Isolating the Roles of Individual Covariates in Reweighting Estimation

Todd E. Elder, John H. Goddeeris, and Steven J. Haider
Michigan State University

10 October 2011

Abstract

A host of recent research has used reweighting methods to analyze the extent to which observable characteristics predict between-group differences in the distribution of an outcome. Much less attention has been paid to using reweighting methods to isolate the roles of individual covariates. We analyze two approaches that have been used in previous studies, and we propose an approach that can be viewed as a generalization of regression-based methods. We illustrate the differences between the methods with Monte Carlo evidence and an empirical analysis of black-white wage differentials among males.

Keywords: reweighting, inverse probability weights, decompositions
JEL: J31, J24, J15, J16

We thank Fabian Lange, Peter Schmidt, Gary Solon, and numerous seminar participants for detailed comments on earlier drafts. A version of this paper was presented at the 8th World Congress of the International Health Economics Association, July 2011.

1. Introduction

Many studies have adopted the use of reweighting methods to analyze the role of observable covariates for "predicting" or "explaining" outcome differences across groups or over time. DiNardo, Fortin, and Lemieux's seminal study (1996; DFL hereafter) developed and applied a reweighting method to assess how changes in the distribution of observed worker characteristics contributed to increases in wage inequality during the 1980s. Subsequently, researchers have applied reweighting estimators to analyze between-group outcome differences in a variety of contexts (see, e.g., Altonji, Bharadwaj, and Lange (2010), Barsky et al. (2002), Biewen (2001), Chiquiar and Hanson (2005), Elder, Goddeeris, and Haider (2011), Firpo, Fortin, and Lemieux (2007), Fortin, Lemieux, and Firpo (2010), and Machado and Mata (2005)).[1]

In all of these applications, a primary focus has been the estimation of overall predicted group differences, which involves assessing how much of the difference in the distributions of outcomes across groups can be predicted by differences in the distributions of all covariates. In contrast, much less attention has been paid to methods for isolating the roles of individual covariates (or other subsets of the full covariate set), although this question is often of interest. For example, one of the key questions raised in DFL was "how much of the increase in wage inequality during the 1980s can be accounted for by changes in the prevalence of unionization?"

In this paper, we examine three approaches to isolating the roles of individual covariates in reweighting estimation; one used by Machado and Mata (2005), one proposed by Fortin, Lemieux, and Firpo (2010; FLF hereafter), and an approach we propose. Our approach can be viewed as a generalization of regression-based methods, where the role of one covariate is examined while holding the other covariates constant. For example, when the underlying model

[1] Reweighting estimators have also been adopted in the program evaluation literature as tools for estimating treatment effects under "selection on observables" assumptions. In that literature, reweighting estimators are commonly referred to as "inverse probability weighting" (IPW) estimators. See Hirano, Imbens, and Ridder (2003) and Wooldridge (2007) for methodological developments in this context.

of outcomes is linear, the implied roles of individual covariates for mean outcome differences based on our approach are asymptotically equivalent to those based on linear regressions, which is not generally true of the other two methods. We demonstrate our methods with Monte Carlo evidence and an illustrative empirical example based on wage differences between black and white men.

2. Reweighting Based on a Full Set of Covariates

In this section, we first introduce our notation for reweighting estimation, and then we introduce an empirical example to focus ideas.

2.1. A Reweighting Framework

Assume that there are two groups, denoted *A* and *B*, and suppose that a researcher is interested in examining the difference between these groups in the density of an outcome *y*. Consider a model in which the outcome *y* in each group is related to a vector of covariates *x*. In order to use a reweighting approach to assess how the group-level difference in the distribution of *x* contributes to the group-level difference in the density of *y*, one could reweight group *A* to have the distribution of *x* found in group *B*.[2] As DFL show, this reweighted density of *y* corresponds to the counterfactual density that would hold if group *A* had group *B*'s distribution of *x* but its own mapping from *x* to *y*. Letting $j_{y|x}$ refer to the group whose mapping from *x* to *y* is used and j_x denote the group whose distribution of *x* is used, the counterfactual density can be written as

[2] One might instead reweight group *B* to match the distribution of *x* found in group *A*, and this alternative choice of reference group would typically lead to different inferences because of differences across groups in the mapping from covariates to outcomes. Because we focus on issues related to differences in the distribution of covariates, for simplicity we maintain that group *A* is the reference group throughout.

$$
\begin{aligned}
& f(y; j_{y|x} = A, j_x = B) \\
& \quad \equiv \int_x f(y \mid j_{y|x} = A, x) dF(x \mid j = B) \\
& \quad = \int_x f(y \mid j_{y|x} = A, x) \psi(x) dF(x \mid j = A), \\
& \text{where } \psi(x) = \frac{dF(x \mid j = B)}{dF(x \mid j = A)}.
\end{aligned} \tag{1}
$$

The equality in (1) assumes a "common support" condition that requires all values of x observed in group B to also be observed in group A; if this condition does not hold, it is impossible to reweight group A to have the distribution of x found in group B. All reweighting methods require some form of a support assumption, an issue we consider in more detail below. The third line of (1) shows that the counterfactual density can be obtained by reweighting group A's population, with weights $\psi(x)$ that depend only on the values of the covariates.

In practice, the weights $\psi(x)$ are usually constructed using a substitution that follows from Bayes' rule:

$$
\psi(x) = \frac{dF(x \mid j = B)}{dF(x \mid j = A)} = \frac{\Pr(j = B \mid x) / \Pr(j = B)}{\Pr(j = A \mid x) / \Pr(j = A)}. \tag{2}
$$

This expression implies that the weights can be calculated from estimated probabilities of group membership conditional on x. Once these weights are obtained, the calculated contribution of between-group differences in all observable characteristics, sometimes referred to as the "aggregate decomposition", is the difference between the actual density of outcome y for group A, $f(y \mid j = A)$, and the counterfactual density $f(y; j_{y|x} = A, j_x = B)$. Hereafter, we denote the reweighted density of y using weights w as $f(y; w \mid j = A)$, so $f(y; j_{y|x} = A, j_x = B)$ may be equivalently rewritten as $f(y; \psi \mid j = A)$. One may calculate any statistic based on this counterfactual distribution and compare it to the analogous statistic based on the factual distribution.

2.2. An Application to the Black-White Male Wage Gap

To focus these ideas, consider the well-known wage gap between black and white males. A large literature has established that some of the gap is predictable based on differences in characteristics such as education, labor market experience, marital status, industry, and occupation.[3] We analyze these black-white wage differences for males aged 25-59 employed in the civilian labor force using the 1 percent PUMS file of the 2000 Census data distributed by IPUMS USA (Ruggles et al. (2010)). We define group *A* to be white males and group *B* to be black males. The outcome measure *y* is the logarithm of average hourly wages (annual earnings divided by usual hours worked per week and by weeks worked last year), and we analyze the roles of five covariates: education, experience, marital status, industry, and occupation. Here and elsewhere in the paper, each covariate is represented by one or more indicator variables, e.g., education is represented by indicators for <12 years, 12 years, 13-15 years, and ≥16years.[4] Given the large sample sizes available in the 2000 PUMS data, we use a 10% random sample and exclude observations with missing data on any of the five covariates or wages. We also exclude observations with hourly wages below $1 and above $3000. Our final analysis sample consists of 213,908 observations for whites and 23,945 observations for blacks.

Table 1 shows our basic sample information. The first two columns in the table show sample means of log wages and each of the covariates, separately by race. Average log wages are .240 higher among whites than among blacks. Whites are also more highly educated, more likely to be married, and more likely to work in relatively high-wage industries and occupations.

[3] Altonji and Blank (1999) provide a detailed review of this literature.

[4] Potential experience is defined as age minus 6 minus the number of years of education (using the algorithm of Angrist et al. (2011) to define years of education) and then grouped into categories of <10 years, 10-14 years, 15-19 years, 20-24 years, 25-29, years, 30-34 years and ≥35 years. Marital status is represented by a binary indicator. Occupation is represented by the five broad categories used in the 2000 Census. Industry is represented by eight categories, which are taken from the twenty sectors defined in the 1997 North American Industry Classification System (see Table 1).

To illustrate how reweighting methods work, the third column presents means of the counterfactual white population in which whites are reweighted to have the black distribution of all covariates. As is readily apparent, full reweighting gives whites very similar characteristics to blacks.[5] Taken together, these characteristics predict about 59 percent of the overall mean log wage gap (= (2.822 – 2.681) / (2.822 – 2.582)).

Of course, if one were only interested in mean differences, then flexibly specified regression models would suffice. An advantage of reweighting methods is that differences in the entire distribution of *y* can be examined. As an illustration, Figure 1 plots the densities of log wages for whites, blacks, and reweighted whites. Reweighting shifts the white density to the left towards the black density, but the magnitude of the shift is not uniform across the support. Specifically, the density appears to shift more in the upper tail than in the lower tail. To see this, note that the reweighted white density essentially matches the black density at values above roughly 3.75, but the reweighted white density lies far from the black density (and relatively close to the unweighted white density) below roughly 1.5. These patterns imply that the covariates can explain more of the black-white differences in the upper tail of the distribution than they can in the lower tail.

Up to this point, we have been considering the joint role of all covariates. However, given that much of the black-white wage gap can be explained by these five variables, an obvious additional question arises: what roles do each of these covariates play in the closing of the wage gap? The rest of the paper focuses on approaches to answering this question.

[5] The reweighted white means are not identical to the black means because the logit model of group membership used to generate weights is not fully saturated, i.e., it includes all of the covariates in Table 1 but not interactions among them.

3. Isolating the Role of Individual Covariates

Before considering how to isolate the role of individual covariates with reweighting methods, we first illustrate what this means in the context of regression. Consider a simple data generating process in which y is linearly related to two binary covariates, denoted x_1 and x_2, and that between-group differences in the mean of y can arise due to differences in the means of those covariates and to an intercept shift. Specifically, suppose

$$y_i = \beta_0 + \beta_d d_i + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i, \tag{3}$$

where d_i equals 1 if individual i is a member of group A and equals zero otherwise, and u_i is an error term that is orthogonal to d_i, x_{1i}, and x_{2i}.[6] In this case, the group difference in the expectation of y can be written as

$$\begin{aligned} &E(y \mid j = A) - E(y \mid j = B) = \\ &\quad \beta_d + \underbrace{\beta_1 \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right]}_{\text{Role of } x_1} + \underbrace{\beta_2 \left[E(x_2 \mid j = A) - E(x_2 \mid j = B)\right]}_{\text{Role of } x_2}, \end{aligned} \tag{4}$$

where β_d represents the component that is unrelated to the two covariates, and the role of each covariate is defined in the typical way. Altonji and Blank's (1999) influential study includes such a decomposition to examine the "differences due to characteristics" between black and white log wages (p. 3159).

