Gap-closing estimators to study categorical inequality that persists under a local intervention to equalize a treatment∗

Ian Lundberg

Department of Sociology and Office of Population Research
Princeton University
Last updated: 29 September 2019

Keywords: causal inference, statistical methods, race, class, gender

*Manuscript in progress. Address correspondence to Ian Lundberg, ilundberg@princeton.edu. I thank Brandon Stewart for comments on an earlier draft and Rebecca Johnson and Simone Zhang for helpful discussions. Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879. Replication code is available at github.com/ilundberg/replication and will be posted on the Harvard Dataverse when the paper is forthcoming for publication.
Title

Gap-closing estimators to study categorical inequality that persists under a local intervention to equalize a treatment

Abstract

Gaps in socioeconomic outcomes by race, gender, and class are central to the study of social stratification and inequality. Studies of these gaps often avoid causal claims because these demographic categories are non-manipulable. Drawing on literature in epidemiology, this paper reviews a causal estimand that may be of interest: the gap across demographic categories which would persist under a local intervention to equalize a treatment. Unlike most causal estimand, this gap-closing estimand involves potential outcomes under a single treatment, instead of two treatment levels. After reviewing the identification assumptions outlined in prior literature, I formalize a double machine learning gap-closing estimator. This estimator uses observed cases in the treatment level of interest to learn a function mapping pre-treatment covariates to the potential outcome under this treatment, and it combines this model with a propensity score model to yield efficient estimates. The paper concludes with implications for practice: a gap-closing estimator directly targets the research goal of greatest substantive interest, thereby providing tools for the rigorous study of categorical inequality across social groups.
1 Introduction

The study of inequality often involves claims about gaps or divides between subgroups. Social scientists often examine the role of certain factors in creating the gaps: the role of education in social mobility, the role of occupational segregation in the perpetuation of the gender pay gap, the role of incarceration in the persistence of racial inequality, etc. In each case, the central question is how a gap would differ if some treatment or institutional arrangement were modified. Because gaps appear descriptive, researchers often tackle these problems without any reference to causal inference: summarize a gap, adjust for covariates hypothesized to play a role, and then report the gap that remains. Yet the degree to which a demographic disparity would close under an intervention is a causal claim: it involves an intervention. Motivated by the causal nature of this claim, this paper links a common demographic goal—how a disparity would change under intervention to a treatment—to a causal estimand formalized in biostatistics (Jackson and VanderWeele, 2018) and to the growing literature in double machine learning for estimation under weak parametric assumptions (Chernozhukov et al., 2018).

Two aspects of this causal estimand must be drawn out. First, the intervention is local. Like many causal estimands, the target of inference involves the expected outcome if a random unit is assigned to a particular treatment. Global claims, such as the degree to which the gender wage gap would close if we equalized the distribution occupations in the population, would require stronger assumptions and structural modeling approaches. Second, the causal estimand involves an intervention to fix a treatment at one level. While we typically think of causal effects as the difference between the outcomes under two treatments, gap-closing estimands involve the difference across groups in the expected outcome under an intervention to set treatment to a single, fixed value. This distinction has clear implications for the design of studies targeting gap-closing estimands: we would like to have a large sample in the treatment level of interest.

A brief overview of the estimation approach clarifies why this data structure is optimal. Denote the demographic gap-defining variable $X$, the treatment $T$, the pre-treatment covariates $L$, and the outcome $Y$. There are two functions to be estimated. First, I estimate an outcome function
\( g(\vec{\ell}, x) = \mathbb{E}(Y \mid \vec{L} = \vec{\ell}, X = x, T = t) \). With a large sample observed in the treatment of interest \( T = t \), the function \( g() \) can be estimated by a flexible machine learning model. Second, I estimate a treatment propensity \( m(\vec{\ell}, x) = \mathbb{P}(T = t \mid X = x, \vec{L} = \vec{\ell}) \). These two estimated functions can yield a doubly-robust augmented inverse probability weighted estimator (Robins and Rotnitzky, 1992), which when applied with machine learning estimators and sample splitting can yield a consistent estimator of the target parameter with valid asymptotic convergence (Chernozhukov et al., 2018).

To illustrate, I explore evidence in the U.S. on a quantity studied in Britain by Laurison and Friedman (2016): the gap in pay by occupational class origins among those who attain a higher managerial or professional class destination. Following Jackson and VanderWeele (2018), I discuss how this descriptive result relates to a causal claim about the degree to which gaps in pay by class origins would close under a particular intervention to class destination.

The paper proceeds in several sections. I first introduce the commonality of gap-closing estimands through examples from sociological research. Second, I review prior work defining the causal estimand at the core of these studies (Jackson and VanderWeele, 2018). I review how this work addresses a fraught problem of how to study immutable characteristics (e.g. race) within a causal framework. In short, one can study the effect of various treatments on racial gaps without any appeal to an intervention on race itself. The resulting estimand is not only scientifically useful but also policy relevant: we would like to know about treatments that can reduce racial inequality. Third, I develop the primary contribution of the paper: the application of double machine learning to produce a gap-closing estimator that researchers can use to study gaps that persist after a local intervention to equalize the treatment. Fourth, I apply the estimator to a substantive example about the class ceiling: the degree to which class origins remain associated with earnings even among those who attain higher managerial and professional occupations (Laurison and Friedman, 2016). Fifth, I briefly discuss the connection between gap-closing estimands and related estimands: conditional gaps (including Blinder-Oaxaca decomposition), treatment effect deviation, controlled mobility, and controlled direct effects. The paper concludes with a discussion.
1.1 Gap-closing estimands are ubiquitous in social science research

Gap-closing estimands are extremely common research goals. Although researchers rarely state their goal as a combination of a descriptive category and a causal treatment, the discussion surrounding published estimates often points directly toward a gap-closing estimand. This is especially true in the study of race, class, and gender.

