Decomposing differences in arithmetic means: a doubly robust estimation approach

Boris Kaiser

Received: 27 November 2013 / Accepted: 22 January 2015 / Published online: 11 April 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract When decomposing differences in average economic outcomes between two groups of individuals, it is common practice to base the analysis on logarithms if the dependent variable is nonnegative. This paper argues that this approach raises a number of undesired issues because decomposition terms have the interpretation of approximate percentage differences in geometric means. Instead, we suggest that the analysis should be based on the arithmetic means of the original dependent variable. We present a flexible parametric decomposition framework that can be used for all types of nonnegative dependent variables. In particular, we derive a propensity score weighted estimator for the aggregate decomposition that is “doubly robust”, that is, consistent under two separate sets of assumptions. A comparative Monte Carlo study illustrates that the proposed estimator performs well in many situations. An application to the union wage gap in the USA finds that the importance of the unexplained union wage premium is much smaller than suggested by the standard log-wage decomposition.

Keywords Oaxaca–Blinder · Decomposition methods · Log-wage decompositions · Doubly robust estimation · Arithmetic and geometric means · Inverse probability weighting

1 Introduction

Decomposition methods are very useful tools to analyse differences in average economic outcomes between groups of individuals. Although a strand of the recent
literature emphasizes the importance of decomposing the entire distribution (e.g. Machado and Mata 2005; Melly 2005), the mean differential remains an important summary statistic in decomposition analysis as it is simple (“one number”), easily understood and sometimes the direct object of interest of policy analysis (e.g. health care expenditures). Differences in means are usually decomposed with the Oaxaca–Blinder method (Oaxaca 1973; Blinder 1973) which splits the gap into a structural effect (due to differences in coefficients) and a composition effect (due to differences in covariates). In many empirical applications where outcomes are continuous and nonnegative, as, for example, wages, the dependent variable is log-transformed before performing the decomposition. While the log-transformation is usually done to allow for convenient estimation, it changes the interpretation of the outcome gap in an important way: the quantity to be decomposed when outcomes are in logs corresponds to a first-order approximation to the percentage difference in geometric means (GM). This quantity cannot be re-transformed to the first moments of the original dependent variable, the arithmetic means (AM).

This paper argues that a decomposition based on the original dependent variable may be preferable to a decomposition based on logarithms on statistical and conceptual grounds. Foremost, if the support includes zero, the GM is theoretically undefined. In empirical studies, the dependent variable is sometimes artificially re-scaled in order to be able to take logarithms, but this distorts the distribution at the lower bound of the support (Santos Silva and Tenreyro 2006). Furthermore, even if there are no zeros in the data, a decomposition of AMs can be preferable for several reasons. First, the log outcome gap is not invariant to changes in the higher-order moments of the distributions even if arithmetic means remain constant (Leslie and Murphy 1997). If, for example, dispersion increases, the log outcome gap will change as well. This is clearly an undesired property for a measure of the difference in average outcomes. Second, the log outcome gap only offers an approximate interpretation, whereas the raw outcome gap offers an exact interpretation. Finally, the AM is the more common and intuitive summary statistic.

In this paper, we suggest modelling the untransformed outcome variable directly to decompose differences in arithmetic means. The leading case is a decomposition of average wages between groups of workers (Firpo et al. 2011, for an overview), but many other outcome variables can be of interest in decomposition analysis, such as health care expenditures (Vargas Bustamante and Chen 2011), financial wealth (Barsky et al. 2002), firm-level productivity (Mueller 2012), revenues (Munn and Hussain 2010), students’ test scores (Krieg and Storer 2006) or counts, such as the number of cigarettes smoked (Bauer et al. 2007). Since the conditional expectation function is thought to be convex in most of these examples, as reflected in the widespread use of log-transformations, a linear outcome model is inappropriate (see, e.g. Barsky et al. 2002). Instead, we propose a flexible nonlinear parametric framework based on an exponential regression model. Following Firpo et al. (2011), we first discuss identification under the ignorability assumptions before proceeding to specification and estimation issues. By contrast, previous papers on nonlinear decompositions (Fairlie 2005; Bauer and Sinning 2008) have largely omitted discussing identification.
As our methodological contribution to the literature, we suggest a new estimation strategy for decomposing differences in average outcomes when outcomes are nonnegative. We propose a “doubly robust” weighted Poisson quasi-maximum-likelihood (WPQML) estimator. The existence of this estimator is briefly mentioned in Wooldridge (2007), but to our knowledge, it has not been studied in detail nor has it been used in empirical work. Quasi-maximum-likelihood (QML) estimators in general are attractive because they only require correct specification of the conditional mean function for consistency and no further distributional assumptions (Gourieroux et al. 1984). The Poisson QML estimator achieves some additional robustness if augmented with appropriate propensity score weights. We show formally that the decomposition can be consistently estimated if either the outcome model or the propensity score model is correctly specified. Therefore, double robustness is a useful property to guard against misspecification. It is important to note, however, that in the case where the validity of the ignorability assumptions is questioned, the same caveat applies as to other estimators that rely on these assumptions (see also Kunze 2008; Huber 2014).

To illustrate how the proposed estimator of the nonlinear decomposition performs in practice, we undertake a small Monte Carlo exercise. We compare the performance of weighted and unweighted QML estimators, linear and quadratic regression, reweighted regression (Firpo et al. 2007) and the nonparametric doubly robust estimator (Rothe and Firpo 2013) under various scenarios with regard to functional form, overlap of covariate distributions and heteroskedasticity. We find that the weighted Poisson QML estimator produces convincing results in many situations. It should therefore prove to be an attractive estimation strategy for applied researchers who wish to perform decompositions of nonnegative outcomes.

In an empirical application to the union wage gap in the USA, we compare the approach of decomposing arithmetic means with the standard Oaxaca–Blinder decomposition based on the log-wage gap (geometric means). The former suggests that the union wage premium is much less important in explaining the wage differential between union and nonunion workers than the latter. In other words, the concept of the mean can have striking implications for the analysis of differences in economic outcomes and therefore requires more careful consideration than it usually receives in empirical research.

The remainder of this paper is organized as follows: Sect. 2 deals with the conceptual and statistical issues of decomposing differences in either AMs or GMs. In Sect. 3, we define the general decomposition framework and present the identification result for the counterfactual of interest. In Sect. 4, we discuss model specification and derive the doubly robust estimator for the decomposition. Section 5 contains the Monte Carlo exercise and Sect. 6 the empirical application. Section 7 briefly covers some extensions to detailed decompositions, endogeneity and sample selection, and Sect. 8 contains some concluding remarks.

1 Other estimators for the aggregate decomposition terms can include any estimator that can consistently estimate the average treatment effect on the treated (ATT) under the ignorability assumptions.
2 Arithmetic and geometric mean differentials

2.1 Analytical framework

The analytical framework used in this paper will be based on the potential outcomes notation popularized by Rubin (1974). While questions of identification have mostly been ignored in the more traditional decomposition literature, the treatment effect framework clarifies the assumptions needed to identify the decomposition terms (Firpo et al. 2011). We begin by defining a large population indexed by \( i = 1, 2, ..., N \) that contains two mutually exclusive and nonempty groups of individuals. Let \( D_i = 1 \) if person \( i \) belongs to group 1, and \( D_i = 0 \) if she belongs to group 0. The potential outcome of person \( i \) if she belongs to group \( g \) is \( y_{ig} \), where \( y_{ig} \in \mathbb{R}_{+}^{+} \forall i \) and \( g \in \{0, 1\} \).

The potential outcomes and the realized outcome are linked by
\[
y_i = D_i y_{i1} + (1 - D_i) y_{i0}.
\]
Hence, we do not observe \( y_{i1} \) for units in group 0 and \( y_{i0} \) for units in group 1, which is why these quantities are “counterfactual”. The notation introduced above will be used throughout this paper.

