NBER WORKING PAPER SERIES

DECOMPOSITION METHODS IN ECONOMICS

Nicole Fortin
Thomas Lemieux
Sergio Firpo

Working Paper 16045
http://www.nber.org/papers/w16045

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2010

We are grateful to Orley Ashenfelter, David Card, Pat Kline, and Craig Riddell for useful comments, and to the Social Sciences and Humanities Research Council of Canada for Research Support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Decomposition Methods in Economics
Nicole Fortin, Thomas Lemieux, and Sergio Firpo
NBER Working Paper No. 16045
June 2010
JEL No. C14,C21,J31,J71

ABSTRACT

This chapter provides a comprehensive overview of decomposition methods that have been developed since the seminal work of Oaxaca and Blinder in the early 1970s. These methods are used to decompose the difference in a distributional statistic between two groups, or its change over time, into various explanatory factors. While the original work of Oaxaca and Blinder considered the case of the mean, our main focus is on other distributional statistics besides the mean such as quantiles, the Gini coefficient or the variance. We discuss the assumptions required for identifying the different elements of the decomposition, as well as various estimation methods proposed in the literature. We also illustrate how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the paper.

Nicole Fortin
Department of Economics
University of British Columbia
#997-1873 East Mall
Vancouver, BC V6T 1Z1
Canada
nifortin@interchange.ubc.ca

Thomas Lemieux
Department of Economics
University of British Columbia
#997-1873 East Mall
Vancouver, BC V6T 1Z1
Canada
and NBER
tlemieux@interchange.ubc.ca

Sergio Firpo
São Paulo School of Economics
sergio.firpo@fgv.br

Contents

1 Introduction

What are the most important explanations accounting for pay differences between men and women? To what extent has wage inequality increased in the United States between 1980 and 2010 because of increasing returns to skill? Which factors are behind most of the growth in U.S. GDP over the last 100 years? These important questions all share a common feature. They are typically answered using decomposition methods. The growth accounting approach pioneered by Solow (1957) and others is an early example of a decomposition approach aimed at quantifying the contribution of labor, capital, and unexplained factors (productivity) to U.S. growth.[1] But it is in labor economics, starting with the seminal papers of Oaxaca (1973) and Blinder (1973), that decomposition methods have been used the most extensively. These two papers are among the most heavily cited in labor economics, and the Oaxaca-Blinder (OB) decomposition is now a standard tool in the toolkit of applied economists. A large number of methodological papers aimed at refining the OB decomposition, and expanding it to the case of distributional parameters besides the mean have also been written over the last three decades.

The twin goals of this chapter are to provide a comprehensive overview of decomposition methods that have been developed since the seminal work of Oaxaca and Blinder, and to suggest a list of best practices for researchers interested in applying these methods.[2] We also illustrate how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the chapter.

At the outset, it is important to note a number of limitations to decomposition methods that are, by and large, beyond the scope of this chapter. As the above examples show, the goal of decomposition methods are often quite ambitious, which means that strong assumptions typically underlie these types of exercises. In particular, decomposition methods inherently follow a partial equilibrium approach. Take, for instance, the question "what would happen to average wages in the absence of unions?" As H. Gregg Lewis pointed out a long time ago (Lewis, 1963, 1986), there are many reasons to believe that eliminating unions would change not only the wages of union workers, but also those of non-union workers. In this setting, the observed wage structure in the non-union sector would not represent a proper counterfactual for the wages observed in the absence of

[1] See also Kendrick (1961), Denison (1962), and Jorgenson and Griliches (1967).

[2] We limit our discussion to so-called "regression-based" decomposition methods where the decomposition focuses on explanatory factors, rather than decomposition methods that apply to additively decomposable indices where the decomposition pertains to population sub-groups. Bourguignon and Ferreira (2005) and Bourguignon, Ferreira, and Leite (2008) are recent surveys discussing these methods.

unions. We discuss these general equilibrium considerations in more detail towards the end of the paper, but generally follow the standard partial equilibrium approach where observed outcomes for one group (or region/time period) can be used to construct various counterfactual scenarios for the other group.

A second important limitation is that while decompositions are useful for quantifying the contribution of various factors to a difference or change in outcomes in an accounting sense, they may not necessarily deepen our understanding of the mechanisms underlying the relationship between factors and outcomes. In that sense, decomposition methods, just like program evaluation methods, do not seek to recover behavioral relationships or "deep" structural parameters. By indicating which factors are quantitatively important and which are not, however, decompositions provide useful indications of particular hypotheses or explanations to be explored in more detail. For example, if a decomposition indicates that differences in occupational affiliation account for a large fraction of the gender wage gap, this suggests exploring in more detail how men and women choose their fields of study and occupations.

Another common use of decompositions is to provide some "bottom line" numbers showing the quantitative importance of particular empirical estimates obtained in a study. For example, while studies after studies show large and statistically significant returns to education, formal decompositions indicate that only a small fraction of U.S. growth, or cross-country differences, in GDP per capita can be accounted for by changes or differences in educational achievement.

Main themes and road map to the chapter

The original method proposed by Oaxaca and Blinder for decomposing changes or differences in the mean of an outcome variable has been considerably improved and expanded upon over the years. Arguably, the most important development has been to extend decomposition methods to distributional parameters other than the mean. For instance, Freeman (1980, 1984) went beyond a simple decomposition of the difference in mean wages between the union and non-union sector to look at the difference in the variance of wages between the two sectors.

But it is the dramatic increase in wage inequality observed in the United States and several other countries since the late 1970s that has been the main driving force behind the development of a new set of decomposition methods. In particular, the new methods introduced by Juhn, Murphy and Pierce (1993) and DiNardo, Fortin and Lemieux (1996) were directly motivated by an attempt at better understanding the underlying factors behind inequality growth. Going beyond the mean introduces a number of important

econometric challenges and is still an active area of research. As a result, we spend a significant portion of the chapter on these issues.

A second important development has been to use various tools from the program evaluation literature to *i*) clarify the assumptions underneath popular decomposition methods, *ii*) propose estimators for some of the elements of the decomposition, and *iii*) obtain formal results on the statistical properties of the various decomposition terms. As we explain below, the key connection with the treatment effect literature is that the "unexplained" component of a Oaxaca decomposition can be interpreted as a treatment effect. Note that, despite the interesting parallel with the program evaluation literature, we explain in the paper that we cannot generally give a "causal" interpretation to the decomposition results.

The chapter also covers a number of other practical issues that often arise when working with decomposition methods. Those include the well known omitted group problem (Oaxaca and Ransom, 1999), and how to deal with cases where we suspect the true regression equation not to be linear.

Before getting into the details of the chapter, we provide here an overview of our main contributions by relating them to the original OB decomposition for the difference in mean outcomes for two groups A and B. The standard assumption used in these decompositions is that the outcome variable Y is linearly related to the covariates, X, and that the error term υ is conditionally independent of X:

$$Y_{gi} = \beta_{g0} + \sum_{k=1}^{K} X_{ik}\beta_{gk} + \upsilon_{gi}, \quad g = A, B, \tag{1}$$

where $\mathbb{E}(\upsilon_{gi}|X_i) = 0$, and X is the vector of covariates ($X_i = [X_{i1}, .., X_{iK}]$). As is well known, the overall difference in average outcomes between group B and A,

$$\widehat{\Delta}_O^{\mu} = \overline{Y}_B - \overline{Y}_A,$$

can be written as:[3]

$$\widehat{\Delta}_O^{\mu} = \underbrace{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)}_{\widehat{\Delta}_S^{\mu} \text{ (Unexplained)}} + \underbrace{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}}_{\widehat{\Delta}_X^{\mu} \text{ (Explained)}}$$

where $\widehat{\beta}_{g0}$ and $\widehat{\beta}_{gk}$ $(k = 1, .., K)$ are the estimated intercept and slope coefficients, respectively, of the regression models for groups $g = A, B$. The first term in the equation is what is usually called the "unexplained" effect in Oaxaca decompositions. Since we mostly focus on wage decompositions in this chapter, we typically refer to this first element as the "wage structure" effect (Δ_S^{μ}). The second component, Δ_X^{μ}, is a composition effect, which is also called the "explained" effect (by differences in covariates) in OB decompositions.

In the above decomposition, it is straightforward to compute both the overall composition and wage structure effects, and the contribution of each covariate to these two effects. Following the existing literature on decompositions, we refer to the overall decomposition (separating Δ_O^{μ} in its two components Δ_S^{μ} and Δ_X^{μ}) as an *aggregate decomposition*. The *detailed decomposition* involves subdividing both Δ_S^{μ}, the wage structure effect, and Δ_X^{μ}, the composition effect, into the respective contributions of each covariate, $\Delta_{S,k}^{\mu}$ and $\Delta_{X,k}^{\mu}$, for $k = 1, .., K$.

The chapter is organized around the following "take away" messages:

A. The wage structure effect can be interpreted as a treatment effect

This point is easily seen in the case where group B consists of union workers, and group A consists of non-union workers. The raw wage gap Δ^{μ} can be decomposed as the sum of the "effect" of unions on union workers, Δ_S^{μ}, and the composition effect linked to differences in covariates between union and non-union workers, Δ_X^{μ}. We can

[3] The decomposition can also be written by exchanging the reference group used for the wage structure and composition effects as follows:

$\widehat{\Delta}_O^{\mu} = \left\{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Ak} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Bk}\right\}$.

Alternatively, the so-called three-fold decomposition uses the same reference group for both effects, but introduces a third interaction term: $\widehat{\Delta}_O^{\mu} = \left\{(\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{K} \overline{X}_{Ak} \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}\right\} + \left\{\sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \left(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}\right)\right)\right\}$. While these various versions of the basic decomposition are used in the literature, using one or the other does not involve any specific estimation issues. For the sake of simplicity, we thus focus on the one decomposition introduced in the text for most of the chapter.

think of the effect of unions for each worker $(Y_{Bi} - Y_{Ai})$ as the individual treatment effect, while Δ_S^μ is the Average Treatment effect on the Treated (ATT). One difference between the program evaluation and decomposition approaches is that the composition effect Δ_X^μ is a key component of interest in a decomposition, while it is a selection bias resulting from confounding factor to be controlled for in the program evaluation literature. By construction, however, one can obtain the composition effect from the estimated treatment effect since $ATT = \Delta_S^\mu$ and $\Delta_X^\mu = \Delta_O^\mu - \Delta_S^\mu$.

Beyond semantics, there are a number of advantages associated with representing the decomposition component Δ_S^μ as a treatment effect:

- The zero conditional mean assumption ($\mathbb{E}(\upsilon|X) = 0$) usually invoked in OB decompositions (as above) is not required for consistently estimating the ATT (or Δ_S^μ). The mean independence assumption can be replaced by a weaker ignorability assumption. Under ignorability, unobservables do not need to be independent or (mean independent) of X as long as their conditional distribution given X is the same in groups A and B. In looser terms, this "selection based on observables" assumption allows for selection biases as long they are the same for the two groups. For example, if unobservable ability and education are correlated, a linear regression of Y on X will not yield consistent estimates of the structural parameters (i.e. the return to education). But the aggregate decomposition remains valid as long as the dependence structure between ability and education is the same in group A and B.

- A number of estimators for the ATT have been proposed in the program evaluation literature including Inverse Probability Weighting (IPW), matching and regression methods. Under ignorability, these estimators are consistent for the ATT (or Δ_S^μ) even if the relationship between Y and X is not linear. The statistical properties of these non-parametric estimators are also relatively well established. For example, Hirano, Imbens and Ridder (2003) show that IPW estimators of the ATT are efficient. Firpo (2007) similarly shows that IPW is efficient for estimating quantile treatment effects. Accordingly, we can use the results from the program evaluation literature to show that decomposition methods based on reweighting techniques are efficient for performing decompositions.[4]

[4]Firpo (2010) shows that for any smooth functional of the reweighted cdf, efficiency is achieved. In other words, decomposing standard distributional statistics such as the variance, the Gini coefficient, or the interquartile range using the reweighting method suggested by DiNardo, Fortin, and Lemieux

- When the distribution of covariates is different across groups, the ATT depends on the characteristics of group B (unless there is no heterogeneity in the treatment effect, i.e. $\beta_{Bk} = \beta_{Ak}$ for all k). The subcomponents of Δ_S^μ associated with each covariate k, $\overline{X}_{Bk}(\beta_{Bk} - \beta_{Ak})$, can be (loosely) interpreted as the "contribution" of the covariate k to the ATT. This helps understand the issues linked to the well-known "omitted group problem" in OB decompositions (see, for example Oaxaca and Ransom, 1999).

B. Going beyond the mean is a "solved" problem for the aggregate decomposition

As discussed above, estimation methods from the program evaluation literature can be directly applied for performing an aggregate decomposition of the gap Δ_O^μ into its two components Δ_S^μ and Δ_X^μ. While most of the results in the program evaluation literature have been obtained in the case of the mean (e.g., Hirano, Imbens and Ridder, 2003), they can also be extended to the case of quantiles (Firpo, 2007) or more general distribution parameters (Firpo, 2010). The IPW estimator originally proposed in the decomposition literature by DiNardo, Fortin and Lemieux (1996) or matching methods can be used to perform the decomposition under the assumption of ignorability. More parametric approaches such as those proposed by Juhn, Murphy and Pierce (1993), Donald, Green, and Paarsch (2000) and Machado and Mata (2005) could also be used. These methods involve, however, a number of assumptions and/or computational difficulties that can be avoided when the sole goal of the exercise is to perform an aggregate decomposition. By contrast, IPW methods involve no parametric assumptions and are an efficient way of estimating the aggregate decomposition.

It may be somewhat of an overstatement to say that computing the aggregate decomposition is a "solved" problem since there is still ongoing research on the small sample properties of various treatment effect estimators (see, for example, Busso, DiNardo, and McCrary, 2009). Nonetheless, performing an aggregate decomposition is relatively straightforward since several easily implementable estimators with good asymptotics properties are available.

C. Going beyond the mean is more difficult for the detailed decomposition

Until recently, no comprehensive approach was available for computing a detailed decomposition of the effect of single covariates for a distributional statistic ν other than

(1996) will be efficient. Note, however, that this result does not apply to the (more complicated) case of the density considered by DiNardo, Fortin, and Lemieux (1996) where non-parametric estimation is involved.

the mean. One popular approach for estimating the subcomponents of Δ_S^ν is Machado and Mata (2005)'s method, which relies on quantile regressions for each possible quantile, combined with a simulation procedure. For the subcomponents of Δ_X^ν, DiNardo, Fortin and Lemieux (1996) suggest a reweighting procedure to compute the contribution of a dummy covariate (like union status) to the aggregate composition effect Δ_X^ν. Altonji, Bharadwaj, and Lange (2007) implemented a generalization of this approach to the case of either continuous or categorical covariates. Note, however, that these latter methods are generally *path dependent*, that is, the decomposition results depend on the order in which the decomposition is performed. Later in this chapter, we show how to make the contribution of the last single covariate path independent in the spirit of Gelbach (2009).

One comprehensive approach, very close in spirit to the original OB decomposition, which is *path independent*, uses the recentered influence function (RIF) regressions recently proposed by Firpo, Fortin, and Lemieux (2009). The idea is to use the (recentered) influence function for the distribution statistic of interest instead of the usual outcome variable Y as the left hand side variable in a regression. In the special case of the mean, the recentered influence function is Y, and a standard regression is estimated, as in the case of the OB decomposition.

More generally, once the RIF regression has been estimated, the estimated coefficients can be used to perform the detailed decomposition in the same way as in the standard OB decomposition. The downside of this approach is that RIF regression coefficients only provide a local approximation for the effect of changes in the distribution of a covariate on the distributional statistics of interest. The question of how accurate this approximation depends on the application at hand.

D. The analogy between quantile and standard (mean) regressions is not helpful

If the mean can be decomposed using standard regressions, can we also decompose quantiles using simple quantile regressions? Unfortunately, the answer is negative. The analogy with the case of the mean just does not apply in the case of quantile regressions.

