Disparity Analysis: A Tale of Two Approaches*

Aleksei Opacic
Harvard University

Lai Wei
The University of Hong Kong

Xiang Zhou
Harvard University

January 17, 2025

Abstract

To understand the patterns and trends of various forms of inequality, quantitative social science research has typically relied on statistical models linking the conditional mean of an outcome variable to a set of explanatory factors. A prime example of this approach is the widely used Kitagawa-Oaxaca-Blinder (KOB) method. By fitting two linear models separately for an advantaged group and a disadvantaged group, the KOB method decomposes the between-group outcome disparity into two parts: a part explained by group differences in a set of background characteristics and an unexplained part often dubbed "residual inequality." In this paper, we explicate, contrast, and extend two distinct approaches to studying group disparities, which we term the *descriptive* approach, as epitomized by the KOB method and its variants, and the *prescriptive* approach, which focuses on how a disparity of interest would change under a hypothetical intervention to one or more manipulable treatments. For the descriptive approach, we propose a generalized nonparametric KOB decomposition that considers multiple (sets of) explanatory variables sequentially. For the prescriptive approach, we introduce a variety of stylized interventions, such as lottery-type and affirmative-action-type interventions that close between-group gaps in treatment. We illustrate the two approaches to disparity analysis by assessing the Black-White gap in college completion, how it is statistically explained by racial differences in demographic and socioeconomic background, family structure, academic performance and behavior, and college selectivity, and the extent to which it would be reduced under hypothetical reallocations of college-goers from different racial and economic backgrounds into different tiers of college --- reallocations that could be targeted by race- or class-conscious admissions policies.

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu. An earlier version of this paper was presented at the 2023 Annual Meeting of the Population Association of America. The authors benefitted from the comments of Geoffrey Wodtke, two anonymous reviewers, and seminar participants at New York University, Princeton University, Harvard University, the University of Chicago, the University of Hong Kong, the Chinese University of Hong Kong, and Seoul National University.

1 Introduction

Social and economic inequalities are a perennial topic of economic, demographic, and sociological research. To understand the patterns and trends of various forms of inequality, quantitative social science research has typically relied on statistical models that link the conditional mean (and, sometimes, other distributional statistics such as the conditional variance) of an outcome of interest, such as logged hourly wage, to a range of explanatory variables, such as education (e.g., Katz and Murphy 1992; Juhn *et al.* 1993), occupation (e.g., Mouw and Kalleberg 2010), and union membership (e.g., DiNardo *et al.* 1996; Western and Rosenfeld 2011). A prime example of this approach is the Kitagawa-Oaxaca-Blinder (KOB) decomposition method (Kitagawa 1955; Blinder 1973; Oaxaca 1973; see also Duncan 1968; Winsborough and Dickinson 1971; Althauser and Wigler 1972).[1] By fitting two linear regression models, one for an advantaged group (e.g., Whites) and one for a disadvantaged group (e.g., African Americans), it decomposes the between-group disparity in the mean outcome into two parts: a part accounted for by group differences in a set of explanatory variables and an unexplained part often dubbed "residual inequality." While the canonical KOB method decomposes the gap in mean outcomes based on linear models, various extensions of the KOB method have been proposed, including decomposition based on nonlinear outcome models (Fairlie, 2005; Bauer and Sinning, 2008), decomposition based on models for group membership (Barsky *et al.* 2002), and decomposition of differences in distributional statistics such as quantiles (Firpo *et al.* 2018) and the variance (Lemieux 2006; Western and Bloome 2009). The KOB decomposition and its extensions have been widely used to study "categorical inequalities" (Tilly 1998) such as disparities across gender (e.g., Petersen and Morgan 1995; Smith-Doerr *et al.* 2019), race (e.g., Mandel and Semyonov 2016; Storer *et al.* 2020), class (e.g., Laurison and Friedman 2016), and sexual orientation (e.g., Mize 2016).

While the KOB method and its extensions are useful for quantifying the contributions of various factors to a disparity of interest, they do so in an accounting or descriptive sense. Specifically,

[1]For a long time, the KOB method was (and is still by many) called the Oaxaca-Blinder or Blinder-Oaxaca decomposition (Blinder 1973; Oaxaca 1973). In recent years, it has been increasingly referred to as the Kitagawa-Oaxaca-Blinder or Kitagawa-Blinder-Oaxaca decomposition, in recognition of Kitagawa (1955), who proposed the same method in a nonparametric framework. We follow this emerging convention and use the KOB acronym.

because the KOB method does not explicitly address potential confounding of the relationship between the explanatory and outcome variables, it does not speak directly to the causal effects of the explanatory variables on the disparity of interest (Huber, 2015; Jackson and VanderWeele, 2018). As such, analyses based on these methods are ill-equipped to inform policy. For example, consider a KOB-based analysis that assesses the extent to which the Black-White gap in college completion rates can be statistically explained by racial differences in socioeconomic background, academic preparation, and college selectivity (Ciocca Eller and DiPrete 2018; Voss *et al.* 2022). While such an analysis may suggest possible causal processes leading to the completion gap, it does not target the causal effect of any of the explanatory variables in question. Consequently, it cannot be used to address policy-relevant questions. From a policy standpoint, we may wish to assess the extent to which the Black-White gap in college completion rates would be reduced if college selectivity were equalized by race, by class, or by both — for example, through race- and/or class-conscious admissions policies. Yet, such questions are not targeted by the KOB method or its current extensions.

To address these and other policy-relevant questions concerning inequality, an emerging line of research has begun to incorporate the language and logic of causal inference into the study of group disparities (VanderWeele and Robinson 2014; Jackson and VanderWeele 2018; Zhou 2019; An and Glynn 2021; Jackson 2021; Lundberg 2022; Yu and Elwert 2022; Zhou and Pan 2023; Park *et al.* 2023, 2024). In this line of research, analysts specify one or more explanatory variables as manipulable treatments and then ask how a disparity of interest would change under a hypothetical intervention to the treatment assignment process. For example, using data from the National Longitudinal Surveys, Jackson and VanderWeele (2018) investigate how racial disparities in wages and other outcomes in the US would change under several hypothetical interventions that equalize childhood SES, test score, or both, between Black and White Americans. Similarly, Zhou (2019) considers how the level of intergenerational income mobility would change under a hypothetical intervention that sets everyone's educational level to "bachelor's degree." Generalizing the quantities of interest examined in Jackson and VanderWeele (2018) and Zhou (2019), Lundberg (2022) proposes the concept of "gap-closing estimand," which "quantifies how much a gap (e.g., incomes by race) would close if we intervened to equalize a treatment (e.g., access to college)" (p. 1). More recently, extending the gap-closing estimand to a setting with sequential treatments, Zhou and Pan

(2023) estimate how much the Black-White earnings gap in the US would be reduced under a series of stylized interventions to college attendance and college completion.

In this paper, we explicate, contrast, and extend these two approaches to studying group disparities, which we term the *descriptive* approach, as epitomized by the KOB decomposition and its variants, and the *prescriptive* approach, which focuses on counterfactual disparities under hypothetical interventions to one or more manipulable treatments. Beyond a systematic review of the two approaches, we propose extensions that allow us to address a broader range of research questions. For the descriptive approach, we propose a generalized KOB decomposition that considers multiple (sets of) explanatory variables sequentially. For the prescriptive approach, we introduce new interventional estimands more akin to real-world policies than those proposed in previous work. We also propose a novel sensitivity analysis technique for gauging potential bias induced by unobserved confounding. Unlike previous approaches to sensitivity analysis that are tailored for a specific intervention (e.g., Park *et al.* 2023), this technique applies more generally to a diverse array of interventional effects. For all estimands in both the descriptive and prescriptive approaches, we present model-free definitions and nonparametric identification results. The nonparametric definition and identification of estimands improves the clarity and transparency of research goals and assumptions (Lundberg *et al.* 2021), and as we will see, opens the door to a range of parametric and nonparametric estimation strategies, including those leveraging modern nonparametric theory and machine learning methods. We also discuss the conditions under which a canonical KOB decomposition based on linear models can be repurposed to assess counterfactual disparities for a limited set of hypothetical interventions, under certain parametric assumptions (see also Jackson and VanderWeele 2018). Finally, we illustrate the two approaches to disparity analysis by considering the Black-White gap in college completion rates in the US. Using the descriptive approach, we explore how the gap is statistically explained by racial differences in demographic and socioeconomic background, family structure, academic performance and behavior, and college selectivity. Using the prescriptive approach, we estimate the extent to which it would be reduced under hypothetical reallocations of college-goers from different racial/economic backgrounds into different tiers of college --- reallocations that could be targeted by race- or class-conscious admissions policies.

2 A Motivating Example

In the United States, Black students are much less likely to complete a bachelor's degree relative to their White peers. Among those who started at a four-year college in 2010, for example, 64% of White students graduated with a bachelor's degree within six years, compared with only 40% of Black students (Jeffrey 2020; see also Snyder *et al.* 2019). The Black-White disparity in college graduation rates has barely changed over the past several decades (Voss *et al.* 2022), and, as documented by previous research, is shaped by a range of factors. These include Black students' disadvantages in socioeconomic background (e.g., Bailey and Dynarski 2011), academic preparation (e.g., Bowen *et al.* 2009), college selectivity (e.g., Reardon *et al.* 2012), as well as academic performance and social engagement during college (Tinto 1994; Braxton *et al.* 1997). In a recent study, Ciocca Eller and DiPrete (2018) examine the relative importance of these factors. Using data from the Educational Longitudinal Study of 2002, the authors apply a nonlinear variant of the KOB decomposition (Fairlie 2005) as well as counterfactual simulations based on several different models of college dropout to quantify the importance of pre-college resources, institutional characteristics, and college experience. These analyses suggest that racial disparities in academic resources prior to and within college constitute the most significant driver of the BA completion gap, and that racial differences in college quality (due to either over- or under-match of Black students) play a much smaller role.

In what folows, we revisit this example using data from the National Longitudinal Survey of Youth, 1997 cohort (NLSY97). Specifically, we illustrate a generalized KOB decomposition that considers four sets of explanatory variables successively: (a) demographic and socioeconomic background, (b) family structure, (c) pre-college academic performance and behavior, and (d) college selectivity. Unlike the model-based decompositions employed by Ciocca Eller and DiPrete (2018), our generalized KOB decomposition is defined nonparametrically, i.e., without reference to a particular linear or logistic regression model. Thus, the estimands and their interpretations are more generic and transparent. Moreover, we highlight that the "contributions" of different variables quantified by the generalized KOB decomposition have few causal interpretations. As such, they are ill-equipped to inform the effectiveness of potential institutional interventions in reducing the Black-White completion gap. For example, they cannot tell us the extent to which

an equalization in college selectivity by race or by class (either marginally or conditional on some metric of "merit") can help reduce the gap. Such questions motivate what we call the *prescriptive* approach to disparity analysis, which we discuss in Section 4.

3 Disparity Analysis for Description

In this section, we discuss the descriptive approach to disparity analysis as characterized by the KOB decomposition and its variants. We first review the canonical KOB method based on linear models (Blinder 1973; Oaxaca 1973) and then a nonparametric version of the KOB method (Barsky *et al.* 2002; Fortin *et al.* 2011). For the nonparametric version, we introduce a doubly robust estimator that offers enhanced protection against model misspecification bias. Next, we consider the case of multiple explanatory variables and contrast two alternative ways of quantifying their respective contributions to the disparity of interest: a simultaneous approach and a sequential approach. We highlight a fully nonparametric formulation of the sequential approach, which we call a generalized KOB/Duncan decomposition. Finally, we illustrate the generalized KOB/Duncan decomposition through a reanalysis of the Black-White gap in college completion.

3.1 The Canonical KOB Method

Let Y denote an outcome of interest, R a binary indicator for group membership, and X an explanatory variable for the group disparity in the mean outcome. To motivate our discussion, let us consider how parental income helps explain the Black-White gap in college completion rates among four-year college goers. In this context, Y denotes college completion, R denotes race, and X denotes parental income. The Black-White gap in college completion rates can be written as $\Delta = \mathbb{E}[Y|R = \text{White}] - \mathbb{E}[Y|R = \text{Black}]$. The goal of the KOB method is to decompose this quantity into two components: a component explained by racial differences in parental income X and an unexplained or "residual" component.

Assuming we have a representative sample of the population (possibly with the aid of sampling weights), the canonical KOB method can be implemented as follows (Blinder 1973; Oaxaca 1973):

1. Compute completion rates separately for Black and White students in the sample, which we denote by $\bar{Y}_B$ and $\bar{Y}_W$;

2. Fit a linear probability model of college completion (Y) on parental income (X) among Black students;

3. Compute the average predicted value of the outcome when we substitute White students' parental incomes for Black students' in the above model, which we denote by $\bar{Y}_B^{\text{pred}}$.

Then, the sample analog of the Black-White completion gap can be decomposed as

$$\widehat{\Delta} = \underbrace{(\bar{Y}_B^{\text{pred}} - \bar{Y}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\bar{Y}_W - \bar{Y}_B^{\text{pred}})}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{1}$$

The first component, $\widehat{\Delta}_{\text{Explained}}$, gauges how the predicted completion rate would differ if we used the parental income distribution of White students versus that of Black students in the linear probability model of step 2. The second component, $\widehat{\Delta}_{\text{Unexplained}}$, gauges the remaining difference between the observed completion rate among White students and the predicted completion rate among Black students under this hypothetical shift of their parental income distribution.

In the above algorithm, we fit a model of college completion among Black students and use it to obtain their predicted completion rate under White students' parental income distribution. Alternatively, we can fit a model of college completion among White students and use this model to obtain White students' predicted completion rate under Black students' parental income distribution. Denoting the latter quantity by $\bar{Y}_W^{\text{pred}}$, we have a different partition of $\widehat{\Delta}$:

$$\widehat{\Delta} = \underbrace{(\bar{Y}_W - \bar{Y}_W^{\text{pred}})}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\bar{Y}_W^{\text{pred}} - \bar{Y}_B)}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{2}$$

Equations (1) and (2) can also be expressed in terms of regression coefficients. If we use $(\hat{\alpha}_B, \hat{\beta}_B)$ and $(\hat{\alpha}_W, \hat{\beta}_W)$ to denote the estimated intercept-slope pairs of the regression models for Black and White students, Equations (1) and (2) can be written as

$$\widehat{\Delta} = \underbrace{\hat{\beta}_B'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_B)'\bar{X}_W}_{\widehat{\Delta}_{\text{Unexplained}}} \tag{3}$$

$$= \underbrace{\hat{\beta}_W'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_B)'\bar{X}_B}_{\widehat{\Delta}_{\text{Unexplained}}}, \tag{4}$$

where $\bar{X}_B$ and $\bar{X}_W$ denote the sample means of parental income among Black and White students, respectively.[2] From these equations, we can see that the two alternative decompositions differ only in the "reference groups" used to gauge $\Delta_{\text{Explained}}$: while the first decomposition uses β_B, the second uses β_W. This difference has motivated many researchers to interpret Equations (1) and (2) in terms of distinct "counterfactuals." Specifically, they would regard $\Delta_{\text{Explained}}$ in Equation (1) as addressing the question: "How much would the Black completion rate rise if their parental incomes rose to those of White students?" Similarly, $\Delta_{\text{Explained}}$ in Equation (2) would be seen as capturing how much the White completion rate would drop if their parental incomes fell to those of Black students (Barsky *et al.* 2002).

However, intuitive as they may seem, such "counterfactual" interpretations of the KOB decomposition entail the implicit and potentially implausible assumption that the slope coefficients β_B and β_W capture the causal effects of the explanatory variable on the outcome (Huber 2015; Jackson and VanderWeele 2018). In our example, this assumption means that the slope coefficient in a simple linear regression of college completion on parental income among Black/White students captures the causal effect of parental income on college completion in this group. Since the relationship between parental income and college completion is likely non-linear and can be confounded by a range of other factors, such as parental education and family structure, this assumption seems unrealistic. More broadly, because the KOB decomposition is typically implemented via an ad hoc regression model that includes multiple predictors and does not explicitly address potential confounding of any of the predictor-outcome relationships, such a counterfactual interpretation cannot be justified in most applications.[3]

For this reason, we consider the KOB method a purely descriptive approach to analyzing group disparities. In other words, we regard $\Delta_{\text{Explained}}$ in Equation (1) or (2) as addressing largely the same question: "How much of the Black-White gap in completion rates can be *statistically* (rather than *causally*) explained by racial differences in parental income?" In fact, if we detach

[2]We use $\hat{\beta}'_B$ and $\hat{\beta}'_W$ to denote the transposes of the slope coefficients $\hat{\beta}_B$ and $\hat{\beta}_W$. When there is only one explanatory variable (e.g., parental income), $\hat{\beta}_B$ and $\hat{\beta}_W$ are scalars, so their tranposes are themselves. When there are multiple explanatory variables, $\hat{\beta}_B$ and $\hat{\beta}_W$ are column vectors, and the transpose operation turns them into row vectors.

[3]The KOB decomposition also does not address the causal effect of group membership itself. See Fortin *et al.* (2011) or An and Glynn (2021) for a detailed discussion.

counterfactual interpretations from the KOB decomposition, we may consider Equations (3) and (4) as merely two special cases of a more general decomposition:

$$\widehat{\Delta} = \underbrace{\hat{\beta}_R'(\bar{X}_W - \bar{X}_B)}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + (\hat{\beta}_W - \hat{\beta}_R)'\bar{X}_W + (\hat{\beta}_R - \hat{\beta}_B)'\bar{X}_B}_{\widehat{\Delta}_{\text{Unexplained}}}, \tag{5}$$

where $\hat{\beta}_R$ are called the *reference* coefficients (Jann 2008). Clearly, Equations (3) and (4) correspond to the cases of $\hat{\beta}_R = \hat{\beta}_B$ and $\hat{\beta}_R = \hat{\beta}_W$, respectively. To make the reference coefficients more "neutral," several other choices have been proposed, such as $\hat{\beta}_R = 0.5\hat{\beta}_B + 0.5\hat{\beta}_W$ (Reimers 1983) and $\hat{\beta}_R = p_B\hat{\beta}_B + p_W\hat{\beta}_W$, where p_B and p_W represent the proportions of the Black and White observations in the sample (Cotton 1988).