2.2 Measures of mean differentials

Before dealing with the technicalities of decomposition analysis, it is sensible to first clearly define the object under study, i.e. the quantity to be decomposed. In most situations, we are interested (amongst other statistics) in the average gap in economic outcomes between two groups, a prominent example being the average wage gap between male and female workers. It seems natural to consider the difference in expected outcomes which is based on population arithmetic means (AM):
\[
\Delta_{AM} = E[y_i | D_i = 1] - E[y_i | D_i = 0]
\]
Of course, the outcome gap may also be expressed in terms of a percentage differential, e.g. \( \%\Delta_{AM} = (E[y_i | D_i = 1] - E[y_i | D_i = 0]) / E[y_i | D_i = 0] \), which requires the choice of a reference group (in this case group 0), but it is free from the underlying units in which outcomes are measured. Alternatively, if the decomposition is based on log outcomes, we have
\[
%\Delta_{GM}^{\text{approx}} \approx E[\ln y_i | D_i = 1] - E[\ln y_i | D_i = 0],
\]
which is the approximate percentage differential in the population geometric means (GM). An exact percentage differential in terms of GMs is given by \( %\Delta_{GM} = \frac{\exp(E[\ln y_i | D_i = 1]) - \exp(E[\ln y_i | D_i = 0])}{\exp(E[\ln y_i | D_i = 0])} \), and the absolute differential in terms of GMs is \( \Delta_{GM} = \exp(E[\ln y_i | D_i = 1]) - \exp(E[\ln y_i | D_i = 0]) \). The approximate nature of (2) comes from a first-order Taylor expansion of \( %\Delta_{GM} \) around \( E[\ln y_i | D_i = 1] - E[\ln y_i | D_i = 0] = 0 \) that becomes more accurate as \( %\Delta_{GM} \) approaches zero.\(^2\)

\(^2\) Of course, the approximation in (2) could be made more exact if higher-order terms are included in the Taylor expansion. For simplicity, write \( z_g \equiv E[\ln y_i | D_i = g] \) for \( g = \{0, 1\} \), then \( %\Delta_{GM} = \exp(z_1 - z_0) - 1 \). A Taylor expansion around \( z_1 - z_0 = 0 \) yields \( (z_1 - z_0) + \frac{1}{2!}(z_1 - z_0)^2 + \frac{1}{3!}(z_1 - z_0)^3 + ... \)
It is notable that in many applications of nonnegative dependent variables, researchers prefer to decompose (2) instead of (1). In particular, this is often the case in studies of wages, where log-linear wage models are estimated by OLS to perform Oaxaca–Blinder-type decompositions. This is the dominant procedure because wages and covariates are usually assumed to have a log-link relationship, but also because it is standard practice to estimate linear models instead of nonlinear models. While the geometric mean interpretation of (2) was noted by Oaxaca (1973), it is not explicitly mentioned in most empirical applications.

The question obviously arises as to how the concepts of AM and GM are related. Due to Jensen’s inequality, we always have that $AM > GM$. However, the relationship between $\Delta AM$ and $\Delta GM$ is ambiguous without distributional assumptions and will depend on the properties of the distributions under study.

### 2.3 Statistical and conceptual issues

Left aside issues of modelling and estimation, which measure of the mean differential should be chosen when performing decompositions? There are several arguments to be made. Foremost, if the support of $y_i$ includes zero, the GM is not defined. Thus, the decomposition of means must be based on AMs for outcomes that may be zero, such as individual health care expenditure, wealth or count variables. Using ad hoc manipulations such as $\ln(1 + y_i)$ to accommodate zeros is inappropriate as it may severely distort the distribution if the fraction of zeros is nonnegligible (cf. Santos Silva and Tenreyro 2006).

Furthermore, even if the support of the dependent variable excludes zero (as is the case with wages), there are a number of arguments why a decomposition of means should be based on AMs instead GMs, or equivalently, on raw outcomes instead of log outcomes. First, as Leslie and Murphy (1997) note, the decomposition based on log outcomes is not invariant to changes in the higher-order moments of the distribution. As a result, even if the AMs of two distributions are the same, the difference in average log outcomes is nonzero if, for example, the dispersion differs across distributions. We illustrate this by two simple examples:

**Example 1** (Leslie and Murphy 1997) Let there be two samples with only two observations each: $S^0 = \{2, 4\}$ and $S^1 = \{1, 5\}$. Then we have $\bar{y}^0 = \bar{y}^1 = 3$ and $\hat{\Delta} AM = 0$, but the log gap equals $\hat{\Delta} approx GM = 0.47$.

**Example 2** Suppose $y^0$ and $y^1$ are log-normally distributed with $E[y^0] = E[y^1] = 1$, but variances are unequal: $V[y^0] = 0.5 < V[y^1] = 1$. In this case, $\Delta AM = 0$, but there is a gap in log means given by $\% \Delta approx GM = E[\ln y^1] - E[\ln y^0] = (-.203) - (-.347) = .144$.\(^3\)

As demonstrated above, if two distributions are identical up to a mean-preserving spread, the log-wage gap picks up the difference in dispersion. For a measure of the difference in means, this is a rather undesired property.

\(^3\) Of course, examples of the converse case, where $% \Delta GM = 0$ and $% \Delta AM \neq 0$, can also be given.
Second, the approximation in (2) becomes inaccurate for large differences in outcomes due to the concavity of the logarithm: for exact GM differentials of 5, 10, 20 and 50 per cent, the “approximation bias” is 0.12, 0.47, 1.77 and 9.45 percentage points, respectively. This is clearly important in practice, since many differentials (e.g. male–female wage gap) are quite substantial in magnitude. In fact, the approximative nature of (2) is only seldom pointed out in empirical studies, and sometimes no reference to the GM interpretation is made (e.g. Darity et al. 1995; Jürges 2002). As a consequence, interpreting log-points as percentage differentials can be misleading. 4

Third, apart from time series contexts, it is customary to report arithmetic means rather than geometric means when summary statistics are presented. Although this does not represent a strong argument in favour of the AM decomposition per se, it nonetheless implies that the AM is usually the statistic of interest.

Finally, a potential weakness of the AM relative to the GM as a measure of central tendency is that it is more strongly affected by the presence of outliers in small samples. 5 Although the GM is indeed more robust, the argument only applies if the two distributions under study are adversely affected by outliers. This is because we are interested in differences in means and not in means per se.

The discussion above provides the motivation for developing a decomposition framework for differences in AMs that can be broadly applied to nonnegative dependent variables.

3 Decomposition framework and identification

As for any decomposition problem, we must define a meaningful counterfactual. In the classic Oaxaca–Blinder decomposition of wages, the counterfactual is implied by the reference wage structure (Cotton 1988; Neumark 1988; Jann 2008). We follow the approach of Firpo et al. (2011) and choose the “simple” counterfactual outcome $\mu_C \equiv E[y_{i0}|D_i = 1]$. This term corresponds to the mean outcome in group 1 that would prevail if their distribution was generated by the conditional expectation function (CEF) of group 0. Since the labelling of group 0 and 1 is arbitrary, the choice of the counterfactual largely depends on the question of interest. However, we find this particular counterfactual useful because it offers a simple interpretation and has a meaningful treatment effect equivalent. 6 As opposed to the counterfactual, the group-

---

4 To avoid the problem described above, the log gap is often interpreted in terms of log-points or “log-dollars” (Black et al. 2006; Edin and Richardson 2002; Kim 2010). While this is formally correct and exact, talking about log-points arguably conveys little meaning in policy analysis.

5 The most robust measure of central tendency is the median. It is unfortunate, however, that a direct extension of OB-type decompositions to the median wage gap is not possible since the law of iterated expectations does not hold for quantiles. See Firpo et al (2011, Section 4) for a discussion of the various approaches to quantile decompositions.

6 In the traditional literature on OB wage decompositions, a number of alternative reference structures (i.e. counterfactuals) have been suggested (Jann 2008), but these alternative measures sometimes imply undesired assumptions on the underlying structural effect. First, if the counterfactual is based on Neumark (1988), the coefficient estimates are likely to suffer from an omitted variable bias (see Jann 2008). Second, if the counterfactual is based on the coefficients from the (population) regression in the pooled sample on the covariates and a dummy for the treated group, the wage structure effect is restricted to be the
specific means are directly observable, i.e. \( E[y_i|D_i = g] = E[y_{ig}|D_i = g] \). As a result, the aggregate AM decomposition is

\[
\Delta_{AM} = (E[y_i|D_i = 1] - E[y_{i0}|D_i = 1]) = \Delta^S + (E[y_{i0}|D_i = 1] - E[y_i|D_i = 0]),
\]

(3)

The first term on the right-hand side is the *structural effect* (also: coefficients effect). In treatment effect terminology, it corresponds to the population average treatment effect on the treated (PATT); see, e.g. Imbens and Wooldridge (2009). The second term is the *composition effect* (also: characteristics effect) and captures the part of the gap due to differences in covariate distributions across groups.

A prominent result from the treatment effects literature (see, e.g. Imbens 2004) states that identification of \( \mu^C \) requires the two so-called ignorability assumptions. The first (and more critical) assumption is \( y_{i0} \perp \perp D_i | X_i \),

(4)

where \( X_i \) is a row vector of covariates with support \( \mathcal{X} \subseteq \mathbb{R}^k \). The condition in (4) is referred to as the conditional independence assumption (CIA) and states that once we control for \( X_i \), the distribution of unobserved characteristics is the same across groups in the absence of treatment. At this point, a critical discussion of the CIA is in order. Firpo et al. (2011) discuss three important situations where the CIA is violated: (i) if there is nonrandom sample selection (e.g. differential selection of men and women into employment); (ii) if \( D_i \) is affected by confounding unobservables (e.g. self-selection of individuals into groups); (iii) if covariates are endogenous in ways that violate the CIA. Huber (2014) shows that standard wage decompositions can be cast into a so-called mediator framework. He argues that in many examples, a causal interpretation is hard to establish because group membership is endogenous and/or observed covariates are mediators (intermediate outcomes) which are themselves confounded by unobservables. Kunze (2008) discusses endogeneity issues in gender wage gap analyses, stressing that human capital variables are often rendered endogenous by unobserved heterogeneity, nonrandom sample selection into work and measurement error. Overall, these studies suggest that a careful assessment of the plausibility of the CIA is of great importance in the context of wage decompositions.