Those who study race frequently seek to explore the degree to which racial inequality would be different under alternative sets of institutional arrangements. Western (2006) studies the role of incarceration in the racial earnings gap among men. Although incarceration substantially harms earnings, it is rare enough that “the difference in earnings between blacks and whites would be reduced only by about 3 percent if the incarceration rate were zero,” (p. 127). This is a gap-closing estimand that involves descriptive components (the initial racial gap and the proportion incarcerated) but turns crucially on a causal effect (the effect of incarceration on the earnings of black and white men). Ciocca Eller and DiPrete (2018) examine the black-white gap in college degree completion. They conclude “if black students matched to colleges in the same way as white students with similar backgrounds, their dropout rate would decrease from 50.4 to 47.5 percent, and the BA attainment rate would increase from about 19 to 20 percent,” (p. 1194). Some studies make explicit the fact that the estimand of interest involves the causal effect of some treatment. Killewald and Bryan (2016) identify the effect of home ownership on wealth and then conduct a simulation to show that “altering their [blacks] homeownership experiences to be comparable to those of whites would substantially narrow race gaps in midlife wealth,” (p. 123). The study of racial gaps quite often turns to the question of how these gaps would differ if an intervention occurred to some policy-amenable variable.

Class gaps also fall in a middle ground between description and causal inference. In the example to which this paper returns repeatedly, Laurison and Friedman (2016) study the earnings of British workers in higher managerial and professional occupations. Within this subgroup, the authors examine the class gap in pay between the intergenerationally stable (those whose parents were in this class) and the upwardly mobile (those whose parents were in a lower occupational
class). This estimand can be framed causally as a gap-closing estimand: the difference in pay between those of lower- and upper-class backgrounds, if we intervened to assign them to upper-class destinations.

Gender inequality likewise begets claims about how gender gaps would be different under alternative conditions. For instance, the gender wage gap within jobs is very small. Petersen and Morgan (1995) motivate this estimand clearly in language that closely resembles a gap-closing estimand: “Suppose sex segregation by occupation, establishment, or occupation-establishment were abolished; what then would the remaining gender relative wages be?” (p. 338). The gender wage gap is purely a descriptive quantity, but the hypothetical intervention of equalizing the distribution across occupations and establishments invokes a causal effect of occupations and establishments on wages. Given persistent gender segregation across occupations with unequal pay (Blau and Kahn, 2017), this gap-closing estimand remains central to the study of gender wage inequality today.

1.2 Causal inference formalizes this research goal

Gap-closing estimands require us to infer the expected outcome that a randomly sampled individual of each group of interest would realize under a particular intervention. Thus, the language of causal inference provides one tool to formalize this goal. This section briefly reviews results by Jackson and VanderWeele (2018) which formalize the notion of disparities that persist under an intervention. I use the potential outcomes notation of Imbens and Rubin (2015). Let \( y_i(t) \) denote the potential outcome that unit \( i \) would realize under an intervention to fix treatment at a value \( T_i = t \). The expected gap that would persist between the outcomes \( Y \) of randomly sampled individuals from demographic groups \( X = 0 \) and \( X = 1 \) under an intervention to fix the treatment

\[ \text{Expected gap} = \mathbb{E}[Y(Y=1) - Y(Y=0)] \]

Jackson and VanderWeele (2018) formalize the identification assumptions using Directed Acyclic Graphs (Pearl, 2000). The use of DAGs requires placing the gap-defining variable \( X \) in the graph, and thus reasoning about the causal effects of this variable on the treatment and the outcome. The potential outcomes notation allows me to formalize assumptions without wading into debates about whether this variable has well-defined causal effects, an issue that has raised questions (Kaufman, 2014; Glymour and Glymour, 2014) for previous papers (VanderWeele and Robinson, 2014) examining the degree to which interventions can close racial gaps.
to $T = t$ can be formalized as follows.

$$
\tau = \frac{1}{\sum_i \mathbb{I}(X_i = 1)} \sum_{i: X_i = 1} y_i(t) - \frac{1}{\sum_i \mathbb{I}(X_i = 0)} \sum_{i: X_i = 0} y_i(t)
$$

A set of assumptions is needed to identify $\tau$ with observable data. The terminology for these assumptions differs across statistics ([Imbens and Rubin, 2015]), computer science ([Pearl, 2000]), economics ([Angrist and Pischke, 2008]), and biostatistics ([Hernán and Robins, 2019]). Because the study of disparities that persist after equalizing a treatment is most developed in biostatistics and epidemiology, I use the terminology from biostatistics which invokes three assumptions: consistency, positivity, and conditional mean independence.

The first required assumption (consistency) links the observable data to the potential outcomes: the observed outcome for unit $i$ is the potential outcome under the treatment assignment that unit $i$ actually experiences.