To understand this point, it is important to recall that the coefficient β in a standard regression has two distinct interpretations. Under the *conditional mean interpretation*, β indicates the effect of X on the conditional mean $\mathbb{E}(Y|X)$ in the model $\mathbb{E}(Y|X) = X\beta$. Using the law of iterated expectations, we also have $\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}(Y|X)] = \mathbb{E}(X)\beta$. This yield an *unconditional mean interpretation* where β can be interpreted as the effect of increasing the mean value of X on the (unconditional) mean value of Y. It is this particular property of regression models, and this particular interpretation of β, which is

used in OB decompositions.

By contrast, only the conditional quantile interpretation is valid in the case of quantile regressions. As we discuss in more detail later, a quantile regression model for the τ^{th} conditional quantile $Q_\tau(X)$ postulates that $Q_\tau(X) = X\beta_\tau$. By analogy with the case of the mean, β_τ can be interpreted as the effect of X on the τ^{th} conditional quantile of Y given X. The law of iterated expectations does not apply in the case of quantiles, so $Q_\tau \neq \mathbb{E}_X[Q_\tau(X)] = \mathbb{E}(X)\beta_\tau$, where Q_τ is the unconditional quantile. It follows that β_τ cannot be interpreted as the effect of increasing the mean value of X on the unconditional quantile Q_τ.

This greatly limits the usefulness of quantile regressions in decomposition problems. Machado and Mata (2005) suggest estimating quantile regressions for all $\tau \in [0,1]$ as a way of characterizing the full conditional distribution of Y given X. The estimates are then used to construct the different components of the aggregate decomposition using simulation methods. Compared to other decomposition methods, one disadvantage of this method is that it is computational intensive.

An alternative regression approach where the estimated coefficient can be interpreted as the effect of increasing the mean value of X on the unconditional quantile Q_τ (or other distributional parameters) has recently been proposed by Firpo, Fortin, and Lemieux (2009). As we mention above, this method provides is one of the few options available for computing a detailed decomposition for distributional parameters other than the mean.

E. Decomposing proportions is easier than decomposing quantiles

A cumulative distribution provides a one-to-one mapping between (unconditional) quantiles and the proportion of observations below this quantile. Performing a decomposition on proportions is a fairly standard problem. One can either run a linear probability model and perform a traditional OB decomposition, or do a non-linear version of the decomposition using a logit or probit model.

Decompositions of quantiles can then be obtained by inverting back proportions into quantiles. Firpo, Fortin and Lemieux (2007) propose doing so using a first order approximation where the elements of the decomposition for a proportion are transformed into elements of the decomposition for the corresponding quantile by dividing by the density (slope of the cumulative distribution function). This can be implemented in practice by estimating recentered influence function (RIF) regressions (see Firpo, Fortin, and Lemieux, 2009).

A related approach is to decompose proportions at every point of the distribution (e.g.

at each percentile) and invert back the whole fitted relationship to quantiles. This can be implemented in practice using the distribution regression approach of Chernozhukov, Fernandez-Val, and Melly (2009).

F. There is no general solution to the "omitted group" problem

As pointed out by Jones (1983) and Oaxaca and Ransom (1999) among others, in the case of categorical covariates, the various elements of Δ_S^μ in a detailed decomposition arbitrarily depend on the choice of the omitted group in the regression model. In fact, this interpretation problem may arise for any covariate, including continuous covariates, that does not have a clearly interpretable baseline value. This problem has been called an identification problem in the literature (Oaxaca and Ransom, 1999, Yun, 2005). But as pointed out by Gelbach (2002), it is better viewed as a conceptual problem with the detailed part of the decomposition for the wage structure effect.

As discussed above, the effect $\beta_{B0} - \beta_{A0}$ for the omitted group can be interpreted as an average treatment effect among the omitted group (group for which $X_k = 0$ for all $k = 1, .., K$). The decomposition then corresponds to a number of counterfactual experiments asking "by how much the treatment effect would change if X_k was switched from its value in the omitted group (0) to its average value $(\overline{X}_{Bk})$"? In cases like the gender wage gap where the treatment effect analogy is not as clear, the same logic applied, nonetheless. For example, one could ask instead "by how much the average gender gap would change if actual experience (X_k) was switched from its value in the omitted group (0) to its average value $(\overline{X}_{Bk})$?"

Since the choice of the omitted group is arbitrary, the elements of the detailed decomposition can be viewed as arbitrary as well. In cases where the omitted group has a particular economic meaning, the elements of the detailed decomposition are more interpretable as they correspond to interesting counterfactual exercises. In other cases the elements of the detailed decomposition are not economically interpretable. As a result, we argue that attempts at providing a general "solution" to the omitted group problem are misguided. We discuss instead the importance of using economic reasoning to propose some counterfactual exercise of interest, and suggest simple techniques to easily compute these counterfactual exercises for any distributional statistics, and not only the mean.

Organization of the chapter

The different methods covered in the chapter, along with their key assumptions and properties are listed in Table 1. The list includes an example of one representative study for each method, focusing mainly on studies on the gender and racial gap (see also Altonji and Blank, 1999), to facilitate comparison across methods. A detailed discussion

of the assumptions and properties follows in the next section. The mean decomposition methodologies comprise the classic OB decomposition, as well as extensions that appeal to complex counterfactuals and that apply to limited depended variable models. The methodologies that go beyond the mean include the classic variance decomposition, methods based on residual imputation, methods based on conditional quantiles and on estimating the conditional distribution, and methods based on reweighting and RIF-regressions.

Since there are a number of econometric issues involved in decomposition exercises, we start in Section 2 by establishing what are the parameters of interest, their interpretation, and the conditions for identification in decomposition methods. We also introduce a general notation that we use throughout the chapter. Section 3 discusses exhaustively the case of decomposition of differences in means, as originally introduced by Oaxaca (1973) and Blinder (1973). This section also covers a number of ongoing issues linked to the interpretation and estimation of these decompositions. We then discuss decompositions for distributional statistics other than the mean in Section 4 and 5. Section 4 looks at the case of the aggregate decomposition, while Section 5 focuses on the case of the detailed decomposition. Finally, we discuss a number of limitations and extensions to these standard decomposition methods in Section 6. Throughout the chapter, we illustrate the "nuts and bolts" of decomposition methods using empirical examples, and discuss important applications of these methods in the applied literature.

2 Identification: What Can We Estimate Using Decomposition Methods?

As we will see in subsequent sections, a large and growing number of procedures are available for performing decompositions of the mean or more general distributional statistics. But despite this rich literature, it is not always clear what these procedures seek to estimate, and what conditions need to be imposed to recover the underlying objects of interest. The main contribution of this section is to provide a more formal theory of decompositions where we clearly define what it is that we want to estimate using decompositions, and what are the assumptions required to identify the population parameters of interest. In the first part of the section, we discuss the case of the aggregate decomposition. Since the estimation of the aggregate decomposition is closely related to the estimation of treatment effects (see the introduction), we borrow heavily from the

identification framework used in the treatment effect literature. We then move to the case of the detailed decomposition where additional assumptions need to be introduced to identify the parameters of interest. We end the section by discussing the connection between program evaluation and decompositions, as well as the more general issue of causality in this context.

Decompositions are often viewed as simple accounting exercises based on correlations. As such, results from decomposition exercises are believed to suffer from the same shortcomings as OLS estimates, which cannot be interpreted as valid estimates of some underlying causal parameters in most circumstances. The interpretation of what decomposition results mean becomes even more complicated in the presence of general equilibrium effects.

In this section, we argue that these interpretation problems are linked in part to the lack of a formal identification theory for decompositions. In econometrics, the standard approach is to first discuss identification (what we want to estimate, and what assumptions are required to interpret these estimates as sample counterparts of parameters of interest) and then introduce estimation procedures to recover the object we want to identify. In the decomposition literature, most papers jump directly to the estimation issues (i.e. discuss procedures) without first addressing the identification problem.[5]

To simplify the exposition, we use the terminology of labor economics, where, in most cases, the agents are workers and the outcome of interest is wages. Decomposition methods can also be applied in a large variety of other settings, such as gaps in test scores between gender (Sohn, 2008), schools (Krieg and Storer, 2006) or countries (McEwan, and Marshall, 2004).

Throughout the chapter, we restrict our discussion to the case of a decomposition for two mutually exclusive groups. This rules out decomposing wage differentials between overlapping groups like Blacks, Whites, and Hispanics, who can be Black or White.[6] In this setting, the dummy variable method (Cain, 1986) with interactions is a more natural way of approaching the problem. Then one can use Gelbach (2009)'s approach, which appeals to the omitted variables bias formula, to compute a detailed decomposition.

The assumption of mutually exclusive groups is not very restrictive, however, since

[5] One possible explanation for the lack of discussion of identification assumptions is that they were reasonably obvious in the case of the original OB decompositions for the mean. The situation is quite a bit more complex, however, in the case of distributional statistics other than the mean. Note also that some recent papers have started addressing these identification issues in more detail. See, for instance, Firpo, Fortin and Lemieux (2007), and Chernozhukov, Fernandez-Val, and Melly (2009).

[6] Alternatively, the overlapping issue can bypassed by excluding Hispanics from the Black and White groups.

most decomposition exercises fall into this category:

ASSUMPTION 1 [***Mutually Exclusive Groups***] *The population of agents can be divided into two mutually exclusive groups, denoted A and B. Thus, for an agent i,* $D_{Ai} + D_{Bi} = 1$*, where* $D_{gi} = \mathbb{1}\{i \text{ is in } g\}$*,* $g = A, B$*, and* $\mathbb{1}\{\cdot\}$ *is the indicator function.*

We are interested in comparing features of the wage distribution for two groups of workers: A and B. We observe wage Y_i for worker i, which can be written as $Y_i = D_{gi}Y_{gi}$, for $g = A, B$, where Y_{gi} is the wage worker i would receive in group g. Obviously, if worker i belongs to group A, for example, we only observe Y_{Ai}.

As in the treatment effect literature, Y_{Ai} and Y_{Bi} can be interpreted as two potential outcomes for worker i. While we only observe Y_{Ai} when $D_{Ai} = 1$, and Y_{Bi} when $D_{Bi} = 1$, decompositions critically rely on counterfactual exercises such as "what would be the distribution of Y_A for workers in group B?". Since we do not observe this counterfactual wage $Y_{A|D_B}$ for these workers, some assumptions are required for estimating this counterfactual distribution.

2.1 Case 1: The Aggregate Decomposition

2.1.1 The overall wage gap and the structural form

Our identification results for the aggregate decomposition are very general, and hold for any distributional statistics.[7] Accordingly, we focus on general distributional measures in this subsection of the chapter.

Consider the case where the distributional statistic of interest is $\nu\left(F_{Y_g|D_s}\right)$, where $\nu\colon \mathcal{F}_\nu \to \mathbb{R}$ is a real-valued functional, and where $\mathcal{F}_\nu$ is a class of distribution functions such that $F_{Y_g|D_s} \in \mathcal{F}_\nu$ if $\left|\nu\left(F_{Y_g|D_s}\right)\right| < \infty$, $g, s = A, B$. The distribution function $F_{Y_g|D_s}$ represents the distribution of the (potential) outcome Y_g for workers in group s. $F_{Y_g|D_s}$ is an observed distribution when $g = s$, and a counterfactual distribution when $g \neq s$.

The overall ν-difference in wages between the two groups measured in terms of the distributional statistic ν is

$$\Delta_O^\nu = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right). \tag{2}$$

[7] Many papers (DiNardo, Fortin, and Lemieux, 1996; Machado and Mata, 2005; Chernozhukov, Fernandez-Val, and Melly, 2009) have proposed methodologies to estimate and decompose entire distributions (or densities) of wages, but the decomposition results are ultimately quantified through the use of distributional statistics. Analyses if the entire distribution look at several of these distributional statistics simultaneously.

The more common distributional statistics used to study wage differentials are the mean and the median. The wage inequality literature has focused on the variance of log wages, the Gini and Theil coefficients, and the differentials between the 90th and 10th percentiles, the 90th and 50th percentiles, and the 50th and 10th percentiles. These latter measures provide a simply way of distinguishing what happens at the top and bottom end of the wage distribution. Which statistic ν is most appropriate depends on the problem at hand.

A typical aim of decomposition methods is to divide Δ_O^ν, the *ν-overall wage gap* between the two groups, into a component attributable to differences in the observed characteristics of workers, and a component attributable to differences in wage structures. In our setting, the wage structure is what links observed characteristics, as well as some unobserved characteristics, to wages.

The decomposition of the overall difference into these two components depends on the construction of a meaningful counterfactual wage distribution. For example, counterfactual states of the world can be constructed to simulate what the distribution of wages would look like if workers had different returns to observed characteristics. We may want to ask, for instance, what would happen if group A workers were paid like group B workers, or if women were paid like men? When the two groups represent different time periods, we may want to know what would happen if workers in year 2000 had the same characteristics as workers in 1980, but were still paid as in 2000. A more specific counterfactual could keep the return to education at its 1980 level, but set all the other components of the wage structure at their 2000 levels.

As these examples illustrate, counterfactuals used in decompositions often consist of manipulating structural wage setting functions (i.e. the wage structure) linking the observed and unobserved characteristics of workers to their wages for each group. We formalize the role of the wage structure using the following assumption:

ASSUMPTION 2 [***Structural Form***] *A worker i belonging to either group A or B is paid according to the wage structure, m_A and m_B, which are functions of the worker's observable (X) and unobservable (ε) characteristics:*

$$Y_{Ai} = m_A\left(X_i, \varepsilon_i\right) \quad \textit{and} \quad Y_{Bi} = m_B\left(X_i, \varepsilon_i\right), \tag{3}$$

where ε_i has a conditional distribution $F_{\varepsilon|X}$ given X, and $g = A, B$.

While the wage setting functions are very general at this point, the assumption implies that there are only three reasons why the wage distribution can differ between

group A and B. The three potential sources of differences are i) differences between the wage setting functions m_A and m_B, ii) differences in the distribution of observable (X) characteristics, and iii) differences in the distribution of unobservable (ε) characteristics. The aim of the aggregate decomposition is to separate the contribution of the first factor (differences between m_A and m_B) from the two others.

When the counterfactuals are based on the alternative wage structure (i.e. using the observed wage structure of group A as a counterfactual for group B)), decompositions can easily be linked to the treatment effects literature. However, other counterfactuals may be based on hypothetical states of the world, that may involve *general equilibrium effects*. For example, we may want to ask what would be the distribution of wages if group A workers were paid according to the pay structure that would prevail if there were no B workers, for example if there were no union workers. Alternatively, we may want to ask what would happen if women were paid according to some non-discriminatory wage structure (which differs from what is observed for either men or women)?

We use the following assumption to restrict the analysis to the first type of counterfactuals.

ASSUMPTION 3 [***Simple Counterfactual Treatment***] *A counterfactual wage structure, m^C, is said to correspond to a simple counterfactual treatment when it can be assumed that $m^C(\cdot,\cdot) \equiv m_A(\cdot,\cdot)$ for workers in group B, or $m^C(\cdot,\cdot) \equiv m_B(\cdot,\cdot)$ for workers in group A.*

It is helpful to represent the assumption using the potential outcomes framework introduced earlier. Consider $Y_{g|D_s}$,where $g = A, B$ indicates the potential outcome, while $s = A, B$ indicates group membership. For group A, the observed wage is $Y_{A|D_A}$, while $Y^C_{B|D_A}$ represents the counterfactual wage. For group B, $Y_{B|D_B}$ is the observed wage while the counterfactual wage is $Y^C_{A|D_B}$. Note that we add the superscript C to highlight counterfactual wages. For instance, consider the case where workers in group B are unionized, while workers in group A are not unionized. The dichotomous variable D_B indicates the union status of workers. For a worker i in the union sector ($D_B = 1$), the observed wage under the "union" treatment is $Y_{B|D_B,i} = m_B(X_i, \varepsilon_i)$, while the counterfactual wage that would prevail if the worker was not unionized is $Y^C_{A|D_B,i} = m^C(X_i, \varepsilon_i) = m_A(X_i, \varepsilon_i)$, $i \in B$. An alternative counterfactual could ask what would be the wage of a non-union worker j if this worker was unionized $Y^C_{B|D_A,j} = m^C(X_j, \varepsilon_j) = m_B(X_j, \varepsilon_j)$, $j \in A$. We note that the choice of which counterfactual to choose is analogous to the choice of reference group

in standard OB decomposition.[8]

What assumption 3 rules out is the existence of another counterfactual wage structure such as $m^*(\cdot)$ that represents how workers would be paid if there were no unions in the labor market. Unless there are no general equilibrium effects, we would expect that $m^*(\cdot) \neq m_A(\cdot)$, and, thus, assumption 3 to be violated.