3.2 A Nonparametric Formulation of the KOB Method

We now present a nonparametric formulation of the KOB method (Barsky *et al.* 2002; Fortin *et al.* 2011), highlighting the distinction between population-level estimands and their estimators. Without loss of generality, we focus on a nonparametric version of Equation (1) and discuss similar generalizations of (2) and (5) in Appendix A.

First, note that Equation (1) involves only three quantities: $\bar{Y}_B$, $\bar{Y}_W$, and $\bar{Y}_B^{\text{pred}}$. The population counterparts of $\bar{Y}_B$ and $\bar{Y}_W$ are simply the conditional means of the outcome for Black and White students, i.e., $\mu_B = \mathbb{E}[Y|R = \text{Black}]$ and $\mu_W = \mathbb{E}[Y|R = \text{White}]$. Since $\bar{Y}_B^{\text{pred}}$ represents the average predicted outcome among Black students under the White parental income distribution, its population counterpart can be written as

$$\mu_B^{\text{pred}} = \int \mathbb{E}[Y|R = \text{Black}, x]dP(x|R = \text{White}), \tag{6}$$

where $P(x|R = \text{White})$ denotes the parental income distribution among White students. In words, μ_B^{pred} represents the conditional mean of college completion given parental income among Black students marginalized over the parental income distribution of White students. Hence, the explained and unexplained components of the KOB decomposition become

$$\Delta_{\text{Explained}} = \mu_B^{\text{pred}} - \mu_B, \tag{7}$$

$$\Delta_{\text{Unexplained}} = \mu_W - \mu_B^{\text{pred}}. \tag{8}$$

To identify Equation (6) nonparametrically, we need to be able to compute $\mathbb{E}[Y|R = \text{Black}, x]$ at all possible values of X among White students. Similarly, to identify a nonparametric version of Equation (2) (see Appendix A), we need to be able to compute $\mathbb{E}[Y|R = \text{White}, x]$ at all possible values of X among Black students. These conditions require that the support of X (i.e., the values of X with a nonzero density) be the same for each racial group, or, equivalently, the conditional probability $\Pr[R = \text{Black}|X = x]$ lie strictly between zero and one (Fortin *et al.* 2011). This requirement is similar to the positivity assumption typically invoked to identify the average causal effect of a binary treatment from observational data.

In the canonical KOB algorithm, if the linear model for $\mathbb{E}[Y|R = \text{Black}, x]$ is correctly specified, then $\widehat{\Delta}_{\text{Explained}}$ and $\widehat{\Delta}_{\text{Unexplained}}$ shown in Equation (3) will be consistent estimates of $\Delta_{\text{Explained}}$ and $\Delta_{\text{Unexplained}}$. However, if $\mathbb{E}[Y|R = \text{Black}, x]$ is a nonlinear function of x, the canonical KOB decomposition will target a different set of population-level quantities. Compared with a model-based definition of $\Delta_{\text{Explained}}$ and $\Delta_{\text{Unexplained}}$, the above definition (Equations 7 and 8) is fully nonparametric, i.e., not depending on any statistical model. In other words, they are still meaningful estimands even if the linear model specified in the canonical procedure is wrong. Thus, the nonparametric formulation decouples the KOB decomposition from parametric models on which it is routinely based, eliminating the model dependency of our estimands (see An and Glynn 2021 for a related discussion).

In the nonparametric decomposition, μ_B and μ_W can be estimated directly by their sample analogs ($\bar{Y}_B$ and $\bar{Y}_W$). To estimate μ_B^{pred}, several strategies can be used. First, note that Equation (6) can be written as an iterated expectation:

$$\mu_B^{\text{pred}} = \mathbb{E}\big[\mathbb{E}[Y|R = \text{Black}, X]|R = \text{White}\big],$$

which suggests a regression-imputation (RI) estimator:

$$\hat{\mu}_B^{\text{pred, RI}} = \hat{\mathbb{E}}\big[\hat{\mathbb{E}}[Y|R = \text{Black}, X]|R = \text{White}\big]. \tag{9}$$

To implement the RI estimator, we first fit a regression model for $\mathbb{E}[Y|R, X]$ and then, for all units, obtain their imputed values of the inner expectation (i.e., $\hat{\mathbb{E}}[Y|R = \text{Black}, X]$) by setting race to be Black. Finally, we estimate the outer expectation by averaging these imputed values only among White students. Here, if the inner expectation $\mathbb{E}[Y|R = \text{Black}, X]$ is estimated through a linear model within the subsample of Black students, $\hat{\mu}_B^{\text{pred, RI}}$ will coincide with $\bar{Y}_B^{\text{pred}}$ in the canonical KOB decomposition. However, the generic form of Equation (9) suggests that $\mathbb{E}[Y|R = \text{Black}, X]$ need not be estimated via a linear model; we can use a logit or probit model for binary outcomes, a Poisson or negative binomial model for count outcomes, or even a complex machine learning model if the explanatory variable X is high-dimensional.

Second, Equation (6) can also be written as the marginal mean of a reweighted outcome:

$$\mu_B^{\text{pred}} = \mathbb{E}\big[\frac{\mathbb{I}(R = \text{Black})}{\Pr[R = \text{White}]} \frac{\Pr[R = \text{White}|X]}{\Pr[R = \text{Black}|X]} Y\big],$$

where $\mathbb{I}(\cdot)$ is an indicator function. The above equation suggests a weighting estimator for μ_B^{pred}:

$$\hat{\mu}_B^{\text{pred, W}} = \frac{1}{n} \sum_i W_i Y_i, \text{ where } W_i = \frac{\mathbb{I}[R_i = \text{Black}]}{\widehat{\Pr}[R_i = \text{White}]} \frac{\widehat{\Pr}[R_i = \text{White}|X_i]}{\widehat{\Pr}[R_i = \text{Black}|X_i]}. \tag{10}$$

Here n is the sample size and $\widehat{\Pr}[R_i = \text{White}]$ is the sample proportion of White students. We can see that the weighting estimator does not involve fitting an outcome model; instead, it involves fitting a model for group membership (in this case, race) given X, which will be used to obtain $\widehat{\Pr}[R_i = \text{Black}|X_i]$ and $\widehat{\Pr}[R_i = \text{White}|X_i]$. The weighting estimator was first proposed by Barsky *et al.* (2002) for analyzing the Black-White wealth gap (see also DiNardo *et al.* 1996). These authors argue that the RI estimator of μ_B^{pred} can be highly sensitive to misspecification of the outcome model, and that the weighting estimator may be more robust because it avoids making functional-form assumptions about the X-Y relationship. However, as we can see from Equation (10), the weighting estimator relies on a correct specification of the group membership model. When this model is misspecified, it can also lead to biased estimates of μ_B^{pred}.

Finally, to mitigate potential bias due to model misspecification, we can combine RI and weight-

ing to form a "doubly robust" (DR) estimator of μ_B^{pred}:

$$\hat{\mu}_B^{\text{pred, DR}} = \frac{1}{n}\sum_{i=1}^{n}\left(W_i(Y_i - \hat{\mathbb{E}}[Y_i|\text{Black}, X_i]) + \frac{\mathbb{I}(R_i = \text{White})}{\widehat{\Pr}[R_i = \text{White}]}\hat{\mathbb{E}}[Y_i|\text{Black}, X_i]\right), \tag{11}$$

where W_i denotes the same weights used in the weighting estimator (10). This estimator is motivated by the efficient influence function (EIF) of μ_B^{pred} in the nonparametric model (see Appendix B for a derivation):

$$\text{EIF}(\mu_B^{\text{pred}}) = \frac{\mathbb{I}(R = \text{Black})}{\Pr[R = \text{White}]}\frac{\Pr[R = \text{White}|X]}{\Pr[R = \text{Black}|X]}(Y - \mathbb{E}[Y|\text{Black}, X]) + \frac{\mathbb{I}(R = \text{White})}{\Pr[R = \text{White}]}(\mathbb{E}[Y|\text{Black}, X] - \mu_B^{\text{pred}}). \tag{12}$$

The DR estimator $\hat{\mu}_B^{\text{pred, DR}}$ is constructed by solving the estimating equation (in terms of μ_B^{pred}) formed by setting the sample mean of the estimated EIF to zero. We can see from Equation (11) that the DR estimator involves fitting both an outcome model (to obtain $\hat{\mathbb{E}}[Y_i|\text{Black}, X_i]$ and $\hat{\mu}_B^{\text{pred, RI}}$) and a group membership model (to obtain W_i). This estimator is doubly robust in the sense that it is consistent if either the outcome model or the group membership model, but not necessarily both, is correctly specified (see Appendix C for a proof of the double-robustness of $\hat{\mu}_B^{\text{pred, DR}}$). The DR estimator (11) is closely related to DR estimators developed for the natural direct effect (NDE) and the average treatment effect on the treated (ATT) in causal inference settings (Tchetgen Tchetgen and Shpitser 2012; Chernozhukov *et al.* 2018). Yet, to the best of our knowledge, it has not been proposed in the context of KOB decompositions.[4]

3.3 The Case of Multiple Explanatory Variables: A Simultaneous Approach

So far, we have focused on the case of one explanatory variable X. Note that X can be vector-valued, in which case the decomposition given by Equations (7-8) does not distinguish the contributions of its individual components. For example, if different components of X are indicators of the same theoretical construct (e.g., socioeconomic status captured by both parental education and parental income), we may want to treat them as a whole. However, researchers often have multiple

[4]Kline (2011) shows that the canonical KOB estimator based on a linear model for $\mathbb{E}[Y|R = \text{Black}, X]$ is also doubly robust in the sense that it is consistent for μ_B^{pred} if either the linear model for $\mathbb{E}[Y|R = \text{Black}, X]$ is correctly specified or if $\frac{\Pr[R=\text{White}|X]}{\Pr[R=\text{Black}|X]}$ is a linear function of X. This condition for double robustness is much more stringent than that for the DR estimator (11).

explanatory variables capturing distinct concepts. For instance, to account for the Black-White completion gap, we may want to isolate the respective contributions of parental income, family structure, academic preparation, and college selectivity. In this and the next subsections, we discuss two alternative strategies for teasing out covariate-specific contributions: a *simultaneous* approach and a *sequential* approach.

To ease exposition, let X denote the column vector $X = (X_1, X_2, \dots X_J)'$, where X_j is the jth (possibly vector-valued) explanatory variable and J denotes the total number of explanatory variables. In the canonical KOB decomposition, the conditional expectation function $\mathbb{E}[Y|R = r, X]$ is estimated via a race-specific linear model with no interaction terms among the X_js. In this case, both the explained and unexplained components can be subdivided into covariate-specific contributions. This can be seen by rewriting Equation (3) as (Blinder 1973; Oaxaca 1973)

$$\widehat{\Delta} = \underbrace{\sum_{j=1}^{J} \hat{\beta}_{B,j}(\bar{X}_{W,j} - \bar{X}_{B,j})}_{\widehat{\Delta}_{\text{Explained}}} + \underbrace{(\hat{\alpha}_W - \hat{\alpha}_B) + \sum_{j=1}^{J} (\hat{\beta}_{W,j} - \hat{\beta}_{B,j})\bar{X}_{W,j}}_{\widehat{\Delta}_{\text{Unexplained}}}. \tag{13}$$

where $\hat{\beta}_{B,j}$ and $\hat{\beta}_{W,j}$ denote the slope coefficients for Black and White students, respectively, and $\bar{X}_{B,j}$ and $\bar{X}_{W,j}$ denote the corresponding sample means of the jth covariate. In this decomposition, $\widehat{\Delta}_{\text{Explained}}$ consists of J components, each representing the contribution of one covariate. $\widehat{\Delta}_{\text{Unexplained}}$, on the other hand, consists of $J+1$ components, one capturing the baseline gap when all the explanatory variables equal zero and the other J components capturing racial differences in the slope coefficients, i.e., the partial effects of X_j on Y. In the KOB literature, Equation (13) is called a *detailed decomposition*, and the detailed components of $\widehat{\Delta}_{\text{Explained}}$ and $\widehat{\Delta}_{\text{Unexplained}}$ are known as the endowment effects and coefficient effects, respectively (e.g., Jann 2008). Because the detailed decomposition quantifies the endowment and coefficient effects of each covariate concurrently, it constitutes a *simultaneous approach* to tackling multiple explanatory variables (see Gelbach 2016 for an alternative formulation of Equation 13 in matrix notation).

While computationally straightforward, the detailed decomposition has several limitations. First, it is not model-free, as its estimands (the endowment and coefficient effects) hinge on the linear and additive specification of the outcome model. In fact, when the outcome model is either nonlinear (e.g., a logit model) or non-additive (e.g., including interaction terms among the

covariates), the overall explained and unexplained components (which can still be estimated via the RI method) cannot be neatly partitioned into covariate-specific contributions in the manner of Equation (13) (see Rothe 2015 for a more involved decomposition that contains both the "direct" contributions of each covariate and the contributions of the covariates' interaction effects and dependence structure). Second, it suffers from the so-called "omitted category" problem (Jones 1983; Oaxaca and Ransom 1999), which refers to the fact that when an explanatory variable, say X_j, is categorical and thus coded as a set of dummy variables, the coefficient effects that compose $\widehat{\Delta}_{\text{Unexplained}}$ in Equation (13) will depend on the choice of the omitted (i.e., reference) category for X_j. This issue arises not only for categorical covariates but also for continuous covariates that do not have an intrinsic baseline value. For example, if a continuous covariate is recentered around its mean before being used to fit the outcome models, the corresponding coefficient effect will differ. As Fortin *et al.* (2011, p. 9) note, "[s]ince the choice of the omitted group is arbitrary, the elements of the detailed decomposition can be viewed as arbitrary as well." In other words, the coefficient effects of the detailed decomposition are not only model-dependent but also indeterminate even under the same model.

Finally, in the detailed decomposition, even the covariate-specific endowment effects — which, unlike the coefficient effects, do not suffer from the omitted group problem — can be hard to interpret, and they may deviate from the researcher's theoretical interests (Lundberg *et al.* 2021). This is especially the case if the covariates are temporally or causally ordered. For example, to explain the Black-White completion gap, we can reasonably expect parental income to be causally prior to academic preparation, and academic preparation causally prior to college selectivity. In this case, the coefficient of parental income in the outcome model reflects its partial effect on college completion net of academic preparation and college selectivity. Given that differences in academic preparation and college selectivity are likely induced by parental income, the contribution of parental income in the detailed decomposition cannot fully capture its influence on the Black-White completion gap. At most, it can be interpreted as a "direct contribution" of parental income, net of its indirect pathways via academic preparation and college selectivity. Similarly, the contribution of academic preparation in the detailed decomposition can at most be interpreted as its "direct contribution" net of its indirect pathway via college selectivity. However, in such cases, the researcher may not be as interested in such "direct contributions" as in the cumulative contributions

of the upstream variables. For example, we may want to assess the overall contribution of Black-White differences in parental income to the Black-White completion gap, whether it stems from the direct influence of parental income during college or from its indirect influence via academic preparation or college selectivity. Unfortunately, this quantity is not targeted by the detailed KOB decomposition. To capture estimands of this type, we need to quantify the contributions of parental income, academic preparation, and college selectivity sequentially rather than simultaneously — a strategy to which we now turn.

3.4 A Generalized KOB/Duncan Decomposition

In an article titled "Inheritance of Poverty or Inheritance of Race?", Duncan (1968) attempted to assess whether the persistence of racial economic inequality across generations in the US is driven by "class" versus "race" effects. Using linear regression models, Duncan conducted a sequence of KOB decompositions for the Black-White income gap, adding one factor at a time: (a) class origin as measured by parental education and occupation; (b) number of siblings; (c) respondent's education; and (d) respondent's occupation. He then quantified the net contribution of each factor by comparing the explained components of every two adjacent models. For example, the difference in $\Delta_{\text{Explained}}$ between the first and second models reflects the net contribution of racial differences in the average number of siblings. Although proposed prior to Blinder (1973) and Oaxaca (1973), Duncan's decomposition can be viewed as a generalization of the KOB decomposition. In what follows, we provide a fully nonparametric formulation of Duncan's sequential approach to decomposing group disparities, which we call a generalized KOB/Duncan decomposition.

Following our earlier notation, let $X_1, X_2, \ldots X_J$ denote J ordered sets of covariates that may help explain a disparity of interest. In our running example, we may consider five ordered sets of covariates: X_1 for demographic and socioeconomic background, X_2 for family structure, X_3 for pre-college measures of academic performance and behavior, X_4 for contextual factors such as peer and school characteristics, X_5 for college selectivity. It might be reasoned that these sets of covariates are causally ordered, although this is not required for purely descriptive interpretations of the sequential decomposition, which we detail below.

We use calligraphic font to denote a "cumulative" set of covariates such that $\mathcal{X}_0 = \varnothing$ and

$\mathcal{X}_j = \{X_1, X_2, \ldots X_j\}$. Define the following iterated expectations:

$$\mu_{B,j}^{\text{pred}} = \mathbb{E}[\mathbb{E}[Y|R = \text{Black}, \mathcal{X}_j]|R = \text{White}], \quad 0 \leq j \leq J. \tag{14}$$

In words, $\mu_{B,j}^{\text{pred}}$ represents the conditional mean of college completion given the covariate set $\mathcal{X}_j = \{X_1, X_2, \ldots X_j\}$ among Black students marginalized over the distribution of these covariates among White students. Then, the Black-White gap in college completion rates can be decomposed as

$$\Delta = \mu_W - \mu_B = \sum_{j=1}^{J} \underbrace{(\mu_{B,j}^{\text{pred}} - \mu_{B,j-1}^{\text{pred}})}_{\Delta_{\text{Explained},X_j}} + \underbrace{\mu_W - \mu_{B,J}^{\text{pred}}}_{\Delta_{\text{Unexplained}}}. \tag{15}$$

Note that since $\mathcal{X}_0 = \varnothing$, $\mu_{B,0}^{\text{pred}} = \mathbb{E}[\mathbb{E}[Y|R = \text{Black}]|R = \text{White}] = \mathbb{E}[Y|R = \text{Black}] = \mu_B$. Here, $\Delta_{\text{Explained},X_j}$ reflects the net contribution of X_j to the completion gap, i.e., the amount of the Black-White completion gap that can be statistically explained by racial differences in X_j after racial differences in $X_1, X_2, \ldots X_{j-1}$ are accounted for. $\Delta_{\text{Unexplained}}$, on the other hand, reflects the amount of the Black-White completion gap that cannot be statistically explained by racial differences in any of the explanatory variables.