**Assumption 1.** (Consistency) $Y_i = y_i(T_i)$ \quad \forall i

The second assumption (positivity) ensures that we can learn from data about each unit’s potential outcome by requiring that each unit has some non-zero probability of being assigned to treatment $T_i = t$.

**Assumption 2.** (Positivity) $P(T_i = t) > 0$ \quad \forall i

The third assumption (conditional mean independence) addresses confounding: assume the mean potential outcome under treatment $t$ for those observed in this treatment is equal to the population-average potential outcome under this treatment, within subgroups of the gap-defining variable $X$ and the pre-treatment covariates $\vec{L}$.

**Assumption 3.** (Conditional mean independence)

$$
\frac{1}{N_{t,x,\ell}} \sum_{i: \{T_i, X_i, \vec{L}_i\} = \{t, x, \ell\}} y_i(t) = \frac{1}{N_{x,\ell}} \sum_{i: X_i = x, \vec{L}_i = \ell} y_i(t) \quad \forall \vec{x}, \vec{\ell}
$$
where \( N_{t,x,\ell} \) denotes the population size with treatment \( t \), gap-defining variable \( x \), and pre-treatment covariates \( \vec{L} \) and \( N_{x,\ell} \) defines the population size summed over all treatments. The assumption is termed conditional mean independence because the conditional mean of \( Y(t) \) given \( \vec{X} \) and \( \vec{L} \) does not depend on the observed treatment value \( T \). This is the same assumption that would be applied to two treatment levels \( t \) and \( t' \) if our aim was to estimate the causal effect of an intervention from \( t \) to \( t' \). Because a gap-closing estimand focuses exclusively on one treatment level \( t \), conditional mean independence only needs to hold for that particular treatment level. The other potential outcomes, which are not the object of study, may be arbitrarily related to treatment assignment without threatening the validity of estimates.

These assumptions identify \( \tau \) by the following proof.

\[
\tau = \frac{1}{N_1} \sum_{i: X_i = 1} y_i(t) - \frac{1}{N_0} \sum_{i: X_i = 0} y_i(t) \tag{2}
\]

Denote \( \theta_1 \) and \( \theta_0 \)

\[
\theta_x = \frac{1}{N_x} \sum_{i: X_i = x} y_i(t) \tag{3}
\]

for each \( x = \{1, 2\} \)

\[
= \sum_{\ell} \frac{N_{x,\ell}}{N_x} \left( \frac{1}{N_{x,\ell}} \sum_{i: X_i = x, \vec{L}_i = \ell} y_i(t) \right) \tag{4}
\]

by conditional mean independence

\[
= \sum_{\ell} \frac{N_{x,\ell}}{N_x} \left( \frac{1}{N_{t,x,\ell}} \sum_{i: T_i = t, X_i = x, \vec{L}_i = \ell} y_i(t) \right) \tag{5}
\]

by consistency

\[
= \sum_{\ell} \frac{N_{x,\ell}}{N_x} \left( \frac{1}{N_{t,x,\ell}} \sum_{i: T_i = t, X_i = x, \vec{L}_i = \ell} Y_i \right) \tag{6}
\]

Positivity guarantees that an infinite sample with \( T_i = t \) would be observed in an infinite sample within the subgroup with \( \vec{X}_i = \vec{x} \) and \( \vec{L}_i = \vec{\ell} \). In the sense that they can be estimated with infinite data, \( \theta_1 \) and \( \theta_0 \) are identified by the causal assumptions. By extension, \( \tau \) is also identified.
1.3 Gap-closing estimands facilitate the study of immutable characteristics

The formal identification assumptions above clarify a key issue with gap-closing estimands: researchers often write as though the treatment variable is the gap-defining variable, when the treatment is actually the variable that we theorize might close the gap. This means that gap-defining estimands are well-defined even for gap-closing variables for which causal inference is difficult or impossible.

For some research topics, the “effect” of the gap-defining variable $X$ is not well-defined. This is particularly true in the study of immutable characteristics such as race, for which the causal effect is a fraught topic (Holland, 2008; Sen and Wasow, 2016). The meaning of an intervention to change one’s race is difficult to fathom, making race challenging for a paradigm that demands no causation without manipulation (Holland, 1986). Those who ascribe to this view may avoid the study of race or study it only in small components, such as by manipulating names on job applications in an audit study (Bertrand and Mullainathan, 2004). Yet manipulations of a single aspect of race do violence to the notion of race as a set of complex and socially constructed categories inherent in one’s identity (Kohler-Hausmann, 2018). To respect the socially construction of race, Kohler-Hausmann (2018) proposes a new framework of constitutive explanation that defines something to have occurred “because of” race only within a set of complex social meanings and moral objections. Gap-closing estimands represent a different resolution to this problem. Instead of abandoning the social construction of race or abandoning the counterfactual causal model, gap-closing estimands sidestep the problem by including all causal and non-causal associations between the gap-defining variable $X$ and the outcome $Y$. The gap is never regarded as a causal effect; it is a descriptive difference that may be different under various interventions to the intervention $T$. While it would be difficult (or impossible) to defend consistency, positivity, and ignorability of the descriptive category $X$, these assumptions may be entirely straightforward for the gap-closing intervention $T$. Gap-closing estimands thus allow the study of race, gender, and class gaps within the framework of causal inference, without requiring us to squeeze these variables into a category of “treatments” where they do not easily fit.
For other research topics, the effect of the gap-defining variable may be defined but difficult to identify. Suppose the gap-defining variable \( X \) is father’s education. It is easy to imagine how one’s life might be different if one’s father had a different degree of education. Yet identifying the effect of father’s education would require data on a large set of confounding variables \( \bar{X} \) that occurred before one’s father completed his degree. In many social science datasets, these variables are simply unavailable because they refer to a generation in the past. If the intervention \( T \) is one’s own completion of college (Hout, 1988; Torche, 2011; Zhou, 2019), there may be substantial data available on confounders of the effect of \( T \) on \( Y \), all of which refer to the focal generation of a given dataset. Likewise, there may be settings in which the intervention is randomly assigned but the gap-defining variable is not. In these settings, the gap-closing estimand (the gap that would persist under a single intervention) is a feasible target of inference. The controlled direct effect (the gap that would persist under an intervention to both the gap-defining and gap-closing variables) is not (see section on related estimands). We can study the gap-defining estimand without defining or identifying the “effect” of the gap-defining variable. Identification requirements focus instead on the gap-closing variable.