3.1. Isolating the Role of Covariates in Reweighting Estimation

Our goal is to isolate the role of a particular covariate in the context of reweighting estimators. Let z denote a particular covariate from the vector of covariates x, and let x_{-z} denote the vector of remaining covariates. We consider three methods for isolating the role of z, two used previously in the literature and a third we propose here.

[6] When the coefficients β_1 and β_2 vary across groups, decompositions of between-group mean differences into the roles attributable to individual covariates are not unique. See Oaxaca (1973) and Blinder (1973) for early discussions of this issue and Elder, Goddeeris, and Haider (2010) for a more recent contribution. The central findings presented below are unaffected if β_1 and β_2 vary across groups, so we impose that they are group-invariant for simplicity.

Method 1 ("First In"). Machado and Mata (2005) use an approach that reweights using only the covariate z.[7] This approach answers the question:

> "What would be group A's density of y if it had group B's distribution of z but its own distribution of x_{-z} conditional on z?"[8]

In the context of reweighting, this approach identifies the counterfactual density

$$
\begin{aligned}
& f(y; j_{y|z,x_{-z}} = A, j_{x_{-z}|z} = A, j_z = B) \\
& \quad \equiv \int_z \int_{x_{-z}|z} f(y \mid j = A, z, x_{-z}) dF(x_{-z} \mid z, j = A) dF(z \mid j = B) \\
& \quad = \int_z \int_{x_{-z}|z} f(y \mid j = A, z, x_{-z}) dF(x_{-z} \mid z, j = A) \psi^{z1}(z, x_{-z}) dF(z \mid j = A), \\
& \text{where } \psi^{z1}(z, x_{-z}) = \frac{dF(z \mid j = B)}{dF(z \mid j = A)}.
\end{aligned} \tag{5}
$$

As is apparent from the weighting function $\psi^{z1}(z, x_{-z})$, this approach involves reweighting the population of group A in order to match the marginal distribution of z observed in group B. One then assesses the role of z by comparing $f(y \mid j = A)$ to $f(y; \psi^{z1} \mid j = A)$.

As FLF point out, however, this approach has a serious drawback: to the extent that z is correlated with elements of x_{-z}, the method attributes to z both the effect of differences between groups in z and the effect of differences in x_{-z} that are correlated with z.[9] As an illustration, assume that z represents marital status, that married workers have greater educational attainment in both groups than do non-married workers, and that workers in group A are both more likely to be married and more highly educated than workers in group B. Then, reweighting workers in group A to match group B's marriage rate – by assigning relatively large weights to non-married

[7] Machado and Mata's approach differs from DFL's in other ways, including the use of quantile regression to estimate parametric models of the mapping from covariates to outcomes. We adapt their approach to a reweighting context for ease of comparison.

[8] From this point forward, we leave implicit that the counterfactual population retains group A's mapping from $\{z, x_{-z}\}$ to y, which will be the case for all of the counterfactuals described below.

[9] FLF write that this method "is invalid as a way of performing the decomposition for the same reason that [a regression-based] decomposition would be invalid if the coefficient used for one covariate was estimated without controlling for the other covariates" (p. 61).

workers and relatively small weights to married workers – also shifts the educational distribution of group *A*. The resulting counterfactual population of workers is both less likely to be married *and* less educated than the actual population of workers in group *A*, so the estimated effect of marital status will also capture some of the effect of education.

Method 2 ("Last In"). FLF propose an approach that is also a variant of standard reweighting methods, but instead is based on excluding the covariate z.[10] This approach answers the question:

> "What would be group *A*'s density of y if it had group *B*'s distribution of z conditional on x_{-z} but its own distribution of x_{-z}?"

which corresponds to the counterfactual density

$$
\begin{aligned}
& f(y; j_{y|z,x_{-z}} = A, j_{x_{-z}} = A, j_{z|x_{-z}} = B) \\
& \quad \equiv \int_{z|x_{-z}} \int_{x_{-z}} f(y \mid j = A, z, x_{-z}) dF(x_{-z} \mid j = A) dF(z \mid x_{-z}, j = B) \\
& \quad = \int_{z|x_{-z}} \int_{x_{-z}} f(y \mid j = A, z, x_{-z}) dF(x_{-z} \mid j = A) \psi^{z2}(z, x_{-z}) dF(z \mid x_{-z}, j = A), \\
& \text{where } \psi^{z2}(z, x_{-z}) = \frac{dF(z \mid x_{-z}, j = B)}{dF(z \mid x_{-z}, j = A)}.
\end{aligned} \tag{6}
$$

The weights in (6) can be calculated by dividing weights produced using the full x vector by weights that use all covariates except z. One then assesses the role of z by comparing $f(y \mid j = A)$ to $f(y; \psi^{z2} \mid j = A)$.

In the context of our example in which z represents marital status, the resulting counterfactual population has the same distribution of education (and all other covariates in x_{-z}) found in group *A*, so this approach does not mistakenly attribute to marital status an effect that is instead due to other characteristics that are correlated with marital status. However, this method has its own limitation: although the counterfactual population has group *B*'s distribution of

[10] This is also the approach that DFL took in answering the question we cite above: "how much of the increase in wage inequality during the 1980s can be accounted for by changes in the prevalence of unionization?" Antecol et al. (2008) also use this approach in assessing the role of occupational sorting in the sexual orientation wage gap.

marital status *conditional* on the other covariates, its marginal distribution of marital status does not match group *B*'s (its marriage rate is higher than that found in group *B*). More generally, the counterfactual population's marginal distribution of z will match that of group *B* in only two cases: when the two groups have identical marginal distributions of x_{-z}, or when z and x_{-z} are independent in group *B*. This is an unappealing feature of Method 2 if the underlying goal – analogous to the goal of a regression-based decomposition – is to assess what would happen to the distribution of outcomes if group *A* had the same distribution of z found in group *B*.

Method 3. We propose a method that is constructed to replicate the marginal distribution of z found in group *B* while retaining the marginal distribution of x_{-z} found in group *A*. This approach corresponds to the following question:

> "What would be group *A*'s density of y if it had group *B*'s marginal distribution of z but its own marginal distribution of x_{-z}?"

For now, we will assume that every value of z observed in group *B* is also observed in group *A* in combination with every value of x_{-z} observed in group *A*:

$$(7) \qquad dF(z, x_{-z} \mid j = A) > 0 \;\; \forall z, x_{-z} \text{ for which } dF(z \mid j = B) > 0 \text{ and } dF(x_{-z} \mid j = A) > 0.$$

If this support condition holds, a set of weights that isolates the role of z is given by

$$(8) \qquad \psi^{z3}(z, x_{-z}) = \frac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)},$$

which produces the counterfactual density

$$\begin{aligned} & f(y; \psi^{z3} \mid j = A) \\ (9) \qquad & \equiv \int_{z|x_{-z}} \int_{x_{-z}} f(y \mid j = A, z, x_{-z}) dF(z \mid x_{-z}, j = A) \psi^{z3}(z, x_{-z}) dF(x_{-z} \mid j = A) \\ & = \int_{z|x_{-z}} \int_{x_{-z}} f(y \mid j = A, z, x_{-z}) dF(z \mid j = B) dF(x_{-z} \mid j = A). \end{aligned}$$

One can then assess the role of z by comparing $f(y \mid j = A)$ to $f(y; \psi^{z3} \mid j = A)$. The $\psi^{z3}(z, x_{-z})$ weights include group *B*'s marginal distribution of z in the numerator, rather than the

conditional (on x_{-z}) distribution of z as in the $\psi^{z2}(z, x_{-z})$ weights. As a result, the $\psi^{z3}(z, x_{-z})$ weights produce a counterfactual distribution that matches group *B*'s marginal distribution of z rather than its conditional distribution. Moreover, in this counterfactual distribution, z and x_{-z} are independent, as is apparent from the last line of (9). See Appendix A1 for proofs of these two claims.

We make two remarks at this point. First, the $\psi^{z3}(z, x_{-z})$ weights are not unique in producing the desired counterfactuals; in Section 4 below, we consider two modifications to these weights that also produce counterfactual populations with the appropriate marginal distributions of z and x_{-z}. Second, unlike in Methods 1 and 2, we cannot use Bayes' rule to rewrite $\psi^{z3}(z, x_{-z})$ as a simple function of the probabilities of group membership. Instead, these weights must be estimated directly. For simplicity of exposition, we use discrete-valued covariates throughout the paper, so we use cell frequencies to construct the weights.[11] We discuss further details regarding how these weights are constructed below.

3.2. Monte Carlo Evidence

We next present Monte Carlo evidence on the performance of the different methods using the data generating process based on (3) above, but with one additional simplification: because the intercept shift is immaterial to these results, we assume $\beta_d = 0$:

(3') $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i.$$

To apply the three reweighting approaches, we augment (3') by a binary group membership equation

(10) $$\Pr(j_i = A \mid x_{1i}, x_{2i}) = \Pr(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + e_i > 0),$$

where e is an error term orthogonal to u, x_1 and x_2.

[11] Estimation of the densities in (8) does not require discrete-valued covariates. See Hall et al. (1999, 2004), Fan and Yim (2004), and Chernozhukov et al. (2010) for recent developments and applications of conditional density estimation for continuous variables.

We consider two different designs to highlight the differences between the approaches:

$$\text{Design 1: } \beta_1 = 0, \beta_2 = 1, \gamma_1 = 1, \gamma_2 = 1$$

$$\text{Design 2: } \beta_1 = 1, \beta_2 = 1, \gamma_1 = 0, \gamma_2 = 1.$$

We evaluate these designs with 200 Monte Carlo replications, each consisting of 10,000 draws of x_1, x_2, u, and e. In both designs, x_1 and x_2 have a correlation coefficient of 0.5.[12]

The first three columns of Table 2 show the results for Design 1. Panel A shows that the means of x_1, x_2, and y each equal .600 in group A and .300 in group B; however, x_1 plays no role in explaining the observed differences in y because $\beta_1 = 0$. Panel B shows the contribution of each covariate to the mean difference in y, with standard errors in parentheses. As shown in the first row of Panel B, the regression-based method based on (4) yields accurate results: the role of the group difference in x_1 is estimated to be very small and insignificantly different from zero. The next three rows in Panel B show the implied roles of x_1 and x_2 based on the three reweighting estimators. Method 1 incorrectly implies that x_1 contributes .148 to the between-group difference in $E(y)$, but both Methods 2 and 3 imply that x_1 contributes roughly zero.