The second assumption required for identification is less critical and defined as

\[
p(X_i) < 1,
\]

(5)

Footnote 6 continued

same for every individual (see Firpo et al. 2011). Third, if the counterfactual is a sample-size weighted average of the two wage structures as in Cotton (1988), the wage structure effect can be shown to equal \( P(D_i = 0) \text{PATT} + P(D_i = 1) \text{PATU} \) under a set of baseline assumptions, with \( \text{PATU} \) being the population average treatment effect on the untreated (see Słoczyński 2012). Clearly, this quantity differs from the population average treatment effect (PATE) unless \( P(D_i = 0) = P(D_i = 1) \) and has therefore no interesting analogue in the treatment effects framework.
where \( p(X_i) \equiv P(D_i = 1|X_i) \) is the propensity score. The condition in (5) is referred to as the common support assumption (CSA) and states that the support of the covariate distribution in group 1 (the treated) must be contained within the support of the covariate distribution in group 0 (the controls). Under the two ignorability assumptions (4) and (5), identification of the counterfactual outcome can be stated as follows (see Imbens 2004):

\[
E[y_{i0}|D_i = 1] = E_X\{E[y_i|X_i, D_i = 0]|D_i = 1\}
\]

To decompose differences in arithmetic means, the main issue that must be addressed is to specify and estimate an appropriate model for \( E[y_i|X_i, D_i = 0] \), which we turn to in the next section.

4 Model specification and estimation

Since researchers are usually also interested in detailed decompositions, parametric models are more practical tools such that we will not discuss nonparametric methods. For the ensuing discussion, we assume that the CEFs can be represented by \( E[y_i|X_i, D_i = g] = \mu(X_i, \beta^g) \) for \( g = \{0, 1\} \) and all \( X_i \in \mathcal{X} \), where \( \beta^g \in \mathbb{R}^k \) is a unique column vector of population coefficients. As a useful starting point for modelling, consider the “generalized” functional form (Wooldridge 1992):

\[
\mu(X_i, \beta^g) = (1 + \lambda X_i \beta^g)^{1/\lambda}, \quad \text{for } g = \{0, 1\},
\]

which encompasses the linear regression model (\( \lambda = 1 \)) and the exponential regression model (\( \lambda \to 0 \)) as special cases. We discuss these two specifications for modelling the dependent variable in the next two subsections.

4.1 A note on the linear model

If the CEF of the outcome is assumed to be linear (\( \lambda = 1 \)), then by the law of iterated expectations, the AM decomposition takes the familiar Oaxaca–Blinder form. Why should we not use a linear model that can be conveniently estimated with OLS? In the case of nonnegative dependent variables, the linear model appears unnatural because it does not restrict the support of the CEF in any way. Given nonnegative outcomes, the true CEF is likely to be nonlinear. While OLS offers the best linear approximation to the CEF, the approximation may still be very poor if the true CEF is, for example, highly convex (Barsky et al. 2002). To see this, consider the stylized example in Fig. 1, where outcomes in both groups are generated by the same CEF (\( \Delta^S = 0 \)), but the covariate

---

7 A detailed decomposition is defined as a procedure that apportions the decomposition terms \( \Delta^S \) and \( \Delta^X \) in (3) into components that can be attributed to each covariate in \( X_i \) (cf. Firpo et al. 2011, p. 26).

8 The uniqueness or identifiability of \( \beta^g \) rules out perfect multicollinearity amongst the covariates.

9 In the language of generalized linear models, the first assumes an identity link between outcomes and covariates and the second assumes a log-link relationship (equivalent to the functional form assumption of the log-linear model).
distributions in the treated group and the control group are different, \( N(0.5, 1) \) and \( N(1, 1) \), respectively, which means \((\Delta^X \neq 0)\). The two group-specific regressions of \( y_i \) on \( X_i \) yield different parameters, meaning that the misspecified model leads to a spurious structural effect \((\hat{\Delta}^S \neq 0)\). This inconsistency persists if ignorability holds. Moreover, Barsky et al. (2002) point out that the performance of the linear estimator becomes particularly poor if there is missing overlap in the covariate distributions due to the linear extrapolation of the misspecified CEF outside the common support.

There are several ways as to how the above-described problems can be alleviated. First, one can try to improve the approximation to the CEF by adding higher-order terms of the covariates. In the case of high-dimensional \( X_i \), however, this may be cumbersome and lead to very noisy estimates. Furthermore, it complicates matters if one is interested in performing detailed decompositions. Second, another alternative is to use the doubly robust reweighted regression as suggested by Firpo et al. (2007). Intuitively, the propensity score adjustment reduces the imbalance between the group-specific covariate distributions and thus reduces the contamination of the structural effect stemming from these imbalances. While reweighting generally increases the robustness of regression with respect to departures from linearity, it may be problematic to rely on the correct specification of the propensity score alone because any misspecification immediately leads to biased estimates.

### 4.2 Nonlinear outcome model

Due to the described weaknesses of the linear model in the case of nonlinear CEFs, a nonlinear specification that is consistent with the dependent variable being non-negative for arbitrary values of \( \beta^g \) appears generally preferable.\(^{10}\) This property is satisfied by the exponential model, which follows from (7) when \( \lambda \to 0 \). The model

\(^{10}\) A decomposition of nonnegative outcomes based on a linear model nonetheless performs well if the data-generating process satisfies the functional form assumption (linearity in parameters).
can be written as
\[ y_{ig} = \exp(X_i \beta^g) + \epsilon_{ig} \quad \text{for } g = \{0, 1\} \quad (8) \]

where \( \epsilon_{ig} \) is an error term, and the assumed CEF is \( \mu(X_i, \beta^g) = \exp(X_i \beta^g) \). The exponential CEF is used in many different models of nonnegative dependent variables. In the case of count data outcomes, it emerges from the Poisson model. In the case of continuous nonnegative outcomes, it follows from the gamma and exponential distributions. The exponential model is also very widely used in practice. In models of health care expenditure, for example, it is the standard functional form assumption (Mullahy 2009). In wage models, the dependent variable used for estimation is usually in logarithmic form which implies the functional form assumed in (8). It is justified theoretically (Mincer 1974) and empirically (Heckman and Polachek 1974; Blackburn 2007). Therefore, the exponential model seems a suitable choice for many contexts.

### 4.2.1 Aggregate decomposition

If the outcome model is (8), the law of iterated expectations implies that the aggregate decomposition defined in (3) can be written as

\[
\Delta_{AM} = E\left[\exp\left(X_i \beta^1\right) | D_i = 1\right] - E\left[\exp\left(X_i \beta^0\right) | D_i = 1\right] \\
+ E\left[\exp\left(X_i \beta^0\right) | D_i = 1\right] - E\left[\exp\left(X_i \beta^0\right) | D_i = 0\right] \quad (9)
\]

The only critical expression is the counterfactual, \( \mu^C = E[\exp(X_i \beta^0)|D_i = 1]\), which is identified if ignorability holds.

### 4.2.2 Doubly robust estimation

To estimate the decomposition in (9), expectations are replaced by sample analogues and population coefficients by \( \hat{\beta}^1 \) and \( \hat{\beta}^0 \), which are estimated coefficients from the two group-specific samples. Estimation can be based on quasi-maximum likelihood (QML) techniques also referred to as estimation of generalized linear models (GLM) in the statistics literature. It is a well-known result by Gourieroux et al. (1984) that QML estimators are consistent regardless of the underlying distribution of the data,

---

11 In (8), we may use a multiplicative or additive error term because we will only need to impose mean independence (Wooldridge 1992).

12 Note that the aggregate decomposition may still be identified if regressors are correlated with unobservables, as long as the conditional distribution of these unobservables is the same across group (i.e. ignorability holds) and the functional form of the outcome model is correctly specified (Firpo et al. 2011, Ch. 6.1.3, p. 91). However, such situations where regressors are endogenous but ignorability still holds are probably rare in practice.
Decomposing differences in arithmetic means... as long as the CEF is correctly specified. In other words, these models can be estimated consistently with minimal distributional assumptions. The difference between the various QML estimators is simply how observations are weighted in the first-order conditions of the maximization problem, leading to different requirements for asymptotic efficiency. Several authors (Manning and Mullahy 2001; Santos Silva and Tenreyro 2006) find that Poisson and gamma QML estimators perform well in many situations, while Gaussian QML (nonlinear least squares) leads to more erratic results, especially in heteroskedastic environments.