2 Estimation procedures for gap-closing estimands

Although called by different names, gap-closing estimands have already existed in the statistics literature (Jackson and VanderWeele, 2018). The novel contribution of the present paper is a link between gap-closing estimands and an estimation framework developed in the setting of causal inference: double machine learning. To build into this connection, I begin by reviewing treatment modeling and outcome modeling approaches and show how these point toward a joint strategy that yields a robust estimator with asymptotic guarantees. Treatment and outcome modeling are each valid approaches to estimate the gap-closing estimand \( \tau \). Suppose we knew the generalized propensity score (Imbens, 2000): the probability of treatment assignment \( T = t \) given
the gap-defining variable $X$ and pre-treatment covariates $\vec{L}$. Denote this function $m(x, \vec{\ell})$.

$$m(x, \vec{\ell}) = P(T = t \mid X = x, \vec{L} = \vec{\ell}) \quad (7)$$

If we knew the propensity function $m()$, we could directly estimate the population-average potential outcome under treatment for subgroup $X = x$ by inverse probability weighting (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983; Robins, 1986).

$$\theta^\text{Treatment}_x = \frac{1}{N} \sum_{i: \{T_i, X_i\} = \{t, x\}} \frac{Y_i}{m(x, \vec{\ell}_i)} \quad (8)$$

Likewise, we would be able to directly estimate the population-average potential outcome under treatment for each subgroup if we knew the outcome function $g()$ within the group to receive treatment $T = t$.

$$g(x, \vec{\ell}) = E(Y \mid T = t, X = x, \vec{L} = \vec{\ell}) \quad (9)$$

In that case, we could directly plug $g()$ into the empirical estimand (Eq. 6) and thus know the potential outcome under treatment $t$ for subgroup $X = x$.

$$\theta^\text{Outcome}_x = \sum_{\ell} \frac{N_{x, \ell}}{N_{t, x, \ell}} g(x, \vec{\ell}) \quad (10)$$

Eq. 8 and 10 can be combined to yield an augmented inverse probability weighted parameter that incorporates both sources of evidence (Robins and Rotnitzky, 1992; Glynn and Quinn).
This parameter can be viewed in two equivalent ways.

\[
\theta_{x}^{\text{AIPW}} = \frac{1}{N} \sum_{i} \left( \frac{\mathbb{I}(T_i = t) Y_i}{m(x, \ell_{i})} - \left( \frac{\mathbb{I}(T_i = t) g(x, \ell_{i})}{m(x, \ell_{i})} - g(x, \ell_{i}) \right) \right)
\]

(11)

\[
= \frac{1}{N} \sum_{i} \left( g(x, \ell_{i}) - \frac{\mathbb{I}(T_i = t) \left( g(x, \ell_{i}) - Y_i \right)}{m(x, \ell_{i})} \right)
\]

(12)

The AIPW parameter is doubly robust: if we either \( g() \) or \( m() \) is correct, then \( \theta_{x}^{\text{AIPW}} = \theta_{x} \).

Suppose the treatment model \( m() \) is correct. In the form of Eq. 11, the inverse probability weighting component equals the truth and the augmentation term equals zero. Suppose the outcome model \( g() \) is correct. In the form of Eq. 12, then the imputed outcome component equals the truth and the augmentation term equals zero. The AIPW parameter is superior to approaches that only model the outcome or only model the treatment because it incorporates both components so that the result is robust to misspecification of either one as long as the other is correct.

Chernozhukov et al. (2018) provide a re-interpretation of the AIPW estimand in terms of bias correction. Focusing on the outcome imputation form (Eq. 12), the augmentation term “corrects” for errors in the imputed outcome. For the units actually observed with treatment \( T_i = t \), the difference between the predicted outcome \( g(x, \ell_{i}) \) and the observed outcome \( Y_i \) provides information about how the model is wrong. The average of these errors would provide a sense of the degree to which the outcome model \( g() \) is biased for the units who actually receive treatment \( T = t \). The quantity of interest, however, averages over the entire population. The denominator \( m(x, \ell_{i}) \) reweights the bias within strata of \( \ell \) to approximate the overall bias of \( g() \) in this target population. The implication of this re-interpretation is that there is something to be gained by sample splitting. A machine learning estimator of \( g() \) will be biased; machine learning estimators accept some bias in order to achieve an optimal bias-variance tradeoff. If we estimate \( g() \) on the same sample on which we aggregate by Eq. 12 then \( g() \) may be biased toward \( Y \) such that an
average involving $g(x, \vec{\ell}_i) - Y_i$ does not yield a good estimate of the bias. Chernozhukov et al. (2018) get around this problem by proposing a cross-fitting estimator.