The differences between the methods can also be observed by examining the sums of the contributions: the Method 1 sum is notably larger than that produced by regression, while the Method 2 sum is notably smaller. These comparisons are not surprising because Method 1 captures both the effect of the variable being changed and the effect of the other covariates that are correlated with that variable, while Method 2 captures only the conditional (on the other covariates) differences between groups in each covariate.[13]

---

[12] We chose values of γ_0 and var(e) in order to produce easily interpretable group-specific means of the covariates and y; the particular values chosen have no effect on the substantive results in this section. In both designs, Pr(x_1 = 1) = Pr(x_2 = 1) = 0.5, u and e are both normally distributed, $\beta_0 = 0$, and var(u) = 1. In Design 1, $\gamma_0 = -0.2$ and var(e) = 3.667, while in Design 2, $\gamma_0 = 0.11$ and var(e) = 1.729.

[13] In light of the variation across methods in the sums of the contributions, one might also analyze the contribution of each covariate as a fraction of the sum of the contributions across both covariates. In Design 1, the relative contributions implied by Method 2 are similar to those implied by regression: Method 2 implies relative roles of x_1 and x_2 of .042 and .958, respectively, while regression implies relative roles of .028 and .972. However, this

The next three columns of Table 2 show results for Design 2, in which x_1 and x_2 have identical causal effects on y. Because $\gamma_1 = 0$ in this design, x_1 is related to group membership only because it is correlated with x_2. Specifically, the between-group mean difference in x_1 (= .550 – .400) is half as large as the corresponding mean difference in x_2 (= .600 – .300) because the simple correlation between x_1 and x_2 is 0.5. As shown in Panel B, the regression-based approach implies that the role of x_1 is roughly half as large as the role of x_2, with x_1 contributing .155 to the between-group difference in $E(y)$. Compared to regression, Method 1 implies larger roles of both x_1 and x_2 (although the relative effect sizes are similar to the regression-based estimates). Method 2 incorrectly implies that x_1 plays no role in explaining group differences in $E(y)$.

To provide further insight into the differences between the three reweighting methods, Panel C shows the counterfactual means for the two covariates when group *A* is reweighted to isolate the role of x_1, i.e., when x_1 plays the role of z. For both designs, Method 3 returns the desired result: the counterfactual mean of x_1 matches the group *B* mean (.300 in design 1, .400 in design 2), and the counterfactual mean of x_2 matches the group *A* mean (.600 in both designs). With Method 1, however, the counterfactual mean of x_2 does not remain at .600, but instead moves towards the group *B* mean of .300 in both designs. As FLF describe, this shift occurs because x_1 and x_2 are correlated, so reweighting group *A* to match group *B*'s distribution of x_1 also shifts the distribution of x_2. As a result, the implied effect of x_1 captures some of the effect of x_2. In contrast, Method 2 ensures that x_2's mean remains at the group *A* mean, but x_1's mean does not fully move to the group *B* mean.

In summary, only Method 3 yields similar estimates to the regression-based approach in both designs. This equivalence is not a specific property of these designs, but a general property

---

adjustment is not a general solution because the differences between the reweighting methods and regression depend on the correlations among the regressors.

of these approaches. We show in Appendix A2 that, under a generalization of the linear outcome model given in (3'), the counterfactual expectation of *y* generated by Method 3 is asymptotically equivalent to the counterfactual expectation generated by linear regression. The implied role of a covariate in explaining between-group mean outcome differences is therefore asymptotically equivalent in the two approaches.

4. Further Issues in Reweighting Based on Method 3

In this section, we discuss the implications of two situations that can arise when applying Method 3: the underlying data generating process might include interactions among covariates, and the underlying support condition given in (7) may not hold.

4.1. Isolating the Roles of Individual Covariates in a Model with Interactions

The goal of our estimation exercise is to isolate the independent roles of individual covariates in between-group differences in *y*. However, when the underlying DGP contains interactions among covariates, there is a well-known ambiguity as to how these "independent roles" should be defined. Specifically, consider a generalization of the linear outcome model given in (3'), where x_1 and x_2 are again binary:

$$y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \beta_{12}(x_{1i} x_{2_i}) + u_i. \tag{11}$$

It is straightforward to show that

$$\begin{aligned} &E(y \mid j = A) - E(y \mid j = B) = a + b + c + d, \\ &\quad \text{where } a = \left[\beta_1 + \beta_{12} E(x_2 \mid j = A)\right] \times \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right] \\ &\qquad b = \left[\beta_2 + \beta_{12} E(x_1 \mid j = A)\right] \times \left[E(x_2 \mid j = A) - E(x_2 \mid j = B)\right] \\ &\qquad c = \beta_{12} \times \left[E(x_1 \mid j = B) - E(x_1 \mid j = A)\right] \times \left[E(x_2 \mid j = A) - E(x_2 \mid j = B)\right] \\ &\qquad\quad - \beta_{12} \times \operatorname{cov}(x_1, x_2 \mid j = B) \\ &\qquad d = \beta_{12} \times \operatorname{cov}(x_1, x_2 \mid j = A). \end{aligned} \tag{12}$$

In this expression, term *a* is the derivative of $E(y)$ with respect to x_1 (evaluated at the mean of x_2 in group *A*), multiplied by the between-group difference in the mean of x_1. As such, it can

reasonably be interpreted as the role of x_1 in explaining the between-group difference in the mean of $E(y)$. Similarly, term b can be interpreted as the role of x_2. However, terms c and d are not naturally assigned to either component, leading to the ambiguity in parsing between the roles of x_1 and x_2.

As we show in Appendix A3, using the $\psi^{z3}(z, x_{-z})$ weights to assess the role of x_1 in explaining the between group-differences in $E(y)$ in (11) gives a result that converges to

$$(13) \quad [\beta_1 + \beta_{12}E(x_2 \mid j = A)] \times [E(x_1 \mid j = A) - E(x_1 \mid j = B)] + \beta_{12}\text{cov}(x_1, x_2 \mid j = A),$$

which equals $a+d$ from (12). The component attributable to x_2 is defined symmetrically as $b+d$. As a result, the term d is attributed to *both* x_1 and x_2, while term c is attributed to neither covariate. It makes little sense to consider d to be as attributable to both variables, and this component need not be small in magnitude.

One approach to dealing with this issue is to eliminate d from (13), leaving only the intuitive term a. Fortunately, this is straightforward to accomplish with alternative weights that still produce the desired counterfactual, i.e., a population with group B's marginal distribution of x_1 and group A's marginal distribution of x_2. Consider reweighting group A using weights defined by $\psi_{AA}^{z3}(x_1, x_2) = \dfrac{dF(x_1 \mid j = A)}{dF(x_1 \mid x_2, j = A)}$. These weights are identical to $\psi^{z3}(x_1, x_2)$ except that group A's distributions are used in both the numerator and denominator. Using $\psi_{AA}^{z3}(x_1, x_2)$ to reweight group A yields a counterfactual population with group A's marginal distributions of x_1 and x_2, but with x_1 orthogonal to x_2. The analog to (13) that these weights produce equals

$$(14) \quad \begin{aligned} &[\beta_1 + \beta_{12}E(x_2 \mid j = A)] \times [E(x_1 \mid j = A) - E(x_1 \mid j = A)] + \beta_{12}\text{cov}(x_1, x_2 \mid j = A) \\ &\quad = \beta_{12}\text{cov}(x_1, x_2 \mid j = A). \end{aligned}$$

Inspection of (13) and (14) suggests that using a weight that involves the difference between $\psi^{z3}(x_1, x_2)$ and $\psi^{z3}_{AA}(x_1, x_2)$ would eliminate the covariance term, and we show in Appendix A3 that this is the case. Specifically, reweighting group *A* by

(15) $$\psi^{z3*}(x_1, x_2) = \psi^{z3}(x_1, x_2) - \psi^{z3}_{AA}(x_1, x_2) + 1$$

produces an analog of (13) that equals

(16) $$\left[\beta_1 + \beta_{12}E(x_2 \mid j = A)\right] \times \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right],$$

which is identical to term *a* in (12).

Like the counterfactual population produced by the ψ^{z3} weights, the ψ^{z3*} weights yield a counterfactual population that has group *B*'s marginal distribution of x_1 and group *A*'s marginal distribution of x_2 (although, unlike the counterfactual produced by the ψ^{z3} weights, x_1 and x_2 are not orthogonal). As we note in Appendix A3, this result can be extended to models with more than two covariates.

In situations in which there are interactions among covariates in their effects on *y* and the covariates are correlated with each other, $\psi^{z3*}(z, x_{-z})$ weights might be preferred to the $\psi^{z3}(z, x_{-z})$ weights. However, given the presence of $-\psi^{z3}_{AA}(x_1, x_2)$ in (15), some of the $\psi^{z3*}(z, x_{-z})$ weights may be negative. While negative weights could be used in the calculation of some objects, such as weighted means, they cannot be used for objects such as weighted densities because they imply that the associated *x* values have negative densities in the counterfactual population.[14] If negative weights arise, an approximate solution is to set them to zero and reduce other weights proportionately.

[14] Kline (2011), for example, does not discard negative weights in calculating a counterfactual mean based on a reweighting estimator.

4.2. Violations of the Support Condition

Up to this point, we have assumed that condition (7) holds. To consider the effects of violations of this condition using the general covariate set $x = \{z, x_{-z}\}$, we now assume that (7) does not hold, so that there exist values of z and x_{-z} for which $dF(z \mid j = B) > 0$ and $dF(x_{-z} \mid j = A) > 0$ but $dF(z \mid x_{-z}, j = A) = 0$. This implies that for some $\{z, x_{-z}\}$ combinations, the weights $\psi^{z3}(z, x_{-z}) = \dfrac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)}$ are undefined. Even though these combinations do not appear in the group A population, implying that the undefined weights do not need to be computed, they still present problems. Moreover, this support condition is stronger than the analogous support condition needed for full reweighting.[15]

To describe the problem and the approach we take, note that the marginal density of z in group j is given by

$$(17) \qquad dF(z \mid j) = \int_{x_{-z}} dF(z, x_{-z} \mid j),$$

and the marginal density of x_{-z} in group j is given by

$$(18) \qquad dF(x_{-z} \mid j) = \int_{z} dF(z, x_{-z} \mid j).$$

To see the problem that arises in applying the $\psi^{z3}(z, x_{-z})$ weights, consider $dF(x_{-z})$ in the reweighted population. This counterfactual density is given by the reweighted version of the right-hand side of (18) in group A:

---

[15] Heuristically, the support condition for full reweighting guarantees that $dF(x \mid j = A)$ is positive whenever $dF(x \mid j = B)$ is positive. When this condition fails, the weights $dF(x \mid j = B) / dF(x \mid j = A)$ are undefined for some values of x because the denominator is zero. When condition (7) fails, the $\psi^{z3}(z, x_{-z})$ weights are undefined for the same reason. The support condition for full reweighting is weaker in the sense that it requires only that a configuration of characteristics that appears in one population (the numerator) also appears in the other (the denominator). In contrast, the support condition in (7) requires that any value of one covariate that appears in one population (the numerator) must appear in combination with every combination of all other covariates in the other population (the denominator). This condition might require the existence of observations with combinations of characteristics that could not appear in either population.

$$
\begin{aligned}
&\int_z \psi^{z3}(z, x_{-z}) dF(z, x_{-z} \mid j = A) \\
&= \int_z \frac{dF(z \mid j = B)}{dF(z, x_{-z} \mid j = A) / dF(x_{-z} \mid j = A)} dF(z, x_{-z} \mid j = A) \\
&= dF(x_{-z} \mid j = A) \times \int_z \left[dF(z \mid j = B) \times 1(dF(z, x_{-z} \mid j = A) > 0) \right].
\end{aligned} \tag{19}
$$

When condition (7) holds, so that $dF(z, x_{-z} j = A) > 0$ for all values of z and x_{-z}, this expression trivially equals $dF(x_{-z} \mid j = A)$ because the integral of $dF(z \mid j = B)$ across all values of z equals 1. When condition (7) does not hold, however, the integral in the second line of (19) is not defined over all values of z, but only over those values of z for which $dF(z, x_{-z} \mid j = A) > 0$. The integral of $dF(z \mid j = B)$ over these values of z is less than one, so the counterfactual $dF(x_{-z})$ is less than its target value of $dF(x_{-z} \mid j = A)$. By similar reasoning, the counterfactual $dF(z)$ is less than its target value of $dF(z \mid j = B)$.