A major advantage of the QML framework with regard to the decomposition in (9) is that doubly robust estimator for the counterfactual mean can be derived. Double robustness broadly refers to an estimator that is consistent under two sets of assumptions. As Wooldridge (2007) briefly mentions, Poisson QML can be made doubly robust if augmented with the appropriate propensity score weighting function. We will refer to this estimator as the WPQML estimator. Denote the probability limit of the WPQML estimator by 
\[ \beta^*_{g} = \text{plim}(\hat{\beta}^ g), \]
where we allow for the possibility that the CEF may be misspecified, i.e. \( E[y_i \mid X_i, D_i = g] \neq \exp(X_i \beta^ g) \) for some \( X_i \in X \). However, the defining property of the WPQML estimator is that \( E[y_i \mid D_i = g] = E[\exp(X_i \beta^ g) \mid D_i = g] \) will always hold by construction. As shown below, this property arises from the first-order conditions associated with this estimator. For the sub-population of interest (group 0), the WPQML estimator solves the sample analogue of the following population maximization problem:

\[
\max_{b^0} \mathbb{E} \left[ \omega(X_i) \left( y_i X_i b^0 - \exp \left( X_i b^0 \right) - y_i \right) \mid D_i = 0 \right],
\]

where \( \omega(X_i) \) is the propensity score weighting function defined below. The corresponding population first-order conditions (FOCs) are:

\[
E \left[ \omega(X_i) \left( y_i - \exp \left( X_i \beta^0 \right) \right) X_i' \mid D_i = 0 \right] = 0. \tag{10}
\]

Since Wooldridge (2007) does not touch upon the WPQML and confines the discussion to the population average treatment effect (PATE), Theorem 1 below shows that WPQML with appropriate weights yields a doubly robust estimator for our counterfactual of interest and thus also for the structural effect \( \Delta^S \) and the characteristics effect \( \Delta^X \) of the decomposition.

**Theorem 1** Define the weighting function by \( \omega(X_i) = \frac{\Lambda(X_i)}{1 - \Lambda(X_i)} \frac{1 - p}{p} \), where \( \Lambda(X_i) \) is a model for \( p(X_i) \equiv P(D_i = 1 \mid X_i) \) and \( p \equiv P(D_i = 1) > 0 \). Assume that (i) we have a random sample from an i.i.d. population (ii) all relevant moments of \( y_i \) and \( X_i \) exist and are finite and (iii) ignorability (assumptions (4) and (5)) holds.

Then, the WPQML estimator is doubly robust in the sense that it identifies the counterfactual of interest through \( E[y_i \mid D_i = 1] = E[\exp(X_i \beta^0) \mid D_i = 1] \), if either the outcome model or the propensity score model is correctly specified (or both, of course). The counterfactual mean can be consistently estimated by the appropriate sample analogue \( \frac{1}{N} \sum_{i=1}^{N_1} \exp(X_i \beta^0_{WPQML}) \).

The proof is given in “Appendix 1.”
The central implication of Theorem 1 is that the counterfactual is identified even if the WPQML estimator does not converge to the true parameters \( \beta^{0*} \neq \beta^0 \) provided that the propensity score model is correctly specified. Therefore, double robustness guards better against misspecification than relying on correct specification of the outcome model alone or the propensity score model alone. Although this property is very appealing, it is important to remember that the result does not automatically extend to the detailed decomposition; if we fail to identify the parameters of the CEF, a detailed decomposition into the contributions of individual covariates to the gap is no longer guaranteed to be consistent.

Due to the nonlinear model and the two-step nature of the estimator, bootstrapping the entire procedure is the more practical alternative to conduct inference than deriving analytical standard errors via nonlinear GMM. Note that propensity score weighting may also come at the cost of an efficiency loss if the CEF is correctly specified, but our Monte Carlo exercise suggests that the opposite can also happen in the case of heteroskedasticity where the conditional variance of the error is proportional to the conditional mean.

## 5 Monte Carlo simulation

To assess and compare the various estimators of the proposed decomposition, we conduct a small simulation study. Specifically, the objective is to test how various estimation methods perform in estimating the decomposition under different data-generating CEFs of the outcome model, different scenarios for the common support of the covariate distributions, as well as heteroskedasticity in the outcome model.

### 5.1 Set-up

The simulation design mimics a potential outcomes framework in which the CEF of group 0 has larger coefficients than the CEF of group 1 (negative structural effect) and group 0 has higher average values of the covariates than group 1 (negative composition effect). To have a realistic data-generating process, we generate both continuous and discrete covariates. The continuous variables are \( X_1, X_2, X_3 \sim N(0, 1) \) with \( \text{Corr}(X_1, X_2) = 0.2, \text{Corr}(X_1, X_3) = 0.1 \) and \( \text{Corr}(X_2, X_3) = -0.1 \). The discrete covariates are two binary dummies with \( P(X_4 = 1) = P(X_5 = 1) = 0.5 \) and \( \text{Corr}(X_4, X_5) = 1/8 \). The assignment of observations to the two groups is based on the following latent variable model:

\[
D^* = X\delta + \xi, \quad \text{with } \xi \sim N(0, 1) \quad \text{and} \quad \delta > 0 \\
D = 1[D^* < \text{Med}(D^*)] \tag{11}
\]

The threshold in the indicator function is equal to the median of the latent variable such that the two groups are of equal size. The group indicator variable is \( D = 1 \) for group 1 (the treatment group) and \( D = 0 \) for group 0 (the control group). The random assignment error \( \xi \) is stochastically independent of the processes determining
outcomes, which means that the CIA is satisfied in our set-up. The parameters \( \delta \) determine how strongly group membership is correlated with covariates, in other words, how strongly the group-specific covariate distributions differ from one another. We consider two specifications for the latent variable equation. In the first one, there is perfect common support. In the second one, common support is small and the overlapping support assumption is violated for a subset of the population; see “Details on the set-up” in Appendix 2 for more details.

To illustrate the consequences of departures from the standard log-link assumption of exponential models, we consider two separate cases for the functional form. Starting from the generalized functional form, \( E[y_g|X] = (1 + \lambda X \beta^g)^{1/\lambda} \) for \( g = \{0, 1\} \), in the first case (the baseline specification) we assume \( \lambda \to 0 \) which reduces to the exponential model. In the second case, we set \( \lambda = 0.5 \). This is a quadratic specification and thus also ensures that outcomes are always nonnegative. The curvature of this specification lies in between the linear and exponential model. It is important to note that we do not consider the linear model as a possible data-generating process because, unlike the two functional forms above, it does not restrict the support of the dependent variable to be nonnegative. Since the outcome distribution is affected by the functional form, we choose different values for \( \beta^g \) such that mean outcomes are of comparable size. See “Details on the set-up” in Appendix 2 for more details on the parameterization of the CEFs.

Finally, the potential outcomes are generated by multiplying the CEF with an error term that is drawn from the log-normal distribution with \( E[\varepsilon_g|x] = E[\varepsilon_g] = 1 \). To test whether the estimators are sensitive to heteroskedasticity, we consider two cases for the conditional variance of the error term. The first is the benchmark case of homoskedasticity, and the second is heteroskedastic, where the conditional variance of the error is positively related to the CEF; see “Details on the set-up” in Appendix 2 for more details.

The quantity associated with uncertainty is the counterfactual mean, which is why the focus lies on estimating \( \mu^C = E[y_0|D = 1] \) in this Monte Carlo exercise. The goal is to have an estimator based on the observed outcomes that is close as possible to the estimates from the “true” estimator \( \hat{\mu}^C_{\text{true}} = \frac{1}{N_1} \sum_{i:D=1} N_1 y_{i0} \). We compare the performance of the following estimators for \( \mu^C \):

i. WPQML: the estimation method is Poisson QML augmented with propensity score weights as described in Sect. 4.2.2. The propensity score is estimated by a logit model.

ii. PQML: unweighted counterpart of WPQML.

iii. GQML: the estimation method is gamma QML; see, e.g. Blackburn (2008).

iv. OLS1: linear regression. The CEF is assumed to be linear in \( X_i \).

v. OLS2: quadratic regression. The vector of covariates includes a second-order polynomial of \( X_i \).

vi. WLS: reweighted linear regression estimator proposed by Firpo et al. (2007). The propensity score is estimated by a logit model.

13 Of course, one could also assume that \( \lambda \) be group-specific. But we do not do that for tractability of the results.
vii. NDR100: the nonparametric doubly robust estimator with optimal bandwidths proposed by Rothe and Firpo (2013). The outcome model is estimated by local linear regression and the propensity score model by local logit regression. Optimal bandwidth calculations are based on a cross-validation procedure. The kernel is Epanechnikov. See “Nonparametric doubly robust estimator (Rothe and Firpo 2013)” in Appendix 2 for more details on the implementation.

viii. NDR50: NDR with half the optimal bandwidth values.

ix. NDR200: NDR with twice the optimal bandwidth values.

We generate samples of size 2000 such that the estimation of the counterfactual is based on 1000 observations. The entire procedure is repeated 10,000 times.

5.2 Results

The results are presented in Table 1. To assess both bias and precision, we report the mean bias in percentage terms relative to the true estimates, as well as the root-mean-squared deviation from the true estimates (referred to as RMSD). For better readability, RMSD is expressed relative to the RMSD of our benchmark, the WPQML estimator. We first turn to the results from the exponential specification in Panel A of Table 1. Under perfect common support, the performance of the QML estimators, WLS and the NDR estimators is generally good. OLS1 is clearly biased, which demonstrates that the best linear approximation to the CEF performs poorly in estimating $\mu^C$ when the true CEF is nonlinear. OLS2 has smaller bias than OLS1 because the higher-order terms of the covariates capture some of the nonlinearities, but its performance is still poor. These findings emphasize the importance of correctly specifying the CEF. Moreover, the reweighted regression (WLS) considerably improves upon unweighted linear regression (OLS1) due to its double robustness property. The numbers for the NDR estimators suggest that undersmoothing is particularly detrimental for the performance of this estimator (especially in terms of precision). Overall, all estimators, except OLS1, OLS2 and NDR50, perform quite well with an average bias of $<1\%$ irrespective of the heteroskedasticity in the data.