2.1.2 Four decomposition terms

With this setup in mind, we can now decompose the overall difference Δ_O^ν into the four following components of interest:

D.1 Differences associated with the return to observable characteristics under the structural m functions. For example, one may have the following counterfactual in mind: What if everything but the return to X was the same for the two groups?

D.2 Differences associated with the return to unobservable characteristics under the structural m functions. For example, one may have the following counterfactual in mind: What if everything but the return to ε was the same for the two groups?

D.3 Differences in the distribution of observable characteristics. We have here the following counterfactual in mind: What if everything but the distribution of X was the same for the two groups?

D.4 Differences in the distribution of unobservable characteristics. We have the following counterfactual in mind: What if everything but the distribution of ε was the same for the two groups?

Obviously, because unobservable components are involved, we can only decompose Δ_O^ν into the four decomposition terms after imposing some assumptions on the joint distribution of observable and unobservable characteristics. Also, unless we make additional separability assumptions on the structural forms represented by the m functions, it is virtually impossible to separate out the contribution of returns to observables from that of unobservables. The same problem prevails when one tries to perform a detailed decomposition in returns, that is, provide the contribution of the return to each covariate separately.

[8] When we construct the counterfactual $Y^C_{g|D_s}$, we choose g to be the reference group and s the group whose wages are "adjusted". Thus counterfactual women's wages if they were paid like men would be $Y^C_{m|D_f}$, although the gender gap example is more difficult to conceive in the treatment effects literature.

2.1.3 Imposing identification restrictions: overlapping support

The first assumption we make to simplify the discussion is to impose a common support assumption on the observables and unobservables. Further, this assumption ensures that no single value of $X = x$ or $\varepsilon = e$ can serve to identify membership into one of the groups.

ASSUMPTION 4 [***Overlapping Support***]*: Let the support of all wage setting factors* $[X', \varepsilon']'$ *be* $\mathcal{X} \times \mathcal{E}$*. For all* $[x', e']'$ *in* $\mathcal{X} \times \mathcal{E}$*,* $0 < \Pr[D_B = 1 | X = x, \varepsilon = e] < 1$.

Note that the overlapping support assumption rules out cases where inputs may be different across the two wage setting functions. The case of the wage gap between immigrant and native workers is an important example where the X vector may be different for two groups of workers. For instance, the wage of immigrants may depend on their country of origin and their age at arrival, two variables that are not defined for natives. Consider also the case of changes in the wage distribution over time. If group A consists of workers in 1980, and group B of workers in 2000, the difference in wages over time should take into account the fact that many occupations of 2000, especially those linked to information technologies, did not even exist in 1980. Thus, taking those differences explicitly into account could be important for understanding the evolution of the wage distribution over time.

The case with different inputs can be formalized as follows. Assume that for group A, there is a $d_A + l_A$ vector of observable and unobservable characteristics $[X'_A, \varepsilon'_A]'$ that may include components not included in the $d_B + l_B$ vector of characteristics $[X'_B, \varepsilon'_B]'$ for group B, where d_g and l_g denote the length of the X_g and ε_g vectors, respectively. Define the intersection of these characteristics by the $d + l$ vector $[X', \varepsilon']'$, which represent characteristics common to both groups. The respective complements, which are group-specific characteristics, are denoted by tilde as $\left[X'_{\widetilde{A}}, \varepsilon'_{\widetilde{A}}\right]'$ and $\left[X'_{\widetilde{B}}, \varepsilon'_{\widetilde{B}}\right]'$, such that $\left[X'_{\widetilde{A}}, \varepsilon'_{\widetilde{A}}\right]' \cup [X', \varepsilon']' = [X'_A, \varepsilon'_A]'$ and $\left[X'_{\widetilde{B}}, \varepsilon'_{\widetilde{B}}\right]' \cup [X', \varepsilon']' = [X'_B, \varepsilon'_B]'$.

In that context, the overlapping support assumption could be restated by letting the support of all wage setting factors $[X'_A, \varepsilon'_A]' \cup [X'_B, \varepsilon'_B]'$ be $\mathcal{X} \times \mathcal{E}$. The overlapping support assumption would then guarantee that, for all $[x', e']'$ in $\mathcal{X} \times \mathcal{E}$, $0 < \Pr[D_B = 1 | [X'_A, X'_B] = x, [\varepsilon'_A, \varepsilon'_B] = e] < 1$. The assumption rules out the existence of the vectors $\left[X'_{\widetilde{A}}, \varepsilon'_{\widetilde{A}}\right]$ and $\left[X'_{\widetilde{B}}, \varepsilon'_{\widetilde{B}}\right]$.

In the decomposition of gender wage differentials, it is not uncommon to have explanatory variables for which this condition does not hold. Black, Haviland, Sanders, and Taylor (2008) and Ñopo (2008) have proposed alternative decompositions based on

matching methods to address cases where they are severe gaps in the common support assumption (for observables). For example, Ñopo (2008) divides the gap into four additive terms. The first two are analogous to the above composition and wage structure effects, but they are computed only over the common support of the distributions of observable characteristics, while the other two account for differences in support.

2.1.4 Imposing identification restrictions: ignorability

We cannot separate out the decomposition terms (D.1) and (D.2) unless we impose some separability assumptions on the functional forms of m_A and m_B. For highly complex nonlinear functions of observables X and unobservables ε, there is no clear definition of what would be the component of the m functions associated with either X or ε. For instance, if X and ε represent years of schooling and unobserved ability, respectively, we may expect the return to schooling to be higher for high ability workers. As a result, there is an interaction term between X or ε in the wage equation $m(X, \varepsilon)$, which makes it hard to separate the contribution of these two variables to the wage gap.

Thus, consider the decomposition term D.1* that combines (D.1) and (D.2):

D.1* Differences associated with the return to observable and unobservable characteristics in the structural m functions.

This decomposition term solely reflects differences in the m functions. We call this decomposition term Δ_S^ν, or the "$\nu-$wage structure effect" on the "$\nu-$overall difference", Δ_O^ν. The key question here is how to identify the three decomposition terms (D.1*), (D.3) and (D.4) which, under assumption 4, fully describe Δ_O^ν?

We denote the decomposition terms (D.3) and (D.4) as Δ_X^ν and Δ_ε^ν , respectively. They capture the impact of differences in the distributions of X and ε between groups B and A on the overall difference, Δ_O^ν. We can now write

$$\Delta_O^\nu = \Delta_S^\nu + \Delta_X^\nu + \Delta_\varepsilon^\nu.$$

Without further assumptions we still cannot identify these three terms. There are two problems. First, we have not imposed any assumption for the identification of the m functions, which could help in our identification quest. Second, we have not imposed any assumption on the distribution of unobservables. Thus, even if we fix the distribution of covariates X to be the same for the two groups, we cannot clearly separate all three

components because we do not observe what would happen to the unobservables under this scenario.

Therefore, we need to introduce an assumption to make sure that the effect of manipulations of the distribution of observables X will not be confounded by changes in the distribution of ε. As we now show formally, the assumption required to rule out these confounding effects is the well-known ignorability, or unconfoundedness, assumption.

Consider a few additional concepts before stating our main assumption. For each member of the two groups $g = A, B$, an outcome variable Y_{ig} and some individual characteristics X_i are observed. Y_g and X have a conditional joint distribution, $F_{Y_g,\mathbf{X}|D_g}(\cdot,\cdot) : \mathbb{R} \times \mathcal{X} \to [0,1]$, and $\mathcal{X} \subset \mathbb{R}^k$ is the support of X.

The distribution of $Y_g|D_g$ is defined using the law of iterated probabilities, that is, after we integrate over the observed characteristics we obtain

$$F_{Y_g|D_g}(y) = \int F_{Y_g|X,D_g}(y|X=x) \cdot \mathrm{d}F_{X|D_g}(x), \quad g = A, B. \tag{4}$$

We can construct a counterfactual marginal wage distribution that mixes the conditional distribution of Y_A given X and $D_A = 1$ using the distribution of $X|D_B$. We denote that counterfactual distribution as $F_{Y_A^C:X=X|D_B}$, which is the distribution of wages that would prevail for group B workers if they were paid like group A workers. This counterfactual distribution is obtained by replacing $F_{Y_B|X,D_B}$ with $F_{Y_A|X,D_A}$ (or $F_{X|D_A}$ with $F_{X|D_B}$) in equation (4) :

$$F_{Y_A^C:X=X|D_B} = \int F_{Y_A|X,D_A}(y|X=x) \cdot \mathrm{d}F_{X|D_B}(x). \tag{5}$$

These types of manipulations play a very important role in the implementation of decomposition methods. Counterfactual decomposition methods can either rely on manipulations of F_X, as in DiNardo, Fortin, and Lemieux (1996), or of $F_{Y|X}$, as in Albrecht et al (2003) and Chernozhukov, Fernandez-Val, and Melly (2009).[9]

Back to our union example, $F_{Y_B|X,D_B}(y|X=x)$ represents the conditional distribution of wages observed in the union sector, while $F_{Y_A|X,D_A}(y|X=x)$ represents the conditional distribution of wages observed in the non-union sector. In the case where $g = B$, equation (4) yields, by definition, the wage distribution in the union sector where we integrate the conditional distribution of wages given X over the marginal distribution of X in the

[9] Chernozhukov, Fernandez-Val, and Melly (2009) discuss the conditions under which the two types of decomposition are equivalent.

union sector, $F_{X|D_B}(x)$. The counterfactual wage distribution $F_{Y_A^C:X=X|D_B}$ is obtained by integrating over the conditional distribution of wages in the non-union sector instead (equation (5)). It represents the distribution of wages that would prevail if union workers were paid like non-union workers.

The connection between these conditional distributions and the wage structure is easier to see when we rewrite the distribution of wages for each group in terms of the corresponding structural forms,

$$F_{Y_g|X,D_g}(y|X=x) = \Pr\left(m_g(X,\varepsilon) \leq y | X=x, D_g=1\right), \quad g=A,B.$$

Conditional on X, the distribution of wages only depends, therefore, on the conditional distribution of ε, and the wage structure $m_g(\cdot)$.[10] When we replace the conditional distribution in the union sector, $F_{Y_B|X,D_B}(y|X=x)$, with the conditional distribution in the non-union sector, $F_{Y_A|X,D_B}(y|X=x)$, we are replacing both the wage structure and the conditional distribution of ε. Unless we impose some further assumptions on the conditional distribution of ε, this type of counterfactual exercise will not yield interpretable results as it will mix differences in the wage structure and in the distribution of ε.

To see this formally, note that unless ε has the **same conditional distribution across groups**, the difference

$$\begin{aligned} & F_{Y_B|D_B} - F_{Y_A^C:X=X|D_B} \qquad (6) \\ =& \int \left(\Pr\left(Y \leq y | X=x, D_B=1\right) - \Pr\left(Y \leq y | X=x, D_A=1\right)\right) \cdot \mathrm{d}F_{X|D_B}(x) \\ =& \int \left(\Pr\left(m_B(X,\varepsilon) \leq y | X=x, D_B=1\right) - \Pr\left(m_A(X,\varepsilon) \leq y | X=x, D_A=1\right)\right) \cdot \mathrm{d}F_{X|D_B}(x) \end{aligned}$$

will mix differences in m functions and differences in the conditional distributions of ε given X.

We are ultimately interested in a functional ν (i.e. a distributional statistic) of the wage distribution. The above result means that, in general, $\Delta_S^\nu \neq \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$. The question is under what additional assumptions will the difference between a statistic from the original distribution of wages and the counterfactual distribution, $\Delta_S^\nu = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$ solely depends on differences in the wage structure? The answer is that under a conditional independence assumption, also known

[10] To see more explicitly how the conditional distribution $F_{Y_g|X,D_g}(\cdot)$ depends on the distribution of ε, note that we can write $F_{Y_g|X,D_g}(y|X=x) = \Pr\left(\varepsilon \leq m_g^{-1}(X,y) | X=x, D_g=1\right)$ under the assumption that $m(\cdot)$ is monotonic in ε (see Assumption 9 introduced below).

as *ignorability of the treatment* in the treatment effects literature, we can identify Δ_S^ν and the remaining terms Δ_X^ν and Δ_ε^ν.

ASSUMPTION 5 [***Conditional Independence/Ignorability***]***:*** *For $g = A, B$, let (D_g, X, ε) have a joint distribution. For all x in $\mathcal{X}$: ε is independent of D_g given $X = x$ or, equivalently, $D_g \perp\!\!\!\perp \varepsilon | X$.*

In the case of the simple counterfactual treatment, the identification restrictions from the treatment effect literature may allow the researcher to give a causal interpretation to the results of the decomposition methodology as discussed in subsection 2.3. The ignorability assumption has become popular in empirical research following a series of papers by Rubin and coauthors and by Heckman and coauthors.[11] In the program evaluation literature, this assumption is sometimes called *unconfoundedness* or *selection on observables*, and allows identification of the treatment effect parameter.

2.1.5 Identification of the aggregate decomposition

We can now state our main result regarding the identification of the aggregate decomposition

PROPOSITION 1 [***Identification of the Aggregate Decomposition***]***:***
Under assumptions 3 (simple counterfactual), 4 (overlapping support), and 5 (ignorability), the overall $\nu - gap$, Δ_O^ν, can be written as

$$\Delta_O^\nu = \Delta_S^\nu + \Delta_X^\nu,$$

where
(i) the wage structure term $\Delta_S^\nu = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})$ solely reflects difference between the structural functions $m_B(\cdot,\cdot)$ and $m_A(\cdot,\cdot)$
(ii) the composition effect term $\Delta_X^\nu = \nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ solely reflects the effect of differences in the distribution of characteristics (X and ε) between the two groups

This important result means that, under the ignorability and overlapping assumptions, we can give a structural interpretation to the aggregate decomposition that is formally linked to the underlying wage setting models, $Y_A = m_A(X, \varepsilon)$ and $Y_B = m_B(X, \varepsilon)$.

[11] See, for instance, Rosenbaum and Rubin (1983, 1984), Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd, (1998).

Note also that the wage structure (Δ_S^ν) and composition effect (Δ_X^ν) terms represent algebraically what we have informally defined by terms D.1* and D.3.

As can be seen from equation (6), the only source of difference between $F_{Y_B|D_B}$ and $F_{Y_A^C:X=X|D_B}$ is the difference between the structural functions $m_B(\cdot)$ and $m_A(\cdot)$. Now note that under assumptions 4 and 5, we have that $\Delta_O^\nu = \Delta_S^\nu + \nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$, where

$$F_{Y_A^C:X=X|D_B} - F_{Y_A|D_A} = \int \Pr(Y \leq y|X=x, D_A=1) \cdot \left(\mathrm{d}F_{X|D_B}(x) - \mathrm{d}F_{X|D_A}(x)\right).$$

Thus, $\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ reflects only changes or differences in the distribution of observed covariates. As a result, under assumptions 4 and 5, we identify Δ_X^ν by $\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})$ and set $\Delta_\varepsilon^\nu = 0$. This normalization makes sense as a result of the conditional independence assumption: no difference in wages will be systematically attributed to differences in distributions of ε once we fix these distributions to be the same given X. Thus, all remaining differences beyond Δ_S^ν are due to differences in the distribution of covariates captured by Δ_X^ν.

