. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(4):104–127.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does job corps work? impact findings from the national job corps study. *American economic review*, 98(5):1864–1886.

Slough, T. (2023). Phantom counterfactuals. *American Journal of Political Science*, 67(1):137–153.

Staff, J. and Mortimer, J. T. (2012). Explaining the motherhood wage penalty during the early occupational career. *Demography*, 49:1–21.

Waldfogel, J. (1997). The effect of children on women's wages. *American sociological review*, pages 209–217.

Western, B. (2002). The impact of incarceration on wage mobility and inequality. *American sociological review*, 67(4):526–546.

Western, B. (2006). Punishment and inequality in america. *Russell Sage Foundation*.

Western, B. and Bloome, D. (2009). Variance function regressions for studying inequality. *Sociological Methodology*, 39(1):293–326.

Winship, C. and Mare, R. D. (1992). Models for sample selection bias. *Annual review of sociology*, 18(1):327–350.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.

A Proof of target stratum size

A.1 General case: No monotonicity assumption

This section provides a formal proof of the set identification results for the proportion in the always-survive principal stratum ($S^0 = S^1 = 1$) discussed in Section 4.3. We show that we can partially identify this quantity through bounds in the most general case without monotonicity or mean

dominance. In this proof, we assume exchangeability and for simplicity assume all probabilities and expectations are conditional on measured covariates $\vec{X}$. We further define abbreviated notation for the conditional probability of membership in each principal stratum.

$$\pi_{11} \equiv \mathrm{P}(S^0 = 1, S^1 = 1) \qquad \text{survive regardless} \tag{50}$$

$$\pi_{10} \equiv \mathrm{P}(S^0 = 0, S^1 = 1) \qquad \text{survive if treated} \tag{51}$$

$$\pi_{01} \equiv \mathrm{P}(S^0 = 1, S^1 = 0) \qquad \text{survive if untreated} \tag{52}$$

$$\pi_{00} \equiv \mathrm{P}(S^0 = 0, S^1 = 0) \qquad \text{never survive} \tag{53}$$

We will use two combinations of principal strata whose sizes are causally identified under conditional exchangeability.

$$\pi_{11} + \pi_{10} = \mathrm{P}(S^1 = 1) \qquad \text{by definitions} \tag{54}$$

$$= \mathrm{P}(S^1 = 1 \mid A = 1) \qquad \text{by exchangeability} \tag{55}$$

$$= \mathrm{P}(S = 1 \mid A = 1) \qquad \text{by consistency} \tag{56}$$

$$\pi_{11} + \pi_{01} = \mathrm{P}(S^0 = 1) \qquad \text{by definitions} \tag{57}$$

$$= \mathrm{P}(S^0 = 1 \mid A = 0) \qquad \text{by exchangeability} \tag{58}$$

$$= \mathrm{P}(S = 1 \mid A = 0) \qquad \text{by consistency} \tag{59}$$

We then solve these equations to get upper bounds on π_{11}.

$$\pi_{11} = \mathrm{P}(S = 1 \mid A = 1) - \pi_{10} \tag{60}$$

$$\leq \mathrm{P}(S = 1 \mid A = 1) \qquad \text{since } \pi_{10} >= 0 \tag{61}$$

$$\pi_{11} = \mathrm{P}(S = 1 \mid A = 0) - \pi_{01} \tag{62}$$

$$\leq \mathrm{P}(S = 1 \mid A = 0) \qquad \text{since } \pi_{01} >= 0 \tag{63}$$

Because these are both valid upper bounds, the upper bound is the minimum of the two. The intuition of these bounds is that the proportion who survive regardless of treatment can be no

larger than the survival rate in each of the treatment conditions.

$$\pi_{11} \leq \min\Big(\mathrm{P}(S=1 \mid A=1), \mathrm{P}(S=1 \mid A=0)\Big) \tag{64}$$

Next, we want to place a lower bound on π_{11}. For this, we begin from the fact that the probabilities of all four strata sum to 1.