If the common support is reduced, we notice a number of changes. First, the performance of the parametric linear estimators (OLS1 and OLS2) further deteriorates considerably. This result is not surprising, since these estimators extrapolate the misspecified CEF to the covariate space of missing overlap. Second, WLS and NDR are now biased and imprecise relative to the QML estimators. NDR50 has small bias because it only relies very little on extrapolation, but this comes at the cost of large variability. Third, the WPQML estimator continues to perform very well and becomes even more precise than PQML under heteroskedasticity because observations with large conditional mean are both more noisy and more likely to be outside the common support. In other words, propensity score weights tend to assign less weight to noisy observations, thus increasing precision. This result is interesting because heteroskedasticity of the form modelled here is often present in real data.

Panel B reports the results where the CEF takes the quadratic form such that the QML estimators are now misspecified and OLS2 is correctly specified. The doubly robust estimators (WPQML, WLS, NDR) continue to perform well alongside OLS2.
### Table 1  Monte Carlo results for aggregate decomposition

| Panel A. Exponential CEF | Hom. BIAS | Hom. RMSD | Het. BIAS | Het. RMSD | Hom. BIAS | Hom. RMSD | Het. BIAS | Het. RMSD |
|--------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| i. WPQML                 | -0.0740   | 1.000     | -0.1601   | 1.000     | -0.0116   | 1.000     | -0.0060   | 1.000     |
| ii. PQML                 | -0.0225   | 1.029     | 0.4708    | 1.008     | 0.1301    | 1.289     | 1.9121    | 1.609     |
| iii. GQML                | -0.0742   | 0.996     | -0.3198   | 0.971     | -0.0115   | 1.000     | 0.0510    | 0.999     |
| iv. OLS1                 | -43.5526  | 2.678     | -43.5446  | 3.866     | -195.2633 | 9.815     | -195.0838 | 25.395    |
| v. OLS2                  | 12.6844   | 2.578     | 12.8229   | 4.319     | 103.1640  | 13.153    | 103.5758  | 43.100    |
| vi. WLS                  | -0.1525   | 1.393     | -0.2158   | 1.094     | -3.2434   | 1.677     | -3.2338   | 1.311     |
| vii. NDR50               | 2.2749    | 7.313     | 1.5694    | 2.806     | -0.0291   | 7.785     | -0.3181   | 4.185     |
| viii. NDR100             | 0.6200    | 1.238     | 0.1940    | 1.039     | -1.4596   | 1.341     | -2.0908   | 1.139     |
| ix. NDR200               | -0.1491   | 1.281     | -0.2476   | 1.110     | -5.2239   | 2.365     | -7.6817   | 2.570     |

| Panel B. Quadratic CEF   | Hom. BIAS | Hom. RMSD | Het. BIAS | Het. RMSD | Hom. BIAS | Hom. RMSD | Het. BIAS | Het. RMSD |
|--------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| i. WPQML                 | -0.3264   | 1.000     | -0.3570   | 1.000     | -1.2072   | 1.000     | -1.2107   | 1.000     |
| ii. PQML                 | 4.6471    | 1.001     | 4.6684    | 1.016     | 16.6375   | 1.119     | 16.8162   | 1.281     |
| iii. GQML                | 1.6510    | 1.091     | 1.4811    | 1.050     | -0.3705   | 1.014     | -0.6337   | 1.052     |
| iv. OLS1                 | -13.5398  | 1.235     | -13.5339  | 1.209     | -53.0401  | 2.220     | -52.9331  | 3.256     |
| v. OLS2                  | -0.0576   | 0.996     | -0.0981   | 1.139     | -0.0120   | 1.901     | -0.1261   | 3.585     |
| vi. WLS                  | -0.4947   | 1.127     | -0.5256   | 1.070     | -4.1751   | 1.449     | -4.1796   | 1.340     |
| vii. NDR50               | -0.1652   | 3.285     | -0.4957   | 2.504     | -3.8845   | 3.068     | -4.1902   | 1.486     |
| viii. NDR100             | -0.6109   | 0.994     | -0.6248   | 1.029     | -4.7378   | 1.458     | -5.0844   | 1.607     |
| ix. NDR200               | -0.5065   | 1.202     | -0.5392   | 1.136     | -6.1427   | 2.605     | -6.3474   | 2.396     |

BIAS refers to the mean percentage deviation from the experimental estimate multiplied by 100. RMSD refers to the root-mean-squared deviation from the experimental estimate and is normalized by the RMSD of the WPQML estimator for ease of comparison. Results are based on 10,000 iterations.
in the scenario with perfect common support because the propensity score weighting ensures consistency. It is only after the overlapping support assumption is violated (small common support) that WLS and NDR also become clearly biased. In contrast, it is surprising that WPQML and also GQML still produce good results with modest bias and good relative precision.

We summarize the main findings from this simulation study as follows. First, as expected, linear regression (OLS1) performs poorly given a nonlinear CEF, and quadratic regression (OLS2) improves only modestly if the CEF is exponential and comes at the expense of more noisy estimates in finite samples. Second, the doubly robust parametric estimators generally produce good results and are usually better than their unweighted counterparts. Third, the nonparametric doubly robust estimator with optimal bandwidths (NDR100) also appears attractive, although the performance appears sensitive to the bandwidth values. A clear drawback of this estimator is that it is computationally much more demanding than the other (parametric) estimators. Finally, comparing results across all the scenarios examined, we find that WPQML produces very convincing results.

6 Application: union wage gap in the USA

We apply the decomposition methods to the union wage gap in the USA. The data set is constructed from the Current Population Survey (CPS) in the year 2012 and contains workers aged between 18 and 64. We define nonunion workers to be the reference group or “controls” \( (D_i = 0) \) and union workers to be the “treatment group” \( (D_i = 1) \).

The dependent variable of interest is hourly wages. Therefore, the counterfactual mean of interest is the average hourly wage union workers would earn if they were paid according to the wage structure of nonunion workers. The set of explanatory variables includes years of education, years of experience (quadratic), dummies for ethnicity, region, marital status, immigration and gender.

The descriptive statistics in Table 2 show a considerable average wage differential between the two groups. Union workers earn 4 dollars more than nonunion workers as measured in terms of AM. The log-wage gap is 0.23, and the corresponding GM wage gap in levels is 4.5 dollars and thus quite smaller than the AM wage gap. Comparing covariates, we see that union workers have more experience and are less likely to live in Southern states but are more likely to be married. For the other explanatory variables, differences seem more modest.

An important lesson from our Monte Carlo exercise is that the decomposition is only sensible for those observations in the common support. We therefore inspect the overlap of the covariate distributions across the two groups. A straightforward way of doing this is to compare propensity score distributions across groups. Figure 2 shows kernel density estimates where the propensity score has been estimated with a logit model using the same set of covariates. We see immediately that estimated propensity scores are bounded well away from unity such that there are no observations in the data that violate the common support assumption in (5).

Table 3 presents the results of the decomposition of average wage differentials.
### Table 2  Descriptive statistics

|                      | Nonunion Mean (SD) | Union Mean (SD) | Difference |
|----------------------|--------------------|----------------|------------|
| Arithmetic mean wage | 21.00 (15.95)      | 24.97 (20.26)  | 3.97       |
| Arithmetic mean of log-wage | 2.843 (0.658) | 3.075 (0.559) | 0.232      |
| Geometric mean wage  | 17.16 (0.658)      | 21.64 (0.559) | 4.48       |
| Experience           | 21.15 (12.69)      | 24.50 (11.72) | 3.35       |
| Years of education   | 13.73 (2.61)       | 14.22 (2.49)  | 0.49       |
| Black                | 0.116 (0.658)      | 0.141 (0.559) | 0.025      |
| Other ethnic minority| 0.089 (0.658)      | 0.074 (0.559) | -0.014     |
| Midwest U.S.         | 0.220 (0.658)      | 0.242 (0.559) | 0.022      |
| Southern U.S.        | 0.392 (0.658)      | 0.189 (0.559) | -0.203     |
| Western U.S.         | 0.215 (0.658)      | 0.285 (0.559) | 0.069      |
| Married              | 0.543 (0.658)      | 0.624 (0.559) | 0.081      |
| Divorced or widowed  | 0.142 (0.658)      | 0.156 (0.559) | 0.014      |
| Foreign-born         | 0.181 (0.658)      | 0.142 (0.559) | -0.039     |
| Female               | 0.492 (0.658)      | 0.454 (0.559) | -0.038     |
| # Observations       | 128,166            | 17,330         |            |

Standard deviations are in parentheses. Sampling weights are used. *Source*: current population survey.