To implement the generalized KOB/Duncan decomposition, we need only estimates of μ_W, μ_B, and $\mu_{B,j}^{\text{pred}}$ for each j, which can then be plugged into Equation (15) to obtain estimates of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$. As in the canonical KOB method, μ_W and μ_B can be directly estimated by their sample analogs. Thus, the key is to estimate $\mu_{B,j}^{\text{pred}}$. Similar to μ_B^{pred} discussed in Section 3.2, $\mu_{B,j}^{\text{pred}}$ can be estimated via RI, weighting, or a doubly robust estimator that combines RI and weighting. First, an RI estimator of $\mu_{B,j}^{\text{pred}}$ can be written as

$$\hat{\mu}_{B,j}^{\text{pred, RI}} = \hat{\mathbb{E}}\big[\hat{\mathbb{E}}[Y|R = \text{Black}, \mathcal{X}_j]|R = \text{White}\big]. \tag{16}$$

Specifically, we first fit a regression model for $\mathbb{E}[Y|R, \mathcal{X}_j]$, and, for all units, obtain their imputed values of the inner expectation (i.e., $\hat{\mathbb{E}}[Y|R = \text{Black}, \mathcal{X}_j]$) by setting race to be Black. Then, we estimate the outer expectation by averaging these imputed values only among White students. By iterating the above steps for all j ($1 \leq j \leq J$), we obtain all RI estimates of $\mu_{B,j}^{\text{pred}}$, which can then be substituted into Equation (15). Note that if all of the inner expectations $\mathbb{E}[Y|R = \text{Black}, \mathcal{X}_j]$

$(1 \leq j \leq J)$ are estimated through a linear model within the subsample of Black students, this procedure will reproduce Duncan's decomposition. In this case, our estimate of $\Delta_{\text{Unexplained}}$ (i.e., $\mu_W - \mu^{\text{pred}}_{B,J}$) will also coincide with the unexplained component from the detailed decomposition (Equation 13).

Second, a weighting estimator of $\mu^{\text{pred}}_{B,j}$ can be written as

$$\hat{\mu}^{\text{pred, W}}_{B,j} = \frac{1}{n}\sum_i W_{ji} Y_i, \quad \text{where } W_{ji} = \frac{\mathbb{I}[R_i = \text{Black}]}{\widehat{\Pr}[R_i = \text{White}]} \frac{\widehat{\Pr}[R_i = \text{White}|\mathcal{X}_{ji}]}{\widehat{\Pr}[R_i = \text{Black}|\mathcal{X}_{ji}]}, \tag{17}$$

where n is the sample size and $\mathcal{X}_{ji}$ denotes the observed covariate values of $\mathcal{X}_j = \{X_1, X_2, \ldots X_j\}$ for individual i. This weighting estimator is similar to the weighting estimator for μ^{pred}_B described in Section 3.2, except that the group membership model is now fit as a function of the covariate set $\mathcal{X}_j$ and that this procedure is repeated for each j.

Both the RI and weighting estimators involve fitting J different models — an outcome model for each $\mathbb{E}[Y|R, \mathcal{X}_j]$ or a group membership model for each $\Pr[R = \text{Black}|\mathcal{X}_j]$. When any of the J models is misspecified, the corresponding estimator of the explained and unexplained components can be biased. From this perspective, both the RI and weighting estimators are prone to model misspecification bias. To mitigate potential bias, we can combine RI and weighting to form a doubly robust estimator of $\mu^{\text{pred}}_{B,j}$:

$$\hat{\mu}^{\text{pred, DR}}_{B,j} = \frac{1}{n}\sum_{i=1}^{n} \big(W_{ji}(Y_i - \hat{\mathbb{E}}[Y_i|R_i = \text{Black}, \mathcal{X}_{ji}]) + \frac{\mathbb{I}(R_i = \text{White})}{\widehat{\Pr}[R_i = \text{White}]} \hat{\mathbb{E}}[Y_i|R_i = \text{Black}, \mathcal{X}_{ji}]\big), \tag{18}$$

where W_{ji} denotes the same weights used in the weighting estimator (17). Similar to Equation (11), $\hat{\mu}^{\text{pred, DR}}_{B,j}$ is derived from the EIF of $\mu^{\text{pred}}_{B,j}$, which is similar to Equation (12) except that X_i is replaced by $\mathcal{X}_{ji}$. We can see that the DR estimator involves fitting both an outcome model for $\mathbb{E}[Y|R, \mathcal{X}_j]$ and a group membership model for $\Pr[R = \text{Black}|\mathcal{X}_j]$. It is doubly robust in the sense that it is consistent if either the outcome model or the group membership model, but not necessarily both, is correctly specified. Moreover, this double robustness of $\hat{\mu}^{\text{pred, DR}}_{B,j}$ holds for each $j \in \{1, 2, \ldots J\}$. Thus, if we use the DR estimator for all of the $\mu^{\text{pred}}_{B,j}$s, the resulting estimators of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$ will be consistent if for each j, either the outcome model or the group membership model is correctly specified. In this regard, the doubly robust method for

the generalized KOB/Duncan decomposition is 2^J-robust, as the number of conditions for the resulting estimates of the explained and unexplained components to be consistent is 2^J (Zhou 2022). Thus, it offers much greater protection from model misspecification bias than the RI and weighting estimators.

Another advantage of the doubly robust estimator $\hat{\mu}_{B,j}^{\text{pred, DR}}$ is its compatibility with the use of machine learning methods for estimating the outcome and group membership models. In this case, our estimates of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$ will be not only robust to model misspecification but also semiparametrically efficient provided that certain mild convergence rate conditions are satisfied (see Zhou [2022], Supplementary Material E). When machine learning methods are used to estimate the outcome and group membership models, it is advisable to use sample splitting to avoid potential bias due to over-fitting. In particular, Chernozhukov *et al.* (2018) introduce a procedure called "cross-fitting," which involves alternate use of different subsamples to estimate nuisance functions[5] and target parameters. In our context, cross-fitting involves the following steps: (a) randomly partition the sample S into K folds: $S_1, S_2 \ldots S_K$; (b) for each k, obtain a fold-specific estimate of $\hat{\mu}_{B,j}^{\text{pred, DR}}$ using only data from S_k ("main sample"), but with the outcome and group membership models learned from the remainder of the sample (i.e., $S \backslash S_k$; "auxiliary sample"); (c) average these fold-specific estimates to form final estimates of $\mu_{B,j}^{\text{pred}}$. This combination of $\hat{\mu}_{B,j}^{\text{pred, DR}}$, machine learning, and cross-fitting is an instance of what Chernozhukov *et al.* (2018) call "double/debiased machine learning" (DML).

For statistical inference, the nonparametric bootstrap can be used for RI, weighting, and DR estimators alike, provided that the outcome and group membership models are both estimated parametrically, e.g., via generalized linear models. When the outcome and group membership models are estimated through machine learning, however, the resulting estimator is no longer a Hadamard differentiable function of the data, making the nonparametric bootstrap not theoretically justified (Wasserman 2006; van der Laan and Rose 2018). Fortunately, when DML is used, we can construct simple analytic standard errors from the estimated EIFs. Specifically, due to the semiparametric efficiency of $\hat{\mu}_{B,j}^{\text{pred, DR}}$ (under cross-fitting and the mild convergence rate conditions

[5] A nuisance function is a function that is not of our primary interest but needed to estimate our target parameters. For the doubly robust estimator (18), the nuisance functions are the outcome and group membership models, i.e., $\mathbb{E}[Y|R, \mathcal{X}_j]$ and $\Pr[R = \text{Black}|\mathcal{X}_j]$.

discussed in Zhou [2022]), its sampling variance can be estimated as $\frac{1}{n}\widehat{\text{Var}}(\widehat{\text{EIF}}(\mu^{\text{pred}}_{B,j}))$. The sampling variances of the corresponding estimators of $\Delta_{\text{Explained},X_j}$ and $\Delta_{\text{Unexplained}}$ can be estimated analogously. For example, the sampling variance of the DML estimator of $\Delta_{\text{Explained},X_j}$ can be estimated as $\frac{1}{n}\widehat{\text{Var}}\big(\widehat{\text{EIF}}(\mu^{\text{pred}}_{B,j}) - \widehat{\text{EIF}}(\mu^{\text{pred}}_{B,j-1})\big)$. The square roots of these expressions yield standard error estimates, which can then be used to construct confidence intervals and compute p-values.

When parametric methods are used to estimate the outcome and group membership models in $\hat{\mu}^{\text{pred, DR}}_{B,j}$, the EIF-based sampling variance estimator $\frac{1}{n}\widehat{\text{Var}}(\widehat{\text{EIF}}(\mu^{\text{pred}}_{B,j}))$ will still be valid if both the outcome and group membership models are correctly specified. However, if one of these models is misspecified, this variance estimator will be inconsistent, even though in this case $\hat{\mu}^{\text{pred, DR}}_{B,j}$ itself is still consistent. In other words, the variance estimator $\frac{1}{n}\widehat{\text{Var}}(\widehat{\text{EIF}}(\mu^{\text{pred}}_{B,j}))$ is not doubly robust. Thus, in practice, we recommend that the EIF-based variance estimator be used only for the DML estimator of $\hat{\mu}^{\text{pred, DR}}_{B,j}$, and that the nonparametric bootstrap be used for the RI, IPW, and DR estimators implemented through parametric models.

3.5 Illustration

We now illustrate the generalized KOB/Duncan decomposition via a reanalysis of the Black-White gap in college completion rates in the US. We use data from a sample of Black and White respondents who had attended a four-year college by age 25 in the NLSY97. Our outcome (Y) is whether the respondent had completed a BA degree by age 29. We consider five sets of explanatory variables: X_1 for demographic and socioeconomic background (gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence), X_2 for family structure (co-residence with both biological parents, presence of a paternal figure), X_3 for pre-college measures of academic performance and behavior (percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18), X_4 for peer and school-level characteristics (college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school), and X_5 for college selectivity (whether the student had attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000 by age 25).

In what follows, we focus on DML estimates of $\mu^{\text{pred}}_{B,j}$, $\Delta_{\text{Explained},X_j}$, and $\Delta_{\text{Unexplained}}$ based on

Equation (18), cross-fitting, and machine learning estimates of the outcome and group membership models.[6] Specifically, we estimate each of the outcome and group membership models using a super learner (van der Laan *et al.* 2007) composed of Lasso and random forests. Because random forests allow for nonlinear and interaction effects, potential bias due to model misspecification is minimized. In keeping with Chernozhukov *et al.* (2018), we use five-fold cross-fitting, meaning that $K = 5$. Moreover, following Cole and Hernán (2008), we stabilize our estimates by censoring all inverse-probability-weights (W_{ji} in Equation 18) at their 1st and 99th percentiles. Standard errors are constructed using the sample variances of the estimated influence functions and adjusted for multiple imputation via Rubin's (1987) method.

In Figure 1, the first and last rows show the observed completion rates for Black and White students in our sample, which are 0.51 and 0.69, respectively. The intermediate rows report our DML estimates of the "adjusted" completion rates for Black students, i.e., $\mu_{B,j}^{\text{pred}}$ for $j = 1, 2, \ldots 5$. We find that the first three sets of the explanatory variables, i.e., demographic and socioeconomic background, family structure, and pre-college academic performance and behavior, account for over 90% of the Black-White gap in college completion. Of them, the first (demographic and socioeconomic background) and the third (academic performance and behavior) sets appear to have the largest explanatory power.

[Figure 1 here]

How would our findings differ if we were to implement the simultaneous KOB decomposition, as discussed in Section 3.3, rather than our proposed sequential approach? To address this question, Figure 2 contrasts estimates obtained under these two approaches. Note that, for the generalized KOB/Duncan decomposition, we now report the specific contributions of each set of covariates (versus Figure 1, which reports the "adjusted" completion rates for each set of covariates).[7] Compared with the simultaneous approach, the generalized KOB/Duncan decomposition yields a larger

[6] We also obtained estimates of $\mu_{B,j}^{\text{pred}}$, $\Delta_{\text{Explained},X_j}$, and $\Delta_{\text{Unexplained}}$ from the RI, weighting, and doubly robust estimators (Equations 16-18) where the outcome and group membership models were both fit via a logistic regression with linear and additive effects of the explanatory variables. The results are qualitatively similar and reported in Appendix G.

[7] To improve the efficiency of our estimates from the simultaneous KOB decomposition, we use slope coefficients obtained from a pooled linear regression model that includes the group indicator as a predictor.

estimate of the contribution of demographic and socioeconomic background, and smaller estimates of the contributions of family structure, academic performance and behavior, contextual factors, and college selectivity. This discrepancy occurs because, for upstream variables, the generalized KOB/Duncan decomposition captures their full influence, while the simultaneous decomposition captures only their direct influence net of the downstream variables. These results demonstrate that the widely used simultaneous KOB decomposition can produce misleading results if the researcher is concerned with the temporal/causal ordering of the explanatory variables in question. In fact, while the simultaneous decomposition suggests that college selectivity explains a non-trivial portion of racial inequality, results from the generalized KOB/Duncan decomposition imply that college selectivity has no additional explanatory power when the upstream variables are accounted for. Finally, we note that the sequential approach and the simultaneous approach yield similar estimates of the unexplained component (both are close to zero). This is expected as the unexplained component has the same interpretation under both approaches — the portion of the Black-White gap in college completion not attributable to racial differences in any of the explanatory variables. The point estimates are not identical, though, because the sequential approach is implemented through DML rather than linear models as in the simultaneous approach (which does not have a nonparametric counterpart).

[Figure 2 here]

4 Disparity Analysis for Prescription

As noted previously, unless the effects of the explanatory variables on the outcome are unconfounded, estimates from the KOB method and its variants have no causal interpretations. In this regard, any policy implications drawn from such analyses could be misguided. For example, in our empirical illustration, we found that, after accounting for other background characteristics, college selectivity does not (statistically) contribute to the Black-White gap in college completion. Given this finding, it might be supposed that the Black-White completion gap cannot be reduced by interventions on college selectivity. This is not necessarily the case, however, as the above finding could arise if either (a) conditional on other background characteristics, Black and White students do not differ in college selectivity, or (b) conditional on race and other background characteristics, there

is no association between college selectivity and college completion. In the first scenario, college selectivity could still have a causal effect on degree completion, in which case an intervention that induces more Black students to attend selective colleges should work toward reducing the gap.

To avoid such misinterpretations of descriptive analyses and to quantify the impacts of different types of hypothetical interventions, we now turn to a more policy-relevant perspective to disparity analysis, which we call the *prescriptive* approach. This emerging research agenda has gained increasing attention in epidemiology and sociology (VanderWeele and Robinson 2014; Jackson and VanderWeele 2018; Jackson 2021; Young *et al.* 2014; Lundberg 2022; Zhou and Pan 2023; Kennedy 2019). In what follows, we first outline several key elements of this approach, including the concept of interventional disparity, its identification assumptions, and several estimation strategies, drawing heavily from Lundberg (2022). We then introduce an original bias formula for assessing the sensitivity of our estimated interventional disparities to unobserved confounding. Next, we introduce some stylized interventions in the context of the Black-White completion gap and illustrate them with the NLSY97 data. Finally, we connect the prescriptive approach to the KOB method by clarifying the conditions under which counterfactual disparities for certain stylized interventions can be evaluated through a simultaneous or sequential KOB decomposition (Jackson 2018; Jackson and VanderWeele 2018).

4.1 Interventional Disparities

In the prescriptive approach, we specify one or more variables as manipulable treatments and ask how the disparity of interest would change under a hypothetical intervention to these treatments. Without loss of generality, consider a binary point-in-time treatment, denoted by A. In our running example, A denotes whether a student attends a selective college. Following Young *et al.* (2014), let A^g and Y^g denote the treatment and outcome variables, respectively, under a hypothetical intervention g. We define *interventional disparity* (Δ^g) as the difference in means between Black and White students in their interventional outcome Y^g (Lundberg 2022):

$$\Delta^g = \mathbb{E}[Y^g|R = \text{White}] - \mathbb{E}[Y^g|R = \text{Black}]. \tag{19}$$

We can similarly define the *gap reduced* as the amount of disparity that would be removed under intervention g: $\Delta - \Delta^g$.

Since Y^g denotes the outcome under a hypothetical intervention, the interventional disparity Δ^g cannot be directly calculated from observational data. In fact, provided that the hypothetical intervention g would change treatment status for some units, Y^g would involve potential outcomes under treatment conditions that are not realized under the status quo. For example, for students who did not attend a selective college but would have had a nonzero chance of attending a selective college under an intervention of interest, their interventional outcomes would involve a counterfactual world in which they attended a selective college, which is not observed.

The above definition of interventional disparities is generic. It does not presuppose the nature of the intervention. In special cases where the intervention fixes treatment status at a prespecified value for all units, Δ^g can be interpreted as a "controlled disparity," akin to the concept of controlled direct effects in causal mediation analysis (Naimi *et al.* 2014a; Howe *et al.* 2018). More generally, the intervention can be stochastic, where the treatment value for each unit is drawn from a probability distribution that may or may not depend on their pretreatment characteristics (Naimi *et al.* 2014b; Jackson and VanderWeele 2018; Nguyen *et al.* 2020; Lundberg 2022). Below we outline a set of assumptions under which an interventional disparity can be identified from observational data.

4.2 Identification and Estimation

Let X denote a set of pretreatment covariates that may confound the treatment-outcome relationship. In our example, X may include the first four sets of explanatory variables described in Section 3.5 (demographic and socioeconomic background, family structure, pre-college academic performance and behavior, and peer and school-level characteristics). In addition, let $Y(a)$ denote the potential outcome associated with treatment status a. An interventional disparity Δ^g can be identified under the following assumptions:

Assumption 1. *Consistency:* $Y = Y(A)$ *and* $Y^g = Y(A^g)$.

Assumption 2. *Unconfoundedness: For any* $a \in \{0, 1\}$, $Y(a) \perp\!\!\!\perp A|X, R$ *and* $Y(a) \perp\!\!\!\perp A^g|X, R$.

Assumption 3. *Positivity:* $0 < \Pr[A = 1|X = x, R = r] < 1$ *for all* x *and* r.