$$\begin{aligned}
1 &= \pi_{11} + \pi_{10} + \pi_{01} + \pi_{00} && (65)\\
&= \underbrace{\pi_{11} + \pi_{10}}_{\text{term 1}} + \underbrace{\pi_{11} + \pi_{01}}_{\text{term 2}} + \pi_{00} - \pi_{11} && \text{add and subtract } \pi_{11} \quad (66)\\
&= \underbrace{\mathrm{P}(S=1 \mid A=1)}_{\text{term 1}} + \underbrace{\mathrm{P}(S=1 \mid A=0)}_{\text{term 2}} + \pi_{00} - \pi_{11} && (67)\\
\pi_{11} &= \mathrm{P}(S=1 \mid A=1) + \mathrm{P}(S=1 \mid A=0) - 1 + \pi_{00} && (68)\\
&\geq \mathrm{P}(S=1 \mid A=1) + \mathrm{P}(S=1 \mid A=0) - 1 && \text{since } \pi_{00} \geq 0 \quad (69)
\end{aligned}$$

In the result above, two limiting cases help to build intuition. For the first limiting case, suppose the survival rate is 100% among the treated and 100% among the untreated, so that everyone survives regardless of treatment. The lower bound on the proportion who survive regardless is (1 + 1 - 1) = 1. For the second limiting case, suppose that the survival rates in the treated and control conditions sum to 1. For example, suppose that 75% of the treated survive and 25% of the untreated survive. In this case, the entire population may consist of treatment-induced-surivors (75% of the population) and control-induced-survivors (25% of the population). It is possible that no one is in the survive-regardless group. In math, the lower limit on the proportion always surviving is (0.75 + 0.25 - 1) = 0.

Finally, we also know $\pi_{11} \geq 0$ since it is a probability. Together with Eq. 69, this gives a lower limit on the proportion always survivors. Paired with Eq. 64, we have partial identification for the proportion who survive regardless of treatment.

$$\max\Big\{0, \mathrm{P}(S=1 \mid A=1) + \mathrm{P}(S=1 \mid A=0) - 1\Big\} \leq \pi_{11} \leq \min\Big(\mathrm{P}(S=1 \mid A=1), \mathrm{P}(S=1 \mid A=0)\Big\} \tag{70}$$

An intuition for the lower bound would be helpful. If $\mathrm{P}(S=1 \mid A=1) + \mathrm{P}(S=1 \mid A=0) > 1$,

then there must be some overlap in the groups that survive under treatment and control. The amount of necessary overlap is exactly $\mathrm{P}(S = 1 \mid A = 1) + \mathrm{P}(S = 1 \mid A = 0) - 1$. If $\mathrm{P}(S = 1 \mid A = 1) + \mathrm{P}(S = 1 \mid A = 0) \leq 1$, then it's possible that there's no overlap (i.e., no one survives regardless of treatment), so the lower bound is 0. For example, If 80% survive under treatment and 70% survive under control, the sum is 150%, meaning at least 50% must survive in both conditions. If 60% survive under treatment and 40% survive under control, the sum is 100%, so it's possible that no one survives in both conditions.

In the data analyzed, we restrict to those who survive ($S = 1$). This implies that we need the proportion always-survivors conditional on survival.

$$\mathrm{P}(S^0 = S^1 = 1 \mid S = 1, A = a) = \frac{\mathrm{P}(S^0 = S^1 = 1 \mid A = a)}{\mathrm{P}(S = 1 \mid A = a)} \qquad \text{def. of cond. prob.} \tag{71}$$

$$= \frac{\mathrm{P}(S^0 = S^1 = 1)}{\mathrm{P}(S = 1 \mid A = a)} \qquad \text{conditional exchangeability} \tag{72}$$

$$= \frac{\pi_{11}}{\mathrm{P}(S = 1 \mid A = a)} \tag{73}$$

Because the denominator can be estimated from data, the bounds on the numerator imply bounds on the fraction.

A.2 Under Monotonicity

We now consider the target stratum size under negative monotonicity ($S^0 \geq S^1$), which in our motherhood example means that motherhood may cause a woman to leave employment but never cause employment. Under negative monotonicity, it holds that $\pi_{10} = 0$. With that, we can point-identify π_{11}:

$$\pi_{11} + \pi_{10} = \mathrm{P}(S = 1 \mid A = 1) \tag{74}$$

$$\Rightarrow \pi_{11} = \mathrm{P}(S = 1 \mid A = 1) \qquad \text{since } \pi_{10} = 0 \tag{75}$$

Now let's determine the proportion of always-survivors conditional on survival for both treat-

ment groups, based on Eq. 73:

$$\mathrm{P}(S^0 = S^1 = 1 | S = 1, A = a) = \begin{cases} 1 & \text{if } a = 1 \\ \frac{\mathrm{P}(S=1|A=1)}{\mathrm{P}(S=1|A=0)} & \text{if } a = 0 \end{cases} \tag{76}$$

This makes intuitive sense under negative monotonicity: when $a = 1$, anyone who survives despite treatment would certainly have survived without it. When $a = 0$, only a fraction of those who survive without treatment would survive with treatment—this fraction equals the ratio of the survival rates.