![Density estimate of the propensity score](image-url)

**Fig. 2** Density estimate of the propensity score
Table 3  Decomposition of the union wage gap in the U.S. (2012)

|                                        | $\hat{\Delta}$ (SE) | $\hat{\Delta}^X$ (SE) | %   | $\hat{\Delta}^S$ (SE) | %   |
|---------------------------------------|---------------------|------------------------|-----|------------------------|-----|
| **Arithmetic mean (AM) decomposition** |                     |                        |     |                        |     |
| i. WPQML                              | 3.977 (0.172)       | 2.604 (0.080)          | 65.49 | 1.372 (0.169)          | 34.51 |
| ii. PQML                              | 3.967 (0.167)       | 2.601 (0.077)          | 65.56 | 1.366 (0.165)          | 34.44 |
| iii. QQML                             | 4.061 (0.164)       | 2.486 (0.071)          | 61.23 | 1.574 (0.161)          | 38.77 |
| iv. OLS1                              | 3.967 (0.167)       | 2.747 (0.068)          | 69.23 | 1.221 (0.164)          | 30.77 |
| v. OLS2                               | 3.967 (0.167)       | 2.764 (0.081)          | 69.67 | 1.203 (0.166)          | 30.33 |
| vi. WLS                               | 4.093 (0.171)       | 3.088 (0.082)          | 75.45 | 1.005 (0.167)          | 24.55 |
| vii. NDR100                           | 3.967 (0.173)       | 2.682 (0.081)          | 67.61 | 1.285 (0.167)          | 32.39 |
| **Geometric mean (GM) decomposition**  |                     |                        |     |                        |     |
| OLS ({$\hat{\Delta}_{GM, \text{approx}}$}) | 0.232 (0.005)      | 0.129 (0.003)          | 55.46 | 0.103 (0.005)          | 44.54 |
| OLS ({$\hat{\Delta}_{GM}$})         | 4.483 (0.111)       | 2.357 (0.059)          | 52.58 | 2.126 (0.102)          | 47.42 |

Bootstrap standard errors (1000 iterations) in parentheses

We do not compute detailed decompositions in order to focus on the essential task of estimating the aggregate decomposition. The top panel shows AM decompositions using various estimators, and the bottom panels shows GM decompositions, i.e. decompositions of the log-wage gap. The QML estimates suggest that about 61–65% of the AM gap is due to different characteristics ($\hat{\Delta}^X$). The remainder is due to different wage structures ($\hat{\Delta}^S$) and often referred to as the average “union wage premium”. If we regard union status as a treatment, $\hat{\Delta}^S$ can also be interpreted as the average treatment effect on the treated (ATT) (Firpo et al. 2011). As discussed in Sect. 3, however, the causal interpretability critically hinges on the validity of the CIA. Several issues may be critical here. First, some relevant confounders may not be controlled for (e.g. additional socio-economic characteristics). Second, if union status $D_i$ has a causal influence on some of the covariates in $X_i$ (e.g. work experience), these covariates must be considered intermediate outcomes. In this case, identification of the decomposition can fail in several scenarios. It is possible, for example, that unobserved socio-economic factors or unobserved motivation jointly affect $X_i$ (work experience) and union status $D_i$. Similarly, these unobservables may also jointly affect $X_i$ (work experience) and wages $Y_i$. In these situations, such covariates are “bad controls”; see Huber (2014) for a detailed discussion of these issues. These type of considerations should always be borne in mind when interpreting decomposition results.

Comparing estimates of the AM union wage premium in Table 3, we note that the OLS estimators (and especially WLS) yield smaller point estimates than the QML estimators and the NDR100 estimator. We therefore argue that the choice of the estimator is important. The fact that all QML estimates are similar (and also close to

---

14 NDR100: the propensity score is estimated by a parametric logit model to reduce the very large computational cost of this estimator. The outcome model is estimated by local linear regression, and the optimal bandwidth values selected by the cross-validation criterion are $h_{opt} = 1.2$ and $\lambda_{opt} = 0.9$. 

Springer
Decomposing differences in arithmetic means

For the GM decomposition based on log-wage regressions, we report results both in logs ($\hat{\Delta}_{GM,\text{approx}}$) and in levels ($\hat{\Delta}_{GM}$). The former corresponds to the standard Oaxaca–Blinder decomposition of log-wages, and the latter is obtained by exponentiating means measured in logs. Of course, the absolute values of AM and GM decompositions cannot be compared directly due to the different concepts of the mean. However, we can compare the shares of the composition effect and the structural effect in the total differential. The GM decomposition suggests that the union wage premium explains about 47% of the average wage gap, while the AM decomposition suggests that it explains only about 35% (WPQML estimate). Based on the bootstrapped joint distribution of the estimators in Table 3, we can test whether this difference between $\% \Delta_{GM}^S$ and $\% \Delta_{AM}^S (\hat{\beta}_{WPQML})$ is significant. The corresponding $t$ test yields a $p$ value of 0.0096, such that the null of no difference can be rejected on the 1% level. Performing the same test for the other estimators of the AM decomposition, we also obtain $p$ values smaller than 0.01 in all cases (except for the GQML, where the $p$ value is 0.0399). We therefore reach quite different conclusions when we base the decomposition on either AMs or GMs.

It is important to note that the discrepancy does not necessarily mean that one of the two decompositions is inconsistently estimated. It may entirely come from the fact that we apply the decomposition to different distributions, with one being a nonlinear (monotone) transformation of the other. Unfortunately, the size of the discrepancy depends on the properties of the distribution and is hard to investigate further. However, it may be instructive to point out that the configuration above, where the relative size of the union wage premium is larger in the log-wage decomposition, for example, also arises in the case where outcomes are log-normally distributed and the inequalities $E[\ln y_i | D_i = 0] < E[\ln y_{i0} | D_i = 1] < E[\ln y_i | D_i = 1]$ and $V[\ln y_{i1}] < V[\ln y_{i0}]$ hold.

Thus, choosing between the two concepts of the mean can matter in practice because we reach different conclusions when we evaluate the relative importance of the union wage premium in explaining the average wage differential.

7 Extentions

We briefly discuss how detailed decompositions can be performed in our framework and how issues of endogenous regressors (that violate ignorability) and sample selection can be addressed.

7.1 Detailed decomposition

An important issue in decomposition analysis is to measure how much individual covariates contribute to the differential in mean outcomes. For example, in wage decompositions, it is interesting to examine how much differences in, say, education contribute to the wage gap. This type of question can be answered by performing detailed decompositions.
Generally, a detailed decomposition of the structural effect is difficult regardless of the dependent variable and the specified outcome model. The reason is that the separate contributions of variables without a natural zero point are not interpretable (see Oaxaca and Ransom 1999). For this reason, detailed decompositions mainly focus on the composition effect.

In general, the detailed decomposition is less straightforward in nonlinear models than in linear models because the individual contributions of covariates are not additively separable. However, several approaches are available to perform detailed decompositions in nonlinear models, as the one discussed in Sect. 4.2. First, Yun (2004) uses a first-order Taylor approximation around the means of the covariates. The advantage is that the decomposition is path independent and satisfies the adding-up property (see Firpo et al. 2011, Chapter 2.2). The drawback is that it does not take into account that differences in higher-order moments (e.g. variances) affect outcomes through the nonlinearity of the CEF. Kaiser (2015) offers a method which is similar in spirit but takes into account the effects of such differences in covariate distributions. Second, Rothe (2012) offers a novel approach based on specifying copula functions to decompose the composition effect. The advantage of this method is that it can be applied to any type of outcome model included the one discussed in this paper. However, the flexibility of the approach comes at the expense of a more complex interpretation of the detailed decomposition terms.

7.2 Endogeneity of the covariates

As discussed in Sect. 3, the structural effect and the composition effect are identified even if some of the regressors in $X_i$ are correlated with unobservables, provided that ignorability still holds. (In contrast, the individual contributions of endogenous covariates are not identified in a detailed decomposition.)

Now consider the case where ignorability is violated due to endogenous covariates. In the exponential model, this problem can generally be addressed with instrumental variables as in the linear model. If an appropriate instrument is available, the parameters of the exponential models in (9) can be consistently estimated by the two-step control function approach described in Wooldridge (2010, Ch. 18), given that the CEF of the outcome model is correctly specified. Consider the case of a single endogenous variable, $y_2$, and instrument(s), $Z_i$, satisfying the usual exclusion restriction. In the first step, the reduced form $y_2 = X_i \alpha_1 + Z_i \alpha_2 + v_2$ is estimated by OLS. In the second step, the parameters of the exponential outcome model can be consistently estimated by PQML with the first-stage residuals $\hat{v}_2$ included as an additional regressor.\footnote{Technically, it is conceivable to use this IV method in combination with WPQML as well, but pursuing this further is beyond the scope of this paper.}

\footnote{Yun (2005) proposes a solution to the problem consisting of an ex-post normalization on the coefficients. However, as Firpo et al. (2011) note, such normalizations clearly come at the cost of interpretability.}
7.3 Sample selection

In the example of wage decompositions, there are often concerns that group 0 and 1 select differently into the labour market. This is innocuous as long as selection is explained by differences in $X_i$, but if selection is due to unobservables, some type of correction procedure is required.