Fortunately, a solution again exists. Consider the last line of (19). For each value of x_{-z} for which at least one value of z does not appear, if one multiplies all weights $\psi^{z3}(z, x_{-z})$ associated with that value of x_{-z} by $\left\{ \int_z \left[dF(z \mid j = B) \times 1(dF(z, x_{-z} \mid j = A) > 0) \right] \right\}^{-1}$, then $dF(x_{-z})$ in the reweighted population will once again equal $dF(x_{-z} \mid j = A)$. Ensuring that all of the target marginal probabilities are matched becomes more complicated when multiple combinations of values of z and x_{-z} are not observed in population A. In Appendix A4, we describe a more complicated algorithm that builds on these ideas to accomplish this task. This algorithm does not guarantee that all weights are non-negative, but as we describe in the Appendix, negative weights are not empirically relevant in the model that we study.

Finally, it is straightforward to identify situations in which support problems are empirically relevant: one can check whether the distributions of the explanatory variables in the

reweighted populations match their target distributions. If the counterfactual population produced by the $\psi^{z3}(z, x_{-z})$ weights does not have the distribution of z found in group B or the distribution of x_{-z} found in group A, then it must be the case that the sample analog of condition (7) does not hold.[16]

5. Applying the Methods to the Black-White Wage Gap

To illustrate how the various methods work in practice, we return to the black-white wage gap example. The last three columns of Table 1 present sample means of counterfactual white populations generated by using Methods 1, 2, and 3 to isolate the role of education, i.e., education plays the role of z. Method 3 is implemented using the $\psi^{z3*}(z, x_{-z})$ weights. The patterns across the three methods mirror those shown above in the Monte Carlo designs. Specifically, Method 3 produces the desired distribution of characteristics: the distribution of education approximately matches that in the black population, while the distributions of experience, occupation, industry, and marital status match those in the white population.[17] Method 1 matches the desired distribution for education, but not for the other covariates. Method 2 matches the desired distribution for the other covariates, but not for education.

Table 3 presents the estimates for the roles of the five covariates in predicting the mean black-white log wage gap based on regression and each of the reweighting methods. According to all methods, the most important covariates are education, marital status and occupation. However, just as we found in the Monte Carlo example, Method 1 produces estimates of

[16] Violation in a sample does not imply that condition (7) fails to hold in the population. Whenever the sample analog does not hold, however, the unadjusted $\psi^{z3}(z, x_{-z})$ weights will not produce a proper counterfactual density.

[17] As noted above, one drawback to using the weights defined in (15), rather than those defined in (8), is that some weights can be negative. The percentages of weights that are negative when education, potential experience, marital status, occupation, and industry play the role of z are 2.6, 0.0, 0.0, 3.2, and 1.0, respectively. We convert all negative weights to zero, causing the reweighted white means to differ slightly from their target values. For example, the reweighted white means of the education variables are not identical to the corresponding black means; reweighted means calculated using the negative weights match their target values exactly.

contributions that are notably larger than those produced by regression, and Method 2 produces estimates that are notably smaller.[18] In contrast, Method 3 produces estimates that are most similar to regression for all covariates (the largest differences from regression are for education and occupation). Similarly, Method 3 produces a sum of contributions that is quite close to that of regression, while Method 1's is considerably larger and Method 2's is considerably smaller.

Of course, the motivation for using reweighting methods is that they can be used to produce entire counterfactual distributions, allowing one to answer questions that are not well-suited to a regression framework. Recall that Figure 1 showed a reweighted white distribution of log wages that was produced using the full set of covariates. Along with this fully reweighted distribution and the unweighted white distribution, Figure 2 shows two additional counterfactual distributions produced by Method 3. The first isolates the role of education, and the second isolates the role of marital status. The curve that isolates the role of education lies nearly halfway between the unweighted and "fully reweighted" curves, reflecting that racial differences in education account for nearly half of the predictable component of the black-white log wage gap. The curve that isolates the role of marital status lies nearly on top of the unweighted white curve, reflecting that marital status has a small (but nonzero) effect throughout the wage distribution.

6. Discussion and Conclusion

We analyze three reweighting methods for isolating the roles of individual covariates in producing between-group differences in outcome distributions. We show that two methods used

[18] We can approximate the Method 2 estimates using modified regression methods. Specifically, we multiply the vector of estimated log wage regression coefficients by the difference between the white means of each covariate and the means predicted for blacks if they had the white means of all other covariates. Similarly, we can approximate the Method 1 estimates by multiplying a vector of estimated log wage regression coefficients by the (unconditional) difference between the black and white means for each covariate, but in this case the vector of coefficients is estimated from a series of regressions of log wages on each covariate in isolation, i.e., 5 regressions in total. The implied roles of the covariates based on these regression-based approaches are similar to those shown in the "Method 2" and "Method 1" columns of the table.

in previous studies answer fundamentally different questions than does a regression-based approach, even when the underlying data generating process is linear and the object of interest is a mean outcome difference. These two reweighting methods also yield unintuitive estimates in the context of two simple Monte Carlo designs. In contrast, the reweighting estimator that we propose is a generalization of the regression-based approach, yielding asymptotically equivalent results for mean outcome differences when the underlying data generating process is linear.

We illustrate our approach using an empirical analysis of log wage differences between black and white adult males. As shown in Table 3, our proposed approach yields results that are similar to those produced by linear regression, while the other methods yield substantially different results. Using the same empirical example, we also show that it is straightforward to graph counterfactual log wage distributions that would result if whites had blacks' marginal distribution of one covariate but their own distribution of all other covariates.

We note two caveats about our approach. First, if a support condition is not satisfied, then the relatively simple weights we initially propose are no longer applicable. However, we present a modification of these weights that achieves the desired properties when the support condition is not satisfied.[19] Second, the simple weights also return results that are not easily interpretable when there are important unmodeled interactions among the covariates. Although we provide a more complicated set of weights to handle this situation, this modification can give rise to negative weights.

Although the results based on Method 3 are the most analogous to the familiar regression-based approach, we note that such results may not always be the object of interest. For example, suppose that one is interested in the effects of maternal age and education, both measured when an infant is born, on the black-white gap in infant mortality. If one wishes to

[19] As noted in Appendix A4, the modified weights are not guaranteed to be non-negative, but none of the weights were negative in the specifications we studied. This remained true if we instead used a 1 percent random sample from the underlying IPUMS data.

estimate the effect on white infant mortality rates of a shift from the white distribution of maternal age to that found among blacks, Method 3 does so in a way that holds the distribution of maternal education constant, as would a regression that included both covariates. However, because very young mothers will necessarily have relatively low educational attainment, it is difficult to imagine a policy change that would shift the share of mothers who are teens but would not also shift the distribution of maternal education. Thus, considering the combined direct and indirect (through education) effects of maternal age may also be of interest for analyses of the potential effects of policies that target the age distribution of mothers.[20]

Taken together, the analytic, Monte Carlo, and empirical results demonstrate that our proposed method for isolating covariates captures the spirit of a regression approach but allows for the flexibility of reweighting. Our method creates counterfactual distributions that shift the marginal distribution of one covariate at a time while holding the distributions of other covariates constant. For researchers interested in using reweighting methods to analyze overall between-group differences in an outcome, the methods developed here provide an attractive approach to further understand the role of each of the individual covariates.

[20] Altonji et al. (2010) reason along these lines in adopting a sequential approach to measuring the role of individual covariates on outcomes. To illustrate their approach, suppose that maternal education and child education are the two covariates of interest. Because maternal education is determined prior to and likely influences child education, it is arguably of interest to consider maternal education first in a sequential decomposition, recognizing that doing so will attribute to maternal education effects on outcomes that arise through correlated changes in child education.

References

Angrist, J. D., S. H. Chen, and J. Song (2011). "Long-Term Consequences of Vietnam-Era Conscription: New Estimates Using Social Security Data." American Economic Review 101(3): 334-38.

Altonji, J., P. Bharadwaj, and F. Lange (2010). "Changes in the Characteristics of American Youth - Implications for Adult Outcomes." Yale Department of Economics Working paper.

Antecol, H., A. Jong, and M. Steinberger (2008). "The Sexual Orientation Wage Gap: The Role Of Occupational Sorting And Human Capital." Industrial and Labor Relations Review 61(4): 518-543.

Barsky R., J. Bound, K.K. Charles, and J.P. Lupton (2002). "Accounting for the Black–White Wealth Gap." Journal of the American Statistical Association 97(459): 663-673.

Biewen, M. (2001). "Measuring the Effects of Socio-Economic Variables on the Income Distribution: An Application to the East German Transition Process." The Review of Economics and Statistics 83(1): 185-190.

Blinder, A. S. (1973). "Wage Discrimination: Reduced Form and Structural Estimates." Journal of Human Resources 8(4): 436-455.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2010). "Inference on Counterfactual Distributions." MIT Department of Economics Working Paper No. 08-16.

Chiquiar, D. and G. H. Hanson (2005). "International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States." Journal of Political Economy 113(2): 239-281.

DiNardo, J., N. M. Fortin and T. Lemieux, (1996). "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." Econometrica 64(5): 1001-44.