Combining these two results, we get

$$\Delta_O^\nu = \left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:X=X|D_B})\right] + \left[\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})\right] = \Delta_S^\nu + \Delta_X^\nu \quad (7)$$

which is the main result in Proposition 1.

When the assumptions 3 (simple counterfactual) and 5 (ignorability) are satisfied, the conditional distribution of Y given X remains invariant under manipulations of the marginal distribution of X. It follows that equation (5) represents a valid counterfactual for the distribution of Y that would prevail if workers in group B were paid according to the wage structure $m_A(\cdot)$. The intuition for this result is simple. Since $Y_A = m_A(X, \varepsilon)$, manipulations of the distribution of X can only affect the conditional distribution of Y_A given X if they either *i*) change the wage setting function $m_A(\cdot)$, or *ii*) change the distribution of ε given X. The first change is ruled out by the assumption of a simple counterfactual treatment (i.e. no general equilibrium effects), while the second effect is ruled out by the ignorability assumption.

In the inequality literature, the invariance of the conditional distribution is often introduced as the key assumption required for $F_{Y_A^C:X=X|D_B}$ to represent a valid counterfactual (e.g. DiNardo, Fortin, Lemieux, 1996, Chernozhukov, Fernandez-Val, and Melly, 2009).

ASSUMPTION 6 [***Invariance of Conditional Distributions***] *The construction of the counterfactual wage distribution for workers of group B that would have prevailed if they were paid like group A workers (described in equation (5)), assumes that the conditional wage distribution* $F_{Y_A|X,D_A}(y|X=x)$ *apply or can be extrapolated for* $x \in \mathcal{X}$*, that is, it remains valid when the marginal distribution* $F_{X|D_B}$ *replaces* $F_{X|D_A}$*.*

One useful contribution of this chapter is to show the economics underneath this assumption, i.e. that the invariance assumption holds provided that there are no general equilibrium effects (ruled out by assumption 3) and no selection based on unobservables (ruled out by assumption 5).

Assumption 6 is also invoked by Chernozhukov, Fernandez-Val, and Melly (2009) to perform the aggregate decomposition using the following alternative counterfactual that uses group B as the reference group. Let $F_{Y_B^C:X=X|D_A}$ be the distribution of wages that would prevail for group A workers under the conditional distribution of wages of group B workers. In our union example, this would represent the distribution of wages of non-union workers that would prevail if they were paid like union workers. Under assumption 6, the terms of the decomposition equation are now inverted:

$$\Delta_O^\nu = \left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_B^C:X=X|D_A})\right] + \left[\nu(F_{Y_B^C:X=X|D_A}) - \nu(F_{Y_A|D_A})\right] = \Delta_X^\nu + \Delta_S^\nu.$$

Now the first term Δ_X^ν is the composition effect and the second term Δ_S^ν the wage structure effect.

Whether the assumption of the invariance of the conditional distribution is likely to be satisfied in practice depends on the economic context. If group A were workers in 2005 and group B were workers in 2007, perhaps assumption 6 would be more likely to hold than if group A were workers in 2007 and group B were workers in 2009 in the presence of the 2009 recession. Thus it is important to provide an economic rationale to justify assumption 6 in the same way the choice of instruments has to be justified in terms of the economic context when using an instrumental variable strategy.

2.1.6 Why ignorability may not hold, and what to do about it

The conditional independence assumption is a somewhat strong assumption. We discuss three important cases under which it may not hold:

1. *Differential selection into labor market.* This is the selection problem that Heckman (1979) is concerned with in describing the wage offers for women. In the case

of the gender pay gap analysis, it is quite plausible that the decisions to participate in the labor market are quite different for men and women. Therefore, the conditional distribution of $(X, \varepsilon) | D_B = 1$ may be different from the distribution of $(X, \varepsilon) | D_B = 0$. In that case, both the observed and unobserved components may be different, reflecting the fact that men participating in the labor market may be different in observable and unobservable ways from women who also participate. The ignorability assumption does not necessarily rule out the possibility that these distributions are different, but it constrains their relationship. Ignorability implies that the joint densities of observables and unobservables for groups A and B (men and women) have to be similar up to a ratio of conditional probabilities:

$$\begin{aligned} f_{X,\varepsilon|D_B}(x, e|1) &= f_{X,\varepsilon|D_B}(x, e|0) \cdot f_{X|D_B}(x|1)/f_{X|D_B}(x|0) \\ &= f_{X,\varepsilon|D_B}(x, e|0) \cdot \left(\frac{\Pr(D_B = 1|X = x)}{\Pr(D_B = 0|X = x)}\right) \cdot \left(\frac{\Pr(D_B = 0)}{\Pr(D_B = 1)}\right). \end{aligned}$$

2. *Self-selection into groups A and B based on unobservables.* In the gender gap example there is no selection into groups, although the consequences of differential selection into the labor market are indeed the same. An example where self-selection based on unobservables may occur is in the analysis of the union wage gap. The conditional independence or ignorability assumption rules out selection into groups based on unobservable components ε beyond X. However, the ignorability assumption does not impose that $(X, \varepsilon) \perp\!\!\!\perp D_B$, so the groups may have different marginal distributions of ε. But if selection into groups is based on unobservables, then the ratio of conditional joint densities will in general depend on the value of e being evaluated, and not only on x, as ignorability requires:

$$\frac{f_{X,\varepsilon|D_B}(x, e|1)}{f_{X,\varepsilon|D_B}(x, e|0)} \neq \left(\frac{\Pr(D_B = 1|X = x)}{\Pr(D_B = 0|X = x)}\right) \cdot \left(\frac{\Pr(D_B = 0)}{\Pr(D_B = 1)}\right).$$

3. *Choice of X and ε.* In the previous case, the values of X and ε are not determined by group choice, although they will be correlated and may even explain the choice of the group. In the first example of the gender pay gap, values of X and ε such as occupation choice and unobserved effort may also be functions of gender 'discrimination'. Thus, the conditional independence assumption will not be valid if ε is a function of D_g, even holding X constant. The interpretation of ignorability here is that given the choice of X, the choice of ε will be randomly determined across

groups. Pursuing the gender pay gap example, fixing X (for example education), men and women would exert the same level of effort. The only impact of anticipated discrimination is that they may invest differently in education.

In Section 6, we discuss several solutions to these problems that have been proposed in the decomposition literature. Those include the use of panel data methods or standard selection models. In case 2 above, one could also use instrumental variable methods to deal with the fact that the choice of group is endogenous. One identification issue we briefly address here is that IV methods would indeed yield a valid decomposition, but only for the subpopulation of compliers.

To see this, consider the case where we have a binary instrumental variable Z, which is independent of (ε, T) conditional on X, where T is a categorical variable which indicates 'type'. There are four possible types: a, n, c and d as described below:

ASSUMPTION 7 [***LATE***]***:*** *For $g = A, B$, let (D_g, X, Z, ε) have a joint distribution in $\{0,1\} \times \mathcal{X} \times \{0,1\} \times \mathcal{E}$. We define T, a random variable that may take on four values $\{a, n, c, d\}$, and that can be constructed using D_B and Z according to the following rule: if $Z = 0$ and $D_B = 0$, then $T \in \{n, c\}$; if $Z = 0$ and $D_B = 1$, then $T \in \{a, d\}$; if $Z = 1$ and $D_B = 0$, then $T \in \{n, d\}$; if $Z = 1$ and $D_B = 1$, then $T \in \{n, c\}$.*
(i) For all x in $\mathcal{X}$: Z is independent of (ε, T).
(ii) $\Pr(T = d | X = x) = 0$.

These are the LATE assumptions from Imbens and Angrist (1994), which allow us to identify the counterfactual distribution of $Y_A^C | X, D_B = 1, T = c$. We are then able to decompose the $\nu-$wage gap under that less restrictive assumption, but only for the population of compliers:

$$\begin{aligned} \Delta^{\nu}_{O|T=c} &= \left[\nu(F_{Y_B|D_B,T=c}) - \nu(F_{Y_A^C:X=X|D_B,T=c})\right] + \left[\nu(F_{Y_A^C:X=X|D_B,T=c}) - \nu(F_{Y_A|D_A,T=c})\right] \\ &= \Delta^{\nu}_{S|T=c} + \Delta^{\nu}_{X|T=c} \end{aligned}$$

2.2 Case 2: The Detailed Decomposition

One convenient feature of the aggregate decomposition is that it can be performed without any assumption on the structural functional forms, $m_g(X, \varepsilon)$, while constraining the distribution of unobserved (ε) characteristics.[12] Under the assumptions of Proposition 1,

[12] Differences in the distribution of the ε are fairly constrained under the ignorability assumption. While the unconditional distribution of ε may differ between group A and B (because of differences in the distribution of X), the conditional distribution of ε has to be the same for the two groups.

the composition effect component Δ_X^ν reflects differences in the distribution of X, while the wage structure component Δ_S^ν reflects differences in the returns to either X or ε.

To perform a detailed decomposition, we need to separate the respective contributions of X or ε in both Δ_S^ν and Δ_X^ν, in addition to separating the individual contribution of each element of the vector of covariates X. Thus, generally speaking, the identification of an interpretable detailed decomposition involves stronger assumptions such as functional form restrictions and/or further restrictions on the distribution of ε, like independence with respect to X and D.

Since these restrictions tend to be problem specific, it is not possible to present a general identification theory as in the case of the aggregate decomposition. We discuss instead how to identify the elements of the detailed decomposition in a number of specific cases. Before discussing these issues in detail, it is useful to state what we seek to recover with a detailed decomposition.

PROPERTY 1 [***Detailed Decomposition***] *A procedure is said to provide a detailed decomposition when it can apportion the composition effect, Δ_X^ν, or the wage structure effect, Δ_S^ν, into components attributable to each explanatory variable:*

1. *The contribution of each covariate X_k to the composition effect, $\Delta_{X_k}^\nu$, is the portion of Δ_X^ν that is only due to differences between the distribution of X_k in groups A and B. When $\Delta_X^\nu = \sum_{k=1}^K \Delta_{X_k}^\nu$, the detailed decomposition of the composition effect is said to* ***add up****.*

2. *The contribution of each covariate X_k to the wage structure effect, $\Delta_{S_k}^\nu$, is the portion of Δ_S^ν that is only due to differences in the parameters associated with X_k in group A and B, i.e. to differences in the parameters of $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$ linked to X_k. Similarly, the contribution of unobservables ε to the wage structure effect, $\Delta_{S_\varepsilon}^\nu$, is the portion of Δ_S^ν that is only due to differences in the parameters associated with ε in $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$.*

Note that unobservables do not make any contribution to the composition effect because of the ignorability assumption we maintain throughout most of the chapter. As we mentioned earlier, it is also far from clear how to divide the parameters of the functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$ into those linked to a given covariate or to unobservables. For instance, in a model with a rich set of interactions between observables and unobservables, it is not obvious which parameters should be associated with a given covariate. As

a result, computing the elements of the detailed decomposition for the wage structure involves arbitrary choices to be made depending on the economic question of interest.

The adding-up property is automatically satisfied in linear settings like the standard OB decomposition, or the RIF-regression procedure introduced in Section 5.2. However, it is unlikely to hold in non-linear settings when the distribution of each individual covariate X_k is changed while keeping the distribution of the other covariates unchanged (e.g. in the case discussed in Section 5.3). In such a procedure "with replacement" we would, for instance, first replace the distribution of X_1 for group A with the distribution of X_1 for group B, then switch back to the distribution of X_1 for group A and replace the distribution of X_2 instead, etc.

By contrast, adding up would generally be satisfied in a sequential (e.g. "without replacement") procedure where we first replace the distribution of X_1 for group A with the distribution of X_1 for group B, and then do the same for each covariate until the whole distribution of X has been replaced. The problem with this procedure is that it would introduce some path dependence in the decomposition since the "effect" of changing the distribution of one covariate generally depends on distribution of the other covariates.

For example, the effect of changes in the unionization rate on inequality may depend on the industrial structure of the economy. If unions have a particularly large effect in the manufacturing sector, the estimated effect of the decline in unionization between, say, 1980 and 2000 will be larger under the distribution of industrial affiliation observed in 1980 than under the distribution observed in 2000. In other words, the order of the decomposition matters when we use a sequential (without replacement) procedure, which means that the property of path independence is violated. As we will show later in the chapter, the lack of path independence in many existing detailed decomposition procedures based a sequential approach is an important shortcoming of these approaches.

PROPERTY 2 [***Path Independence***] *A decomposition procedure is said to be path independent when the order in which the different elements of the detailed decomposition are computed does not affect the results of the decomposition.*

A possible solution to the problem of path dependence suggested by Shorrocks (1999) consists of computing the marginal impact of each of the factors as they are eliminated in succession, and then average these marginal effects over all the possible elimination sequences. He calls the methodology the Shapley decomposition, because the resulting formula is formally identical to the Shapley value in cooperative game theory. We return to these issues later in the chapter.

2.2.1 Nonparametric identification of structural functions

One approach to the detailed decomposition is to identify the structural functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$, and then use the knowledge of these structural forms to compute various counterfactuals of interest. For example, one could look at what happens when all the parameters of $m_A(\cdot,\cdot)$ pertaining to education are switched to their values estimated for group B, while the rest of the $m_A(\cdot,\cdot)$ function remains unchanged.

For the purpose of identifying the structural functions $m_A(\cdot,\cdot)$ and $m_B(\cdot,\cdot)$, neither ignorability nor LATE assumptions are very helpful. Stronger assumptions invoked in the literature on nonparametric identification of structural functions (e.g. Matzkin, 2003, Blundell and Powell, 2007, and Imbens and Newey, 2009) have to be used instead:

ASSUMPTION 8 [***Independence***]***:*** *For $g = A, B$, $X \perp\!\!\!\perp \varepsilon | D_g$.*

ASSUMPTION 9 [***Strict Monotonicity in the Random Scalar ε***] *For $g = A, B$ and for all values x in $\mathcal{X}$, ε is a scalar random variable and $m_g(X, \varepsilon)$ is strictly increasing in ε.*

With these two additional assumptions we can write, for $g = A, B$, the functions $m_g(\cdot,\cdot)$ using solely functionals of the joint distribution of (Y, D_g, X). We can assume without loss of generality that $\varepsilon | D_g \sim U[0,1]$, because i) we observe the conditional distributions of $X|D_g$, and ε is a scalar random variable independent of X given D_g. Once we have identified the functions $m_g(\cdot,\cdot)$ for $g = A, B$, we can construct the counterfactual distribution of $F_{Y_A^C : X = X|D_B}$ and compute any distributional statistic of interest.[13]

Note, however, that the monotonicity assumption is not innocuous in the context of comparisons across groups. If there was only one group of workers, the monotonicity assumption would be a simple normalization. With more than one group, however, it requires that the same unobservable variable has positive returns for all groups of workers, which in some settings may not be plausible, though this is automatically satisfied in additively separable models.

There are various reasons why this assumption may be problematic in practice. Empirical wage distributions exhibit many flat spots because of heaping or minimum wage effects. For example, if group A and B corresponded to two different years or countries

[13] This monotonicity assumption can also be found in the works of Matzkin (2003), Altonji and Matzkin (2005), Imbens and Newey (2009), and Athey and Imbens (2006).

with different minimum wages, the monotonicity assumption would not be satisfied.[14] The monotonicity assumption would also break down in the presence of measurement error in wages since the wage residual would now mix measurement error and unobserved skills. As a result, the same amount of unobserved skills would not guarantee the same position in the conditional distribution of residuals in the two groups.

In most labor economics applications, assuming that unobservables are independent of the covariates is a strong and unrealistic assumption. Thus, the identification of the structural functions comes at a relatively high price. The milder assumption of ignorability allows us to identify Δ_S^ν and Δ_X^ν. With full independence, we can go back and identify more terms. In fact, because we obtain an expression for Δ_S^ν, we can construct detailed decompositions by fixing deterministically the values of some covariates while letting other to vary.