Split the sample randomly into folds $f = 1, \ldots, F$. Let $g(-f)$ and $m(-f)$ denote treatment and outcome functions estimated on the subsample of units not in fold $f$. Let $f_i$ denote the fold membership of each unit $i$. The double-debiased machine learning estimator adapts the AIPW estimator so that the predictions for each unit $i$ are based on fits $g(-f_i)$ and $m(-f_i)$ fit on samples that exclude unit $i$.

$$
\theta^\text{Debiased}_x = \frac{1}{N} \sum_i \left( g(-f_i)(x, \vec{\ell}_i) - \frac{\mathbb{I}(T_i = t) \left( g(-f_i)(x, \vec{\ell}_i) - Y_i \right)}{m(-f_i)(x, \vec{\ell}_i)} \right) \tag{13}
$$

2.1 Estimation from a sample

The preceding arguments have assumed that $g()$ and $m()$ are known and that the sum can be taken over the entire population. However, we often work only with a sample of units (denoted $S_i = 1$) which we use to learn estimates $\hat{g}$, $\hat{m}$ and $\hat{\theta}_x$. The literature on causal inference and machine learning often assumes a simple random sample (Athey and Imbens, 2016; Chernozhukov et al., 2018). Yet the standard surveys used in social science research often involve weighted samples. This section focuses on estimation for a probability sample in which sample inclusion is independent of $T$ and $Y$ given the pre-treatment covariates $X$ and $\vec{L}$, with known probabilities of inclusion.

**Assumption 4.** (Probability sample) $\mathbb{P}(S_i = 1) = f(X_i, \vec{L}_i) = \pi_i \quad \forall \ i$

Because sample inclusion is a function of predictor variables only, we can ignore sample design when estimating the treatment model $m()$ and the outcome model $g()$. Weights are neces-
sary in the step that converts these models to an estimate \( \hat{\theta}_x \) of the target parameter \( \theta_x \).

\[
\hat{\theta}_x^{\text{Debiased}} = \frac{1}{\sum_{i:S_i=1} \frac{1}{\pi_i}} \sum_{i:S_i=1} \frac{1}{\pi_i} \left( \hat{g}(-f_i)(x, \ell_i) - \mathbb{I}(T_i = t) \left( \hat{g}(-f_i)(x, \ell_i) - Y_i \right) - \hat{m}(-f_i)(x, \ell_i) \right)
\]

(14)

3 Empirical example: Class ceiling in pay

Laurison and Friedman (2016) explore the log incomes of British workers who attain higher managerial and professional occupations. Among this high-attainment subgroup, the mean log income is still lower for those whose parents held working-class occupations. The authors coin the term “class ceiling” for this intriguing descriptive result. I extend this result to a related estimand which is causal: what gap in pay by class origin would persist if we randomly assigned individuals to higher managerial and professional occupational destinations? This gap-closing estimand more directly relates to a theoretical claim about whether one’s own occupational attainment can break one free from the constraints of social origins. I assess this new question in the U.S. context. To highlight the concept of a gap-closing estimand, I keep other aspects of the empirical example as simple as possible (e.g. only four control variables). The purpose of this example is to make a gap-closing estimand concrete, not to provide the final word on the class ceiling in the United States. Results indicate that a gap-closing estimand points toward a substantively different research goal with a different point estimate from the regression coefficients that lie at the core of a traditional approach.

3.1 Sample

I analyze data from the pooled 1975–2018 General Social Survey (GSS, Smith et al. 2018).\(^2\) The advantage of the GSS is that it has consistently asked questions about the respondent’s income and the income of the father figure with whom the respondent lived at age 16, making it a standard

\(^2\)This excludes the 1972 and 1973 waves, in which respondent incomes were not recorded, and the 1975 wave, in which the sample design variables used to calculate standard errors are unavailable.
In most years, the GSS drew a sample of U.S. households selected with equal probabilities. To focus on this core sampling design which was consistent across years, I exclude the oversamples of black households which were added to the design in 1982 and 1987. Because the analysis focuses on respondent-level variables and only one respondent per household was interviewed, those residing in larger households had lower probabilities of selection. These weights are incorporated into the analysis according to Eq. 14. Because I pool across years, I adjust the weight variable to give equal total weight to each survey year in the analytic sample. Pooling across years yields a sufficiently large sample to illustrate the key differences between the gap-closing estimand and other estimands of potential interest; future research might explore how the gap-closing estimand has changed over time.

I restrict to those in the prime working ages of 30–45 (N = 19,398), with valid reports for the occupation of the father figure with whom they lived at age 16 (N = 15,773), who report their own current occupation (N = 15,352), and for whom each of these occupations maps into the Erikson-Goldthorpe-Protocarero class schema (N = 14,725, see variable construction below). Because those who do not complete high school are very unlikely to attain higher managerial or professional employment, I restrict to those who report at least a high school degree (N = 13,166). Finally, I restrict to an analytic sample that reports a non-zero personal income (N = 10,704).