Elder, T., J. Goddeeris, and S. Haider (2010). "Unexplained Gaps and Oaxaca-Blinder Decompositions." Labour Economics 17(1): 284-290.

Elder, T., J. Goddeeris, and S. Haider (2011). "A Deadly Disparity: A Unified Assessment of the Black-White Mortality Gap." The B.E. Journal of Economic Analysis & Policy 11(1) (Contributions).

Fan, J. and Yim, T. H. (2004). "A Crossvalidation Method for Estimating Conditional Densities." Biometrika 91(4): 819-834.

Firpo, S., N. M. Fortin, and T. Lemieux (2007). "Decomposing Wage Distributions using Recentered Influence Functions Regressions." University of British Columbia Department of Economics Working paper.

Fortin, N., T. Lemieux, and S. Firpo (2010). "Decomposition Methods in Economics." National Bureau of Economic Research Working Paper No. 16045.

Gelbach, J. B. (2009). "When Do Covariates Matter? And Which Ones, and How Much?" Working paper.

Hall, P., J. Racine, J. and Q. Li (2004). "Cross-validation and the Estimation of Conditional Probability Densities." Journal of the American Statistical Association 99: 1015-1026.

Hall, P., R. C. L. Wolff, and Q. Yao (1999). "Methods for Estimating a Conditional Distribution Function." Journal of the American Statistical Association 94: 154-163.

Hirano K., H. Imbens, and G. Ridder (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." Econometrica 71(4): 1161-89.

Kline, P. (2011). "Oaxaca-Blinder as a Reweighting Estimator." American Economic Review Papers and Proceedings 101(3): 532-37.

Machado, J.A.F., and J. Mata (2005). "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression." Journal of Applied Econometrics 20(4): 445-465.

Oaxaca, R. (1973). "Male-Female Wage Differentials in Urban Labor Markets." International Economic Review 14(3): 693-709.

Ruggles S. J., T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek (2010). Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota.

Wooldridge, J. (2007). "Inverse Probability Weighted Estimation for General Missing Data Problems." Journal of Econometrics 141(2): 1281-1301.

Table 1: Unweighted and Weighted Sample Means, 2000 Census

	Unweighted		Reweighting Method			
				(z = Education)		
	Whites	Blacks	Full	Method 1	Method 2	Method 3
Log Hourly Wage	2.822	2.582	2.681	2.738	2.785	2.770
Years of education						
<12	.085	.127	.128	.127	.111	.124
12	.318	.405	.404	.405	.364	.401
13-15	.299	.310	.310	.310	.314	.316
16+	.298	.158	.158	.158	.211	.158
Occupation						
Management/ Professional	.335	.193	.192	.249	.335	.336
Service	.092	.176	.176	.101	.091	.091
Sales/Office	.160	.160	.159	.154	.160	.160
Farming/Fishing/ Forestry	.011	.008	.008	.013	.011	.011
Construction/ Maintenance	.185	.150	.150	.221	.184	.185
Production/ Transportation	.218	.314	.315	.262	.219	.218
Married	.713	.582	.582	.707	.714	.714

Table 1 (continued.): Unweighted and Reweighted Sample Means, 2000 Census

	Unweighted		Reweighting Method			
				(z = Education)		
	Whites	Blacks	Full	Method 1	Method 2	Method 3
Industry						
Agriculture / Mining / Construction	.165	.117	.117	.191	.165	.165
Manufacturing	.215	.202	.202	.231	.215	.215
Trade	.147	.122	.122	.153	.147	.147
Transportation and Warehousing	.063	.102	.103	.070	.063	.063
FIRE	.083	.075	.075	.069	.083	.083
Management / Administration	.087	.084	.083	.074	.087	.087
Education / Health Care	.152	.192	.193	.123	.152	.152
Arts and Other Services	.086	.105	.105	.091	.086	.086
Potential Experience						
<10	.126	.128	.125	.108	.126	.126
10-14	.147	.164	.162	.146	.147	.147
15-19	.163	.181	.180	.163	.163	.163
20-24	.174	.174	.175	.176	.174	.174
25-29	.163	.149	.151	.162	.163	.163
30-34	.126	.110	.112	.126	.126	.126
35+	.101	.093	.094	.119	.101	.101

Notes: The means in the column labeled as "Full Reweighting Method" are based on weights contructed from a logit model of group membership as a linear function of all of the covariates. N = 213,908 for whites and reweighted whites; N = 23,945 for blacks.

Table 2: Monte Carlo Evidence on Methods for Isolating the Role of Individual Covariates

	Design 1 $\beta_1 = 0, \beta_2 = 1, \gamma_1 = 1, \gamma_2 = 1$			Design 2 $\beta_1 = 1, \beta_2 = 1, \gamma_1 = 0, \gamma_2 = 1$		
	x_1	x_2	y	x_1	x_2	y
A: Unweighted sample means:						
Group A	.600	.600	.600	.550	.600	1.150
	(.002)	(.006)	(.013)	(.006)	(.006)	(.015)
Group B	.300	.300	.300	.400	.300	.700
	(.008)	(.008)	(.018)	(.008)	(.009)	(.024)
B: Contribution of covariates to mean difference in y						
Regression	.009	.298		.155	.305	
	(.009)	(.014)		(.012)	(.014)	
Method 1	.148	.293		.227	.443	
	(.008)	(.005)		(.011)	(.010)	
Method 2	.007	.152		.007	.223	
	(.005)	(.008)		(.010)	(.011)	
Method 3	.007	.297		.154	.303	
	(.012)	(.018)		(.014)	(.018)	
C: Reweighted group A sample means ($z = x_1$):						
Method 1	.300	.463		.400	.523	
	(.008)	(.009)		(.008)	(.008)	
Method 2	.461	.600		.539	.600	
	(.008)	(.008)		(.008)	(.006)	
Method 3	.300	.600		.400	.600	
	(.008)	(.006)		(.008)	(.006)	

Notes: Numbers given in parentheses are standard errors based on 200 replications of samples of 10,000 observations.

Table 3: Contribution of Covariates to the Black-White Log Wage Gap

		Reweighting Method		
	Regression	1	2	3
Education	.068	.084	.039	.055
	(.001)	(.001)	(.001)	(.001)
Experience	.005	.004	.005	.005
	(.001)	(.001)	(.001)	(.001)
Marital Status	.022	.035	.016	.021
	(.001)	(.001)	(.001)	(.001)
Occupation	.045	.083	.018	.052
	(.001)	(.001)	(.001)	(.001)
Industry	.005	-.001	.002	.007
	(.001)	(.001)	(.001)	(.001)
Sum of Contributions	.143	.206	.080	.140

Notes: Reweighting methods are described in the text. In each cell, the top number is the estimated contribution of the row covariate, and the number in parentheses is its standard error. For the reweighting methods, standard errors are estimated from 200 bootstrap replications. N = 213,908 for whites and reweighted whites in all cases.

Figure 1

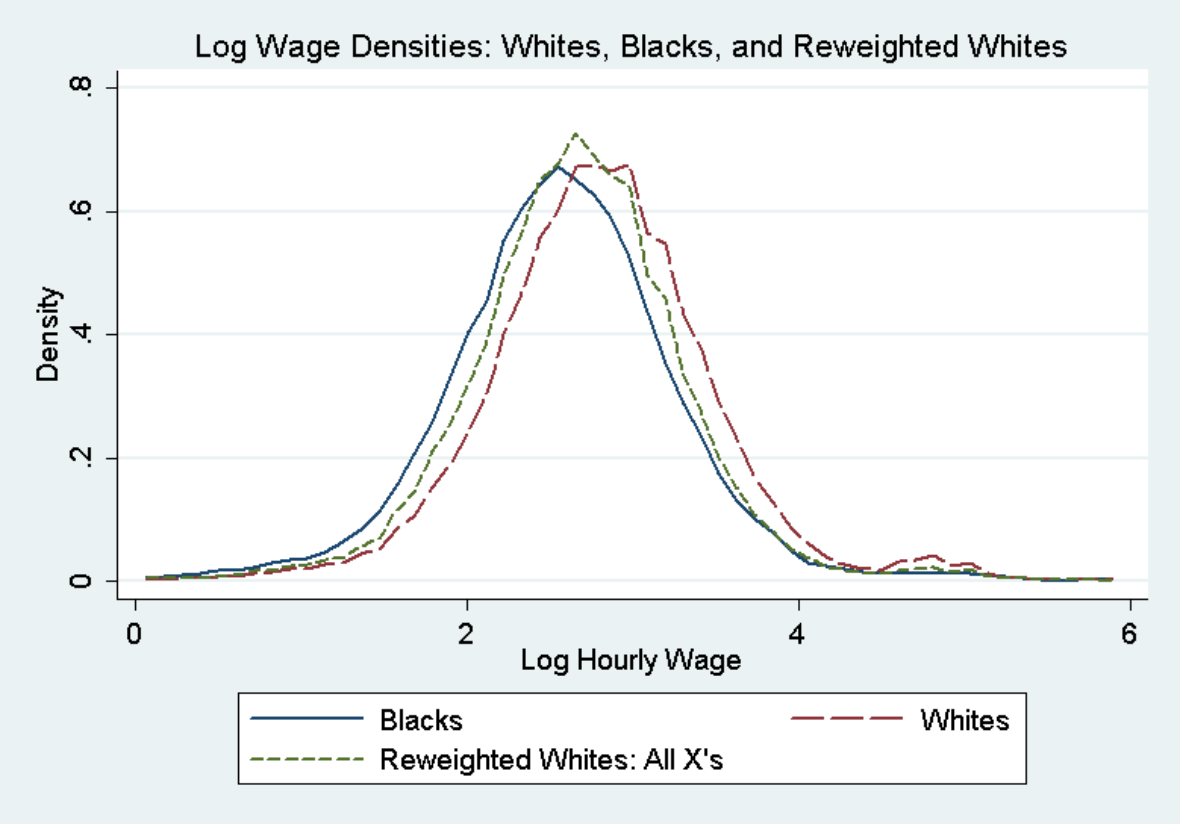
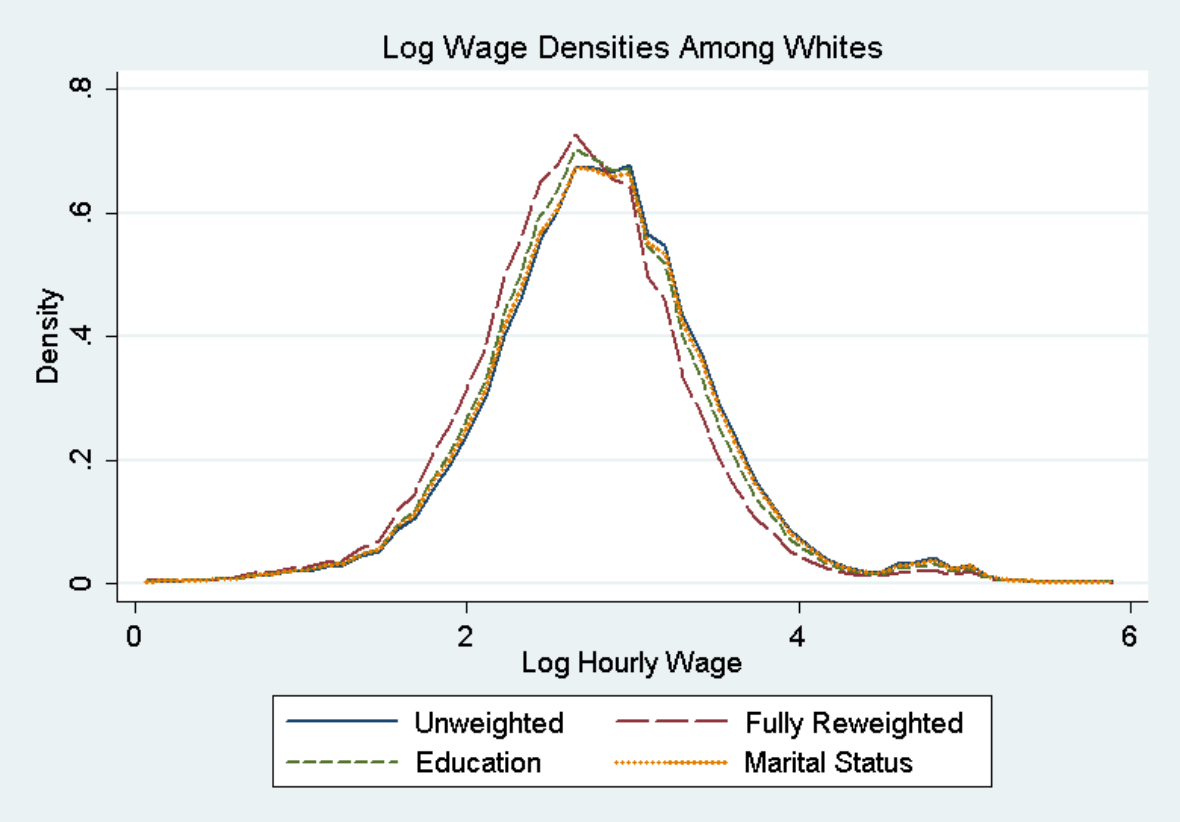