2.2.2 Functional form restrictions: decomposition of the mean

A more common approach used in the decomposition literature consists of imposing functional form restrictions to identify the various elements of a detailed decomposition. For instance, detailed decompositions can be readily computed in the case of the mean using the assumptions implicit in Oaxaca (1973) and Blinder (1973). The first assumption is additive linearity of the $m_g(\cdot,\cdot)$ functions. The linearity assumption is also commonly used in quantile-based decomposition methodologies, such as Albrecht et al. (2003), Machado and Mata (2005), and Melly (2006). The linearity assumption allows for heteroscedasticity due, for example, to the fact that the variance of unobservables increases as educational attainment increases.

ASSUMPTION 10 [***Additive Linearity***] *The wage structure, m_A and m_B, are linear additively separable functions in the worker's observable and unobservable characteristics:*

$$Y_{gi} = m_g(X_i, \varepsilon_i) = X_i\beta_g + \upsilon_{ig}, \qquad g = A, B.$$

where $\upsilon_{ig} = h_g(\varepsilon_i)$.

The second assumption implicit in the OB procedure is that the conditional mean of υ_{ig} is equal to zero:

[14] The rank pairing of two outcome variables Y_A and Y_B will be disrupted if as the rank of Y_A remains the same because at a mass point corresponding to the minimum wage, while the rank of Y_B continues to increase in the absence of minimum wage at the rank. Heckman, Smith, and Clements (1997) consider the case of mass points at zero, but the case of multiple mass points is much more difficult.

ASSUMPTION 11 [***Zero Conditional Mean***]: $\mathbb{E}[\upsilon_g|X, D_B] = 0$.

Under mean independence, we have that for $g = A, B$, $\mathbb{E}[Y_g|D_g = 1] = \mathbb{E}[X|D_g = 1]\beta_g$ and therefore we can write the mean counterfactual $\mu(F_{Y_A^C:X=X|D_B})$ as $\mathbb{E}[X|D_B = 1]\beta_A$. Therefore,

$$\Delta_S^\mu = \mathbb{E}[X|D_B = 1](\beta_B - \beta_A) \quad \text{and} \quad \Delta_X^\mu = (\mathbb{E}[X|D_B = 1] - \mathbb{E}[X|D_B = 0])\beta_A.$$

2.2.3 Functional form restrictions: more general decompositions

Under Assumption 11, the error term conveniently drops out of the decomposition for the mean. For more general distributional statistics such as the variance, however, we need more assumptions about the distribution of unobservables to perform a detailed decomposition. If we add the following assumptions on the conditional wage variance and on the function of the unobservables υ_{ig}, we can separate out the wage structure effects of observables and unobservables.

ASSUMPTION 12 [***Constant Returns to Unobservables***]: *For* $g = A, B$, $\upsilon_g = \sigma_g\varepsilon$.

ASSUMPTION 13 [***Homoscedasticity***]: *For* $g = A, B$, $Var[\varepsilon|X, D_g = 1] = 1$.

Under these two additional assumptions, we can identify σ_g, and interpret it as the price of unobservables.[15] Assumption 10 (additive linearity) then allows us to separate out returns to observable and unobservable factors, and to separately identify the contribution of observable and unobservable factors to the wage structure effect. Note that because of the zero conditional mean assumption, only the observable factors influence mean wages.

More formally, consider the counterfactual wage, $Y_A^{C,1}$, for group B workers where the return to unobservables is set to be as in group A,[16]

$$Y_A^{C,1} = X\beta_B + \sigma_A\varepsilon. \tag{8}$$

Under the assumptions 5, and 9 to 13, we can divide the wage structure effect into a component linked to unobservables, $\Delta_{S,\sigma}^\nu$, and a component linked to observables, $\Delta_{S,\beta}^\nu$,

[15] Note that it is possible to relax the homoskedasticity assumption while maintaining the assumption of a single price of unobservables σ_g, as in Chay and Lee (2000). We do not follow this approach here to simplify the presentation.

[16] Note that we depart somewhat from our previous notation, as $Y_A^{C,1}$ retains some components of the structural form of group B, which will disappear in $Y_A^{C,3}$ below.

as follows

$$\Delta_S^\nu = \underbrace{\left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^{C,1}:X=X|D_B})\right]}_{\Delta_{S,\sigma}^\nu} + \underbrace{\left[\nu(F_{Y_A^{C,1}:X=X|D_B}) - \nu(F_{Y_A^C:X=X|D_B}\right]}_{\Delta_{S,\beta}^\nu}.$$

The above assumptions correspond to those implicitly used by Juhn, Murphy and Pierce (1991) in their influential study on the evolution of the black-white wage gap.[17] While it is useful to work with a single "price" of unobservables σ_g, doing so is not essential for performing a detailed decomposition. Juhn, Murphy, and Pierce (1993) [JMP] use a weaker set of assumptions in their influential study of wage differentials over time that we now discuss in more detail.

JMP propose a residual imputation procedure that relies on the key assumption that the rank of worker i in the distribution of υ_A is the same as in the distribution of υ_B, conditional on X. This procedure enables them to perform a decomposition even when the function $h_g(\cdot)$ used to define the regression residual $\upsilon_g = h_g(\varepsilon)$ is not linear (non-linear skill pricing). Since the (conditional) rank of the residual υ_g normalized on a $[0,1]$ scale is simply the cumulative distribution $F_{\upsilon_B|X}(\cdot)$ evaluated at that point, conditional rank preservation can be stated as follows in our context:

ASSUMPTION 14 [***Conditional Rank Preservation***]***:*** *For all individual i, we have $\tau_{Ai}(x_i) = \tau_{Bi}(x_i)$, where $\tau_{Ai}(x_i) = F_{\upsilon_A|X}(\upsilon_{Ai}|X = x_i)$ and $\tau_{Bi}(x_i) = F_{\upsilon_B|X}(\upsilon_{Bi}|X = x_i)$ are the rankings of the residuals υ_{Ai} and υ_{Bi} in their respective conditional distributions.*

Under this assumption, if individual i in group A observed at rank $F_{\upsilon_A|X}(\upsilon_{iA}|X = x_i)$ were in group B instead, he/she would remain at the same rank in the conditional distribution of residuals for that other group (and vice versa). Conditional rank preservation is a direct consequence of the assumptions of ignorability (Assumption 5) and monotonicity (Assumption 9). Under ignorability, the distribution of ε given X does not depend on group membership. Since $\upsilon_A = h_A(\varepsilon)$ and $\upsilon_B = h_B(\varepsilon)$, the rank of υ_A and υ_A in their respective distributions is the same as the rank of ε, provided that $h_A(\cdot)$ and $h_B(\cdot)$ are monotonic.

Note that the assumption of rank preservation is substantially stronger than ignorability. For instance, consider the case where ε is a vector of two ability measures:

[17] See Blau and Kahn (1992, 2003) for an application of the methodology to the study of gender wage differentials across countries .

cognitive ability and manual ability. If cognitive ability is more valued under the wage structure $m_A(\cdot)$ than under the wage structure $m_B(\cdot)$, the ranking of workers in the A and B distributions will be different, which means that neither monotonicity nor rank preservation will hold. But provided that the conditional distribution of cognitive and manual ability given X is the same for groups A and B, ignorability holds, which means that the aggregate decomposition is still identified.

We explain how to implement the JMP procedure in practice in Section 4.3. Compared to the procedure described above to construct the counterfactual wage, $Y_A^{C,1} = X\beta_B + \sigma_A\varepsilon$, the difference is that an imputed residual from the group A distribution is used instead of $\sigma_A\varepsilon$. The idea is to replace the residual υ_{Bi} with rank $\tau_{Bi}(x_i)$ in the conditional distribution of residuals with an imputed residual

$$\upsilon_{Ai}^{C,2} = F^{-1}_{\upsilon_A|X}(\tau_{Bi}(x_i), x_i). \tag{9}$$

The resulting counterfactual wage for group B workers,

$$Y_{Ai}^{C,2} = X\beta_B + \upsilon_{Ai}^{C,2}, \tag{10}$$

can then be used to compute the following two elements of the decomposition:

$$\Delta^{\nu}_{S,\sigma} = \nu(F_{Y_B|D_B}) - \nu(F_{Y_A^{C,2}:X=X|D_B}) \quad \text{and} \quad \Delta^{\nu}_{S,\beta} = \nu(F_{Y_A^{C,2}:X=X|D_B}) - \nu(F_{Y_A^{C}:X=X|D_B}).$$

One important implementation issue we discuss in Section 4.3 is how to impute residuals conditional on X. This is an important limitation of JMP's procedure that can be addressed in a number of ways. One popular approach is to use conditional quantile regressions to allow for different returns to observables that vary along the conditional wage distribution. This approach was proposed by Machado and Mata (2005) and reexamined by Albrecht et al. (2003) and Melly (2005). It relies on the assumption that the conditional distribution of $Y_g|X, D_g$, is completely characterized by the collection of regression quantiles $\{\beta_{g,\tau}; \tau \in (0,1)\}$.

ASSUMPTION 15 [***Heterogenous Returns to Observables***]*: For* $g = A, B$, $Y_{gi} = X_i\beta_{g,\tau} + h_{g,\tau}(\varepsilon_i)$.

ASSUMPTION 16 [***Complete Collection of Linear Conditional Quantiles***]*: For* $g = A, B$, *and* $\forall\tau \in (0,1)$ $\tau = \Pr\left(Y_g \leq x\beta_{g,\tau}|X = x, D_g = 1\right)$.

The above assumptions plus ignorability allow the decomposition of Δ_O^ν into Δ_S^ν and Δ_X^ν . Note that because $\tau = F_{Y_g|X,D_g}\left(x\beta_{g,\tau}|X=x\right)$ for all τ, we are fully parameterizing the conditional distribution of $Y_g|X, D_g$ by $\beta_{g,\tau}$ using all $\tau \in (0,1)$. Thus, once one inverts the conditional quantile to obtain a conditional c.d.f., one can apply equation (4) and (5) to compute an actual or counterfactual distribution.

Many other decomposition methods have been proposed to deal with parametric and nonparametric identification of conditional distribution functions. We have discussed the JMP procedure, as well as extensions to the case of conditional quantiles, as a way of illustrating the kind of assumptions required for identifying detailed decompositions of general distributional statistics. The general message is that more stringent assumptions have to be imposed to perform a detailed decomposition instead of an aggregate decomposition. The same general message would apply if we had discussed the identification of other decomposition procedures such as (to cite a few examples) Donald, Green and Paarsch (2000), Fortin and Lemieux (1998), Melly (2005), Chernozhukov, Fernandez-Val and Melly (2009), and Rothe (2009) instead.

Finally, it is also possible to relax some of the above assumptions provided that other assumptions are used instead. For instance, if one fixes the prices of unobservables to be the same across groups, say to a unit price, then $\Delta_{S,\sigma}^\nu$ reflects in fact changes in the distribution of unobservables. In that case, ignorability does not hold, but because of linearity and zero conditional mean assumptions we can identify the parameter β's. The difference between $\left(Q_{B,\tau}\left(X_i\right) - X_i\beta_B\right)$ and $\left(Q_{A,\tau}\left(X_i\right) - X_i\beta_A\right)$ is interpreted as differences in the τ-quantile of the conditional distribution of ε given X across groups B and A ($Q_{g,\tau}\left(X\right)$ is the τ-quantile of the conditional distribution of Y for group g). Let us state the following normalization assumption,

ASSUMPTION 17 [***Unit Price to Unobservables***]**:** *For $g = A, B$, $\upsilon_g = \sigma_g\varepsilon = \varepsilon$.*

The overall wage gap can then be decomposed as follows

$$\begin{aligned}
\Delta_O^\nu &= \Delta_S^\nu + \Delta_\varepsilon^\nu + \Delta_X^\nu \\
&= \underbrace{\left[\nu(F_{Y_B|D_B}) - \nu(F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B})\right]}_{\Delta_S^\nu} + \underbrace{\left[\nu(F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B}) - \nu(F_{Y_A^C:X=X|D_B})\right]}_{\Delta_\varepsilon^\nu} \\
&\quad + \underbrace{\left[\nu(F_{Y_A^C:X=X|D_B}) - \nu(F_{Y_A|D_A})\right]}_{\Delta_X^\nu}. \qquad (11)
\end{aligned}$$

Because of assumptions 10, 12 and 17, we now have $Y_A = X\beta_A + \varepsilon$ and $Y_B = X\beta_B + \varepsilon$. The first difference Δ_S^ν, corresponds to differences in β's only; the second difference is due to differences in

$$F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B} - F_{Y_A^C:X=X|D_B},$$

which are explained by differences in the conditional distribution of ε given X across groups B and A. Thus, an easy way to obtain that difference is to construct a counterfactual

$$Y_{Ai}^{C,3} = X_i\beta_A + (Y_{Bi} - X_i\beta_B), \tag{12}$$

and to replace $F_{Y_A^C:(X,\varepsilon)=(X,\varepsilon)|D_B}$ with $F_{Y_A^{C,3}:(X,\varepsilon)=(X,\varepsilon)|D_B}$ given that they will be equivalent under the above functional form assumptions.

Finally, the difference Δ_X^ν can be obtained as a residual difference. However, under the maintained assumptions it shall reflect only differences in the marginal distributions of X.

2.3 Decomposition terms and their relation to causality and the treatment effects literature.

We end this section by discussing more explicitly the connection between decompositions and various concepts introduced in the treatment effects literature. As it turns out, when the counterfactuals are based on hypothetical alternative wage structures, they can be easily linked to the treatment effects literature. For example: What if group A workers were paid according to the wage structure of group B? What if all workers were paid according to the wage structure of group A?

Define the overall average treatment effect (ATE) as the difference between average wages if everybody were paid according to the wage structure of group B and average wages if everybody were paid according to the wage structure of group A. That is:

$$ATE = \mathbb{E}[Y_B] - \mathbb{E}[Y_A],$$

where switching a worker of from "type A" to "type B" is thought to be the "treatment".

We also define the average treatment effect on the treated (ATT) as the difference between actual average wages of group B workers and average wages if group B workers were paid according to the pay structure of group A. That is:

$$ATT = \mathbb{E}[Y_B|D_B = 1] - \mathbb{E}[Y_A|D_B = 1].$$

These treatment effects can be generalized to other functionals or statistics of the wage distribution. For example, define ν-TE, the ν-treatment effect, as

$$\nu\text{-}TE = \nu\left(F_{Y_B}\right) - \nu\left(F_{Y_A}\right),$$

and its version applied to the subpopulation of "treated", ν-TT as

$$\nu\text{-}TT = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_B}\right).$$

The distributions F_{Y_B}, F_{Y_A} and $F_{Y_A|D_B}$ are not observed from data on (Y, D_B, X).[18] Following the treatment effects literature, we could in principle identify these parameters if "treatment" was randomly assigned. This is hardly the case, at least for our examples, and one needs extra identifying restrictions. In fact, we note that ignorability and common support assumptions (which together are termed strong ignorability after Rosenbaum and Rubin, 1983) are sufficient to guarantee identification of the previous parameters. For example under strong ignorability, for $g = A, B$

$$\begin{aligned} F_{Yg}(y) &= \mathbb{E}\left[F_{Y|X,D_g}(y|X)\right], \\ F_{Y_A|D_B}(y) &= \mathbb{E}\left[F_{Y|X,D_A}(y|X)\,|D_B = 1\right]. \end{aligned}$$

Under ignorability, it follows that $F_{Y_A|D_B} \sim F_{Y_A^C:X=X|D_B}$. Then $\Delta_S^\nu = \nu$-TT and $\Delta_X^\nu = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right) - (\nu\text{-}TT)$. Reweighting methods, as discussed by DiNardo, Fortin and Lemieux (1996), Hirano, Imbens and Ridder (2003), Firpo (2007, 2010) have implicitly or explicitly assumed strong ignorability to identify specific ν-treatment effects.