Assumption 1 (*consistency*) states that a unit's observed outcome equals its potential outcome associated with the observed treatment status under both the status quo and the intervention of interest. The first part of the consistency assumption ($Y = Y(A)$) is commonly known as the stable unit treatment value assumption (SUTVA; Rubin 1986) in the causal inference literature. It implies that for any unit and any $a \in \{0, 1\}$, $Y(a)$ is fixed, thus ruling out the possibility that a treatment has multiple versions, each corresponding to a different potential outcome. It also implies that the potential outcomes for any unit do not depend on treatment status of other units. Thus, there should be no interference between units. The second part of the consistency assumption ($Y^g = Y(A^g)$) additionally requires that the potential outcomes $Y(0)$ and $Y(1)$ be unaffected by the intervention g. In other words, the intervention is not allowed to change a unit's outcome other than through changing its treatment status. This assumption would be violated if the intervention affects a unit's potential outcome either "directly" (i.e., via pathways other than changing A) or via other units' treatment status (i.e., interference). For example, when analyzing the interventional effect associated with an expansion in selective college attendance, Assumption 1 might be violated if an increasing prevalence of selective college attendance leads to a dilution of resources that each student receives, which may in turn lower the effect of selective college attendance on degree completion. Assumption 2 (*unconfoundedness*) means that among units with the same race and covariate values, treatment assignment is as-if random, i.e., independent of potential outcomes, under both the status quo and the intervention. Finally, Assumption 3 (*positivity*) requires that treatment assignment under the status quo is probabilistic at all possible values of race and the covariates.

From Equation (19), we can see that to estimate the interventional disparity Δ^g, it suffices to identify and estimate the mean interventional outcome for each group r and then take the difference in our estimates between Blacks versus Whites. To simplify notation, let μ_r^g denote the mean interventional outcome for group r such that $\mu_B^g = \mathbb{E}[Y^g|R = \text{Black}]$ and $\mu_W^g = \mathbb{E}[Y^g|R = \text{White}]$. Under Assumptions 1-3, μ_r^g can be identified as

$$\mu_r^g = \mathbb{E}_{X|R=r}\mathbb{E}_{A^g|R=r,X}\mathbb{E}[Y|R = r, X, A = A^g] \tag{20}$$

$$= \mathbb{E}[\frac{p^g(A|R, X)}{p(A|R, X)}Y|R = r], \tag{21}$$

where $p^g(\cdot|r,x)$ denotes the probability mass function (PMF) of the interventional treatment A^g given race and the covariates. Since this PMF is the defining characteristic of an intervention, we assume that it is known (i.e., not an unknown quantity that needs to be estimated from data). A proof of Equations (20-21) is given in Appendix D.

Given the above identification result, Lundberg (2022) shows that μ_r^g can be evaluated through regression-imputation, weighting, and a doubly robust estimator. First, Equation (20) suggests a regression-imputation (RI) estimator:

$$\hat{\mu}_r^{g,\text{RI}} = \frac{1}{n_r} \sum_{i:R_i=r} \underbrace{\sum_{a=0,1} p^g(a|R_i,X_i)\hat{\mathbb{E}}[Y_i|R_i,X_i,A_i=a]}_{\text{predicted outcome for unit } i \text{ under the intervention}}, \tag{22}$$

where n_r denotes the number of individuals in racial group r in the sample. To implement the RI estimator, we first fit a model for the conditional mean of the outcome given race, the covariates, and the treatment. Then, for each unit, we obtain two predicted outcomes, one under treatment and one under control, and compute a weighted average of them where the weights are the conditional probabilities of treatment and control given race and the covariates under the intervention. This weighted average serves as a predicted outcome for this unit under the intervention. Finally, we average these predicted outcomes for each racial group r. In our example, the difference between $\hat{\mu}_B^{g,\text{RI}}$ and $\hat{\mu}_W^{g,\text{RI}}$ constitutes an RI estimator of the interventional disparity Δ^g.

By contrast, Equation (21) suggests a weighting estimator of μ_r^g:

$$\hat{\mu}_r^{g,\text{W}} = \frac{1}{n_r} \sum_{i:R_i=r} W_i Y_i, \quad \text{where } W_i = \frac{p^g(A_i|R_i,X_i)}{\hat{p}(A_i|R_i,X_i)}. \tag{23}$$

Here, the weight W_i is the ratio of the probability of receiving the observed treatment under the intervention ($p^g(A_i|R_i,X_i)$) to the estimated probability of receiving the observed treatment under the status quo ($\hat{p}(A_i|R_i,X_i)$). The weighting estimator can be viewed as a generalization of the inverse-probability-weighted (IPW) estimator for the mean of a potential outcome, say $\mathbb{E}[Y(a)]$, where the numerator is an indicator function for whether the observed treatment A equals a (see Hernan and Robins 2023 for an introduction to IPW estimators). To understand this connection, note that under Assumptions 1-3, $\mathbb{E}[Y(a)]$ is simply the mean outcome under a deterministic

intervention that sets everyone's treatment status to a. Also, note that the weighting estimator (23) is quite different from the weighting estimator (10) for the KOB decomposition, where we need a group membership model rather than a treatment model.

RI and weighting can be combined to form a doubly robust estimator of μ_r^g:

$$\hat{\mu}_r^{g,\mathrm{DR}} = \frac{1}{n_r} \sum_{i:R_i=r} \big\{ \hat{Y}_i^g + W_i(Y_i - \hat{\mathbb{E}}[Y|R_i, X_i, A_i]) \big\}, \tag{24}$$

where W_i denotes the same weights used in the weighting estimator (23) and $\hat{Y}_i^g = \sum_{a=0,1} p^g(a|R = r, X_i)\hat{\mathbb{E}}[Y|R_i, X_i, A = a]$ denotes the individual-level predicted outcome under the intervention as used in the RI estimator (22). Similar to the doubly robust estimators of μ_B^{pred} in the context of the KOB decomposition, this estimator is motivated by the efficient influence function (EIF) of μ_r^g in the nonparametric model (see Appendix E for a derivation):

$$\mathrm{EIF}(\mu_r^g) = \frac{\mathbb{I}(R=r)}{\Pr[R=r]} \Big(\frac{p^g(A|R,X)}{p(A|R,X)} \big(Y - \mathbb{E}[Y|R,X,A]\big) + \sum_a \mathbb{E}[Y|R,X,a] p^g(a|R,X) - \mu_r^g \Big). \tag{25}$$

The DR estimator (24) is constructed by solving the estimating equation (in terms of μ_r^g) formed by setting the sample mean of the estimated EIF to zero.

Assuming that the interventional distribution $p^g(\cdot|r,x)$ is known, this estimator is doubly robust — it is consistent if either the outcome model or the treatment model is correctly specified (see Lundberg 2022 for a proof of the double-robustness). This assumption requires some qualification. On the one hand, we rarely have a fully specified intervention a priori, and often consider a stylized intervention where $A^g|R, X$ is a function of the observed distribution of (R, X, A). To give an example, consider an intervention that sets the treatment probability for Black students to be the same as that for Whites. The latter is still a population quantity, and therefore needs to be estimated from sample data. In this sense, the interventional distribution $p^g(\cdot|r,x)$ is also unknown, albeit estimable. On the other hand, real-world policies are rarely calibrated using data from the true population distribution, meaning that the actual interventional distributions being specified and implemented are also approximations of their "population counterparts." Thus, we consider it reasonable to use our sample data to formulate an intervention, and, then, by treating it as fixed, evaluate its impact on our disparity of interest.

As with the doubly robust estimator of μ_B^{pred} for the KOB decomposition, $\hat{\mu}_r^{g,\text{DR}}$ is also congenial to the use of DML, meaning that the outcome and treatment models can be estimated through flexible machine learning methods. Provided that cross-fitting is used and that appropriate convergence rate conditions are satisfied for the machine learning estimates of the outcome and treatment models, $\hat{\mu}_r^{g,\text{DR}}$ will be $\sqrt{n}$-consistent, asymptotically normal, and semiparametric efficient. In this case, its sampling variance can be estimated as $\frac{1}{n}\widehat{\text{Var}}(\widehat{\text{EIF}}(\mu_r^g))$, and the sampling variance of the corresponding estimator of Δ^g can be estimated as $\frac{1}{n}\widehat{\text{Var}}\big(\widehat{\text{EIF}}(\mu_B^g) - \widehat{\text{EIF}}(\mu_W^g)\big)$. The square roots of these expressions lead to standard error estimates, which can then be used to construct confidence intervals and compute p-values. Otherwise, if parametric procedures are used to fit the outcome and treatment models, the nonparametric bootstrap will be a more prudent approach to inference.

4.3 Sensitivity Analysis

In practice, due to either data limitations or measurement error, it is unlikely that a researcher is able to capture all confounders of the treatment-outcome relationship to make Assumption 2 (unconfoundedness) hold. If there exists an unobserved confounder that affects both the treatment and the outcome, then the estimated interventional disparity may be biased. Sensitivity analysis is a popular approach to assessing potential bias induced by unobserved confounding. Below we introduce a simple formula for sensitivity analysis that links the bias in the estimated interventional disparity (Δ^g) to the associations between the unobserved confounder, the treatment, and the outcome.

Let U denote an unobserved confounder, such as extracurricular excellence, that affects both selective college attendance and BA completion. Suppose that the following simplifying assumptions hold: (i) U is binary; (ii) the difference in the prevalence of U between selective- and non-selective college goers given race and the covariates, i.e., $\Pr[U = 1|x, r, A = 1] - \Pr[U = 1|x, r, A = 0]$, does not depend on the covariates; (iii) the difference in BA completion probability between students with and without U given race, the covariates, and treatment status, i.e., $\mathbb{E}[Y|r, x, a, U = 1] - \mathbb{E}[Y|r, x, a, U = 0]$, does not depend the covariates or treatment status. Under these assumptions, the bias for the estimated interventional mean ($\mathbb{E}[Y^g|R = r]$) can be written as (see Appendix F)

$$\alpha_r \beta_r (\Pr[A^g = 1|R = r] - \Pr[A = 1|R = r]), \tag{26}$$

where $\alpha_r = \Pr[U = 1|x, r, A = 1] - \Pr[U = 1|x, r, A = 0]$ (a measure of U-A association), and $\beta_r = \mathbb{E}[Y|r, x, a, U = 1] - \mathbb{E}[Y|r, x, a, U = 0]$ (a measure of U-Y association). Note that this bias formula is agnostic about the causal ordering between race, the pretreatment covariates, and the unobserved confounder. It can be applied whether U is causally prior to R and X or potentially affected by these variables.

If we further assume that $\alpha_{\text{White}} = \alpha_{\text{Black}} = \alpha$ and that $\beta_{\text{White}} = \beta_{\text{Black}} = \beta$, the bias of the estimated interventional disparity can be written as

$$\text{bias}(\Delta^g) = \alpha\beta C, \tag{27}$$

where $C = (\Pr[A^g = 1|\text{White}] - \Pr[A = 1|\text{White}]) - (\Pr[A^g = 1|\text{Black}] - \Pr[A = 1|\text{Black}])$, i.e., the Black-White difference in how the intervention affects treatment prevalence. Similarly, the bias of the estimated "gap reduced" $(\Delta - \Delta^g)$ can be written as

$$\text{bias}(\Delta - \Delta^g) = -\alpha\beta C. \tag{28}$$

Since C is directly estimable from the observed data given a user-specified intervention, it is possible to draw a contour plot showing bias-adjusted estimates of $\Delta - \Delta^g$ (or Δ^g) at different possible values of (α, β). In such a plot, we can identify how strong the unobserved confounder would need to be, in terms of its association with the treatment (α) and the outcome (β), to reduce our estimated $\Delta - \Delta^g$ to zero. In addition, we can use observed covariates to suggest plausible values for α and β. For example, if we have an observed binary confounder $Z \in X$, we can fit a linear model of Y on X, R, and A, whose coefficient on Z will provide a plausible value of β. In the meantime, we can fit a linear model of Z on R, A, and other components of X, whose coefficient on A will provide a plausible value of α. By combining these plausible values of α and β, we can assess the amount of bias that would result if an unobserved variable "worked exactly like" Z in confounding the treatment-outcome relationship. We illustrate these techniques in Section 4.5.

Park *et al.* (2023) proposed a set of alternative bias formulas for conducting sensitivity analysis for estimated interventional disparities (see also Park *et al.* 2024). On the one hand, the bias formulas presented by Park et al. are more general as they do not presuppose the nature of the unobserved confounder U or any form of homogeneity in the U-A or U-Y association. As such,

they remain valid without the simplifying assumptions (i), (ii), and (iii). On the other hand, Park *et al.*'s bias formulas are tailored for a specific intervention, namely, a stochastic intervention under which treatment assignment among one group (e.g., Black students) is a random draw from the marginal distribution of treatment among another group (e.g., White students). By contrast, the bias formulas (26-28) are agnostic about the form of the intervention. As such, they can be applied to a variety of stylized interventions (see the next subsection) without modification. Moreover, due to the simplifying assumptions, our bias formulas involve only a few basic sensitivity parameters, making them much easier to use compared with general formulas that place no restrictions on the pattern of unobserved confounding.[8]

4.4 Some Stylized Interventions

We now introduce some stylized interventions that could be used to guide real-world policies for reducing disparities. In keeping with our running example, we consider a family of interventions that equalize college selectivity across race or class background, either marginally or conditionally on some metric of merit or demonstrated ability. Considering the limited availability of seats in selective colleges, we focus on interventions that maintain the share of students attending selective colleges under the status quo. Interventions of this type are also less likely to induce violations of the consistency assumption (Lundberg 2022). However, by maintaining treatment prevalence at its status quo level, all of these interventions inevitably improve treatment probability for some segments of the population at the expense of others, which could lead to varying degrees of political contention and resistance. Interventions that change treatment prevalence, such as an expansion of selective college enrollments, could also be analyzed through our framework (e.g., Zhou and Pan 2023).

First, let us consider a stochastic intervention under which treatment assignment becomes a random draw from the marginal distribution of treatment. For example, if 20% of all Black and White students currently attend a selective college, then whether a student (Black or White) attends a selective college under this intervention will be governed by a random draw from a

[8]Park *et al.* (2023) also presented simplified versions of their general bias formulas under a set of linear models for the outcome Y, treatment A, and the unobserved confounder U. These simplified formulas are also tailored for their intervention of interest — a stochastic intervention that sets the marginal distribution of treatment for one group to be identical to that for another.

Bernoulli distribution with probability 0.2. Such an intervention was first introduced by Jackson and VanderWeele (2018) and considered also in Lundberg (2022). Its associated PMF of treatment can be written as

$$\Pr[A^g = 1|R, X] = \Pr[A = 1]. \tag{29}$$

Under this intervention, the prevalence of treatment will be equalized between Black and White students. However, at the individual level, treatment becomes a complete lottery, and pre-college characteristics such as academic performance are no longer predictive of whether a student attends a selective college or not. Thus, this intervention might be called *lottery-type marginal equalization*, as it equalizes the marginal distribution of treatment across race via a lottery-type process.

Given that marginal equalization involves a substantial redistribution of opportunity, it could be difficult to achieve due to political resistance. In fact, previous research suggests that complete lottery-type interventions such as the one outlined above would be unpopular with both admissions offices and the general public because they completely dismiss the role of merit (Carnevale and Rose 2013). In light of these considerations, we could also consider less intrusive interventions that equalize treatment not necessarily marginally, but conditional on one or more prespecified covariates. In health disparity research, for example, Jackson (2018) considered how the mortality disparity between the Māori and Europeans in New Zealand would change after intervening to set the smoking behavior of the Māori to the same distribution observed among the Europeans within levels of socioeconomic status (see also Nguyen *et al.* 2020 and Jackson 2021 for similar types of interventions). In the context of selective college attendance, we might consider an intervention that targets racial parity among students with the same high school GPA. This can be operationalized, for example, through the following post-intervention PMF:

$$\Pr[A^g = 1|R, X] = w \Pr[A = 1|R = \text{Black}, Z] + (1 - w) \Pr[A = 1|R = \text{White}, Z], \tag{30}$$

where $Z \in X$ denotes high school GPA. We can see that under this intervention, the probability of treatment given R and X is solely a function of high school GPA (Z), which is defined by a weighted average of the conditional probabilities of treatment given high school GPA for Black

students and for White students. Here, the parameter w can be used to control the conditional distribution tailored for the intervention. In particular, if w equals zero, equation (30) would be akin to the intervention considered in Jackson (2018). By contrast, Nguyen *et al.* (2020) considered an interventional distribution that is an even mixture of the observed treatment distributions between two groups, which would be equivalent to equation (30) with $w = 0.5$.

In our context, to maintain the prevalence of treatment at its pre-intervention level, we propose that w be chosen such that $\Pr[A^g = 1] = \Pr[A = 1]$. This can be achieved by solving the following linear equation in terms of w:

$$\frac{1}{n}\sum_i \left(w\hat{\Pr}[A_i = 1|R_i = \text{Black}, Z_i] + (1-w)\hat{\Pr}[A_i = 1|R_i = \text{White}, Z_i]\right) = \hat{\Pr}[A_i = 1],$$

where n is the sample size, $\hat{\Pr}[A_i = 1]$ is an estimated proportion of selective college goers, and $\hat{\Pr}[A_i = 1|R_i = \text{Black}, Z_i]$ and $\hat{\Pr}[A_i = 1|R_i = \text{White}, Z_i]$ are predicted probabilities of attending a selective college for individual i that can be obtained by fitting a treatment model given race and high school GPA. By doing so, we achieve racial parity in selective college attendance conditional on high school GPA while keeping the overall share of students attending selective colleges unchanged. Under this intervention, treatment is no longer a random draw from a fixed Bernoulli distribution for everyone. Nonetheless, it is still a random draw within levels of high school GPA. We call this intervention *lottery-type conditional equalization,* as it forces the conditional distribution given high school GPA to be identical between Black and White students via a lottery-type intervention.

Lundberg (2022) discusses a potential tradeoff between equity and credibility in the design of interventions for reducing disparities. This tradeoff can be seen by comparing the two stylized interventions outlined previously. To the extent that equity implies unconditional parity in treatment access across groups, lottery-type marginal equalization may be viewed as most equitable as it gives everyone equal access to treatment. Yet, it is not highly credible as it eliminates the roles of all factors currently affecting treatment, some of which may be morally or legally justified (see Jackson 2021 for a discussion of several alternative value judgments about equity). By contrast, interventions targeting conditional equalization are more credible, since they allow certain factors to maintain their influence in the treatment assignment process. If these factors are correlated with group membership, however, such interventions may not be able to equalize treatment marginally,

compromising their equity. In what follows, we introduce an alternative type of intervention that could be envisioned to achieve equity without an excessive loss of credibility.