Figure 2



Appendix

A1. Properties of the Method 3 Weights

The first claim in the text is that the $\psi^{z3}(z, x_{-z})$ weights produce a counterfactual distribution that matches group *B*'s marginal distribution of *z*. To see this, first note that $dF(z \mid j = A)$ can be written as follows:

$$\text{(A1)} \qquad dF(z \mid j = A) = \int_{x_{-z}} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A).$$

Consider $dF(z)$ in the counterfactual population, which is given by the reweighted version of the right-hand side of (A1):

$$\begin{aligned} &\int_{x_{-z}} dF(z \mid x_{-z}, j = A) \psi^{z3}(z, x_{-z}) dF(x_{-z} \mid j = A) \\ &= \int_{x_{-z}} dF(z \mid x_{-z}, j = A) \frac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)} dF(x_{-z} \mid j = A) \\ &= \int_{x_{-z}} dF(z \mid j = B) dF(x_{-z} \mid j = A) \\ &= dF(z \mid j = B). \end{aligned} \qquad \text{(A2)}$$

The second equality in (A2) holds because condition (7) guarantees that the $\psi^{z3}(z, x_{-z})$ weights are defined for every value of *z* that appears in group *B* and every value of x_{-z} that appears in group *A*. This establishes the claim.

The second claim is that the $\psi^{z3}(z, x_{-z})$ weights produce a counterfactual distribution that matches group *A*'s marginal distribution of x_{-z}. By the same logic underlying (A2) above, $dF(x_{-z})$ in the counterfactual population can be written as

$$
\begin{aligned}
&\int_z dF(x_{-z} \mid z, j=A)\psi^{z3}(z, x_{-z})dF(z \mid j=A) \\
&\quad = \int_z dF(x_{-z} \mid z, j=A)\frac{dF(z \mid j=B)}{dF(z \mid x_{-z}, j=A)}dF(z \mid j=A) \\
&\quad = \int_z dF(x_{-z} \mid z, j=A)\frac{dF(z \mid j=B)dF(x_{-z} \mid j=A)}{dF(x_{-z} \mid z, j=A)dF(z \mid j=A)}dF(z \mid j=A) \qquad \text{(A3)} \\
&\quad = \int_z dF(z \mid j=B)dF(x_{-z} \mid j=A) \\
&\quad = dF(x_{-z} \mid j=A),
\end{aligned}
$$

where the third line follows from the second by applying Bayes' rule to $dF(z \mid x_{-z}, j=A)$. This establishes the claim.

The third claim is that z and x_{-z} are independent in the counterfactual population. Consider the counterfactual $dF(z \mid x_{-z})$:

$$
\begin{aligned}
dF(z \mid x_{-z}, j=A)\psi^{z3}(z, x_{-z}) &= dF(z \mid x_{-z}, j=A)\frac{dF(z \mid j=B)}{dF(z \mid x_{-z}, j=A)} \qquad \text{(A4)} \\
&= dF(z \mid j=B),
\end{aligned}
$$

which equals the counterfactual $dF(z)$ according to (A2). This establishes the claim.

A2. The Asymptotic Equivalence of Regression and Method 3

Consider a generalization of the linear outcome model given in (3'),

(A5) $$y = \beta_0 + \beta_1 z + \beta_2 x_{-z} + u_i,$$

where z is a covariate (or set of covariates) whose role is to be isolated, x_{-z} is the vector of other covariates, and u is an error term that is orthogonal to z, x_{-z}, and the group membership indicator j. The group difference in the expectation of y is given by

(A6) $$E(y \mid j = A) - E(y \mid j = B) = \underbrace{\beta_1 \left[E(z \mid j = A) - E(z \mid j = B)\right]}_{\text{Role of } z} + \underbrace{\beta_2 \left[E(x_{-z} \mid j = A) - E(x_{-z} \mid j = B)\right]}_{\text{Role of } x_{-z}},$$

where the roles of z and x_{-z} implied by regression are defined in the usual way.

Now consider reweighting based on the $\psi^{z3}(z, x_{-z})$ weights, which are defined at all relevant values of z and x_{-z} if condition (7) holds. The role of z is given by the difference between the actual and reweighted expectation of y in group A:

(A7)
$$\begin{aligned}
&E(y \mid j = A) - E(\psi^{z3}(z, x_{-z}) y \mid j = A) \\
&\quad = \int_{x_{-z}} \int_{z} \left\{E(y \mid z, x_{-z}, j = A) - E(\psi^{z3}(z, x_{-z}) y \mid z, x_{-z}, j = A)\right\} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A) \\
&\quad = \int_{x_{-z}} \int_{z} \left\{E(y \mid z, x_{-z}, j = A) - \psi^{z3}(z, x_{-z}) E(y \mid z, x_{-z}, j = A)\right\} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A) \\
&\quad = \int_{x_{-z}} \int_{z} \left\{ \begin{array}{l} E(y \mid z, x_{-z}, j = A) - \\ \qquad \dfrac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)} E(y \mid z, x_{-z}, j = A) \end{array} \right\} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A) \\
&\quad = \int_{x_{-z}} \int_{z} E(y \mid z, x_{-z}, j = A) \left\{dF(z \mid x_{-z}, j = A) - dF(z \mid j = B)\right\} dF(x_{-z} \mid j = A).
\end{aligned}$$

The second equality in (A7) holds by the law of iterated expectations, the third equality holds by the definition of $\psi^{z3}(z, x_{-z})$, and the fourth equality holds assuming that the support condition in

(7) is satisfied.[21] Making use of the fact that $E(y \mid z, x_{-z}, j = A) = \beta_0 + \beta_1 z + \beta_2 x_{-z}$, the last equality in (A7) can be rewritten as follows:

$$
\begin{aligned}
& E(y \mid j = A) - E(\psi^{z3}(z, x_{-z}) y \mid j = A) \\
& \quad = \int_{x_{-z}} \int_{z} \beta_0 \{dF(z \mid x_{-z}, j = A) - dF(z \mid j = B)\} dF(x_{-z} \mid j = A) \\
& \quad + \int_{x_{-z}} \int_{z} \beta_1 z \{dF(z \mid x_{-z}, j = A) - dF(z \mid j = B)\} dF(x_{-z} \mid j = A) \\
& \quad + \int_{x_{-z}} \int_{z} \beta_2 x_{-z} \{dF(z \mid x_{-z}, j = A) - dF(z \mid j = B)\} dF(x_{-z} \mid j = A).
\end{aligned} \tag{A8}
$$

Consider the second line in (A8):

$$
\begin{aligned}
& \int_{x_{-z}} \int_{z} \beta_0 \{dF(z \mid x_{-z}, j = A) - dF(z \mid j = B)\} dF(x_{-z} \mid j = A) \\
& \quad = \beta_0 \left[\int_{x_{-z}} \int_{z} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A) - \int_{x_{-z}} \int_{z} dF(z \mid j = B) dF(x_{-z} \mid j = A) \right] \\
& \quad = \beta_0 \left[\int_{x_{-z}} \int_{z} dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A) - \int_{x_{-z}} \left[\int_{z} dF(z \mid j = B) \right] dF(x_{-z} \mid j = A) \right] \\
& \quad = \beta_0 \left[1 - \int_{x_{-z}} dF(x_{-z} \mid j = A) \right] \\
& \quad = \beta_0 [1 - 1] \\
& \quad = 0.
\end{aligned} \tag{A9}
$$

The third equality in (A9) follows because the integral of $dF(z \mid x_{-z}, j = A) dF(x_{-z} \mid j = A)$ across all values of z and x_{-z} equals 1, as does the integral of $dF(z \mid j = B)$ across all values of z. Similarly, the fourth equality follows because the integral of $dF(x_{-z} \mid j = A)$ across all values of x_{-z} equals one.