It is interesting to see how the choice of the reference or base group is related to the treatment effects literature. Consider the treatment effect parameter for the non-treated, ν-TNT:

$$\nu\text{-}TNT = \nu\left(F_{Y_B|D_A}\right) - \nu\left(F_{Y_A|D_A}\right).$$

Under strong ignorability, we have $F_{Y_B|D_A}(\cdot) = \mathbb{E}\left[F_{Y|X,D_B}(\cdot|X)\,|D_B = 0\right] = F_{Y_B^C:X=X|D_A}(\cdot)$. Thus, in this case, $\Delta_S^\nu = \nu$-TNT and $\Delta_X^\nu = \nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right) - (\nu\text{-}TNT)$.

We could also consider other decompositions, such as:

$$\nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_A|D_A}\right) = \nu\text{-}TE + \left(\nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_B}\right)\right) + \left(\nu\left(F_{Y_A}\right) - \nu\left(F_{Y_A|D_A}\right)\right),$$

[18] Only $F_{Y_{B|D_B}}$ and $F_{Y_{A|D_A}}$ are observed.

where F_{Y_B} includes the actual wages of group B workers and the counterfactual wages of group A workers if they were are paid like group B workers, and conversely for F_{Y_A}. In this case, the wage structure effect is ν-TE, while the composition effect is the sum $\left(\nu\left(F_{Y_B|D_B}\right) - \nu\left(F_{Y_B}\right)\right) + \left(\nu\left(F_{Y_A}\right) - \nu\left(F_{Y_A|D_A}\right)\right)$.[19]

The above discussion reveals that the reference group choice problem is just a matter of choosing a meaningful counterfactual. There will be no right answer. In fact, we see that analogously to the treatment effects literature, where treatment effect parameters are different from each other because they are defined over distinct subpopulations, the many possible ways of performing decompositions will reflect the reference group that we want to emphasize.

We conclude this section by discussing briefly the relationship between causality, structural parameters and decomposition terms. In this section, we show that the decomposition terms do not necessarily rely on the identification of structural forms. Whenever we can identify those structural functions linking observable and unobservable characteristics to wages, we benefit from being able to perform counterfactual analysis that we may not be able to do otherwise. However, that comes at the cost of having to impose either strong independence assumptions, as in the case of nonparametric identification, or restrictive functional form assumptions plus some milder independence assumption (mean independence, for instance) between observables and unobservables within each group of workers.

If we are, however, interested in the aggregate decomposition terms Δ_X^ν and Δ_S^ν, we saw that a less restrictive assumption is sufficient to guarantee identification of these terms. Ignorability is the key assumption here as it allows fixing the conditional distribution of unobservables to be the same across groups. The drawback is that we cannot separate out the wage structure effects associated with particular observable and unobservable characteristics.

The treatment effect literature is mainly concerned with causality. Under what conditions can we claim that although identifiable under ignorability, Δ_S^ν may have a causal interpretation? The conditions under which we could say that Δ_S^ν is a causal parameter are very stringent and unlikely to be satisfied in general cases. There are two main reasons for that, in our view.

First, in many cases, "treatment" is not a choice or a manipulable action. When decomposing gender or race in particular, we cannot conceive workers choosing which

[19] We note that this last decomposition corresponds, in the OB context, to the so-called three-fold decomposition presented in footnote 3.

group to belong to.[20] They may have different labor market participation behavior, which is one case where ignorability may not hold, as discussed in subsection 2.1.6. However, workers cannot choose treatment. Thus, if we follow, for example, Holland (1986)'s discussion of causality, we cannot claim that Δ_S^{ν} is a causal parameter.

A second reason for failing to assign causality to the pay structure effect is that most of the observable variables considered as our X (or unobservables ε) are not necessarily pre-treatment variables.[21] In fact, X may assume different values as a consequence of the treatment. In the treatment effects literature, a confounding variable X may have different distributions across treatment groups. But that is not a direct action of the treatment. It should only be a selection problem: People who choose to be in a group may have a different distribution of X relative to people who choose to be in the other group. When X is affected by treatment, we cannot say that controlling for X we will obtain a causal parameter. In fact, what we will obtain is a partial effect parameter, netted from the indirect effect through changes in X.

3 Oaxaca-Blinder – decompositions of mean wages differentials

In this section, we review the basics of OB decompositions, discussing at length some thorny issues related to the detailed decomposition. We also address alternative choices of counterfactuals, including the case of the pooled regression that uses a group membership dummy to obtain a measure of the aggregate wage structure effect. We introduce a reweighted-regression decomposition as an attractive alternative when the linearity of the conditional mean as a function of the covariates is questionable. Finally, we briefly discuss the extensions of OB decompositions to limited dependent variable models, which carry some of the issues, such as path dependence, that will surface in methods that go beyond the mean.

3.1 Basics

Despite its apparent simplicity, there are many important issues of estimation and interpretation in the classic OB decomposition. The goal of the method is to decompose

[20] The union/non-union wage gaps or private/public sector wage gaps are more amenable to choice.

[21] Note that some analyses (e.g. Neal and Johnson, 1996) take great care to focus on pre-market variables.

differences in mean wages, μ, across two groups. The wage setting model is assumed to be linear and separable in observable and unobservable characteristics (Assumption 10):

$$Y_g = X\beta_g + \upsilon_g, \qquad \text{for} \quad g = A, B. \tag{13}$$

where $\mathbb{E}[\upsilon_g|X] = 0$ (Assumption 11). Letting $D_B = 1$ be an indicator of group B membership, and taking the expectations over X, the overall mean wage gap Δ_O^μ can be written as

$$\begin{aligned}
\Delta_O^\mu &= \mathbb{E}[Y_B|D_B = 1] - \mathbb{E}[Y_A|D_B = 0] \\
&= \mathbb{E}[\mathbb{E}(Y_B|X, D_B = 1)|D_B = 1] - \mathbb{E}[\mathbb{E}(Y_A|X, D_B = 0)|D_B = 0] \\
&= (\mathbb{E}\left[X|D_B = 1\right]\beta_B + \mathbb{E}\left[\upsilon_B|D_B = 1\right]) - (\mathbb{E}\left[X|D_B = 0\right]\beta_A + \mathbb{E}\left[\upsilon_A|D_B = 0\right])
\end{aligned}$$

where $\mathbb{E}\left[\upsilon_A|D_B = 0\right] = \mathbb{E}\left[\upsilon_B|D_B = 1\right] = 0$. Adding and subtracting the average counterfactual wage that group B workers would have earned under the wage structure of group A, $\mathbb{E}\left[X|D_B = 1\right]\beta_A$, the expression becomes

$$\begin{aligned}
\Delta_O^\mu &= \mathbb{E}\left[X|D_B = 1\right]\beta_B - \mathbb{E}\left[X|D_B = 1\right]\beta_A + \mathbb{E}\left[X|D_B = 1\right]\beta_A - \mathbb{E}\left[X|D_B = 0\right]\beta_A \\
&= \mathbb{E}\left[X|D_B = 1\right](\beta_B - \beta_A) + (\mathbb{E}\left[X|D_B = 1\right] - \mathbb{E}\left[X|D_B = 0\right])\beta_A \\
&= \qquad\qquad \Delta_S^\mu \qquad\qquad + \qquad\qquad \Delta_X^\mu.
\end{aligned}$$

Replacing the expected value of the covariates $\mathbb{E}\left[X|D_B = d\right]$, for $d = 0, 1$, by the sample averages $\overline{X}_g$, the decomposition is estimated as

$$\begin{aligned}
\widehat{\Delta}_O^\mu &= \overline{X}_B\widehat{\beta}_B - \overline{X}_B\widehat{\beta}_A + \overline{X}_B\widehat{\beta}_A - \overline{X}_A\widehat{\beta}_A && (14) \\
&= \overline{X}_B\left(\widehat{\beta}_B - \widehat{\beta}_A\right) + \left(\overline{X}_B - \overline{X}_A\right)\widehat{\beta}_A && (15) \\
&= \qquad\quad \widehat{\Delta}_S^\mu \qquad\quad + \qquad\quad \widehat{\Delta}_X^\mu. && (16)
\end{aligned}$$

The first term in equation (15) is the wage structure effect, $\widehat{\Delta}_S^\mu$, while the second term is the composition effect, $\widehat{\Delta}_X^\mu$. Note that in cases where group membership is linked to some immutable characteristics of the workers, such as race or gender, the wage structure effect has also been called the "unexplained" part of the wage differentials or the part due to "discrimination".

The OB decomposition is very easy to use in practice. It is computed by plugging in the sample means and the OLS estimates $\widehat{\beta}_g$ in the above formula. Various good

implementations of the procedure are available in existing software packages.[22] Table 2 displays the various underlying elements of the decomposition in the case of the gender wage gap featured in O'Neill and O'Neill (2006) using data from the NLSY79. The composition effect is computed as the difference between the male and female means reported in column (1) multiplied by the male coefficients reported in column (2).[23] The corresponding wage structure effect is computed as the difference between the male and female coefficients reported in columns (2) and (3). The results are reported in column (1) of Table 3. The composition effect accounts for 0.197 (0.018) log points out of the 0.233 (0.015) average log wage gap between men and women in 2000. When the male wage structure is used as reference, only an insignificant 0.036 (0.019) part of the gap (the wage structure effect) is left unexplained.

Because of the additive linearity assumption, it is easy to compute the various elements of the detailed decomposition. The wage structure and composition effects can be written in terms of sums over the explanatory variables

$$\widehat{\Delta}_S^{\mu} = (\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \sum_{k=1}^{M} \overline{X}_{Bk}(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak}), \tag{17}$$

$$\widehat{\Delta}_X^{\mu} = \sum_{k=1}^{M} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}, \tag{18}$$

where $(\widehat{\beta}_{B0} - \widehat{\beta}_{A0})$ represents the omitted group effect, and where $\overline{X}_{gk}$ and $\widehat{\beta}_{gk}$ represent the k^{th} element of $\overline{X}_g$ and $\widehat{\beta}_g$, respectively. $\left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\beta}_{Ak}$ and $\overline{X}_{Bk}(\widehat{\beta}_{Bk} - \widehat{\beta}_{Ak})$ are the respective contributions of the k^{th} covariate to composition and wage structure effect. Each element of the sum $\widehat{\Delta}_S^{\mu}$ can be interpreted as the contribution of the difference in the returns to the k^{th} covariate to the total wage structure effect, evaluated at the mean value of X^k. Whether or not this decomposition term is economically meaningful depends on the choice of the omitted group, an issue we discuss in detail in Section 3.2 below.[24]

Similar to O'Neill and O'Neill (2006), Table 3 reports the contribution of single variables and groups of variables to composition (upper panel) and wage structure effects (lower panel). Life-time work experience 'priced' at the male returns to experience stands out as the factor with the most explanatory power (0.137 out of 0.197, or 69%) for com-

[22] The empirical applications of the OB procedure in this chapter use Jann (2008) procedures in Stata.

[23] As is common in the gender pay gap literature, we begin with the counterfactual that use group B (males) as the reference group. In column (3) of Table 3, we present the decomposition that corresponds to equation (15), that is uses group A (females) as the reference group.

[24] In particualr, see the discussion of the case of *scalable* or *categorical* variables below.

position effects. The wage structure effects are not significant in this example, except for the case of industrial sectors which we discuss below.

Because regression coefficients are based on partial correlations, an OB decomposition that includes all K explanatory variables of interest satisfies the property of path independence (Property 2). Note, though, that a sequence of Oaxaca-Blinder decompositions, each including a subset of the K variables, would suffer from path dependence, as pointed out by Gelbach (2009). Despite these attractive properties, there are some important limitations to the standard OB decomposition that we now address in more detail.

3.2 Issues with detailed decompositions: choice of the omitted group

There are many relevant economic questions that can be answered with the detailed decomposition of the composition effect $\widehat{\Delta}_X^{\mu}$ in equation (18). For example, what has been the contribution of the gender convergence in college enrollment to the gender convergence in average pay? There are also some important questions that are based on the detailed decomposition of the wage structure effect $\widehat{\Delta}_S^{\mu}$. For example, consider the related "swimming upstream" query of Blau and Kahn (1997). To what extent have the increases in the returns to college slowed down the gender convergence in average pay? Or, to what extent has the decline in manufacturing and differences in industry wage premia contributed to that convergence?

Some difficulties of interpretation arise when the explanatory variables of interest are categorical (with more than two categories, or more generally, in the case of scalable variables, such as test scores) and do not have an absolute interpretation. In OB decompositions, categorical variables generate two problems. The first problem is that categorical or scalable variables do no have a natural zero, thus the reference point has to be chosen arbitrarily. The conventional practice is to omit one category which becomes the reference point for the other groups. This generates some interpretation issues even in the detailed decomposition of the composition effect.

Returning to our NLSY example, assume that the industry effects can captured by four dummy variables, $ind1$ to $ind4$, for the broad sectors: *i)* primary, construction, transportation & utilities, *ii)* manufacturing, *iii)* education and health services & public administration, and *iv)* other services. Consider the case where $ind1$ is the omitted category, $\beta_{g,ind1} = 0$, and denote by $\beta_{g,indk}$ the coefficients from the wage regression, as

in column (2) of Table 2. Denote by $\beta'_{g,indk}$ the coefficients of a wage regression where $ind3$ is the omitted category, $\beta'_{g,ind3} = 0$, as in column (4) of Table 2, so that $\widehat{\beta}'_{g,ind4} = \widehat{\beta}_{g,ind4} - \widehat{\beta}_{g,ind3}$ [0.066=0.007-(-0.059)]. In our example, given the large difference in the coefficients of manufacturing between columns (2) and (4) of Table 2, this could mistakenly lead one to conclude that the effect of the underrepresentation of women in the manufacturing sector has an effect three times as large (0.237-0.120)×0.093 in one case (education and health omitted) as (0.237-0.120)×0.034 in the other case (primary omitted). In the first case, the underrepresentation of women in the manufacturing sector is 'priced' at the relative returns in the manufacturing versus the education and health sector, while in the other it is 'priced' at the relative returns in the manufacturing versus the primary sector.[25]

Note, however, that the overall effect of 0.017 (0.006) of gender differences in industrial sectors on the gender wage gap, is the same in columns (1) and (2) of Table 3. To simplify the exposition, consider the special case where industrial sectors are the only explanatory factors in the wage regression. It follows that the composition effect,

$$\widehat{\Delta}_X^{\mu} = \sum_{k=1}^{4} \left[\overline{X}_{B,indk} - \overline{X}_{A,indk}\right] \widehat{\beta}_{A,indk}, \tag{19}$$

is unaffected by the choice of omitted category.[26]

The second problem with the conventional practice of omitting one category to identify the coefficients of the remaining categories is that in the unexplained part of the decomposition one cannot distinguish the part attributed to the group membership (true "unexplained" captured by the difference in intercepts) from the part attributed to differences in the coefficient of the omitted or base category.[27] These difficulties with the detailed decomposition of the unexplained part component were initially pointed by Jones (1983) who argued that "this latter decomposition is in most applications arbitrary and uninterpretable" (p.126). Pursuing the example above, the effect of industry wage differ-

[25] This interpretation issue also arises in other applications that use categorical variables, notably the inter-industry wage differentials literature. In this literature, following the seminal Krueger and Summers (1988) paper on inter-industry wage differentials, the standard practice is to express industry differentials as deviations from an employment-share weighted mean, a well-defined average.

[26] In the first regression, the composition effect is given by $\sum_{k\neq 1}\left(\overline{X}_{B,indk} - \overline{X}_{A,indk}\right)\widehat{\beta}_{A,indk}$, and in the second regression, $\sum_{k\neq 3}\left(\overline{X}_{B,indk} - \overline{X}_{A,indk}\right)\widehat{\beta}'_{A,indk} = \sum_{k\neq 3}\left(\overline{X}_{B,indk} - \overline{X}_{B,indk}\right)[\widehat{\beta}_{A,indk} - \widehat{\beta}_{A,ind1}] = \sum_{k\neq 1}\left(\overline{X}_{A,indk} - \overline{X}_{A,indk}\right)\widehat{\beta}_{A,indk}$ because $\sum_{k\neq 3}\overline{X}_{g,indk} = 1 - \overline{X}_{B,ind1}$, $g = A, B$.