Similarly, we can easily get the same quantity under positive monotonicity case:

$$\mathrm{P}(S^0 = S^1 = 1 | S = 1, A = a) = \begin{cases} \frac{\mathrm{P}(S=1|A=0)}{\mathrm{P}(S=1|A=1)} & \text{if } a = 1 \\ 1 & \text{if } a = 0 \end{cases} \tag{77}$$

B Derivation of bias in naive comparison under Monotonicity

To formally see why naive comparisons of wages between employed mothers and non-mothers can be misleading, we derive the bias term explicitly. For simplicity, we focus on the case under monotonicity (that motherhood never increases employment) and suppress conditioning on observed covariates $\vec{X}$ in our notation, though this conditioning would be necessary in practice for identification in observational setting. We assume monotonicity in the direction that motherhood does not increase employment: $S^0 \geq S^1$, meaning a woman who would be employed as a mother would also be employed as a non-mother.

Let $\hat{\tau}_{naive}$ be the difference in mean wages between employed mothers and employed non-mothers. Note that under monotonicity, employed non-mothers represent a mixture of the always-employed stratum and those employed only when not mothers, hence $\mathrm{P}(S^0 = 1) = \mathrm{P}(S^0 = S^1 =$

$1) + \mathrm{P}(S^0 = 1, S^1 = 0)$.

$$\begin{aligned}
\hat{\tau}_{naive} &= \mathbf{E}(Y \mid S = 1, A = 1) - \mathbf{E}(Y \mid S = 1, A = 0) && (78)\\
&= \mathbf{E}(Y^1 \mid S^1 = 1) - \mathbf{E}(Y^0 \mid S^0 = 1) && \text{by consistency} \quad (79)\\
&= \mathbf{E}(Y^1 \mid S^0 = S^1 = 1) - \mathbf{E}(Y^0 \mid S^0 = 1) && \text{by monotonicity} \quad (80)\\
&= \mathbf{E}(Y^1 \mid S^0 = S^1 = 1)\\
&\quad - \Bigg(\mathbf{E}(Y^0 \mid S^0 = S^1 = 1)\frac{\mathrm{P}(S^0 = S^1 = 1)}{\mathrm{P}(S^0 = 1)}\\
&\quad + \mathbf{E}(Y^0 \mid S^0 = 1, S^1 = 0)\frac{\mathrm{P}(S^0 = 1, S^1 = 0)}{\mathrm{P}(S^0 = 1)}\Bigg), && (81)
\end{aligned}$$

where the last equality is by conditioning on S^1 for the second term. The causal estimand, the average causal effect among the employed-regardless, we seek to identify is:

$$\tau = \mathbf{E}(Y^1 - Y^0 \mid S^0 = S^1 = 1).$$

Thus, the bias in the naive comparison can be written as:

$$\begin{aligned}
Bias(\hat{\tau}_{naive}, \tau) &= \mathbf{E}(\hat{\tau}_{naive} - \tau) && (82)\\
&= \underbrace{[\mathbf{E}(Y^0 \mid S^0 = S^1 = 1) - \mathbf{E}(Y^0 \mid S^0 = 1, S^1 = 0)]}_{(1)} \underbrace{\frac{\mathrm{P}(S^0 = 1, S^1 = 0)}{\mathrm{P}(S^0 = 1)}}_{(2)} && (83)
\end{aligned}$$

This bias term shows why the naive comparison may not recover the causal effect even under monotonicity. The bias is the product of two quantities: (1) the difference in potential wages under non-motherhood between the always-employed and those who would be employed only as non-mothers, and (2) the proportion of employed non-mothers who would leave upon motherhood. The sign of the bias term depends on whether women who maintain employment through motherhood have higher or lower potential wages, on average, compared to those who would leave employment upon motherhood.

This derivation also speaks to the implication of mean dominance. Mean dominance would assume that women who maintain employment through motherhood have higher potential wages even under non-motherhood compared to those who would exit employment upon becoming moth-

ers. Under mean dominance, our bias term becomes surely positive since $E(Y^0 \mid S^0 = S^1 = 1) - E(Y^0 \mid S^0 = 1, S^1 = 0) \geq 0$, suggesting that naive comparisons would understate the size of the motherhood wage penalty.