Next consider the third line in (A8):

[21] If the support condition (7) is not satisfied, the weights $\dfrac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)}$ are undefined for some combinations of z and x_{-z}, so that the integral in the fourth line of (A7) will not be defined for all values of z and x_{-z}. As a result, the fourth equality will not hold.

$$\int_{x_{-z}}\int_{z}\beta_1 z\{dF(z\mid x_{-z}, j=A)-dF(z\mid j=B)\}dF(x_{-z}\mid j=A)$$

$$=\beta_1\left[\int_{x_{-z}}\int_{z} z\{dF(z\mid x_{-z}, j=A)-dF(z\mid j=B)\}dF(x_{-z}\mid j=A)\right]$$

(A10)
$$=\beta_1\left[\int_{x_{-z}}\int_{z} z dF(z\mid x_{-z}, j=A)dF(x_{-z}\mid j=A)-\int_{x_{-z}}\left[\int_{z} z dF(z\mid j=B)\right]dF(x_{-z}\mid j=A)\right]$$

$$=\beta_1\left[E(z\mid j=A)-\int_{x_{-z}}\left[E(z\mid j=B)\right]dF(x_{-z}\mid j=A)\right]$$

$$=\beta_1\left[E(z\mid j=A)-E(z\mid j=B)\right],$$

where the last equality again uses the fact that the integral of $dF(x_{-z}\mid j=A)$ over all values of x_{-z} equals one.

Finally, consider the fourth line in (A8):

$$\int_{x_{-z}}\int_{z}\beta_2 x_{-z}\{dF(z\mid x_{-z}, j=A)-dF(z\mid j=B)\}dF(x_{-z}\mid j=A)$$

$$=\beta_2\left[\int_{x_{-z}}\int_{z} x_{-z}\{dF(z\mid x_{-z}, j=A)-dF(z\mid j=B)\}dF(x_{-z}\mid j=A)\right]$$

(A11)
$$=\beta_2\left[\int_{x_{-z}} x_{-z} dF(x_{-z}\mid j=A)\left[\int_{z} dF(z\mid x_{-z}, j=A)-dF(z\mid j=B)\right]\right]$$

$$=\beta_2\left[\int_{x_{-z}} x_{-z} dF(x_{-z}\mid j=A)\times[1-1]\right]$$

$$=0,$$

where the third equality follows because $\int_{z} dF(z\mid x_{-z}, j=A)=\int_{z} dF(z\mid j=B)=1.$

Together, (A9), (A10), and (A11) show that $E(y\mid j=A)-E(\psi^{z3}(z,x_{-z})y\mid j=A)$ can be written as $\beta_1\left[E(z\mid j=A)-E(z\mid j=B)\right]$, which is equivalent to the role of z implied by a linear regression, as shown in (A6).

A3. Verifying Expressions (13), (15) and (16):

1) *Expression (13)*:

We wish to show that if condition (7) is satisfied and the $\psi^{z3}(x_1, x_2)$ weights are used to reweight group *A*, then the role of x_1 in explaining the between-group differences in $E(y)$ in (11) converges to (13). We reproduce (13) here:

(A12) $\left[\beta_1 + \beta_{12}E(x_2 \mid j = A)\right] \times \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right] + \beta_{12}\,\mathrm{cov}(x_1, x_2 \mid j = A).$

The estimated role of x_1 converges to $E(y \mid j = A) - E(y^c)$, where $E(y^c)$ is the counterfactual expectation of *y* based on the reweighted population. Using the results of Appendix A1,

(A13) $E(y^c) = \beta_0 + \beta_1 E(x_1 \mid j = B) + \beta_2 E(x_2 \mid j = A) + \beta_{12} E(x_1 \mid j = B) E(x_2 \mid j = A).$

Because

(A14)
$$\begin{aligned} & E(y \mid j = A) \\ & \quad = \beta_0 + \beta_1 E(x_1 \mid j = A) + \beta_2 E(x_2 \mid j = A) + \beta_{12} E(x_1 x_2 \mid j = A) \\ & \quad = \beta_0 + \beta_1 E(x_1 \mid j = A) + \beta_2 E(x_2 \mid j = A) \\ & \qquad + \beta_{12}\left[\mathrm{cov}(x_1, x_2 \mid j = A) + E(x_1 \mid j = A) E(x_2 \mid j = A)\right], \end{aligned}$$

it follows that

(A15)
$$\begin{aligned} E(y \mid j = A) - E(y^c) = & \left[\beta_1 + \beta_{12} E(x_2 \mid j = A)\right] \times \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right] \\ & + \beta_{12}\,\mathrm{cov}(x_1, x_2 \mid j = A), \end{aligned}$$

which establishes the claim.

2) *Expression (15)*:

Defining ψ^{z3*} as

(A16) $\psi^{z3*}(x_1, x_2) = \psi^{z3}(x_1, x_2) - \psi^{z3}_{AA}(x_1, x_2) + 1,$

we claim that using these weights to reweight group *A* yields a counterfactual population that has group *B*'s marginal distribution of x_1 and group *A*'s marginal distribution of x_2. We assume that

condition (7) holds and that all values of $\psi^{z3*}(x_1, x_2)$ are non-negative, so that all (x_1, x_2) combinations have positive probability in the reweighted population.

The weight $\psi^{z3*}(x_1, x_2)$ can be written as

$$\text{(A17)} \quad \psi^{z3*}(x_1, x_2) = \frac{dF(x_1 \mid j = B) - dF(x_1 \mid j = A) + dF(x_1 \mid x_2, j = A)}{dF(x_1 \mid x_2, j = A)}.$$

Because condition (7) holds, the denominator of (A17) is always nonzero for any value of x_2 that appears in group *A*. Consider the counterfactual density $dF^c(x_1, x_2)$ produced by reweighting group *A* using $\psi^{z3*}(x_1, x_2)$. If all the weights are non-negative as assumed, this is a proper probability distribution if $\int_{x_1}\int_{x_2} dF^c(x_1, x_2) = 1$. Using (A16) and the fact that

$$\frac{dF(x_1, x_2)}{dF(x_1 \mid x_2)} = dF(x_2),$$

$$\begin{aligned} &\int_{x_1}\int_{x_2} dF^c(x_1, x_2) \\ &= \int_{x_2} dF(x_2 \mid j = A) \int_{x_1} \left[dF(x_1 \mid j = B) - dF(x_1 \mid j = A)\right] + \int_{x_1}\int_{x_2} dF(x_1, x_2 \mid j = A) \\ &= 0 + 1 = 1. \end{aligned} \quad \text{(A18)}$$

Define $dF^c(x_1)$ and $dF^c(x_2)$ to be the marginal densities in this counterfactual distribution. Then,

$$\begin{aligned} dF^c(x_1) &= \int_{x_2} \left[\psi^{z3*}(x_1, x_2) \times dF(x_1, x_2 \mid j = A)\right] \\ &= \int_{x_2} \left[\left(dF(x_1 \mid j = B) - dF(x_1 \mid j = A) + dF(x_1 \mid x_2, j = A)\right) \times dF(x_2 \mid j = A)\right] \\ &= dF(x_1 \mid j = B) - dF(x_1 \mid j = A) + \int_{x_2} \left[dF(x_1 \mid x_2, j = A) \times dF(x_2 \mid j = A)\right] \\ &= dF(x_1 \mid j = B). \end{aligned} \quad \text{(A19)}$$

Similarly,

$$
\begin{aligned}
dF^{c}(x_2) &= \int_{x_1} \left[\psi^{z3*}(x_1, x_2) \times dF(x_1, x_2 \mid j = A)\right] \\
&= \int_{x_1} \left\{\left[dF(x_1 \mid j = B) - dF(x_1 \mid j = A) + dF(x_1 \mid x_2, j = A)\right] dF(x_2 \mid j = A)\right\} \\
&= dF(x_2 \mid j = A),
\end{aligned} \tag{A20}
$$

where the last equality follows because the three terms inside the square brackets each integrate to 1 over x_1. This establishes the claim.

3) *Expression (16)*:

We now consider the estimated role of x_1 implied by the $\psi^{z3*}(x_1, x_2)$ weights when y is generated by (3'). We continue to assume that condition (7) holds. Expression (A13) shows the counterfactual expectation of y using the $\psi^{z3}(x_1, x_2)$ weights. The counterfactual expectation of y based on reweighting using the $\psi^{z3}_{AA}(x_1, x_2)$ weights, denoted $E(y^{c}_{AA})$, is similar, except that only marginal distributions from group A appear:

$$
E(y^{c}_{AA}) = \beta_0 + \beta_1 E(x_1 \mid j = A) + \beta_2 E(x_2 \mid j = A) + \beta_{12} E(x_1 \mid j = A) E(x_2 \mid j = A). \tag{A21}
$$

Now suppose that the $\psi^{z3*}(x_1, x_2)$ weights are used, and let $E(y^{c*})$ denote the resulting counterfactual expectation. Because $\psi^{z3*}(x_1, x_2) = \psi^{z3}(x_1, x_2) - \psi^{z3}_{AA}(x_1, x_2) + 1$, $E(y^{c*}) = E(y^{c}) - E(y^{c}_{AA}) + E(y \mid j = A)$, so (A13) and (A21) together imply that the role of x_1, $E(y \mid j = A) - E(y^{c*})$, converges to

$$
\begin{aligned}
&E(y \mid j = A) - E(y^{c}) + E(y^{c}_{AA}) - E(y \mid j = A) \\
&\quad = -E(y^{c}) + E(y^{c}_{AA}) \\
&\quad = -\beta_1 E(x_1 \mid j = B) - \beta_2 E(x_2 \mid j = A) - \beta_{12} E(x_1 \mid j = B) E(x_2 \mid j = A) \\
&\qquad + \beta_1 E(x_1 \mid j = A) + \beta_2 E(x_2 \mid j = A) + \beta_{12} E(x_1 \mid j = A) E(x_2 \mid j = A) \\
&\quad = \left[\beta_1 + \beta_{12} E(x_2 \mid j = A)\right] \times \left[E(x_1 \mid j = A) - E(x_1 \mid j = B)\right],
\end{aligned} \tag{A22}
$$

which is identical to expression (16).

We show in an unpublished appendix that these results can be extended in a natural way to models with more than two covariates and any possible pattern of interactions among the covariates. Specifically, letting $q(z)$ index the unique values that the vector z can take, the use of $\psi^{z3^*}(x_1, x_2)$ weights implies a role of z that converges to $\sum_q \alpha(q)[\Pr(q \mid j = A) - \Pr(q \mid j = B)]$, where $\alpha(q)$ is the average derivative of $E(y)$ with respect to $\Pr(q)$ in group A. Expression (16) (and the last line of (A22)) is a special case of this more general result.

A4. Adjustments for Violations of the Support Condition

To describe our approach for addressing violations of the support condition given by (7), we first define some notation for various counts of observations. Let $p(x_{-z})$ be an index of the possible values of x_{-z} ($p = 1 \ldots P$), and let $q(z)$ be an index of the possible values of z ($q = 1 \ldots Q$) (note that we now explicitly treat x_{-z} and z as discrete and finite-valued). For convenience, we will suppress the arguments of p and q and will express the weights as functions of q and p rather than of z and x_{-z}. Then, define

N_{pq}^{j} as the number of observations with characteristics (p, q) in group j,
$N_{.q}^{j}$ as the number of observations with characteristics q in group j ($\equiv \sum_{p} N_{pq}^{j}$)
$N_{p.}^{j}$ as the number of observations with characteristics p in group j ($\equiv \sum_{q} N_{pq}^{j}$)
$N_{..}^{j}$ as the number of observations in group j ($\equiv \sum_{p}\sum_{q} N_{pq}^{j}$).