Specifically, consider an intervention that would equalize treatment across race while preserving the dependency of treatment on all other pretreatment characteristics. Thus, in our example, individuals more likely to attend a selective college under the status quo — due to either more advantaged social backgrounds, stronger academic qualifiations, or both — would still be more likely to attend a selective college under the intervention. This can be achieved through an intervention that multiplies a person's odds of treatment (given all background characteristics) by a race-specific constant such that the prevalence of treatment will be the same between Black and White students and also the same as its overall prevalence under the status quo. Under this intervention, the PMF of treatment can be written as

$$\Pr[A^g = 1|\text{Black}, x] = \frac{e^{\delta_B} \Pr[A = 1|\text{Black}, x]}{1 - \Pr[A = 1|\text{Black}, x] + e^{\delta_B} \Pr[A = 1|\text{Black}, x]}, \tag{31}$$

$$\Pr[A^g = 1|\text{White}, x] = \frac{e^{\delta_W} \Pr[A = 1|\text{White}, x]}{1 - \Pr[A = 1|\text{White}, x] + e^{\delta_W} \Pr[A = 1|\text{White}, x]}, \tag{32}$$

where δ_B and δ_W are race-specific constants chosen such that $\Pr[A^g = 1|\text{Black}] = \Pr[A^g = 1|\text{White}]$=$\Pr[A = 1]$. In practice, they can be computed numerically as solutions to the following nonlinear equations:

$$\frac{1}{n_{\text{Black}}} \sum_{i:R_i=\text{Black}} \frac{e^{\delta_B} \hat{\Pr}[A_i = 1|R_i, X_i]}{1 - \hat{\Pr}[A_i = 1|R_i, X_i] + e^{\delta_B} \hat{\Pr}[A_i = 1|R_i, X_i]} = \hat{\Pr}[A_i = 1],$$

$$\frac{1}{n_{\text{White}}} \sum_{i:R_i=\text{White}} \frac{e^{\delta_W} \hat{\Pr}[A_i = 1|R_i, X_i]}{1 - \hat{\Pr}[A_i = 1|R_i, X_i] + e^{\delta_W} \hat{\Pr}[A_i = 1|R_i, X_i]} = \hat{\Pr}[A_i = 1],$$

where n_{Black} and n_{White} represent the numbers of Black and White students in the sample, $\hat{\Pr}[A_i = 1]$ is an estimated proportion of selective college goers, and $\hat{\Pr}[A_i = 1|R_i, X_i]$ are estimated probabilities of attending a selective college under the status quo that can be obtained by fitting a treatment model given race and the pretreatment covariates.

To see why the above intervention preserves the dependency of A on X (conditional on R), it

is helpful to rewrite Equations (31-32) in terms of log-odds:

$$\log \frac{\Pr[A^g = 1|r, x]}{\Pr[A^g = 0|r, x]} = \delta_r + \log \frac{\Pr[A = 1|r, x]}{\Pr[A = 0|r, x]}.$$

In each racial group, this intervention changes everyone's log-odds of treatment by the same amount (δ_r), but without altering the way it depends on X. In other words, the association between pre-college characteristics and selective college attendance within each racial group, as measured by the odds ratio $\frac{\Pr[A=1|r,x]/\Pr[A=0|r,x]}{\Pr[A=1|r,x']/\Pr[A=0|r,x']}$ (where x and x' represent two different levels of X), is unchanged under the intervention. The functional form characterizing Equations (31-32) was referred to as an incremental propensity score intervention (IPSI) by Kennedy (2019), who considered the impact of a uniform change in everyone's log-odds of treatment on the mean outcome (i.e., $\mathbb{E}[Y^g]$). Here, the IPSI is applied differentially to Black and White students to achieve racial parity in treatment (see Zhou and Pan 2023 for a recent application). Since this intervention is race-conscious but does not alter the roles of other pretreatment characteristics, it is more akin to real-world affirmation action policies than lottery-type equalization. For this reason, we might call it "affirmation-action-type (AA-type) equalization."[9] Notably, by its functional form, AA-type equalization does not change treatment status for units whose treatment probability under the status quo is zero or one. Hence, Assumption 3 (positivity) is not required to identify its effects (Kennedy 2019).

Due to the contested state of race-conscious admissions policies, scholars have also considered the extent to which racial disparities in college attendance, completion, and later-life outcomes can be reduced by class-based affirmative action — as either a substitute for or supplement to race-based affirmative action (e.g., Carnevale and Rose 2013, Alon 2015). In our framework, "AA-type

[9]Equations (31) and (32) can be interpreted as a stochastic intervention that sets each unit's treatment status to be a random draw from a Bernoulli distribution with probability $\Pr[A^g = 1|R, X]$ (Kennedy 2019). Alternatively, it can be interpreted as an intervention under which each unit's treatment status A_i^g is determined by a latent index model $A_i^g = \mathbb{I}(\delta_{R_i} + g(X_i) - V_i \geq 0)$, where $g(\cdot)$ is an unspecified function and V_i is a unit-specific unobserved variable that follows a standard logistic distribution (Zhou and Opacic 2022). In other words, it can be interpreted as a deterministic intervention conditional on factors that are both observed (X_i) and unobserved (V_i) to the researcher (but observed to decision makers). Under either interpretation, AA-type equalization preserves the association of all pretreatment characteristics — apart from group membership — with treatment status. Thus, we use the term "AA-type equalization" to distinguish it from the lottery-type interventions discussed previously.

equalization by class" can be operationalized through the following intervention:

$$\Pr[A^g = 1|r, x] = \frac{e^{\delta_0+\delta_1 s}\Pr[A = 1|r, x]}{1 - \Pr[A = 1|r, x] + e^{\delta_0+\delta_1 s}\Pr[A = 1|r, x]},$$

where $S \in X$ is an indicator of class background, such as parental income rank. Among individuals with the same class background, this intervention changes everyone's log-odds of treatment by the same amount ($\delta_0 + \delta_1 s$), without altering its dependence on other components of X. Thus, it preserves the association of selective college attendance with other pre-college characteristics (on the odds ratio scale) within each level of class background. Here, δ_0 and δ_1 are chosen such that (a) the prevalence of treatment under the intervention is the same as that under the status quo, i.e., $\Pr[A^g = 1] = \Pr[A = 1]$, and (b) treatment is no longer correlated with class background, i.e., $\mathrm{Cor}[A^g, S] = 0$. The first of these conditions ensures that the overall share of students attending selective colleges will stay the same, while the second ensures equalization of treatment across students from different class backgrounds. However, to the extent that Black and White students with the same class background may still differ in other predictors of selective college attendance, such as high school GPA, this intervention does not ensure racial parity in treatment. In practice, the parameters δ_0 and δ_1 can be computed numerically as solutions to the following system of equations:

$$\begin{aligned}
\frac{1}{n}\sum_i \frac{e^{\delta_0+\delta_1 S_i}\hat{\Pr}[A_i = 1|R_i, X_i]}{1 - \hat{\Pr}[A_i = 1|R_i, X_i] + e^{\delta_0+\delta_1 S_i}\hat{\Pr}[A_i = 1|R_i, X_i]} &= \hat{\Pr}[A_i = 1], \\
\frac{1}{n}\sum_i S_i \cdot \frac{e^{\delta_0+\delta_1 S_i}\hat{\Pr}[A_i = 1|R_i, X_i]}{1 - \hat{\Pr}[A_i = 1|R_i, X_i] + e^{\delta_0+\delta_1 S_i}\hat{\Pr}[A_i = 1|R_i, X_i]} &= \big(\frac{1}{n}\sum_i S_i\big)\hat{\Pr}[A_i = 1],
\end{aligned}$$

where n is the sample size, $\hat{\Pr}[A_i = 1]$ is an estimated proportion of selective college goers, and $\hat{\Pr}[A_i = 1|R_i, X_i]$ are estimated probabilities of attending a selective college under the status quo that can be obtained from a treatment model given race and the pretreatment covariates.

To achieve both economic and racial parity in treatment while again maintaining the dependency of A on X, we could also consider "AA-type equalization by class and race." This can be achieved

through the use of race-specific δ_0 and δ_1 parameters:

$$\Pr[A^g = 1|\text{Black}, x] = \frac{e^{\delta_{0B}+\delta_{1B}s}\Pr[A = 1|\text{Black}, x]}{1 - \Pr[A = 1|\text{Black}, x] + e^{\delta_{0B}+\delta_{1B}s}\Pr[A = 1|\text{Black}, x]},$$
$$\Pr[A^g = 1|\text{White}, x] = \frac{e^{\delta_{0W}+\delta_{1W}s}\Pr[A = 1|\text{White}, x]}{1 - \Pr[A = 1|\text{White}, x] + e^{\delta_{0W}+\delta_{1W}s}\Pr[A = 1|\text{White}, x]},$$

where δ_{0B}, δ_{1B}, δ_{0W}, and δ_{1W} are chosen such that (a) the prevalence of treatment will be the same between Black and White students and also the same as its overall prevalence under the status quo, i.e., $\Pr[A^g = 1|\text{Black}] = \Pr[A^g = 1|\text{White}] = \Pr[A = 1]$, and (b) treatment is uncorrelated with class background within each racial group, i.e., $\text{Cor}[A^g, S|\text{Black}] = \text{Cor}[A^g, S|\text{White}] = 0$. Since conditions (a) and (b) involve four equality constraints (each of them involves two), the four parameters δ_{0B}, δ_{1B}, δ_{0W}, and δ_{1W} are uniquely identified and can also be computed numerically via a corresponding system of equations. Similar to the previous two AA-type interventions, this intervention preserves the association of selective college attendance with pre-college characteristics (on the odds ratio scale) among individuals with the same race and class background.

4.5 Illustration

We now illustrate the prescriptive approach to disparity analysis by assessing how much of the Black-White gap in college completion rates would be reduced under the different stylized interventions outlined above. We use the same sample of the NLSY97 data as described in Section 3.5. However, we now treat college selectivity as a treatment and all other explanatory variables (demographic and socioeconomic background, family structure, academic performance and behavior, and peer and school-level characteristics) as pretreatment covariates. In what follows, we focus on DML estimates of the gap reduced, i.e., $\Delta - \Delta^g$, based on Equation (24), cross-fitting, and machine learning estimates of the outcome and treatment models.[10] As in Section 3.5, we use five-fold cross-fitting, estimate each of the outcome and treatment models using a super learner composed of Lasso and random forests, and stabilize our estimates by censoring all inverse-probability-weights (W_i in Equation 24) at their 1st and 99th percentiles. Standard errors are constructed using the sample

[10]We also obtained estimates of the gap reduced from the RI, weighting, and doubly robust estimators based on Equations (22-24) where the outcome and treatment models were both fit via a logistic regression with linear and additive effects of the explanatory variables. The results are qualitatively similar and reported in Appendix H.

variances of the estimated influence functions and adjusted for multiple imputation via Rubin's (1987) method.

Figure 3 shows the estimated reductions in the Black-White completion gap under the five stylized interventions described in Section 4.4: lottery-type marginal equalization, lottery-type conditional equalization (given high-school GPA), AA-type equalization, AA-type equalization by class, AA-type equalization by class and race. In AA-type equalization by class and AA-type equalization by class and race, class background is measured by the percentile rank of parental income. We see that different interventions would reduce the gap to varying degrees. The largest reductions are associated with AA-type equalization by race and AA-type equalization by class and race, in which case the gap would drop by about 2.6 percentage points (from a baseline of 18 percentage points to 15.4). Lottery-type marginal equalization might reduce the gap by a similar amount, although its estimated effect is not statistically significant due to a larger standard error. By contrast, AA-type equalization by class only would have a much smaller and statistically insignificant effect on the Black-White completion gap. This finding is consistent with previous research showing that economic affirmation action would not improve racial parity in *selective college attendance* in the first place (Carnevale and Rose 2013; Alon 2015). Finally, the estimated reduction associated with lottery-type conditional equalization (given high school GPA) is also relatively small and statistically insignificant. Overall, our analyses suggest that interventions that target racial parity in selective college attendance can still lead to appreciable reductions in the Black-White completion gap, despite the fact that little of the observed gap can be statistically explained by racial differences in selective college attendance after other characteristics are accounted for (Figure 1).

[Figure 3 here]

In order to assess the robustness of the above findings to potential violations of Assumption 2 (*Unconfoundedness*), we implement our sensitivity analysis as discussed in Section 4.3. For illustrative purposes, we focus on our estimate of the gap reduced ($\Delta - \Delta^g$) under AA-type equalization, which is about 0.026. In this case, because $\Pr[A^g = 1|\text{White}] = \Pr[A^g = 1|\text{Black}]$, the parameter C in the bias formula (28) reduces to $\Pr[A = 1|\text{Black}] - \Pr[A = 1|\text{White}]$, which, according to our data, is about -0.18. Hence, the amount of bias associated with given values of (α, β)

can be approximated as $0.18\alpha\beta$, and the corresponding bias-adjusted estimate can be written as $0.026 - 0.18\alpha\beta$, where α and β represent the strengths of the U-A and U-Y relationships conditional on race and other covariates, as defined earlier. Figure 4 shows contours of the bias-adjusted estimates associated with a range of potential values of α and β. To benchmark these contours, we also show the values of (α, β) that would arise if the unobserved binary confounder U behaved like one of two observed binary confounders of the A-Y relationship: *urban residence* and *peers' college expectation.*[11] We can see that the original estimate (0.026) can be explained away by unobserved confounding only when both α and β are much larger than the sizes implied by these observed confounders. For example, the bias-adjusted estimate would reduce to zero only when both the α and β values of the unobserved confounder are about 10 times those of *peers' college expectation.*

[Figure 4 here]

4.6 KOB Decompositions and Interventional Disparities

Throughout the preceding discussion, we have stressed the distinct assumptions, estimands, and estimators associated with the descriptive and prescriptive approaches to disparity analysis. In general, unless the explanatory variables are unconfounded, estimates from the KOB method and its variants possess no causal interpretations and, therefore, should not be used to inform interventional disparities of any kind. Nonetheless, in cases where a KOB decomposition appropriately adjusts for potential confounding of a predictor-outcome relationship, the explained portion associated with the predictor can be causally meaningful. In the context of linear models, Jackson and VanderWeele (2018) establish a set of conditions under which certain interventional disparities can be recovered from KOB decompositions. In what follows, we highlight two key results from this discussion, which correspond to the simultaneous and sequential KOB decompositions, respectively.

First, in a detailed KOB decomposition (Equation 13), if the causal effect of a predictor X_j on the outcome is unconfounded given race and all other predictors (i.e., $Y(x_j) \perp\!\!\!\perp X_j | R, X \backslash X_j$ for any x_j), then the explained portion associated with X_j could have an interventional interpretation. Specifically, suppose that Assumptions 1-3 hold for X_j (treating $X \backslash X_j$ as pretreatment covariates)

[11] Among all observed binary confounders of the A-Y relationship in our data, *urban residence* and *peers' college expectation* are the two strongest according to their implied values of $\alpha\beta$. *Peers' college expectation* is a dummy variable denoting whether at least 75% of the respondent's peers expected college in 1997.

and that the linear model for the conditional expectation $\mathbb{E}[Y|R = \text{Black}, X]$ is correctly specified. Under these assumptions, the explained portion associated with X_j in a detailed KOB decomposition can be interpreted as the gap reduced under a stochastic intervention that sets X_j for each Black person to be a random draw from the marginal distribution of X_j among Whites. In other words, under the above assumptions,

$$\beta_{B,j}(\bar{X}_{W,j} - \bar{X}_{B,j}) = \Delta - \Delta^g,$$

where g represents an intervention under which $X_j^g|R = \text{Black}$ is a random draw from $X_j|R = \text{White}$. This intervention is similar to the aforementioned lottery-type marginal equalization except that (a) it modifies the treatment assignment process only among Blacks and (b) it does not preserve the marginal distribution of X_j in the population.

We can revisit our estimates of the detailed KOB decomposition (Figure 2) in light of the above results. Given the temporal structure of the predictors involved, the unconfoundedness assumption seems unrealistic for any variable other than college selectivity (X_5), as college selectivity is likely a post-treatment variable for any other predictor-outcome relationship and thus should not have been adjusted for. If Assumptions 1-3 hold for college selectivity and if the linear model for $\mathbb{E}[Y|R = \text{Black}, X]$ is correctly specified, then the gap explained by college selectivity in the detailed KOB decomposition can be interpreted as the degree to which the Black-White gap in college completion would be reduced if we intervened to set college selectivity among Black students to be a random draw from its distribution among Whites. Our results suggest that such an intervention could reduce the completion gap by two percentage points, a magnitude similar to that associated with lottery-type marginal equalization (see Figure 3).

Second, in a *sequential* KOB decomposition (Equation 15), if the causal effect of a predictor X_j on the outcome is unconfounded given race and $\mathcal{X}_{j-1} = (X_1, X_2, \ldots X_{j-1})$ (i.e., $Y(x_j) \perp\!\!\!\perp X_j|R, \mathcal{X}_{j-1}$ for any x_j), then the explained portion associated with X_j could have an interventional interpretation. Specifically, suppose that (a) Assumptions 1-3 hold for X_j (treating $\mathcal{X}_{j-1}$ as pretreatment covariates), (b) for each group $R = r$, the effect of X_j on the outcome does not vary by $\mathcal{X}_{j-1}$, and (c) the disparity in the conditional distribution of X_j given $\mathcal{X}_{j-1}$ between Blacks and Whites does not vary by $\mathcal{X}_{j-1}$. Under these assumptions, the explained portion asso-

ciated with X_j in a sequential KOB decomposition can be interpreted as the gap reduced under a stochastic intervention that, for each Black person, sets X_j to be a random draw from the conditional distribution of X_j given $\mathcal{X}_{j-1}$ among Whites (see Jackson 2018 for a related discussion). In other words, under the above assumptions,

$$\Delta_{\text{Explained},X_j} = \Delta - \Delta^g,$$

where g represents an intervention under which $X_j^g|R = \text{Black}, \mathcal{X}_{j-1}$ is a random draw from $X_j|R = \text{White}, \mathcal{X}_{j-1}$. This intervention differs from the aforementioned lottery-type conditional equalization in that equalization of treatment is now conditional on the full set of potential confounders of the X_j-Y relationship. In our example, if we consider the gap explained by college selectivity in a sequential decomposition, it could be interpreted as the degree to which the gap would reduce if we intervened to set college selectivity among Black students to be a random draw from its distribution among Whites within levels of all pre-college characteristics. Our results in Figure 2 suggest that such an intervention is unlikely to induce a meaningful reduction in the Black-White gap in college completion rates.