### 3.2 Variables

I operationalize class by whether an individual holds an occupation in Class I of the Erikson-Goldthorpe-Portocarero (EGP) schema (Erikson et al., 1979), which Laurison and Friedman (2016) highlight as the U.S. parallel to the U.K. National Statistics Socioeconomic Classification used in their study. Class I occupations are a broad class of higher managerial and professional jobs including managers, engineers, scientists, and lawyers. The GSS harmonizes occupations according to 2010 Census codes, and I categorize these into EGP classes according to a crosswalk provided in a GSS methodological memo (Morgan, 2017).
The gap-defining variable is whether one’s father held a Class I occupation. The gap-closing intervention is whether one personally attains a Class I occupation. To identify the effect of the gap-closing estimand, I condition on a set of confounding variables \( \vec{L} \) that are likely to affect attainment of a Class I occupation and also affect income directly. Because the aim is not to estimate the effect of the gap-defining variable, these variables may be consequences of father’s occupation with no consequence. I condition on race (white, black, other), sex, age, and highest degree attained (high school, junior college, bachelor’s, or a graduate degree).

### 3.3 Identification

Identification in this setting is similar to the general discussion of Sec. 1.2. The causal effect of occupation (and thus the gap-closing estimand) is identified if the population-average potential outcome under control is equal for those observed in the active treatment (higher managerial and professional occupations) and those not observed in the active treatment, within each subgroup of pre-treatment covariates. In this example, pre-treatment covariates include father’s occupational class, race, sex, age, and highest degree attained. Identification of causal effects in mobility research is always challenging, and this assumption is unlikely to hold perfectly. The aim of this paper is to introduce estimation procedures for gap-closing estimands, and it is left for future work to re-estimate them in settings where the identifying assumption may be particularly plausible.

### 3.4 Estimation

I estimate the outcome function \( g \) and the treatment propensity function \( m \) by OLS and by a random forest fit with the \texttt{ranger} package in R, estimated via \( F = 5\)-fold cross-fitting and double machine learning as discussed in Section 1.2. The OLS model is likely to perform

---

3The category of higher managerial and professional occupations is an aggregate category which actually contains many treatments (occupations) that may lead to different potential outcomes. The estimand in this example is expected outcome realized under a random draw from the distribution of higher managerial and professional occupations among those with covariates that match the covariates of the unit in question. Lundberg (2019) terms this estimand an aggregate effect weighted by the conditional prevalence of treatments.
well even in small samples because it involves a few parameters, but it offers no guarantees if the parametric form of the model is incorrect. The random forest, on the other hand, can flexibly adapt to any functional form but leans more heavily on the data and may have poorer performance in very small samples. To report uncertainty, I estimate $V(\hat{\tau})$ by the empirical variance of 100 bootstrap iterations and construct a 95% confidence interval by a normal approximation centered at $\hat{\tau}$ with the estimated variance $\hat{V}(\hat{\tau})$.

3.5 Results

Fig. 1 presents results. Those whose father held an EGP-Class I occupation attain log incomes that are 0.28 higher than those whose father did not (A). This gap attenuates to 0.16 log income points among those who are observed to have themselves attained a Class I occupation (B). It is, however, difficult to interpret (B) in terms of a hypothetical intervention because those who attain high class destinations are positively selected in numerous ways. Once we address observed selection, our estimate (C) for the gap-closing estimand is as large as the descriptive gap (0.28). Although the gap by social origins is smaller among those who attain high class destinations, this can be entirely explained by selection on observed characteristics. The results provide a stronger version of a central claim of [Laurison and Friedman] (2016): high occupational attainment does little, if anything at all, to erase income gaps by class origins.

4 Related estimands

Gap-closing estimands are closely related to several other common estimands: conditional gaps, controlled mobility, treatment effect deviation, and controlled direct effects. The intended audience of this section is specialists who may wonder how the estimand in this study differs from those of related studies.
Fig. 1. Attainment a high occupational class only slightly closes the gap in pay by class origins in the U.S. Result illustrates a conceptual idea from Laurison and Friedman (2016). Data come from the pooled 1975–2018 General Social Survey. Incomes are in constant 1986 dollars, logged for analysis. The descriptive gap is the mean difference in the log incomes of those from Class I vs. other origins. The gap given a high class destination is the gap between those with Class I vs. other origins, among those who personally attain a Class I destination. The gap-closing estimand targets the gap in pay that we would expect if we randomly sampled someone of Class I origins and someone of lower-class origins and assigned them both to a Class I occupational destination, without changing their other covariates.

4.1 Conditional gap estimands

A common empirical approach examines the gap in the outcome $Y$ across the gap-defining variable $X$ unconditionally and then conditional on a set of other variables. In our running example, Laurison and Friedman (2016, Table 3) restrict the sample to those who attain higher-level managerial and professional occupations $T$ and then present the coefficient on class origin $X$ in a series of models that condition on increasing sets of covariates $\bar{L}$. The authors conclude that “the upwardly mobile...still face a class pay gap that persists even when adjusting for a range of factors believed to affect earnings,” (Laurison and Friedman, 2016, p. 680). This conclusion is valid but remains in the realm of description, and is therefore distinct from a gap-closing estimand.