When the sample analog of condition (7) is satisfied, the weights are constructed empirically as

$$\text{(A23)} \quad \hat{\psi}^{z3}(q,p) = \frac{N_{.q}^{B} / N_{..}^{B}}{N_{pq}^{A} / N_{p.}^{A}}.$$

When condition (7) is not satisfied, we partition the values of x_{-z} into two sets, denoting the set of values of p for which there is at least one observation in group A for each value of z (i.e., values of p such that $\min_q(N_{pq}^{A}) > 0$) as "$p\bullet$" and the remaining values of p as "$p\circ$" (i.e., those values of p such that $\min_q(N_{pq}^{A}) = 0$)). To create a counterfactual distribution with the appropriate marginal distributions, $p\bullet$ must not be empty. Then, the adjusted weights are

$$\text{(A24)} \quad \hat{\psi}^{z3}(q,p) = \frac{N_{.q}^{B} / N_{..}^{B}}{N_{pq}^{A} / N_{p.}^{A}} \times f_1(p) \times f_2(q) - \begin{cases} \dfrac{f_3(p) \times f_4(q) \times N_{..}^{A} / N_{pq}^{A}}{\sum_q [f_4(p) \times 1(N_{pq}^{A} > 0)]} & \text{if } p \in p\circ \\ \dfrac{f_3(p) \times f_5(q) \times N_{..}^{A} / N_{pq}^{A}}{\sum_q f_5(q)} & \text{if } p \in p\bullet, \end{cases}$$

where

$$f_1(p) = \frac{N^B_{..}}{\sum_q [N^B_{.q} \times 1(N^A_{pq} > 0)]}$$

$$f_2(q) = \frac{N^A_{..}}{\sum_p [N^A_{p.} \times 1(N^A_{pq} > 0)]}$$

(A25)
$$f_3(p) = \frac{N^A_{p.} \times f_1(p)}{N^A_{..}} \times \sum_q \left[\frac{N^B_{.q}}{N^B_{..}} \times [f_2(q) - 1] \times 1(N^A_{pq} > 0) \right]$$

$$f_4(q) = \frac{N^B_{q.} \times f_2(q)}{N^B_{..}} \times \sum_p \left[\frac{N^A_{p.}}{N^A_{..}} \times [f_1(p) - 1] \times 1(N^A_{pq} > 0) \right]$$

$$f_5(q) = f_4(q) \left[1 - \sum_{p^\circ} \left(\frac{f_3(p)}{\sum_z f_4(q) \times 1(N^A_{pq} > 0)} \right) \right].$$

When condition (7) is not satisfied, using the (A23) weights multiplied by $f_1(p)$ would produce a distribution with group *A*'s distribution of x_{-z}; similarly, using the (A23) weights multiplied by $f_2(q)$ would produce a distribution with group *B*'s distribution of *z*. Multiplying by $f_1(p) \times f_2(q)$, as we do in (A24), produces weights that are on average too large, necessitating further adjustments involving $f_3(p)$, $f_4(q)$, and $f_5(q)$. For all values of x_{-z} in p°, the adjustments to the weights shown in (A24) assure that group *A*'s marginal probabilities are matched. The adjustments for values of x_{-z} in $p^\bullet$ shown in (A24) take account of what has been done for x_{-z} in p°, and assure not only that group *A*'s marginal probabilities are matched for x_{-z} in $p^\bullet$, but also that group *B*'s marginal probabilities are matched for all values of *z*.

While these adjustments for violation of condition (7) do not guarantee that all weights are non-negative, in our experience this issue has not been empirically relevant. In the empirical example we describe in Section 5, the adjustments produce no negative weights. They also produce no negative weights if we instead take a 1 percent random sample from the underlying data, so that we are left with 21,497 observations and 7,560 possible unique values of the five covariates.

An Illustrative Example

To see how the adjustments described above work in practice, suppose that z and x_{-z} can each take only two values and that the relative frequencies in group *B* are as follows:

Table A1: Group *B* Distribution Case 1			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/4	1/4	1/2
$x_{-z} = 1$	1/4	1/4	1/2
Total	**1/2**	**1/2**	1

First consider a case in which condition (7) is satisfied, with the following relative frequencies in group *A*:

Table A2: Group *A* Distribution Case 1			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/6	1/6	**1/3**
$x_{-z} = 1$	1/2	1/6	**2/3**
Total	2/3	1/3	1

These tables highlight in bold the marginal distribution of z in group *B* and the marginal distribution of x_{-z} in group *A*, which we want to match in the counterfactual distribution. Using the notation used in the main text, if we use $\psi^{z3}(z, x_{-z}) = \dfrac{dF(z \mid j = B)}{dF(z \mid x_{-z}, j = A)}$ to reweight group *A*, the weight applied to the ($z = 0$, $x_{-z} = 0$) cell would be $\dfrac{\Pr(z = 0 \mid j = B)}{\Pr(z = 0 \mid x_{-z} = 0, j = A)} = \dfrac{1/2}{\left[\dfrac{1/6}{1/3}\right]} = 1$, so that the counterfactual relative frequency would be 1×1/6 = 1/6. Filling in the other entries (by multiplying the group *A* relative frequency by the relevant weight), we find that all of the counterfactual marginal probabilities are correct:

Table A3: Group *A* Distribution Reweighted by $\psi^{z3}(z, x_{-z})$ Case 1			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/6 × 1=1/6	1/6 × 1 = 1/6	**1/3**
$x_{-z} = 1$	1/2 × 2/3 =1/3	1/6 × 2 = 1/3	**2/3**
Total	**1/2**	**1/2**	1

Note that z and x_{-z} are independent in the counterfactual distribution: $\Pr(x_{-z} = 0 \mid z)$ does not vary by z, and $\Pr(z = 0 \mid x_{-z})$ does not vary by x_{-z}. This will always be true if condition (7) is satisfied and the $\psi^{z3}(z, x_{-z})$ weights are used.

Now consider a situation in which condition (7) is not satisfied, which means that there is at least one "empty cell". Suppose the group *B* distribution of $\{z, x_{-z}\}$ is given by

Table A4: Group *B* Distribution Case 2			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/3	1/3	2/3
$x_{-z} = 1$	1/3	0	1/3
Total	**2/3**	**1/3**	1

and the group *A* distribution is given by

Table A5: Group *A* Distribution Case 2			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/5	1/5	**2/5**
$x_{-z} = 1$	3/5	0	**3/5**
Total	4/5	1/5	1

Thus, each group has an empty cell at ($z = 1$, $x_{-z} = 1$). If we reweight the nonempty cells of group *A* using $\psi^{z3}(z, x_{-z})$, we obtain the following counterfactual relative frequencies:

Table A6: Group *A* Distribution Reweighted by $\psi^{z3}(z, x_{-z})$ Case 2			
	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/5 × 4 = 4/15	1/5 × 2/3 = 2/15	**2/5**
$x_{-z} = 1$	3/5 × 2/3 = 2/5	0	2/5
Total	**2/3**	2/15	4/5

The $z = 0$ column and $x_{-z} = 0$ row, with no empty cells, have the correct totals, in that the frequency of the $z = 0$ column matches that in Table A4 and the frequency of the $x_{-z} = 0$ row matches that in Table A5. However, the $z = 1$ column and $x_{-z} = 1$ row totals are too small. Note that the sum of the joint probabilities is 4/5; the "missing probability", 1/5, is what weighting by $\psi^{z3}(z, x_{-z})$ would have placed in cell ($z = 1$, $x_{-z} = 1$) if there were any such observations in group *A*.[22]

Comparing Tables A5 and A6, it is easy to see that one could "fix" the $x_{-z} = 1$ row by multiplying the weight used for ($z = 0$, $x_{-z} = 1$) by 3/2, the ratio of the desired row total to the actual total. However, the $z = 0$ column total would no longer be correct. Similarly, it is easy to fix the second column by multiplying the weight used for ($z = 1$, $x_{-z} = 0$) by 5/2, but the row 1 total would no longer be correct.

To get a sense of how our algorithm works by applying it to this example, we begin by performing *both* of the above adjustments (i.e., multiplying by both $f_1(p)$ and $f_2(q)$), which yields the following matrix:

Table A7: Group *A* Distribution Reweighted by f_1 and f_2
Case 2

	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/5 × 4/3 = 4/15	1/5 × 5/3 = 1/3	3/5
$x_{-z} = 1$	3/5 × 1 = 3/5	0	**3/5**
Total	13/15	**1/3**	6/5

After this step, in general some row and column totals may be correct, but the remaining totals will be too large. In this case, only the $z = 1$ column and $x_{-z} = 1$ row totals are correct: the $z = 1$

[22] As an informal way to see why reweighting by $\psi^{z3}(z, x_{-z})$ fails in this case, note that this scheme can only be successful when implementing it will make the distributions of z and x_{-z} independent. When some combinations of z and x_{-z} do not appear in group *A*, the reweighted distributions of z and x_{-z} cannot be independent, regardless of the weighs used.

column total matches that in Table A4, while the $x_{-z} = 1$ row total matches that in Table A5. Our next step is to reduce weights if necessary in any rows with empty cells to correct the totals in those rows. In this case there is only one such row ($x_{-z} = 1$), and no further adjustment to it is needed, but that will not typically be true in more complex cases.

Once the rows with empty cells have correct totals, our last step is to remove the excess weight from the row (or rows) with no empty cells, in such a way that the row total(s) and all column totals are correct. Because the other rows now have correct totals and the grand totals are the same across rows and columns, this is always possible. It is easy to see what must be done in this example: only cell ($z = 0$, $x_{-z} = 0$) must be changed, because the second column is already correct; this adjustment corresponds to the term involving $f_3(p)$, $f_4(q)$, and $f_5(q)$ in (A23). The final result is as follows:

Table A8: Group *A* Distribution Reweighted by (A23) Case 2

	$z = 0$	$z = 1$	Total
$x_{-z} = 0$	1/5 × 1/3 = 1/15	1/5 × 5/3 = 1/3	**2/5**
$x_{-z} = 1$	3/5 × 1 = 3/5	0	**3/5**
Total	**2/3**	**1/3**	1

In this reweighted population, the marginal distribution of z matches that found in Table A4, while the marginal distribution of x_{-z} matches that found in Table A5, exactly as intended.