[27] Actually, problems arise when they are more than two categories, Blinder (1973, footnote 13) and Oaxaca (2007) correctly point out that in the case of a binary dummy variable, these problems do not occur.

entials on the gender wage gap is be given by the right-hand side sums in the following expressions

$$\widehat{\Delta}_S^{\mu} = \left[(\widehat{\beta}'_{B0} + \widehat{\beta}'_{B,ind1}) - (\widehat{\beta}'_{A0} + \widehat{\beta}'_{A,ind1})\right] + \sum_{k\neq 1} \overline{X}_{B,indk} \left(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}\right) \quad (20)$$

$$\widehat{\Delta_S^{\mu}}' = \left[(\widehat{\beta}_{B0} + \widehat{\beta}_{B,ind3}) - (\widehat{\beta}_{A0} + \widehat{\beta}_{A,ind3})\right] + \sum_{k\neq 3} \overline{X}_{B,indk} \left(\widehat{\beta}'_{B,indk} - \widehat{\beta}'_{A,indk}\right), \quad (21)$$

where $\widehat{\beta}_{g0} = \widehat{\beta}'_{g0} + \widehat{\beta}'_{g,ind1}$ and $\widehat{\beta}'_{g0} = \widehat{\beta}_{g0} + \widehat{\beta}_{g,ind3}$, $g = A, B$. The overall wage structure effect is the same irrespective of the omitted category $\widehat{\Delta}_S^{\mu} = \widehat{\Delta_S^{\mu}}'$, as shown in the last row of column (1) and (2) of Table 3. However, the overall effect of differences in the returns to industrial sectors, given by the right hand side sums with either choice of omitted group, -0.092 (0.033) in column (1) and 0.014 (0.028) in column (2), are different because different parts of the effect is hidden in the intercepts [0.128 (0.213) in column (1) and 0.022 (0.212) in column (2)].[28]

This invariance issue has been discussed by Oaxaca and Ransom (1999), Gardeazabal and Ugidos (2004), and Yun (2005, 2008), who have proposed tentative solutions to it. These solutions impose some normalizations on the coefficients to purge the intercept from the effect of the omitted category, either by transforming the dummy variables before the estimation, or by implementing the restriction, $\sum_k \beta_{g,indk} = 0$, $g = A, B$, via restricted least squares.[29] Yun (2005) imposes the constraint that the coefficient on the first category equals the unweighted average of the coefficients on the other categories, $\beta_{g,ind1} = -\sum_{k\neq 1} \beta_{g,indk}/K$ along with $\sum_{k=1}^{K} \beta_{g,indk} = 0$. While these restrictions may appear to solve the problem of the omitted group, as pointed out by Yun (2008) "some degree of arbitrariness in deriving a normalized equation is unavoidable" (p.31). For example, an alternative restriction on the coefficients, that goes back to Kennedy (1986), could be a weighted sum, $\sum_k w_k \beta_{gk} = 0$, where the weights w_k reflect the relative frequencies of the categories in the pooled sample. The coefficients would then reflect deviations from the overall sample mean.

The pitfall here is that the normalizations proposed by Gardeazabal and Ugidos (2004) and Yun (2005) may actually leave the estimation and decomposition without a simple meaningful interpretation. Moreover, these normalizations will likely be sample

[28] This problem is different from a "true" identification problem which arises when multiple values of a parameter of interest are consistent with a given model and population.

[29] As pointed by Gardeazabal and Ugidos (2004), such restrictions can have some disturbing implications. In the case of educational categories, it rules out an outcome where group B members would earn higher returns than group A members for all levels of education.

specific and preclude comparisons across studies. By contrast, in the case of educational categories, the common practice of using high school graduates as the omitted category allows the comparison of detailed decomposition results when this omitted category is comparable across studies.

Invariance of the detailed decomposition with respect to the choice of omitted category may appear to be a desirable property, but it is actually elusive and should not come at the expense of interpretability. There is no quick fix to the difficult choice of the appropriate omitted category or base group which is actually exacerbated in procedures that go beyond the mean. To mimic the case of continuous variables, one may argue that an education category such as less than high school that yields the smallest wage effect should be the omitted one, but this category may vary more across studies than the high school category. Issues of internal logic have to be balanced with comparability across studies.

Another way of reporting the results of counterfactual experiments, proposed in the context of the gender wage gap by industry, is to report the wage structure effects for each k category by setting $\overline{X}_{g,indk} = 1$ and $\overline{X}_{g,indl} = 0$ for $l \neq k$ in the expression (20) for the total wage structure effect

$$\widehat{\delta}_g^{\mu}(indk) = (\widehat{\beta}_{B0} - \widehat{\beta}_{A0}) + \left(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}\right) + \sum_{j=1}^{J} \overline{X}_{gj} \left(\widehat{\beta}_{Bj} - \widehat{\beta}_{Aj}\right) \quad k = 1, \ldots, 4, \quad (22)$$

in a case where there are other explanatory variables, X_j, $j = 1, J$.[30] Initially, such expressions included only the first two terms, the intercept and the effect of the category k (Fields and Wolf, 1995). Later, Horrace and Oaxaca (2001) added the wage structure effect associated with the other variables. This allows one to compare the effect of wage structure on gender wage differentials by category while controlling for other explanatory variables X_j, $j = 1, \ldots, J$ in a way that is invariant to the choice of omitted category.[31] In columns (1) and (2) of Table 3, the wage structure effect associated with variables other than industrial sectors is essentially zero, and the $\widehat{\delta}_A^{\mu}(indk)$ can be computed as the difference between the male and female coefficients in columns (2) and (3) of Table 2 plus the 0.128 difference in the constant, yielding values of 0.128, 0.022, 0.004, and 0.048

[30] In the gender wage gap literature, when the reference wage structure is the male wage structure (group B) the means among women $\overline{X}_{Aj}$ will be used in equation (22).

[31] It is indeed easy to see that $\widehat{\delta_g^{\mu}}'(indk) = [(\widehat{\beta}_{B0} + \widehat{\beta}_{B,ind1}) - (\widehat{\beta}_{A0} + \widehat{\beta}_{A,ind1})] + [(\widehat{\beta}_{B,indk} - \widehat{\beta}_{B,ind1}) - (\widehat{\beta}_{A,indk} - \widehat{\beta}_{A,ind1})] + \sum_{j=1}^{J} \overline{X}_{gj} \left(\widehat{\beta}_{Bj} - \widehat{\beta}_{Aj}\right) = \widehat{\delta}^{\mu}(indk)$.

for industries 1 through 4, respectively. Horrace and Oaxaca (2001) also proposed to ex-post normalize the effects of each category with respect to the maximum categorical effect.

One disadvantage of decomposition terms like $\widehat{\delta}_g^{\mu}(indk)$ relative to the usual components of the detailed decomposition is that they do not sum up to the overall wage structure effect. As a result, just looking at the magnitude of the $\widehat{\delta}_g^{\mu}(indk)$ terms gives little indication of their quantitative importance in the decomposition. We propose a normalization to help assess the proportion of the total wage structure effect which can be attributed to a category k given that a proportion $\overline{X}_{g,indk}$ of group g workers belongs to that category, and that is also invariant to the choice of omitted category. The normalization uses the fact that the weighted sum of the $\widehat{\delta}_g^{\mu}(indk)$, $k = 1, \ldots, 4$ (that is, including the omitted category), is equal to the total wage structure effect, so that the proportional effect $\%\widehat{\delta}_{S,g}^{\mu}(indk)$ of category k in the total wage structure can be computed as[32]

$$\%\widehat{\delta}_S^{\mu}(indk) = \frac{\widehat{\delta}_g^{\mu}(indk)\overline{X}_{g,indk}}{\widehat{\Delta}_S^{\mu}} \quad \text{because} \quad \widehat{\Delta}_S^{\mu} = \sum_{k=1}^{4} \widehat{\delta}_g^{\mu}(indk)\overline{X}_{g,indk}. \tag{23}$$

In our empirical example, with group B as the reference group, this expression is computed using female averages, thus $\%\widehat{\delta}_S^{\mu}(indk)$ will tell us the proportion of the total wage structure effect that can be attributed to industrial category k given the proportion of women in each category. The numbers are 0.308 for primary, 0.074 for manufacturing, 0.040 for education and health, and 0.578 for other services. Despite being underrepresented in the manufacturing sector, because women's returns to manufacturing jobs are relatively high, the share of the unexplained gap attributable to that factor turns out not to be that large.

3.3 Alternative choices of counterfactual

On the one hand, the choice of a simple counterfactual treatment is attractive because it allows us to use the identification results from the treatment effects literature. On the other hand, these simple counterfactuals may not always be appropriate for answering the economic question of interest. For instance, the male wage structure may not represent the appropriate counterfactual for the way women would be paid in absence of labor market discrimination. If the simple counterfactual does not represent the appropriate

[32] The $\widehat{\delta_g^{\mu}}(indk)$ for the omitted category is simply the first and last components of equation (22) since $(\widehat{\beta}_{B,indk} - \widehat{\beta}_{A,indk}) = 0$ for that category.

treatment, it may be more appropriate to posit a new wage structure. For example, in the case of the gender pay gap, typically propositions (Reimers, 1983; Cotton, 1988; Neumark, 1988; Oaxaca and Ransom, 1994) have use a weighted average expression $\beta^* = \Omega\beta_A + (I - \Omega)\beta_B$, where $\Omega = I$ corresponds to $\beta^* = \beta_A$, $\Omega = 0$ corresponds to $\beta^* = \beta_B$, and where $\Omega = \omega \cdot I$ could reflect a weighting corresponding to the share of the two groups in the population. Another popular choice is the matrix $\Omega^* = (\mathbf{X}_B^\intercal \mathbf{X}_B + \mathbf{X}_A^\intercal \mathbf{X}_A)^{-1} \mathbf{X}_B^\intercal \mathbf{X}_B$ which captures the sample variation in the characteristics of group A and B workers.[33] The decomposition is then based on the triple differences:

$$\begin{aligned} \widehat{\Delta}_O^\mu &= \left(\overline{X}_B - \overline{X}_A\right)\widehat{\beta}^* + \left[\overline{X}_B\left(\widehat{\beta}_B - \widehat{\beta}^*\right) + \overline{X}_A\left(\widehat{\beta}^* - \widehat{\beta}_A\right)\right] \\ &= \qquad \widehat{\Delta}_X^\mu \qquad + \qquad\qquad \widehat{\Delta}_S^\mu. \end{aligned}$$

Table 3 shows that in the NLSY example, the gender gap decomposition is substantially different when either the female wage structure (column 3) or the weighted sum of the male and female wage structure (column 4) is used as the reference wage structure. Typically (as in Bertrand and Hallock (2001) for example), with the female wage structure as reference, the explained part of the decomposition (composition effect) is smaller than with the male wage structure as reference. Indeed, evaluated at either female 'prices' or average of male and female 'prices', the total unexplained (wage structure) effect becomes statistically significant.

An alternative measure of "unexplained" differences (see Cain, 1986) in mean wages between group A and group B workers is given by the coefficient δ of the group membership indicator variable D_B in the wage regression on the pooled sample, where the coefficients of the observed wage determination characteristics are constrained to be the same for both groups:

$$\mathbb{E}[Y_i|X, D_B] = \alpha_0 + X_i\beta^{**} + \delta D_{Bi}, \tag{24}$$

where the vector of observed characteristics X_i excludes the constant. It follows that,

$$\begin{aligned} \Delta_O^\mu &= \mathbb{E}[Y_i|X, D_B = 1] - \mathbb{E}[Y_i|X, D_B = 0] \\ &= (\alpha_0 + \mathbb{E}[X_i|D_B = 1]\beta^{**} + \delta) - (\alpha_0 + \mathbb{E}[X_i|D_B = 0]\beta^{**}) \\ &= \left(\mathbb{E}[X_i|D_B = 1] - \mathbb{E}[X_i|D_B = 0)\right)\beta^{**} + \delta = \Delta_X^\mu + \Delta_S^\mu, \end{aligned}$$

where $\delta = \Delta_S^\mu$. As noted by Fortin (2008), this "regression-compatible" approach is

[33] $\mathbf{X}_A$ and $\mathbf{X}_B$ are the matrices of covariates (of dimension $N_A \times k$ and $N_B \times k$) for groups A and B, respectively.

preferable to the one based on a pooled regression that omits the group membership variable (as in Neumark (1988) and Oaxaca and Ransom (1994)), because in the latter case the estimated coefficients are biased (omitted variable bias). Note, however, that this counterfactual corresponds to the case where the group membership dummy is thought to be sufficient to purge the reference wage structure from any group membership effect, an assumption that is maintained in the common practice of using the group membership dummy in a simple regression to assess its effect. The detailed decomposition is obtained using the above triple differences decomposition.[34]

The results of this decomposition, reported in Column (5) of Table 3, are found to be closest to the one using the female coefficients in column (3), but this is not necessarily always the case. Notice that the magnitude of the total unexplained wage log wage gap 0.092 (0.014) log points corresponds to the coefficient of the female dummy in column (5) of Table 2.

3.4 Reweighted-regression decompositions

A limitation of OB decompositions, discussed by Barsky et al. (2002), is that they may not provide consistent estimates of the wage structure and composition effect when the conditional mean is a non-linear function. Barsky et al. (2002) look at the role of earnings and other factors in the racial wealth gap. They argue that a standard OB decomposition is inadequate because the wealth-earnings relationship is non linear, and propose a more flexible approach instead.

Under the linearity assumption, the average counterfactual wage that group B workers would have earned under the wage structure of group A is equal to $\mathbb{E}\left[X_B|D_B=1\right]\cdot\beta_A$, and is estimated as the product $\overline{X}_B\widehat{\beta}_A$, a term that appears in both the wage structure and composition effect in equation (15). However, when linearity does not hold, the counterfactual mean wage will not be equal to this term.

One possible solution to the problem is to estimate the conditional expectation using non-parametric methods. Another solution proposed by Barsky et al. (2002) is to use a (non-parametric) reweighting approach as in DiNardo, Fortin and Lemieux (1996) to perform the decomposition. One drawback of this decomposition method discussed later in the chapter is that it does not provide, in general, a simple way of performing a detailed decomposition. In the case of the mean, however, this drawback can be readily addressed by estimating a regression in the reweighted sample.

[34] This "pooled" decomposition is easily implemented using the option "pooled" in Jann (2008) "oaxaca" procedure in Stata 9.2.

To see this, let $\Psi(X_i)$ be the reweighting function, discussed in section 4.5, that makes the characteristics of group A workers similar to those of group B workers. The counterfactual coefficients β_A^C and the counterfactual mean $\overline{X}_A^C$, are then estimated as:[35]

$$\begin{aligned}\widehat{\beta}_A^C &= \left(\sum_{i\in A}\widehat{\Psi}(X_i)\cdot X_i\cdot X_i^{\intercal}\right)^{-1}\cdot\sum_{i\in A}\widehat{\Psi}(X_i)\cdot Y_{Ai}\cdot X_i \\ \overline{X}_A^C &= \sum_{i\in A}\widehat{\Psi}(X_i)\cdot X_i,\end{aligned}$$

where $plim(\overline{X}_A^C) = plim(\overline{X}_B) = \mathbb{E}(X|D_B = 1)$.[36] If the conditional expectation of Y given X was linear, both the weighted and unweighted regressions would yield the same consistent estimate of β_A., i.e. we would have $plim(\widehat{\beta}_A^C) = plim(\widehat{\beta}_A) = \beta_A$. When the conditional expectation is not linear, however, the weighted and unweighted estimates of β_A generally differ since OLS minimizes specification errors over different samples.[37]

Consider the "reweighted-regression" decomposition of the overall wage gap $\widehat{\Delta}_{O,R}^{\mu}$, where

$$\begin{aligned}\widehat{\Delta}_{O,R}^{\mu} &= \left(\overline{X}_B\widehat{\beta}_B - \overline{X}_A^C\widehat{\beta}_A^C\right) + \left(\overline{X}_A^C\widehat{\beta}_A^C - \overline{X}_A\widehat{\beta}_A\right) \\ &= \qquad \widehat{\Delta}_{S,R}^{\mu} \qquad + \qquad \widehat{\Delta}_{X,R}^{\mu}.\end{aligned}$$

The composition effect $\widehat{\Delta}_{X,R}^{\mu}$ can be divided into a pure composition effect $\widehat{\Delta}_{X,p}^{\mu}$ using the wage structure of group A, and a component linking to the specification error in the

[35] When considering covariates X, we use the subscript g to denote the group whose characteristics are "adjusted" with reweighting.