In sum, under appropriate identification assumptions, both the simultaneous and sequential variants of the KOB method could be used to assess interventional disparities. Compared to our proposed estimators of interventional disparities, the KOB approach is relatively easy to use with continuous treatments and can also be readily implemented with standard software. Nonetheless, this approach can only be used to evaluate disparities under a limited set of highly specific interventions. Moreover, it hinges on the validity of additional modeling assumptions such as linearity and effect homogeneity. By contrast, the framework outlined in Sections 4.1-4.5 is more general, allowing us to define and identify disparities under a diverse array of user-defined interventions and estimate them without strong modeling assumptions. Despite their lack of policy implications, we do not seek to discourage the use of KOB decompositions. Instead, our hope is to empower future practitioners of the KOB method both by clarifying the interpretive scope of this tool, both descriptively as well as prescriptively, and by relaxing the modeling assumptions needed to apply it.

5 Concluding Remarks

In this article, we have explicated, contrasted, and extended two distinct approaches to disparity analysis. In the descriptive approach, all estimands are a function of *observed data*: they do not involve counterfactuals. As with a scatter plot, a variance-covariance analysis (ANOVA), or a linear regression model, a descriptive analysis of disparity is a useful tool for summarizing our data. In some cases, combined with additional identification and modeling assumptions, it may also provide useful indications of how a disparity would change under a particular intervention. Yet, in general, results from a descriptive analysis only address a very limited set of policy-relevant questions. This is true regardless of what kinds of statistical models have been used to decompose the disparity. By contrast, the prescriptive approach directly interrogates the extent to which a disparity of interest would change under a given intervention. Such an approach fosters more rigorous, and less ad hoc, discussions of the potential impacts of different policies.

For both approaches, we demonstrate that one can define and identify all relevant estimands nonparametrically, i.e., without reference to any statistical model. For example, while the canonical KOB decomposition is routinely implemented via linear regression models, the underlying estimands can be defined in an entirely model-free manner. The model-free definition and identification of estimands improves the clarity of research goals and the transparency of the associated assumptions. Moreover, as we have shown, it opens the door to a range of flexible estimation strategies, including regression-imputation, weighting, and doubly robust methods. In particular, the doubly robust estimator can be combined with flexible machine learning models to produce estimators with desirable statistical properties. To be sure, the separation of the definition of estimands from their identification and estimation should be a goal not only for disparity analysis but for all quantitative social science research (Lundberg *et al.* 2021).

For both the descriptive and prescriptive approaches, we have suggested several extensions that allow us to examine a broader range of research questions than those enabled by extant methods. For example, our proposed generalized KOB/Duncan decomposition enables researchers to assess the cumulative contributions of temporally ordered covariates successively, regardless of whether these contributions are "direct" or instead "flow through" subsequent explanatory variables. Extant methods, such as the simultaneous KOB decomposition, are not geared toward this goal. For the

prescriptive approach, leveraging the generality of identification and estimation results, we have introduced a range of interventional effects that capture a broader set of policy objectives than existing work. For interventional disparities, we have also proposed and illustrated an original method for assessing the robustness of our estimates to potential unobserved confounding.

The prescriptive approach we discussed can be extended in several ways. First, while the point-in-time interventions we consider align closely with the way most real-world policies are formulated, researchers may be interested in assessing the impact of a bundle of hypothetical interventions implemented at multiple points of an individual's life course. For example, a policy designed to equalize admission rates in postsecondary education may concur with an intervention designed to equalize rates of college completion. Researchers may then want to assess the impact of this dual set of policies on a range of later-life outcomes. Second, while we have focused on "categorical inequalities" across disparate groups, many axes of inequality may be best characterized as lying on a continuum. A prime example is intergenerational income (im)mobility, where the "group" indicator is parental income or a monotone transformation of it. Extending our proposed framework to assess interventional disparities by a continuous characteristic is an important avenue for future research.

References

Alon, S. (2015) *Race, Class, and Affirmative action*. Russell Sage Foundation.

Althauser, R. P. and Wigler, M. (1972) Standardization and component analysis. *Sociological Methods & Research*, **1**, 97–135.

An, W. and Glynn, A. N. (2021) Treatment effect deviation as an alternative to blinder–oaxaca decomposition for studying social inequality. *Sociological Methods & Research*, **50**, 1006–1033.

Bailey, M. J. and Dynarski, S. M. (2011) Gains and gaps: Changing inequality in us college entry and completion. Tech. rep., National Bureau of Economic Research.

Barsky, R., Bound, J., Charles, K. K. and Lupton, J. P. (2002) Accounting for the black–white wealth gap: a nonparametric approach. *Journal of the American Statistical Association*, **97**, 663–673.

Bauer, T. K. and Sinning, M. (2008) An extension of the blinder–oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis*, **92**, 197–206.

Blinder, A. S. (1973) Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 436–455.

Bowen, W. G., Chingos, M. M. and McPherson, M. S. (2009) *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton, NJ: Princeton University Press.

Braxton, J. M., Hirschy, A. S. and McClendon, S. A. (1997) Understanding and reducing college student departure. In *ASHE-ERIC Higher Education Report*, vol. 30. San Francisco, CA: Wiley Periodicals.

Carnevale, A. P. and Rose, S. (2013) Socioeconomic status, race/ethnicity, and selective college admissions.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**, C1–C68.

Ciocca Eller, C. and DiPrete, T. A. (2018) The paradox of persistence: Explaining the black-white gap in bachelor's degree completion. *American Sociological Review*, **83**, 1171–1214. URL `https://journals.sagepub.com/doi/abs/10.1177/0003122418808005https://journals.sagepub.com/doi/pdf/10.1177/0003122418808005`.

Cole, S. R. and Hernán, M. A. (2008) Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, **168**, 656–664.

Cotton, J. (1988) On the decomposition of wage differentials. *The Review of Economics and Statistics*, 236–243.

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

Duncan, O. D. (1968) Inheritance of poverty or inheritance of race? *On Understanding Poverty*, 85–110.

Fairlie, R. W. (2005) An extension of the blinder-oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social measurement*, **30**, 305–316.

Firpo, S. P., Fortin, N. M. and Lemieux, T. (2018) Decomposing wage distributions using recentered influence function regressions. *Econometrics*, **6**, 28.

Fortin, N., Lemieux, T. and Firpo, S. (2011) Decomposition methods in economics. In *Handbook of Labor Economics*, vol. 4, 1–102. Elsevier.

Gelbach, J. B. (2016) When do covariates matter? and which ones, and how much? *Journal of Labor Economics*, **34**, 509–543.

Hernan, M. A. and Robins, J. M. (2023) *Causal inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hines, O., Dukes, O., Diaz-Ordaz, K. and Vansteelandt, S. (2022) Demystifying statistical learning based on efficient influence functions. *The American Statistician*, **76**, 292–304.

Howe, C. J., Dulin-Keita, A., Cole, S. R., Hogan, J. W., Lau, B., Moore, R. D., Mathews, W. C., Crane, H. M., Drozd, D. R., Geng, E. *et al.* (2018) Evaluating the population impact on racial/ethnic disparities in hiv in adulthood of intervening on specific targets: A conceptual and methodological framework. *American journal of epidemiology*, **187**, 316–325.

Huber, M. (2015) Causal pitfalls in the decomposition of wage gaps. *Journal of Business & Economic Statistics*, **33**, 179–191.

Jackson, J. W. (2018) On the interpretation of path-specific effects in health disparities research. *Epidemiology*, **29**, 517–520.

Jackson, J. W. (2021) Meaningful causal decompositions in health equity research: Definition, identification, and estimation through a weighting framework. *Epidemiology (Cambridge, Mass.)*, **32**, 282.

Jackson, J. W. and VanderWeele, T. J. (2018) Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology*, **29**, 825.

Jann, B. (2008) The blinder–oaxaca decomposition for linear regression models. *The Stata Journal*, **8**, 453–479.

Jeffrey, W. (2020) Crossing the finish line? a review of college completion inequality in the united states by race and class. *Sociology Compass*, **14**, e12787.

Jones, F. L. (1983) On decomposing the wage gap: a critical comment on blinder's method. *The Journal of human resources*, **18**, 126–130.

Juhn, C., Murphy, K. M. and Pierce, B. (1993) Wage inequality and the rise in returns to skill. *Journal of Political Economy*, **101**, 410–442.

Katz, L. F. and Murphy, K. M. (1992) Changes in relative wages, 1963–1987: Supply and demand factors. *The Quarterly Journal of Economics*, **107**, 35–78.

Kennedy, E. H. (2019) Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, **114**, 645–656.

Kitagawa, E. M. (1955) Components of a difference between two rates. *Journal of the American Statistical Association*, **50**, 1168–1194.

Kline, P. (2011) Oaxaca-blinder as a reweighting estimator. *American Economic Review*, **101**, 532–537.

Laurison, D. and Friedman, S. (2016) The class pay gap in higher professional and managerial occupations. *American Sociological Review*, **81**, 668–695.

Lemieux, T. (2006) Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, **96**, 461–498.

Lundberg, I. (2022) The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. *Sociological Methods & Research*, **0**, 00491241211055769. URL `https://doi.org/10.1177/00491241211055769`.

Lundberg, I., Johnson, R. and Stewart, B. M. (2021) What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, **86**, 532–565.

Mandel, H. and Semyonov, M. (2016) Going back in time? gender differences in trends and sources of the racial pay gap, 1970 to 2010. *American Sociological Review*, **81**, 1039–1068.

Mize, T. D. (2016) Sexual orientation in the labor market. *American Sociological Review*, **81**, 1132–1160.

Mouw, T. and Kalleberg, A. L. (2010) Occupations and the structure of wage inequality in the united states, 1980s to 2000s. *American Sociological Review*, **75**, 402–431.

Naimi, A. I., Moodie, E. E., Auger, N. and Kaufman, J. S. (2014a) Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*, **25**, 292–299.

Naimi, A. I., Moodie, E. E., Auger, N. and Kaufman, J. S. (2014b) Stochastic mediation contrasts in epidemiologic research: interpregnancy interval and the educational disparity in preterm delivery. *American journal of epidemiology*, **180**, 436–445.

Nguyen, T. Q., Schmid, I. and Stuart, E. A. (2020) Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*.

Oaxaca, R. (1973) Male-female wage differentials in urban labor markets. *International Economic Review*, 693–709.

Oaxaca, R. L. and Ransom, M. R. (1999) Identification in detailed wage decompositions. *Review of Economics and Statistics*, **81**, 154–157.

Park, S., Kang, S., Lee, C. and Ma, S. (2023) Sensitivity analysis for causal decomposition analysis: Assessing robustness toward omitted variable bias. *Journal of Causal Inference*, **11**, 20220031.

Park, S., Qin, X. and Lee, C. (2024) Estimation and sensitivity analysis for causal decomposition in health disparity research. *Sociological Methods & Research*, **53**, 571–602.

Petersen, T. and Morgan, L. A. (1995) Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. *American Journal of Sociology*, **101**, 329–365.

Reardon, S. F., Baker, R. and Klasik, D. (2012) Race, income, and enrollment patterns in highly selective colleges, 1982-2004.

Reimers, C. W. (1983) Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics*, 570–579.

Rothe, C. (2015) Decomposing the composition effect: the role of covariates in determining between-group differences in economic outcomes. *Journal of Business & Economic Statistics*, **33**, 323–337.

Rubin, D. B. (1986) Comment: Which ifs have causal answers. *Journal of the American statistical association*, **81**, 961–962.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley and Sons.

Smith-Doerr, L., Alegria, S., Husbands Fealing, K., Fitzpatrick, D. and Tomaskovic-Devey, D. (2019) Gender pay gaps in us federal science agencies: An organizational approach. *American Journal of Sociology*, **125**, 534–576.

Snyder, T. D., De Brey, C. and Dillow, S. A. (2019) Digest of education statistics 2017, nces 2018-070. *National Center for Education Statistics*.

Storer, A., Schneider, D. and Harknett, K. (2020) What explains racial/ethnic inequality in job quality in the service sector? *American Sociological Review*, **85**, 537–572.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012) Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, **40**, 1816.

Tilly, C. (1998) *Durable Inequality*. University of California Press.

Tinto, V. (1994) *Leaving College: Rethinking the Causes and Cures of Student Attrition*. Chicago, IL: University of Chicago Press, 2 edn.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**.

van der Laan, M. J. and Rose, S. (2018) *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer.

VanderWeele, T. J. and Robinson, W. R. (2014) On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, **25**, 473.

Voss, K., Hout, M. and George, K. (2022) Persistent inequalities in college completion, 1980–2010. *Social Problems*.

Wasserman, L. (2006) *All of Nonparametric Statistics*. Springer Science & Business Media.

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

Western, B. and Rosenfeld, J. (2011) Unions, norms, and the rise in us wage inequality. *American Sociological Review*, **76**, 513–537.

Winsborough, H. and Dickinson, P. (1971) Components of negro-white income differences. *Age*, **25**, 35–44.

Young, J. G., Hernán, M. A. and Robins, J. M. (2014) Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic Methods*, **3**, 1–19.

Yu, A. and Elwert, F. (2022) Causal decomposition of group disparities: The role of within-group heterogeneity. In *PAA 2022 Annual Meeting*. PAA.

Zhou, X. (2019) Equalization or selection? reassessing the "meritocratic power" of a college degree in intergenerational income mobility. *American Sociological Review*, **84**, 459–485.

Zhou, X. (2022) Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**, 794–821.

Zhou, X. and Opacic, A. (2022) Marginal interventional effects. *arXiv preprint arXiv:2206.10717*.

Zhou, X. and Pan, G. (2023) Higher education and the black-white earnings gap. *American Sociological Review*, **88**, 154–188.

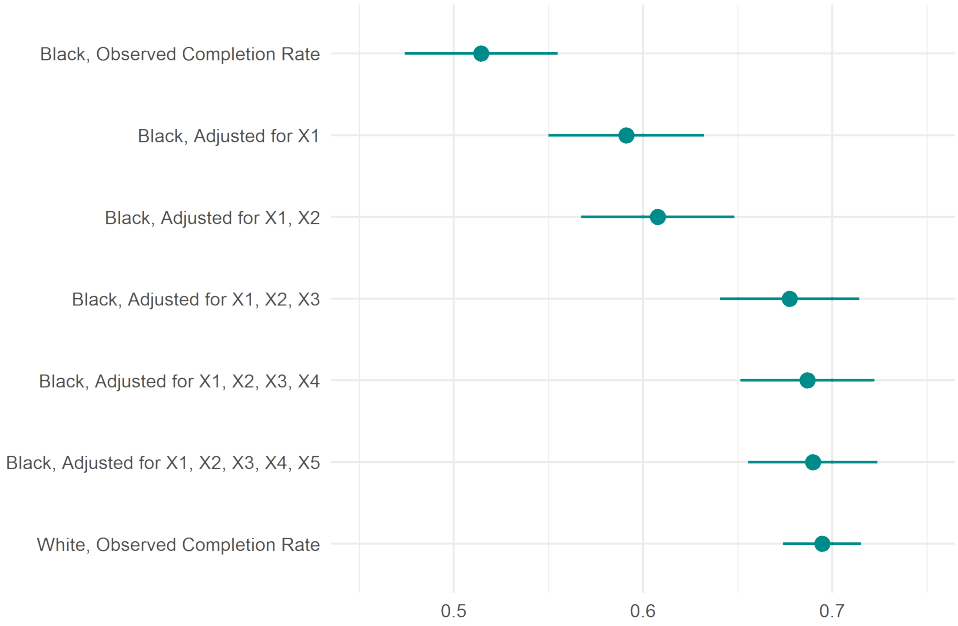


Figure 1: DML estimates of observed and adjusted completion rates.

Note: Data are from the NLSY97. X_1 denotes demographic and socioeconomic background (gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence), X_2 denotes family structure (co-residence with both biological parents, presence of a paternal figure), X_3 denotes pre-college measures of academic performance and behavior (percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18), X_4 denotes peer and school-level characteristics (college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school), and X_5 denotes college selectivity (whether the student attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000). In the DML estimation, both the outcome and group membership models are fitted via a super learner composed of Lasso and random forests. Line ranges show 95% Wald-type confidence intervals, which are based on standard errors constructed from the sample variances of the estimated efficient influence functions.

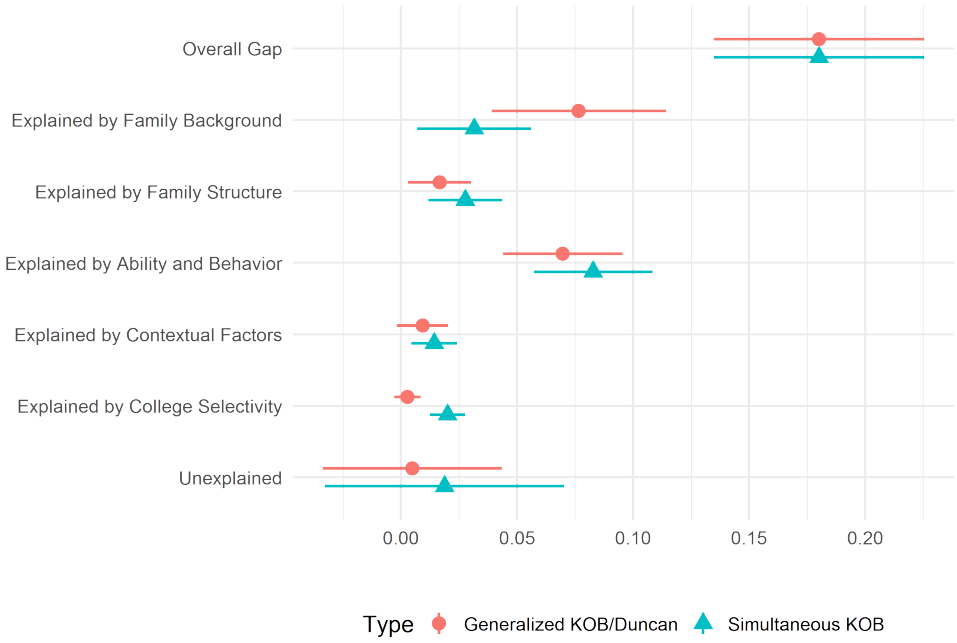


Figure 2: Estimates of the explained and unexplained components of the Black-White gap in college completion under the generalized KOB/Duncan decomposition and the simultaneous KOB decomposition.