In the log-linear model, a popular choice is the two-step approach due to Heckman (1976), which requires a joint normality assumption between the errors in the selection model and the outcome model. Terza (1998) shows how this two-step approach can be extended to the exponential outcome model. The weakness of this procedure is the same as in the linear model; the coefficient estimates are inconsistent if the joint normality assumption is violated. In addition, credible identification usually requires an exclusion restriction.

Less restrictive methods can be used when panel data are available, and confounding unobservables are thought to be time invariant. For individuals where outcomes are observed at least once, missing outcomes in the other periods can be imputed. For example, Melly and Santangelo (2013) assume that individuals’ position in the conditional outcome distribution is time invariant. The advantage of such nonparametric imputation procedures is that distributional assumptions on outcome and selection models are not required.

8 Conclusions

This paper has argued that the standard Oaxaca–Blinder decomposition (Oaxaca 1973; Blinder 1973) based on logarithms brings about statistical problems because it is invalid if the data includes zero and because the log differential is not invariant to changes in higher-order moments of the distribution. Furthermore, the approach also raises conceptual issues because the log differential is an approximate percentage difference in geometric means, which is not an intuitive quantity and thus difficult to interpret.

We have therefore suggested modelling the original dependent variable directly to base the analysis on differences in arithmetic means. At the core of our study, we have proposed a new doubly robust estimation strategy based on a propensity score weighted Poisson quasi-maximum-likelihood (WPQML) estimator. This flexible parametric framework is suitable in many contexts of nonnegative outcomes and easy to implement in practice. Our Monte Carlo study has shown that the proposed estimator performs well in many circumstances when compared to a range of competing estimators. The WPQML estimator produces good results even under a misspecified outcome model or missing overlap in covariate distributions.

In our empirical application to the union wage gap in the USA, we find that quite different conclusions are reached depending on the decomposition approach taken. The arithmetic mean decomposition suggests that the union wage premium is much less important in explaining the union wage gap than the geometric mean decomposition, which is usually used in empirical studies. We conclude that the measure of the mean
can have important implications for decomposition analysis and therefore deserves more careful consideration than it usually receives in empirical research.

Acknowledgments I am grateful to three anonymous referees, Michael Gerfin, Blaise Melly, Stefan Boes, Kaspar Wüthrich, Steven Stillman, and various seminar participants for helpful comments.

Appendix 1: Proof of Theorem 1

Part I. Consistency. The uniform law of large numbers implies \( \hat{\beta}_{\text{WPQML}} \overset{p}{\to} \beta^0 \) such that

\[
\frac{1}{N} \sum_{i:D_i=1} ^N \exp(X_i \hat{\beta}_{\text{WPQML}}) \overset{p}{\to} E[\exp(X_i \beta^0)|D_i = 1].
\]

Part II: Assume the CEF of the outcome model implied by WPQML is correct, i.e. \( E[y_i|X_i, D = 0] = \exp(X_i \beta^0) \forall X_i \in \mathcal{X} \), where \( \beta^0 \in \mathbb{R}^k \) is the vector of “true” coefficients. By applying the law of iterated expectations to the first-order conditions in (10), it follows immediately that \( \beta^0* = \beta^0 \). Assumptions (i)–(iii) then imply

\[
E[y_i|D_i = 1] = E[\exp(X_i \beta^0*)|D_i = 1].
\]

Part III: Assume the model for the propensity score is correct in the sense that \( \lim [\hat{\Lambda}(X_i)] = p(X_i) \). Since \( X_i \) includes a constant, the first-order conditions in (10) imply

\[
E[\omega(X_i)(E[y_i|X_i, D = 0] - \exp(X_i \beta^0*))|D_i = 0] = 0.
\]

Manipulating this expression, we obtain

\[
E[\omega(X_i)E[y_i|X_i, D_i = 0]|D_i = 0] = E[\omega(X_i) \exp(X_i \beta^0*)|D_i = 0]
\]

\[
\int E[y_i|X_i = z] \omega(z) dF_X|D = 0(z) = \int \exp(Z \beta^0*) \omega(z) dF_X|D = 0(z)
\]

\[
\int E[y_i|X_i = z] dF_X|D = 1(z) = \int \exp(Z \beta^0*) dF_X|D = 1(z)
\]

\[
E[y_i|D_i = 1] = E[\exp(X_i \beta^0*)|D_i = 1]
\]

On the LHS, the second line follows from assumption (iii). On both sides, the third line follows from assumption (iii) and Bayes’ rule which implies that \( \omega(X_i) = \frac{p(X_i)}{1-p(X_i)} \frac{1-p}{p} dF_{X|D=1}(X_i) \). The fourth line follows from the law of iterated expectations. \( \square \)

Appendix 2: Monte Carlo simulation

Details on the set-up

In the case of perfect common support, we set \( \delta = 0.2 \) for all slope parameters such that the covariate distributions are very similar across groups. If we estimate a logit model for the propensity score, \( p(X_i) \), more than 99.5% of the predicted values lie between 0.1 and 0.9. Hence, there is (virtually) perfect common support. (We appeal to these
 Decomposing differences in arithmetic means...

Fig. 3  Kernel densities of estimated propensity scores

Table 4  Parameterization of the outcome model

| Coefficients | Exponential CEF | Quadratic CEF |
|--------------|----------------|---------------|
|              | $\beta^1$       | $\beta^0$     | $\beta^1$     | $\beta^0$ |
| Constant     | 2               | 2             | 5             | 5         |
| $X_1, X_2, X_3, X_4, X_5$ | 0.5          | 0.6           | 2             | 3         |

particular threshold values because the findings of Crump et al. (2009) suggest that discarding observations outside this interval is a good approximation to the optimal cut-off rule.) In the second specification, we set $\delta = 0.4$ for all slope parameters to induce large differences in the group-specific covariate distributions. The value is chosen such that roughly 10% of the predicted propensity score values (as estimated from a logit model) are outside the interval $[0.1, 0.9]$. Hence, this represents a scenario in which the common support assumption is severely violated. For illustration, Fig. 3 shows kernel densities of the estimated propensity scores for the two assignment rules based on a random Monte Carlo sample. Comparing counterfactual estimates from these two assignment rules will allow us to assess the impact of the common support assumption on the performance of the estimators.

The CEFs in the simulation exercise are parameterized as shown in Table 4 below. The slope coefficients in $\beta^0$ are larger than the coefficients in $\beta^1$ to induce a structural effect between group 0 and group 1. The coefficients in the quadratic CEF model are larger than in the exponential CEF model because the curvature in the former is less pronounced. The values of the coefficients are chosen such that mean outcomes are of similar size across the two specifications.

In the homoskedastic case, we set $V[\varepsilon_g|x] = V[\varepsilon_g] = 0.5$, which implies $V[y_g|x] \propto E[y_g|x]^2$. Note that this is equivalent to assuming homoskedasticity in the log-linear regression model (Santos Silva and Tenreyro 2006). In the heteroskedastic case, the error variance is proportional to the CEF, $V[\varepsilon_g|x] \propto E[y_g|x]$, which implies $V[y_g|x] \propto E[y_g|x]^3$. In the latter case, we scale the distribution of the error term such
Table 5  Optimal bandwidths of NDR estimator by specification

| Panel | Exponential CEF | Quadratic CEF |
|-------|-----------------|---------------|
|       | Perfect common support | Small common support |
| Outcome model | $\hat{h}_{opt}$ | $\hat{\lambda}_{opt}$ | $\hat{h}_{opt}$ | $\hat{\lambda}_{opt}$ | $\hat{h}_{opt}$ | $\hat{\lambda}_{opt}$ |
| Hom. | 2.04 | 0.35 | 405.90 | 0.83 |
| Het. | 2.45 | 0.53 | 405.90 | 0.83 |
| Hom. | 1.79 | 0.30 | 502.30 | 0.89 |
| Het. | 2.17 | 0.63 | 502.30 | 0.89 |

Bandwidth values in the table are averages across 10 Monte Carlo iterations. Optimal bandwidth calculations are based on a cross-validation procedure. The bandwidth $h$ pertains to the continuous covariates and bandwidth $\lambda$ to the discrete covariates. The outcome model is estimated by local linear regression and the propensity score model by local logit regression. Estimation is based on the Epanechnikov kernel $V[\varepsilon_g] = 0.5$. This enables us to analyse the impact of heteroskedasticity while holding the overall dispersion of the error distribution fixed.

Nonparametric doubly robust estimator (Rothe and Firpo 2013)

The implementation of this estimator is similar to the description in Rothe and Firpo (2013). Using logit or probit for the propensity score model ensures that the predicted values are always in the interval $[0,1]$. To define the local neighbourhood in the five-dimensional covariate space of our data-generating process, we employ a product kernel as suggested by Racine and Li (2004) that allows us to combine the different types of covariates. The kernel weight of observation $(y_i, X_i)$ when estimating the regression function at the covariate values $X = c$ is:

$$K(X_i - c) = \prod_{k=1}^3 \frac{1}{h} W \left( \frac{X_{ki} - c_k}{h} \right) \cdot \prod_{k=4}^5 \lambda^{I(X_{ki} \neq c_k)},$$

where $W(\cdot)$ is the Epanechnikov kernel and $h$ and $\lambda$ are bandwidth parameters. The first component is the usual product kernel for the continuous covariates ($X_1, X_2, X_3$). The larger the bandwidth $h$, the more smoothing is performed. The second component measures the congruence in the values of the dummy covariates ($X_4, X_5$). For example, if $\lambda = 0$, only observations in the same covariate cell (as defined by the values of the dummy variables) are assigned positive weight. If $\lambda = 1$, observations across all cells obtain the same weight. For intermediate values of $\lambda$, observations in different cells obtain positive but less weight than observations in the same cell.
Decomposing differences in arithmetic means...