Fig. 2 reproduces this result using GSS data. Estimates (A) and (B) are as before: the gap in pay by class origins taken marginally and then conditional on attaining a high class destination. A researcher might correctly argue that (B) could be misleading: it fails to adjust for selection into a high class destination. Because occupational destination is a consequence of both class origin $X$
and confounding variables \( \vec{L} \) such as education, those who break free from disadvantaged origins to attain a high class destination despite disadvantaged origins may be especially positively selected on the confounding \( \vec{L} \). Formally, class destination \( T \) is a collider \([Elwert and Winship, 2014]\) and conditioning on \( T \) induces an association between \( X \) and \( \vec{L} \) (Fig. 3 Panel C).

A researcher might therefore add controls for other variables that cause occupational attainment: race, sex, age, and education. Adding these covariates reduces the coefficient on class origin to 0.10 log points, with a confidence interval that includes 0. One might therefore be tempted to conclude that class origins have only small consequences once class destinations are accounted for. Yet this would miss the primary substantive conclusion from the gap-closing estimand of Fig. 1: an intervention to class destination would hardly reduce the gap in pay by class origins at all. The conditional gap estimator of Fig. 2 blocks an important pathway through \( \vec{L} \) that would still create a gap in pay by class origins even under an intervention to class destinations. This is illustrated in the DAG in Fig. 3 Panel C. A fully conditional estimator may provide a useful descriptive quantity, and it may be informative (under very strong assumptions) about the class gap that would persist under a grand intervention to equalize all the variables in both \( T \) and \( \vec{L} \). It nevertheless does not answer the research question formalized in a gap-closing estimand.

Conditional gap estimands that adjust for many covariates appear frequently in sociological research in the form of Blinder-Oaxaca decompositions. Originally proposed by \([Blinder, 1973]\) and \([Oaxaca, 1973]\), this technique (or the extension to generalized linear models by \([Fairlie, 2005]\)) appears in examinations of inequality over numerous gap-defining variables in sociology: class origins \([Laurison and Friedman, 2016]\), race \([Ciocca Eller and DiPrete, 2018]\), disability \([Shandrea, 2018]\), sexual orientation \([Mize, 2016]\), and gender \([Weisshaar, 2017]\), to name a few. These decompositions inherit all the problems of fully conditional estimands. They tell us about the descriptive gap among those who are equal on all observed covariates, but without additional assumptions they say nothing about what gap would persist under an intervention.

The interpretations authors give to Blinder-Oaxaca decompositions, however, frequently reach beyond the descriptive claim to a causal goal. It is common to attribute “sources” of the gap
or to reason about what gap would persist if covariates were equal. For example, Mize (2016, p. 1144) writes “even if bisexual women had the same education, were in similar professions, of a similar age, and equivalent on all other characteristics—they would still earn less than heterosexual women.” This claim invokes a causal quantity: what would happen under an alternative state of the world. In that sense, the goal stated in the text is closer in spirit to a gap-closing estimand than to the descriptive quantity estimated by a Blinder-Oaxaca decomposition. This goal, however, is far more grand than a gap-closing estimand. It seems to require an intervention to education, profession, age, and all other characteristics at once. Further, it appears to involve an intervention to these variables for all individuals simultaneously, creating challenges for inference about generalized equalibrium effects. Rather than drawing sweeping conclusions from a descriptive decomposition, I propose that researchers use gap-closing estimands to formally identify the effect of a specific

Fig. 2. Misleading result that simply adds covariates to the regression model. Result illustrates a conceptual idea from Laurison and Friedman (2016). Data come from the pooled 1975–2018 General Social Survey. Incomes are in constant 1986 dollars, logged for analysis. (A) The descriptive gap is the mean difference in the log incomes of those from Class I vs. other origins. (B) The gap given a high class destination is the gap between those with Class I vs. other origins, among those who personally attain a Class I destination. (C) The fully conditional gap adds controls for race, sex, age, and education. None of the results in this figure show how wide the class gap would be if we intervened to raise people to high-class occupations. For that research question, a gap-closing estimand is needed.
Fig. 3. **A gap-closing estimand is distinct from other common estimands.** Solid and dotted blue edges contribute to the association between $X$ and $Y$ included in each estimand. Dashed gray edges do not contribute to the association included in the estimand. Boxes indicate conditioning. Dotted edges denote dependence induced by conditioning. A gap-closing estimand (A) captures the relationship between a gap-defining variable $X$ and an outcome $Y$ that would persist if we intervened to set an intervention $T$ to a particular value. It is distinct from the gap (B) conditional on the gap-closing variable $T$ because conditioning on $T$ induces an association between $X$ and $\bar{L}$. It is distinct from the gap (C) conditional on $T$ and $\bar{L}$ because this gap blocks the path $X \rightarrow \bar{L} \rightarrow Y$ by which the gap-defining $X$ would be related to $Y$ even after an intervention to $T$. It is distinct from a controlled direct effect (D) which excludes the non-causal association between $X$ and $Y$.

intervention to one variable, with transparency about the required identification assumptions.

### 4.2 Treatment effect deviation

[An and Glynn (2019)](An and Glynn, 2019) propose an extension of Blinder-Oaxaca decomposition to a causal framework. Their discussion differs from the present paper in one critical respect: they conceptualize the gap-defining variable as a treatment which could be manipulated. The authors show that a Blinder-Oaxaca decomposition in this setting decomposes a difference between the two treatment arms into a selection component (endowment effects) and a causal component (coefficient effects), under a linear parametric assumption. To move away from the parametric assumption, they pro-
pose a new estimand—treatment effect deviation—which explores how the estimated causal effect of a gap-defining variable changes when a given pre-treatment covariate is omitted from the model or the matching procedure. Treatment effect deviation is an important contribution that may be especially useful when the gap-defining variable is a manipulable treatment, but it does not apply to the setting of a difference across non-manipulable groups that is causally modified by some intervention. For this setting, a gap-closing estimand is necessary.