Note: Data are from the NLSY97. *Family background* denotes a broad set of variables concerning demographic and socioeconomic background, including gender, race, ethnicity, age in 1997, parental education, parental income, parental assets, rural residence, southern residence; *family structure* includes co-residence with both biological parents, presence of a paternal figure; *academic performance and behavior* includes percentile score on the ASVAB test, high school GPA, an index of substance use, an index of delinquency, whether the respondent had any children by age 18; *contextual factors* include college expectation among peers and three dummy variables denoting whether the respondent ever had property stolen at school, was ever threatened at school, and was ever in a fight at school; *college selectivity* is a dummy variable for whether the student attended one of the "most competitive," "highly competitive," and "very competitive" colleges according to Barron's Profile of American Colleges 2000. In the DML estimation of the generalized KOB/Duncan decomposition, both the outcome and group membership models are fitted via a super learner composed of Lasso and random forests. Line ranges show 95% Wald-type confidence intervals. For the generalized KOB/Duncan decomposition, standard errors are constructed from the sample variances of the estimated efficient influence functions. For the simultaneous KOB decomposition, standard errors are constructed using the nonparametric bootstrap (with 250 replications).

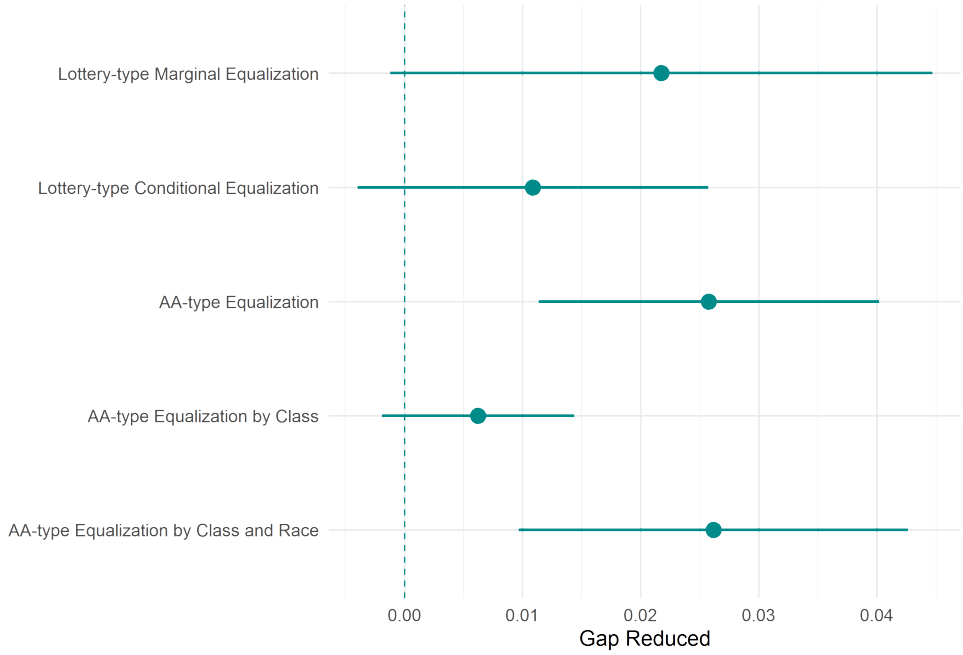


Figure 3: DML estimates of the gap reduced ($\Delta - \Delta^g$) under different stylized interventions to selective college attendance.
Note: Data are from the NLSY97. Pretreatment covariates include demographic and socioeconomic background, family structure, academic performance and behavior, and peer- and school-level characteristics, as detailed in the descriptive analysis. In the DML estimation, both the outcome and treatment models are fitted via a super learner composed of Lasso and random forests. Line ranges show 95% Wald-type confidence intervals, which are based on standard errors constructed from the sample variances of the estimated efficient influence functions.

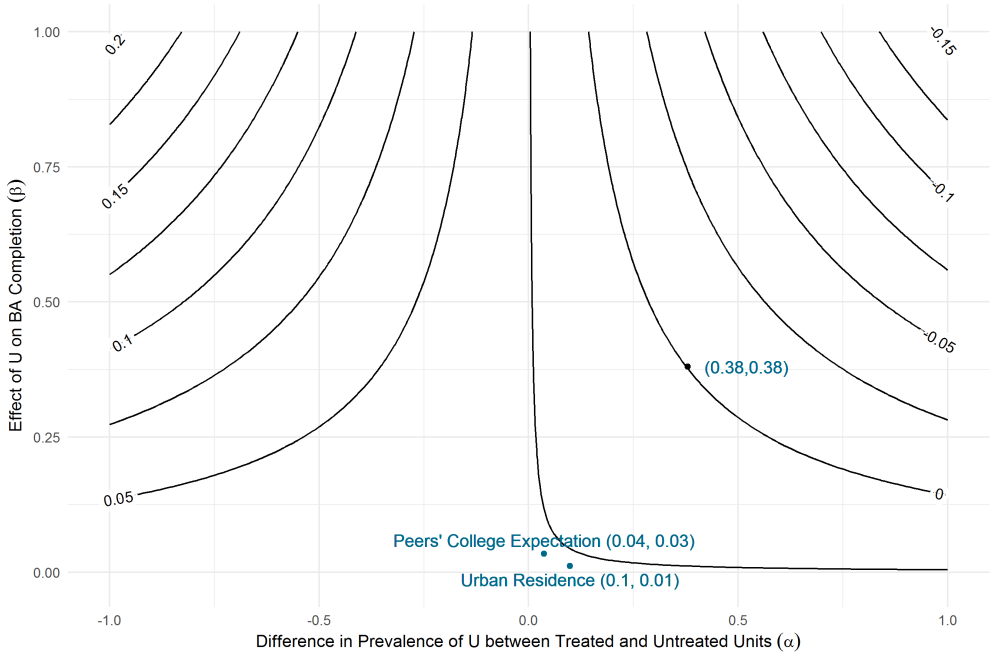


Figure 4: Contour plot showing bias-adjusted estimates of the gap reduced ($\Delta - \Delta^g$) under AA-type equalization as a function of (α, β). The two annotated points show values of (α, β) if the unobserved confounder U behaved like one of two observed binary confounders in the NLSY97.

A Nonparametric Formulations of Equations (2) and (5)

The decomposition in the form of Equation (2) involves fitting an outcome model among White students, which is then used to obtain the average predicted outcome among White students under the Black covariate distribution, $\bar{Y}_W^{\text{pred}}$. The population counterpart of $\bar{Y}_W^{\text{pred}}$ can be written as

$$\mu_W^{\text{pred}} = \int \mathbb{E}[Y|R = \text{White}, x] dP(x|R = \text{Black}), \tag{33}$$

where $P(x|R = \text{Black})$ denotes the covariate distribution among Black students. Consequently, the explained and unexplained components of the KOB decomposition become

$$\Delta_{\text{Explained}} = \mu_W - \mu_W^{\text{pred}}, \tag{34}$$

$$\Delta_{\text{Unexplained}} = \mu_W^{\text{pred}} - \mu_B. \tag{35}$$

The explained component is the difference between the average observed outcome of White students and the average predicted outcome of White students when their covariate distribution is adjusted to that of Black students. The unexplained component is the difference between the average predicted outcome of White students after the adjustment and the average observed outcome of Black students.

The more general form of the canonical KOB decomposition shown in Equation (5) can also be defined nonparametrically. Consider a weighted average of the conditional expectations of the outcome given the covariates for Black students and for White students, $\mu_R(X) = q_B \mathbb{E}[Y|R = \text{Black}, X] + q_W \mathbb{E}[Y|R = \text{White}, X]$, where $q_B + q_W = 1$. The function $\mu_R(X)$ can be seen as a composite conditional expectation function (CEF). The X-Y relationship encoded in this composite CEF is analogous to the reference coefficients β_R in Equation (5). Accordingly, a nonparametric counterpart of Equation (5) can be written as

$$\Delta_{\text{Explained}} = \int \mu_R(x) \big(dP(x|R = \text{White}) - dP(x|R = \text{Black}) \big), \tag{36}$$

$$\Delta_{\text{Unexplained}} = \int (\mu_W(x) - \mu_R(x)) dP(x|R = \text{White}) + \int (\mu_R(x) - \mu_W(x)) dP(x|R = \text{Black}). \tag{37}$$

Clearly, this decomposition reduces to Equations (7-8) when $q_B = 1$ and Equations (34-35) when $q_W = 1$.

B Efficient Influence Function of μ_B^{pred}

To simplify notation, let us recode Black as 1 and White as 0 such that μ_B^{pred} can be written as

$$\mu_B^{\text{pred}} = \int \mathbb{E}[Y|R=1,x]dP(x|R=0).$$

We now derive the efficient influence function (EIF) of μ_B^{pred} in the nonparametric model through a perturbation approach introduced by Hines *et al.* (2022). First, rewrite μ_B^{pred} as a functional of the true observed data distribution $\mathcal{P}$: $\psi(\mathcal{P}) \equiv \mu_B^{\text{pred}}$. Then, consider a one-dimensional parametric submodel $\mathcal{P}_t$ characterized by the following density function:

$$p_t(r,x,y) = (1-t)p(r,x,y) + t\delta_{\tilde{r},\tilde{x},\tilde{y}}(r,x,y),$$

where $p(r,x,y)$ is the true density function of (R,X,Y), and $\delta_{\tilde{r},\tilde{x},\tilde{y}}(r,x,y)$ is the Dirac delta function with respect to a hypothetical data point $(\tilde{r},\tilde{x},\tilde{y})$. Hines *et al.* (2022) show that the EIF at $(\tilde{r},\tilde{x},\tilde{y})$ can be calculated as the derivative of $\psi(\mathcal{P}_t)$ with respect to t at the truth, i.e.,

$$\partial_t \psi(\mathcal{P}_t) = \frac{d\psi(P_t)}{dt}\Big|_{t=0}.$$

Following the chain and quotient rules for derivatives and the Bayes rule, we have

$$\begin{aligned}
\partial_t \psi(\mathcal{P}_t) &= \partial_t \int \mathbb{E}_t[Y|R=1,x]p_t(x|R=0)dx \\
&= \int \big(\partial_t \mathbb{E}_t[Y|R=1,x]\big)p(x|R=0)dx + \int \mathbb{E}[Y|R=1,x]\big(\partial_t p_t(x|R=0)\big)dx \\
&= \int \frac{\delta_{\tilde{r},\tilde{x}}(1,x)}{p(1,x)}\big(\tilde{y} - \mathbb{E}[Y|R=1,x]\big)p(x|R=0)dx + \int \mathbb{E}[Y|R=1,x]\frac{\delta_{\tilde{r}}(0)}{p(0)}\big(\delta_{\tilde{x}}(x) - p(x|R=0)\big)dx \\
&= \int \frac{\delta_{\tilde{r}}(1)\delta_{\tilde{x}}(x)}{p(1|x)p(x)}\big(\tilde{y} - \mathbb{E}[Y|R=1,x]\big)\frac{p(0|x)p(x)}{p(0)}dx + \frac{\delta_{\tilde{r}}(0)}{p(0)}\int \mathbb{E}[Y|R=1,x]\big(\delta_{\tilde{x}}(x) - p(x|R=0)\big)dx \\
&= \delta_{\tilde{r}}(1)\int \frac{\delta_{\tilde{x}}(x)p(0|x)}{p(1|x)p(0)}\big(\tilde{y} - \mathbb{E}[Y|R=1,x]\big)dx + \frac{\delta_{\tilde{r}}(0)}{p(0)}\int \mathbb{E}[Y|R=1,x]\big(\delta_{\tilde{x}}(x) - p(x|R=0)\big)dx \\
&= \frac{\delta_{\tilde{r}}(1)p(0|x)}{p(0)p(1|x)}\big(\tilde{y} - \mathbb{E}[Y|R=1,\tilde{x}]\big) + \frac{\delta_{\tilde{r}}(0)}{p(0)}\big(\mathbb{E}[Y|R=1,\tilde{x}] - \psi(\mathcal{P})\big).
\end{aligned}$$

Thus, the EIF of μ_B^{pred} at (R,X,Y) is

$$\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{\Pr[R=0|X]}{\Pr[R=1|X]}(Y - \mathbb{E}[Y|R=1,X]) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}(\mathbb{E}[Y|R=1,X] - \mu_B^{\text{pred}}).$$

Alternatively, the EIF of μ_B^{pred} can be found by noticing the connection of μ_B^{pred} to causal mediation analysis. Given a binary exposure E, a mediator M, an outcome Y, and a set of baseline

confounders X, Tchetgen Tchetgen and Shpitser (2012) define the "mediation functional" as

$$\theta = \iint \mathbb{E}[Y|E=1, M=m, X=x] dP(m|E=0, X=x) dP(x).$$

These authors then show that in the nonparametric model for (X, E, M, Y), the EIF of θ is

$$\begin{aligned}
\text{EIF}(\theta) =& \frac{\mathbb{I}(E=1)p(M|E=0,X)}{\Pr[E=1|X]p(M|E=1,X)}\big(Y - \mathbb{E}[Y|X,M,E=1]\big) \\
&+ \frac{\mathbb{I}(E=0)}{\Pr[E=0|X]}\big(\mathbb{E}[Y|X,M,E=1] - \mathbb{E}\big[\mathbb{E}[Y|X,M,E=1]\big|E=0,X\big]\big) \\
&+ \mathbb{E}\big[\mathbb{E}[Y|X,M,E=1]\big|E=0,X\big] - \theta
\end{aligned}$$

Note that when $X = \varnothing$, θ is isomorphic to μ_B^{pred} (by viewing E as R and M as X). Thus, the EIF of μ_B^{pred} can be obtained by setting X to $\varnothing$ in the above expression and then renaming E as R and M as X, which gives

$$\begin{aligned}
\text{EIF}(\mu_B^{\text{pred}}) =& \frac{\mathbb{I}(R=1)p(X|R=0)}{\Pr[R=1|X]p(X|R=1)}\big(Y - \mathbb{E}[Y|R=1,X]\big) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big(\mathbb{E}[Y|R=1,X] - \mu_B^{\text{pred}}\big) \\
=& \frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{\Pr[R=0|X]}{\Pr[R=1|X]}(Y - \mathbb{E}[Y|R=1,X]) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}(\mathbb{E}[Y|R=1,X] - \mu_B^{\text{pred}}).
\end{aligned}$$

C Double Robustness of $\hat{\mu}_B^{\mathbf{pred},DR}$

To simplify notation, let us recode Black as 1 and White as 0, and let $\mu_R(X)$ and $\pi(X)$ denote the true outcome and group membership models

$$\begin{aligned}\mu_R(X) &\equiv \mathbb{E}[Y|R,X],\\ \pi(X) &\equiv \Pr[R=1|X].\end{aligned}$$

The estimator $\hat{\mu}_B^{\text{pred},DR}$ can then be written as

$$\hat{\mu}_B^{\text{pred},DR} = \frac{1}{n}\sum_{i=1}^{n}\big(\frac{\mathbb{I}(R_i=1)}{\widehat{\Pr}[R_i=0]}\frac{1-\hat{\pi}(X_i)}{\hat{\pi}(X_i)}(Y_i-\hat{\mu}_1(X_i)) + \frac{\mathbb{I}(R_i=0)}{\widehat{\Pr}[R_i=0]}\hat{\mu}_1(X_i)\big),$$

where $\hat{\mu}_1(\cdot)$ and $\hat{\pi}(\cdot)$ denote the estimated outcome and group membership models.

Let (μ_1^*, π^*) denote the probability limits of $(\hat{\mu}_1, \hat{\pi})$. A first-order Taylor expansion of $\hat{\mu}_B^{\text{pred},DR}$ yields

$$\hat{\mu}_B^{\text{pred},DR} = \frac{1}{n}\sum_{i=1}^{n}\big(\frac{\mathbb{I}(R_i=1)}{\Pr[R_i=0]}\frac{1-\pi^*(X_i)}{\pi^*(X_i)}(Y_i-\mu_1^*(X_i)) + \frac{\mathbb{I}(R_i=0)}{\Pr[R_i=0]}\mu_1^*(X_i)\big) + o_p(1).$$

Hence it suffices to show

$$\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1^*(X)) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1^*(X)\big] = \mu_B^{\text{pred}}$$

whenever $\mu_1^* = \mu_1$ or $\pi^* = \pi$. Consistency then follows from the law of large numbers.

First, if $\mu_1^* = \mu_1$, we have

$$\begin{aligned}
&\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1^*(X)) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1^*(X)\big]\\
=&\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1(X)) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1(X)\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1(X)) + \frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1(X)\big|X\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1(X))\big|X\big] + \mathbb{E}_X\big[\mathbb{E}\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1(X)\big|X\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{\pi(X)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1(X))\big|X,R=1\big] + \mathbb{E}_X\big[\mathbb{E}\frac{\Pr[R=0|X]}{\Pr[R=0]}\mu_1(X)\big|X,R=0\big]\\
=&\mathbb{E}_X\big[\frac{\pi(X)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}\mathbb{E}[(Y-\mu_1(X))\big|X,R=1]\big] + \int\mu_1(x)\frac{\Pr[R=0|x]}{\Pr[R=0]}p(x)dx\\
=&\mathbb{E}_X\big[\frac{\pi(X)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}\underbrace{\big(\mu_1(X)-\mu_1(X)\big)}_{=0}\big] + \underbrace{\int\mu_1(x)p(x|R=0)dx}_{=\mu_B^{\text{pred}}}
\end{aligned}$$

$$=\mu_B^{\text{pred}}.$$

Second, if $\pi^* = \pi$, we have

$$\begin{aligned}
&\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi^*(X)}{\pi^*(X)}(Y-\mu_1^*(X))+\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1^*(X)\big]\\
=&\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}(Y-\mu_1^*(X))+\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\mu_1^*(X)\big]\\
=&\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}Y+\big(\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big)\mu_1^*(X)\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}Y+\big(\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big)\mu_1^*(X)\big|X\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}Y\big|X\big]+\mathbb{E}_X\mathbb{E}\big[\big(\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big)\mu_1^*(X)\big|X\big]\\
=&\mathbb{E}_X\mathbb{E}\big[\frac{1-\pi(X)}{\Pr[R=0]}Y|X,R=1\big]+\mathbb{E}_X\Big[\mu_1^*(X)\cdot\mathbb{E}\big[\big(\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big)\big|X\big]\Big]\\
=&\mathbb{E}_X\Big[\frac{1-\pi(X)}{\Pr[R=0]}\cdot\mathbb{E}\big[Y|X,R=1\big]\Big]+\mathbb{E}_X\Big[\mu_1^*(X)\cdot\mathbb{E}\big[\big(\frac{\mathbb{I}(R=1)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{\mathbb{I}(R=0)}{\Pr[R=0]}\big)\big|X\big]\Big]\\
=&\int\mu_1(x)\frac{\Pr[R=0|x]}{\Pr[R=0]}p(x)dx+\mathbb{E}_X\Big[\mu_1^*(X)\cdot\big(\frac{\pi(X)}{\Pr[R=0]}\frac{1-\pi(X)}{\pi(X)}-\frac{1-\pi(X)}{\Pr[R=0]}\big)\Big]\\
=&\underbrace{\int\mu_1(x)p(x|R=0)dx}_{=\mu_B^{\text{pred}}}+\mathbb{E}_X\Big[\mu_1^*(X)\cdot\underbrace{\big(\frac{1-\pi(X)}{\Pr[R=0]}-\frac{1-\pi(X)}{\Pr[R=0]}\big)}_{=0}\Big]\\
=&\mu_B^{\text{pred}}.
\end{aligned}$$

Thus, $\hat{\mu}_B^{\text{pred},DR}$ is doubly robust — consistent when either the outcome model or the group membership model is correctly specified.