As in Racine and Li (2004), the same bandwidth $h$ is applied to all continuous covariates. Otherwise, the computational burden imposed by cross-validation and estimation becomes prohibitively large. To make the single bandwidth more attractive, we orthogonalize the matrix with the continuous covariates prior to the estimation procedure. Specifically, we use the QR decomposition solving $X = QR$, where $X$ is the original data matrix, $Q$ is the orthogonalized data matrix, and $R = \text{Chol}((\text{Var}(X))')$ is an upper triangular matrix obtained from the classical Cholesky decomposition of the covariance matrix. As a result, we can use the variables contained in $Q$, whose covariance matrix corresponds to the identity matrix. Since all transformed variables are uncorrelated between themselves and have unit variance, we have drastically reduced the need to specify multiple bandwidths.

The optimal bandwidth values are computed separately for each specification because the curvature of the outcome model, the common support and the error variance (may) have an effect on the optimal amount of smoothing. The values of the optimal bandwidths as obtained from a cross-validation procedure are shown in Table 5. We see that the estimator for the optimal propensity score (PS) model is essentially parametric as implied by the very large bandwidth values. For the estimator of the outcome model, we note that more smoothing is required in the case of heteroskedasticity. Likewise, the exponential model requires more smoothing than the quadratic model because the curvature of the former is more pronounced. The bandwidths in Table 5 are used in each iteration of the the Monte Carlo exercise.

References

Barsky R, Bound J, Charles KK, Lupton JP (2002) Accounting for the black–white wealth gap. J Am Stat Assoc 97(459):663–673. http://www.jstor.org/stable/3085702

Bauer T, Sinning M (2008) An extension of the Blinder–Oaxaca decomposition to nonlinear models. AStA Adv Stat Anal 92(2):197–206. doi: 10.1007/s10182-008-0056-3

Bauer T, Göhlmann S, Sinning M (2007) Gender differences in smoking behavior. Health Econ 16(9):895–909. doi: 10.1002/hec.1259

Black D, Haviland A, Sanders S, Taylor L (2006) Why do minority men earn less? A study of wage differentials among the highly educated. Rev Econ Stat 88(2):300–313. http://www.jstor.org/stable/40042996

Blackburn ML (2007) Estimating wage differentials without logarithms. Labour Econ 14(1):73–98. doi: 10.1016/j.labeco.2005.04.005

Blackburn ML (2008) Are union wage differentials in the united states falling? Ind Rel J Econ Soc 47(3):390–418. doi: 10.1111/j.1468-232X.2008.00526.x

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8(4):436–455. http://www.jstor.org/stable/144855

Cotton J (1988) On the decomposition of wage differentials. Rev Econ Stat 70(2):236–243. http://www.jstor.org/stable/1928307

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96(1):187–199. doi: 10.1093/biomet/asn055

Darity W, Guilkey D, Winfrey W (1995) Ethnicity, race, and earnings. Econ Lett 47(3):401–408. doi: 10.1016/0165-1765(94)00560-O

Edin PA, Richardson K (2002) Swimming with the tide: solidary wage policy and the gender earnings gap. Scand J Econ 104(1):49–67. doi: 10.1111/1467-9442.00271

Fairlie RW (2005) An extension of the Blinder–Oaxaca decomposition technique to logit and probit models. J Econ Soc Meas 30(4):305–316
Firpo S, Fortin N, Lemieux T (2007) Decomposing wage distributions using recentered influence function regressions. University of British Columbia, Vancouver
Firpo S, Fortin N, Lemieux T (2011) Decomposition methods in economics. Handb Labor Econ 4:1–102. doi:10.1016/S0169-7218(11)00407-2
Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: theory. Econometrica 52(3):681–700. http://www.jstor.org/stable/1913471
Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In: Berg SV (ed) Annals of economic and social measurement, vol 5. NBER, Cambridge, pp 475–492
Heckman JJ, Polacheck S (1974) Empirical evidence on the functional form of the earnings-schooling relationship. JASA 69(346):350–354. http://www.jstor.org/stable/2285656
Huber M (2014) Causal pitfalls in the decomposition of wage gaps. Discussion paper, University of St. Gallen
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86(1):4–29. doi:10.1162/003465304323023651
Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47(1):5–86. http://www.jstor.org/stable/27647134
Jann B (2008) The Blinder–Oaxaca decomposition for linear regression models. Stata J 8(4):453–479
Jürges H (2002) The distribution of the German public–private wage gap. Labour 16(2):347–381. doi:10.1111/1467-9914.00199
Kaiser B (2015) Detailed decompositions in nonlinear models. Appl Econ Lett 22(1):25–29. doi:10.1080/13504851.2014.907469
Kim CH (2010) Decomposing the change in the wage gap between white and black men over time, 1980–2005: an extension of the Blinder–Oaxaca decomposition method. Sociol Methods Res 38(4):619–651. doi:10.1177/0049124110366235
Krig IM, Storer P (2006) How much do students matter? Applying the Oaxaca decomposition to explain determinants of adequate yearly progress. Contemp Econ Policy 24(4):563–581. doi:10.1093/cepm/byl003
Kunze A (2008) Gender wage gap studies: consistency and decomposition. Empir Econ 35(1):63–76. doi:10.1007/s00181-007-0143-4
Leslie D, Murphy P (1997) Measuring discrimination by decomposing earnings functions. Appl Econ Lett 4(2):117–120. doi:10.1080/758526708
Machado JA, Mata J (2005) Counterfactual decomposition of changes in wage distributions using quantile regression. J Appl Econ 20(4):445–465. doi:10.1002/jae.788
Manning WG, Mullahy J (2001) Estimating log models: to transform or not to transform? J Health Econ 20(4):461–494. doi:10.1016/S0167-6296(01)00086-8
Melly B (2005) Decomposition of differences in distribution using quantile regression. Labour Econ 12(4):577–590. doi:10.1016/j.labeco.2005.05.006
Melly B, Santangelo G (2013) The evolution of the gender wage gap: 1968–2008. Working paper
Mincer JA (1974) Schooling, experience, and earnings. NBER Books, Cambridge
Mueller S (2012) Works councils and establishment productivity. ILR Rev 65:880–975
Mullahy J (2009) Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. Med Care 47:S104–S108. doi:10.1097/MLR.0b013e31819593
Munn IA, Hussain A (2010) Factors determining differences in local hunting lease rates: insights from Blinder–Oaxaca decomposition. Land Econ 86(1):66–78
Neumark D (1988) Employers’ discriminatory behavior and the estimation of wage discrimination. J Hum Resour 23(3):279–295. http://www.jstor.org/stable/145830
Oaxaca RL (1973) Male–female wage differentials in urban labor markets. Int Econ Rev 14(3):683–709. http://www.jstor.org/stable/2525981
Oaxaca RL, Ransom MR (1999) Identification in detailed wage decompositions. Rev Econ Stat 81(1):154–157. doi:10.1162/003465399767923908
Racine J, Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. J Econom 119(1):99–130. doi:10.1016/S0304-4076(03)00157-X
Rothe C (2012) Decomposing the composition effect. Working paper, Columbia University
Rothe C, Firpo S (2013) Semiparametric estimation and inference using doubly robust moment conditions. IZA discussion paper

898 B. Kaiser
Decomposing differences in arithmetic means...

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688–701. doi: 10.1037/h0037350
Santos Silva JM, Tenreyro S (2006) The log of gravity. Rev Econ Stat 88(4):641–658. doi: 10.1162/rest.88.4.641
Słoczyński T (2012) New evidence on linear regression and treatment effect heterogeneity. University of Warsaw, Warsaw
Terza JV (1998) Estimating count data models with endogenous switching: sample selection and endogenous treatment effects. J Econ 84(1):129–154. doi: 10.1016/S0304-4076(97)00082-1
Vargas Bustamante A, Chen J (2011) Health expenditure dynamics and years of us residence: analyzing spending disparities among latinos by citizenship/nativity status. BMC Health Serv Res 47(2):794–818. doi:10.1111/j.1475-6773.2011.01278.x
Wooldridge JM (1992) Some alternatives to the box–cox regression model. Int Econ Rev 33(4):935–955. http://www.jstor.org/stable/2527151
Wooldridge JM (2007) Inverse probability weighted estimation for general missing data problems. J Econ 141(2):1281–1301. doi:10.1016/j.jeconom.2007.02.002
Wooldridge JM (2010) Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge
Yun MS (2004) Decomposing differences in the first moment. Econ Lett 82(2):275–280. doi:10.1016/j.econlet.2003.09.008
Yun MS (2005) A simple solution to the identification problem in detailed wage decompositions. Econ Inq 43(4):766–772. doi:10.1093/ei/cbi053