4.3 Controlled mobility

Zhou (2019) proposes an estimand—controlled mobility—which can be interpreted as a special case of gap-closing estimands, albeit with some issues which I discuss. The paper focuses on the role of education in intergenerational income mobility, following substantial prior literature (Hout, 1988; Torche, 2011). The gap is the income differential in adulthood among those who begin with different family incomes in childhood, and the intervention is completion of a college degree. The controlled mobility estimand is defined as follows.

Controlled mobility...reflects the degree of mobility we would observe among college graduates if, given parental income, college graduation did not depend on other predictors of adult income (i.e., after we control for selection processes that may confound the causal effect of a college degree on intergenerational mobility). (Zhou, 2019, p. 466).

This definition makes an important contribution by highlighting the need to adjust for confounders of the college-income association. Yet it falls short of the precision possible in gap-closing estimands in three key respects.

First, while the text statement invokes a causal effect, the mathematical statement (Eq. 4 in the original paper) defines this quantity without potential outcomes or do-operators, thus not invoking any formal framework of causal inference. Instead, the estimand is defined by the conditional independence assumption in the text: the quantity that would result “if, given parental
income, college graduation did not depend on other predictors of adult income,” (Zhou, 2019, p. 466).

Second, the author goes on to state that the estimand “might also reflect the degree of intergenerational mobility we would observe in a utopian world of ‘college for all’,” (Zhou, 2019, p. 466). This statement confuses a general equilibrium result with a local equilibrium result. The general equilibrium notion of “college for all” requires strong assumptions that other aspects of the system would not change in this massive intervention. Placing gap-closing estimands within the formal language of causal inference makes it clear that the evidence more strongly supports the usual local equilibrium claim: the difference in intergenerational mobility we would expect for an individual selected at random from the population who is assigned to college vs. not.

Third, the definition of controlled mobility by a particular conditional independence assumption points toward Zhou’s (2019) regression-with-residuals approach as the only estimation strategy to learn this quantity. Gap-closing estimands, by contrast, place the research goal squarely within the domain of causal inference and therefore open the door to numerous estimation strategies developed for the estimation of causal effects.

To summarize, Zhou (2019) makes an important contribution by proposing controlled mobility as a theoretically-motivated estimand and a tool to estimate it. This paper simply points out that controlled mobility is one example of a broader class of gap-closing estimands that can be made even more precise through the formal notation and results of causal inference.

4.4 Controlled direct effects

Gap-closing estimands are similar to controlled direct effects (Pearl, 2001; Robins, 2003; Acharya et al., 2016; Zhou and Wodtke, 2019). A controlled direct effect (Fig. 3 Panel D) is best understood in the context of a hypothetical experiment. It is the expected difference in an outcome $Y$ if we intervene to assign someone to treatment $X = 1$ or control $X = 0$ while intervening to hold a post-treatment intervention at some fixed value $T = t$. Using the notation $Y(x, t)$ to denote
the potential outcome under intervention to set \( X = x \) and \( T = t \),

\[
\text{CDE}(t) = \frac{1}{N} \sum_{i=1}^{N} \left( Y_i(1, t) - Y_i(0, t) \right)
\]

(15)

A gap-closing estimand, by contrast, does not invoke a hypothetical intervention to the variable \( X \) over which the gap is defined; potential outcomes are defined over the treatment only (Eq. [1]). This is an essential difference because the CDE requires identification of the effect of \( X \) on \( Y \), thereby losing applicability to settings in which \( X \) is immutable.\(^4\)

5 Discussion

This paper presents a gap-closing estimator by which sociologists can target inquiry toward the descriptive gap in an outcome \( Y \) between pre-existing groups \( X \) that would persist if we intervened to assign random individuals to a fixed value of some gap-closing intervention \( T \). This estimator links a problem common in sociology and demography to an estimand developed in biostatistics (Jackson and VanderWeele 2018) and an estimator from machine learning (Chernozhukov et al. 2018), thereby empowering social scientists with a flexible model to explicitly target the quantity most relevant to theories about demographic disparities.

If sociologists applied gap-closing estimators in the study of social inequality, this would substantially change the language with which we talk about race, class, gender, and other immutable characteristics. Instead of discussing the “effect” or “influence” of these variables (in scare quotes), we can formalize our research goal in the precise language of causal inference: the degree to which an intervention to some other treatment \( T \) would close the social origin, race, or gender gap in some outcome of interest. We would no longer hide behind vague phrases like “the role of \( T \)” and would instead acknowledge the causal component of our research questions. This shift would improve rhetorical clarity and also contribute to the policy-relevance of sociological

\(^4\) An and Glynn (2019) discuss controlled direct effects as one possible interpretation of Blinder-Oaxaca decompositions in which the group membership has a causal effect and covariates are consequences of the gap-defining variable, though they note that this interpretation requires very strong assumptions.
research, since policymakers can clearly understand that the research implies that an intervention to the variable studied might plausibly reduce group-based inequality. Gap-closing estimands provide a language for clarity about the target of statistical inference, thereby promoting transparent research and new estimators that may directly target the quantities we most want to know.
References