D Identification of $\mathbb{E}[Y^g|R=r]$ under Assumptions 1-3

$$
\begin{aligned}
\mathbb{E}[Y^g|R=r] &\underbrace{=}_{A.1} \mathbb{E}[Y(A^g)|R=r] \\
&= \int \mathbb{E}[Y(A^g)|R=r,x]dP(x|R=r) \\
&= \int \mathbb{E}[\sum_a Y(a)\mathbb{I}(A^g=a)|R=r,x]dP(x|R=r) \\
&= \int \big(\sum_a \mathbb{E}[Y(a)\mathbb{I}(A^g=a)|R=r,x]\big)dP(x|R=r) \\
&= \int \big(\sum_a \mathbb{E}[Y(a)\mathbb{I}(A^g=a)|R=r,x,A^g=a]\Pr[A^g=a|R=r,x]\big)dP(x|R=r) \\
&= \int \big(\sum_a \mathbb{E}[Y(a)|R=r,x,A^g=a]\Pr[A^g=a|R=r,x]\big)dP(x|R=r) \\
&\underbrace{=}_{A.2} \int \big(\sum_a \mathbb{E}[Y(a)|R=r,x]\Pr[A^g=a|R=r,x]\big)dP(x|R=r) \\
&\underbrace{=}_{A.2} \int \big(\sum_a \mathbb{E}[Y(a)|R=r,x,A=a]\Pr[A^g=a|R=r,x]\big)dP(x|R=r) \\
&\underbrace{=}_{A.1} \int \big(\sum_a \mathbb{E}[Y|R=r,x,A=a]\Pr[A^g=a|R=r,x]\big)dP(x|R=r) \\
&= \mathbb{E}_{X|R=r}\mathbb{E}_{A^g|R=r,X}\mathbb{E}[Y|R=r,X,A=A^g] \\
&= \int \big(\sum_a \frac{p^g(a|R=r,x)}{p(a|R=r,x)}\mathbb{E}[Y|R=r,X=x,A=a]p(a|R=r,x)\big)dP(x|R=r) \\
&= \mathbb{E}[\frac{p^g(A|R,X)}{p(A|R,X)}Y|R=r].
\end{aligned}
$$

E Efficient Influence Function of μ_r^g

Note that μ_r^g can be written as

$$\mu_r^g = \int \big(\sum_a \mathbb{E}[Y|r,x,a]p^g(a|r,x)\big)p(x|r)dx.$$

We now derive the efficient influence function (EIF) of μ_r^g in the nonparametric model through a perturbation approach introduced by Hines *et al.* (2022). First, rewrite μ_r^g as a functional of the true observed data distribution $\mathcal{P}$: $\psi(\mathcal{P}) \equiv \mu_r^g$. Then, consider a one-dimensional parametric submodel $\mathcal{P}_t$ characterized by the following density function:

$$f_t(r,x,a,y) = (1-t)p(r,x,a,y) + t\delta_{\tilde{r},\tilde{x},\tilde{a},\tilde{y}}(r,x,a,y),$$

where $p(r,x,a,y)$ is the true density function of (R,X,A,Y), and $\delta_{\tilde{r},\tilde{x},\tilde{a},\tilde{y}}(r,x,a,y)$ is the Dirac delta function with respect to a hypothetical data point $(\tilde{r},\tilde{x},\tilde{a},\tilde{y})$. Hines *et al.* (2022) show that the EIF at $(\tilde{r},\tilde{x},\tilde{a},\tilde{y})$ can be calculated as the derivative of $\psi(\mathcal{P}_t)$ with respect to t at the truth, i.e.,

$$\partial_t \psi(\mathcal{P}_t) = \frac{d\psi(P_t)}{dt}\Big|_{t=0}.$$

Following the chain and quotient rules for derivatives and the Bayes rule, we have

$$\begin{aligned}
\partial_t\psi(\mathcal{P}_t) =& \partial_t \int \big(\sum_a \mathbb{E}[Y|r,x,a]p^g(a|r,x)\big)p(x|r)dx \\
=& \int \Big(\big(\sum_a \partial_t \mathbb{E}_t[Y|r,x,a]p^g(a|r,x)\big)p(x|r) + \big(\sum_a \mathbb{E}[Y|r,x,a]p^g(a|r,x)\big)\partial_t p_t(x|r)\Big)dx \\
=& \int \Big(\big(\sum_a \frac{\delta_{\tilde{r},\tilde{x},\tilde{a}}(r,x,a)}{p(r,x,a)}(\tilde{y} - \mathbb{E}[Y|r,x,a])p^g(a|r,x)\big)p(x|r) \\
&+ \big(\sum_a \mathbb{E}[Y|r,x,a]p^g(a|r,x)\big)\frac{\delta_{\tilde{r}}(r)}{p(r)}\big(\delta_{\tilde{x}}(x) - p(x|r)\big)\Big)dx \\
=& \frac{\delta_{\tilde{r}}(r)}{p(r)}\Big(\int \big(\sum_a \frac{\delta_{\tilde{x},\tilde{a}}(x,a)}{p(a|r,x)p(x|r)}(\tilde{y} - \mathbb{E}[Y|r,x,a])p^g(a|r,x)\big)p(x|r)dx \\
&+ \int \big(\sum_a \mathbb{E}[Y|r,x,a]p^g(a|r,x)\big)\big(\delta_{\tilde{x}}(x) - p(x|r)\big)dx\Big) \\
=& \frac{\delta_{\tilde{r}}(r)}{p(r)}\Big(\frac{p^g(\tilde{a}|r,\tilde{x})}{p(\tilde{a}|r,\tilde{x})}\big(\tilde{y} - \mathbb{E}[Y|r,\tilde{x},\tilde{a}]\big) + \sum_a \mathbb{E}[Y|r,\tilde{x},a]p^g(a|r,\tilde{x}) - \mu_r^g\Big).
\end{aligned}$$

Thus, the EIF of μ_r^g at (R,X,A,Y) is

$$\frac{\mathbb{I}(R=r)}{\Pr[R=r]}\Big(\frac{p^g(A|R,X)}{p(A|R,X)}\big(Y - \mathbb{E}[Y|R,X,A]\big) + \sum_a \mathbb{E}[Y|R,X,a]p^g(a|R,X) - \mu_r^g\Big)$$

F Derivation of Bias Formulas for Sensitivity Analysis

Assume we have a binary unobserved confounder, U, for the treatment-outcome relationship. We further assume that (i) $\alpha_r \stackrel{\Delta}{=} \Pr[U=1|x,r,A=1] - \Pr[U=1|x,r,A=0]$ does not depend on x and that (ii) $\beta_r \stackrel{\Delta}{=} \mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]$ does not depend on x or a.

Consider first the post-intervention mean, $\mathbb{E}[Y^g|R=r]$, which is identified as

$$\mathbb{E}[Y^g|R=r] = \int \Big[\sum_{a=0,1} p^g(a|r,x)\big(\mathbb{E}[Y|x,r,a,U=1]\Pr[U=1|x,r] + \mathbb{E}[Y|x,r,a,U=0]\Pr[U=0|x,r]\big)\Big] dP(x|r).$$

By contrast, our estimator of $\mathbb{E}[Y^g|R=r]$ converges to

$$\begin{aligned}\tilde{\mathbb{E}}[Y^g|R=r] = \int \Big[\sum_{a=0,1} p^g(a|r,x)\big(&\mathbb{E}[Y|x,r,a,U=1]\Pr[U=1|x,r,a] \\ &+ \mathbb{E}[Y|x,r,a,U=0]\Pr[U=0|x,r,a]\big)\Big] dP(x|r).\end{aligned}$$

Taking the difference of the above two quantities, we have

$$\begin{aligned}&\text{bias}(\mathbb{E}[Y^g|R=r]) \\ =&\int \Big[\sum_{a=0,1} p^g(a|r,x)\big(\mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]\big)\big(\Pr[U=1|x,r,a] - \Pr[U=1|x,r]\big)\Big] dP(x|r) \\ =&\int \Big[\sum_{a=0,1} p^g(a|r,x)\big(\mathbb{E}[Y|r,x,a,U=1] - \mathbb{E}[Y|r,x,a,U=0]\big) \\ &\big(\Pr[U=1|x,r,a] - \Pr[U=1|x,r,1-a]\big)p(1-a|x,r)\Big] dP(x|r)\end{aligned}$$

Then, applying assumptions (i) and (ii), we have

$$\begin{aligned}&\text{bias}(\mathbb{E}[Y^g|R=r]) \\ =&\int \big[\sum_{a=0,1} p^g(a|r,x)(-1)^{1-a}\alpha_r\beta_r p(1-a|x,r)\big] dP(x|r) \\ =&\alpha_r\beta_r \int \big(\Pr[A=0|x,r]\Pr[A^g=1|x,r] - \Pr[A=1|x,r]\Pr[A^g=0|x,r]\big) dP(x|r) \\ =&\alpha_r\beta_r \int \big(\Pr[A^g=1|x,r] - \Pr[A=1|x,r]\big) dP(x|r) \\ =&\alpha_r\beta_r\big(\Pr[A^g=1|r] - \Pr[A=1|r]\big).\end{aligned}$$

Since the interventional disparity (Δ^g) is defined as $\mathbb{E}[Y^g|R=\text{White}] - \mathbb{E}[Y^g|R=\text{Black}]$, $\text{bias}(\Delta^g)$ can be written as:

$$\begin{aligned}\text{bias}(\Delta^g) =&\alpha_{\text{White}}\beta_{\text{White}}\big(\Pr[A^g = 1|\text{White}] - \Pr[A = 1|\text{White}]\big)\\ &- \alpha_{\text{Black}}\beta_{\text{Black}}\big(\Pr[A^g = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big).\end{aligned}$$

Similarly, the bias of the estimated gap reduced $(\Delta - \Delta^g)$ can be written as:

$$\begin{aligned}\text{bias}(\Delta - \Delta^g) =& - \alpha_{\text{White}}\beta_{\text{White}}\big(\Pr[A^g = 1|\text{White}] - \Pr[A = 1|\text{White}]\big)\\ &+ \alpha_{\text{Black}}\beta_{\text{Black}}\big(\Pr[A^g = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big).\end{aligned}$$

If we further assume that $\alpha_{\text{White}} = \alpha_{\text{Black}} = \alpha$ and $\beta_{\text{White}} = \beta_{\text{Black}} = \beta$, these bias formulas will reduce to

$$\text{bias}(\Delta^g) = \alpha\beta C,$$

and

$$\text{bias}(\Delta - \Delta^g) = -\alpha\beta C,$$

where $C = \big(\Pr[A^g = 1|\text{White}] - \Pr[A = 1|\text{White}]\big) - \big(\Pr[A^g = 1|\text{Black}] - \Pr[A = 1|\text{Black}]\big)$.

G Parametric Estimates of the Generalized KOB Decomposition

In the main text, we estimated the generalized KOB decomposition using the doubly robust (DR) estimator for $\mu_{B,j}^{\text{pred}}$ combined with cross-fitting and machine learning estimates of the outcome and group membership models. The use of machine learning estimators in tandem with cross-fitting and the DR estimator is desirable because it leads to an instance of what Chernozhukov *et al.* (2018) call "double/debiased machine learning" (DML) — a procedure that is robust, efficient, and has theoretically justified analytical standard errors. As mentioned in the main text, it is also possible to obtain estimates of the generalized KOB decomposition with parametric models applied to the regression-imputation (RI), weighting, or DR estimator. In this case, the DR estimator enjoys the "double-robustness" property described in the main text.

We have implemented these alternative estimators for our running example. The results are reported in Figures G1 and G2. In general, the results obtained under different estimation procedures are similar, although we see that the weighting estimator is generally less efficient than the parametric RI and DR estimators. As a combination of RI and weighting, the DR estimates tend to lie between estimates from the other two methods.

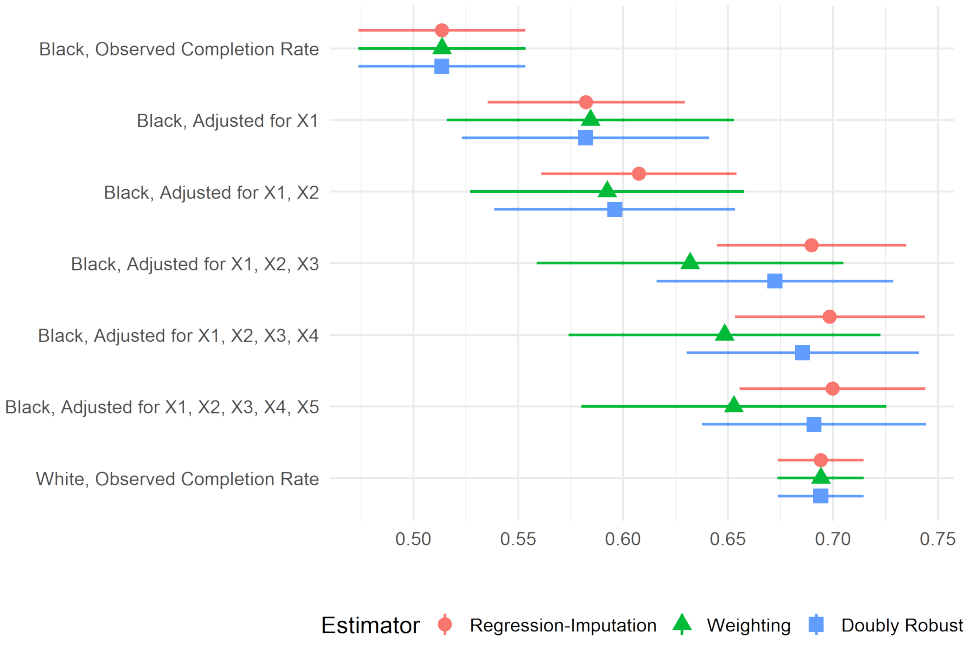


Figure G1: Parametric estimates of observed and adjusted completion rates. The outcome and group membership models are both fitted with logistic regression. Line ranges show 95% Wald-type confidence intervals, which are based on standard errors constructed from the nonparametric bootstrap (with 250 replications).



Figure G2: Parametric estimates of the explained and unexplained components of the Black-White gap in college completion. The outcome and group membership models are both fitted with logistic regression. Line ranges show 95% Wald-type confidence intervals, which are based on standard errors constructed from the nonparametric bootstrap (with 250 replications).

H Parametric Estimates of the Gap Reduced ($\Delta - \Delta^g$)

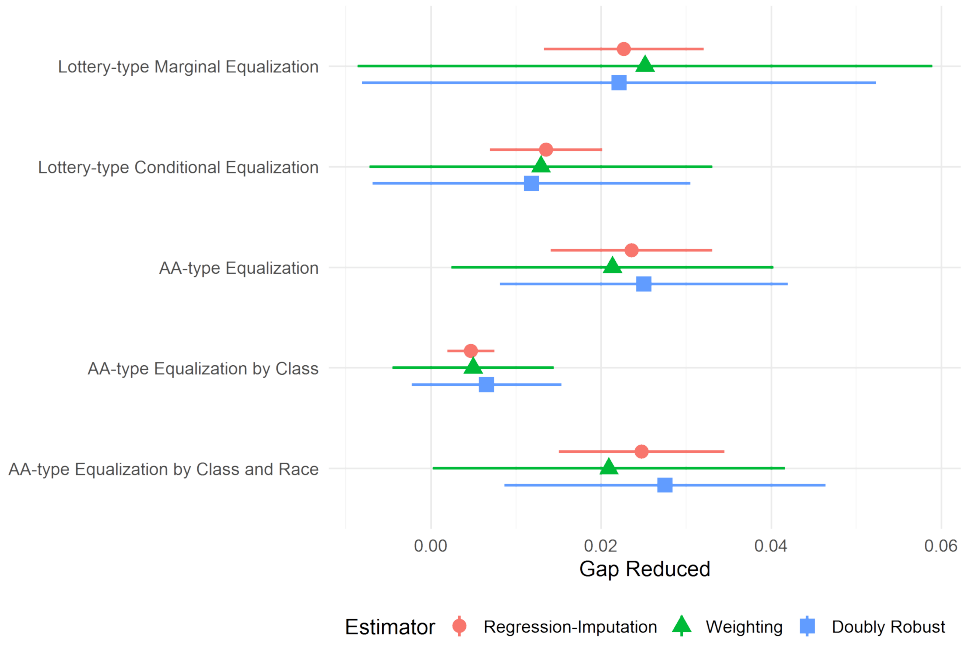


Figure H3: Parametric estimates of the gap reduced under different stylized interventions to selective college attendance. The outcome and treatment models are both fitted with logistic regression. Line ranges show 95% Wald-type confidence intervals, which are based on standard errors constructed from the nonparametric bootstrap (with 250 replications).