[36] We show in Section 4 that the reweighting factor $\Psi(X)$ is defined as the ratio of the marginal distributions of X for groups B and A, $\Psi(X) = dF_{X_B}(X)/dF_{X_A}(X)$. As a result, the reweighted distribution of X for group A should be the same as the original distribution of X in group B. This implies that the mean value of X in the reweighted sample, $\overline{X}_A^C$, should be the same as the mean value of X for group B, $\overline{X}_B$.

[37] When the conditional expectation is non-linear, the OLS estimate of β can be interpreted as the one which minimizes the square of the specification error $\mathbb{E}(Y|X) - X\beta$ over the distribution of X. Since the expected value of the OLS estimate of β depends on the distribution of X, differences in β over two samples may either reflect true underlying differences in the conditional expectation (i.e. in the wage structure), or "spurious" differences linked to the fact that the distribution of X is different in the two samples. For example, if $\mathbb{E}(Y|X)$ is convex in X, the expected value of β will tend to grow as the distribution of X shifts up, since the relationship between Y and X gets steeper as X becomes larger.

linear model, $\widehat{\Delta}^{\mu}_{X,e}$:

$$\begin{aligned}\widehat{\Delta}^{\mu}_{X,R} &= \left(\overline{X}^{C}_{A} - \overline{X}_{A}\right)\widehat{\beta}_{A} + \overline{X}^{C}_{A}\left[\widehat{\beta}^{C}_{A} - \widehat{\beta}_{A}\right] \\ &= \widehat{\Delta}^{\mu}_{X,p} + \widehat{\Delta}^{\mu}_{X,e}.\end{aligned}$$

The wage structure effect can be written as

$$\begin{aligned}\widehat{\Delta}^{\mu}_{S,R} &= \overline{X}_{B}\left(\widehat{\beta}_{B} - \widehat{\beta}^{C}_{A}\right) + \left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\beta^{C}_{A}} \\ &= \widehat{\Delta}^{\mu}_{S,p} + \widehat{\Delta}^{\mu}_{S,e}\end{aligned}$$

and reduces to the first term $\widehat{\Delta}^{\mu}_{S,p}$ as the reweighting error $\widehat{\Delta}^{\mu}_{S,e}$ goes to zero in large samples ($plim\left(\overline{X}_{B} - \overline{X}^{C}_{A}\right) = 0 \Rightarrow plim(\widehat{\Delta}^{\mu}_{S,e}) = 0$).

The reweighted-regression decomposition is similar to the usual OB decomposition except for two small differences. The first difference is that the wage structure effect is based on a comparison between $\widehat{\beta}_{B}$ and the weighted estimate $\widehat{\beta}^{C}_{A}$ instead of the usual unweighted estimate $\widehat{\beta}_{A}$. As discussed in Firpo, Fortin, and Lemieux (2007), this ensures that the difference $\widehat{\beta}_{B} - \widehat{\beta}^{C}_{A}$ reflects true underlying differences in the wage structure for group A and B, as opposed to a misspecification error linked to the fact that the underlying conditional expectation is non-linear. Note that is also useful to check whether the reweighting error $\left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\beta}^{C}_{A}$ is equal to zero (or close to zero), as it should be when the reweighting factor $\widehat{\Psi}(X)$ is consistently estimated.

The other difference relative to the OB decomposition is that the composition effects consists of a standard term $\left(\overline{X}^{C}_{A} - \overline{X}_{A}\right)\widehat{\beta}_{A}$ plus the specification error $\overline{X}^{C}_{A}\left[\widehat{\beta}^{C}_{A} - \widehat{\beta}_{A}\right]$. If the model was truly linear, the specification error term would be equal to zero. Computing the specification error is important, therefore, for checking whether the linear model is well specified, and adjusting the composition effect in the case where the linear specification is found to be inaccurate.

In the case where the conditional expectation $\mathbb{E}(Y_i|X_i, D = d)$ is estimated non-parametrically, a whole different procedure would have to be used to separate the wage structure into the contribution of each covariate. For instance, average derivative methods could be used to estimate an effect akin to the β coefficients used in standard decompositions. Unfortunately, these methods are difficult to use in practice, and would not be helpful in dividing up the composition effect into the contribution of each individual covariate.

On a related note, Kline (2009) points out that the standard OB decomposition can

be interpreted as a reweighting estimator where the weights have been linearized as a function of the covariates. This suggests that the procedure may actually be more robust to departures from linearity than what has been suggested in the existing literature. Since the procedure is robust to these departures and remains the method of choice when linearity holds, Kline (2009) points out that it is "doubly robust" in the sense of Robins, Rotnizky, and Zhao (1994) and Egel, Graham, and Pinto (2009).

3.5 Extensions to limited dependent variable models

OB decompositions have been extended to cases where the outcome variable is not a continuous variable. To mention a few examples, Gomulka and Stern (1990) study the changes over time in labor force participation of women in the United Kingdom using a probit model. Even and Macpherson (1990) decomposes the male-female difference in the average probability of unionization, while Doiron and Riddell (1994) propose a decomposition of the gender gap in unionization rate based on a first order Taylor series approximation of the probability of unionization. Fitzenberger et al. (2006) used a probit model to decompose changes over time in the rate of unionization in West and East Germany. Fairlie (1999; 2005) discuss the cases of the racial gaps in self-employment and computer ownership. Bauer and Sinning (2008) discuss the more complicated cases of a count data model, for example where the dependent variable is the number of cigarettes smoked by men and women (Bauer, Göhlmann, and Sinning, 2007), and of the truncated dependent variable, where for example the outcome of interest is hours of work.

In the case of a limited dependent variable Y, the conditional expectation of Y is typically modelled as a non-linear function in X, $\mathbb{E}(Y_g|X;\beta_g) = G(X;\beta_g)$. For example, if Y is a dichotomous outcome variable $(Y = 0, 1)$ and $Y_g^* = X\beta_g + \upsilon_g$ is a latent variable which is linear in X, it follows that $\mathbb{E}(Y_g|X;\beta_g) = G(X\beta_g)$ where $G(\cdot)$ is the PDF of υ_g. When υ_g follows a standard normal distribution, we have a standard probit model and $G(\cdot) = \Phi(\cdot)$. More generally, under various assumptions regarding the functional form G and/or the distribution of the error terms υ_g, the models are estimated by maximum likelihood.

Because $\mathbb{E}(Y_g|D_g = 1) = \mathbb{E}\left[\mathbb{E}(Y_g|X;\beta_g)|D_g = 1\right] = \mathbb{E}\left[G(X;\beta_g)|D_g = 1\right] \neq G\big(\mathbb{E}[X|D_g = 1];\beta_g\big)$, the decomposition cannot simply be computed by plugging in the estimated β's and the mean values of X's, as in the standard OB decomposition. Counterfactual conditional expectations have to be computed instead, and averaged across observations. For example, if group A is thought to be the reference group, $\mathbb{E}(Y_B|D_A = 1) =$

$\mathbb{E}\left[G(X;\beta_B)|D_A=1\right]$ will be the counterfactual conditional expectation of Y_B that would prevail if the coefficients of the determinants of self-employment (for example) for group B were the same as for group A. This involves computing predicted (i.e. expected) values based on the estimated model for group B, $G(X;\beta_B)$, over all observations in group A, and averaging over these predicted values.

The mean gap between group B and group A is then decomposed as follows

$$\begin{aligned}
\Delta_O^\mu &= \mathbb{E}(Y_B|D_B=1)] - \mathbb{E}(Y_A|D_A=1) \\
&= \mathbb{E}\left[G(X;\beta_B)|D_B=1\right] - \mathbb{E}\left[G(X;\beta_A)|D_A=1\right] \\
&= \left(\mathbb{E}\left[G(X;\beta_B)|D_B=1\right] - \mathbb{E}\left[G(X;\beta_A)|D_B=1\right]\right) \\
&\qquad + \left(\left[\mathbb{E}\left[G(X;\beta_A)|D_B=1\right] - \mathbb{E}\left[G(X;\beta_A)|D_A=1\right]\right) \\
&= \Delta_S^\mu + \Delta_X^\mu,
\end{aligned}$$

into a component that attributes differences in the mean outcome variable to differences in the characteristics of the individuals, and a component that attributes these differences to differences in the coefficients.

The same difficult issues in the appropriate choice of counterfactuals persist for more general non-linear models. In addition, extra care has to be taken to verify that the sample counterfactual conditional expectation lies within the bounds of the limited dependent variable. For example, Fairlie (1999) checks that average self-employment for Blacks predicted from the White coefficients is not negative.

The non-linear decomposition may perform better than the linear alternative (linear probability model, LPM) when the gap is located in the tails of the distribution or when there are very large differences in the explanatory variables, which effects would remain unbounded in a LPM. On the other hand, there are many challenges in the computation of detailed decompositions for non-linear models. Because of non-linearity, the detailed decomposition of the two components into the contribution of each variable, even if the decomposition was linearized using marginal effects, would not add up to the total. Gomulka and Stern (1990) and Fairlie (2005) have proposed alternative methodologies based on a series of counterfactuals, where the coefficient of each variable is switched to reference group values in sequence. In the latter cases, the decomposition will be sensitive to the order of the decomposition, that is will be path dependent. We discuss these issues further in the context of the decompositions of entire distributions in Section 5.

3.6 Statistical inference

OB decompositions have long been presented without standard errors. More recently, Oaxaca and Ransom (1998), followed by Greene (2003, p. 53-54), have proposed approximate standard errors based the delta method under the assumption that the explanatory variables were fixed.[38] A more modern approach where, as above, (Y, X) are stochastic was suggested and implemented by Jann (2005). In cases where the counterfactuals are not a simple treatment, or where a non-linear estimator is used, boostrapping the entire procedure may prove to be the practical alternative.

4 Going beyond the Mean - Distributional Methods

Developing new decomposition methods for distributional statistics other than the mean has been an active research area over the last 15 years. In this section, we discuss a number of procedures that have been suggested for decomposing general distributional statistics. We focus on the case of the aggregate decomposition, though some of the suggested methods can be extended to the case of the detailed decomposition, which we discuss in section 5. We begin by looking at the simpler case of a variance decomposition. The decomposition is obtained by extending the classic analysis of variance approach (based on a between/within group approach) to a general case with covariates X. We then turn to new approaches based on various "plugging in" methods such as JMP's residual imputation method and Machado and Mata (2005)'s conditional quantile regression method. Finally, we discuss methods that focus on the estimation of counterfactuals for the entire distribution. These methods are either based on reweighting or on the estimation of the conditional distribution.

Most of this recent research was initially motivated by the dramatic growth in earnings inequality in the United States. Prior to that episode, the literature was considering particular summary measures of inequality such as the variance of logs and the Gini coefficient. For instance, Freeman (1980, 1984) looks at the variance of log wages in his influential work on the effect of unions on wage dispersion. This research establishes that unions tend to reduce wage dispersion as measured by the variance of log wages. Freeman shows that despite the inequality-enhancing effect of unions on the between-group component of inequality, the overall effect of unions is to reduce inequality because of the even larger effect of unions on within-group inequality.

[38] This corresponds to an experimental setting where, for example, regression analysis was used to assess the impact of various soils and fertilizers (X) on agricultural yields Y.

One convenient feature of the variance is that it can be readily decomposed into a within- and between-group component. Interestingly, related work in the inequality literature shows that other measures such as the Gini or Theil coefficient are also decomposable into a within and between group component.[39]

Note that the between vs. within decomposition is quite different in spirit from the aggregate or detailed OB decomposition discussed in the previous section. There are advantages and disadvantages to this alternative approach. On the positive side, looking at between- and within-group effects can help understand economic mechanisms, as in the case of unions, or the sources of inequality growth (Juhn, Murphy, and Pierce, 1993).

On the negative side, the most important drawback of the between vs. within decomposition is that it does not hold in the case of many other interesting inequality measures such as the interquartile ranges, the probability density function, etc. This is a major shortcoming since looking at what happens where in the distribution is important for identifying the factors behind changes or differences in distributions. Another drawback of the between vs. within approach is that it does not provide a straightforward way of looking at the specific contribution of each covariate, i.e. to perform a detailed decomposition. One final drawback is that with a rich enough set of covariates the number of possible groups becomes very large, and some parametric restrictions have to be introduced to keep the estimation problem manageable.

In response to these drawbacks, a new set of approaches have been proposed for performing aggregate decompositions on any distributional statistic. Some approaches such as Juhn, Murphy, and Pierce (1993), Donald, Green, and Paarsch (2000), and Machado and Mata (2005) can be viewed as extensions of the variance decomposition approach where the whole conditional distribution (instead of just the conditional variance) are estimated using parametric approaches. Others such as DiNardo, Fortin, and Lemieux (1996) completely bypass the problem of estimating conditional distributions and are, as such, closer cousins to estimators proposed in the program evaluation literature.

4.1 Variance decompositions

Before considering more general distributional statistics, it is useful the recall the steps used to obtain the standard OB decomposition. The first step is to assume that the conditional expectation of Y given X is linear, i.e. $\mathbb{E}(Y|X) = X\beta$. This follows directly from the linearity and zero conditional mean assumptions (Assumptions 10 and 11) in-

[39] See, for instance, Bourguignon (1979), Cowell (1980), and Shorrocks (1980, 1984).

troduced in Section 2. Using the law of conditional expectations, it then follows that the unconditional mean is $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X)\beta$. This particular property of the mean is then used to compute the OB decomposition.

In light of this, it is natural to think of extending this type of procedure to the case of the variance. Using the analysis of variance formula, the unconditional variance of Y can be written as:[40]

$$\begin{aligned} Var(Y) &= \mathbb{E}[Var(Y|X)] + \mathbb{E}\left\{[\mathbb{E}(Y|X) - \mathbb{E}(Y)]^2\right\} \\ &= \mathbb{E}[Var(Y|X)] + \mathbb{E}\left\{[X\beta - \mathbb{E}(X)\beta]^2\right\} \\ &= \mathbb{E}[Var(Y|X)] + \beta' Var(X)\beta, \end{aligned}$$

where the expectations are taken over the distribution of X. The first component of the equation is the within-group component (also called residual variance), while the second component is the between-group component (also called regression variance). Writing $Var(Y|X, D_g = 1) \equiv \mathrm{v}_g(X)$, $g = A, B$, we can write the difference in variances across groups B and A as

$$\Delta_O^V = \mathbb{E}\left[\mathrm{v}_B(X)|D_B = 1\right] - \mathbb{E}\left[\mathrm{v}_A(X)|D_B = 0\right] + \beta_B' Var\left[X|D_B = 1\right]\beta_B - \beta_A' Var\left[X|D_B = 0\right]\beta_A.$$

A few manipulations yield $\Delta_O^V = \Delta_X^V + \Delta_S^V$, where

$$\Delta_X^V = \left\{\mathbb{E}\left[\mathrm{v}_A(X)|D_B = 1\right] - \mathbb{E}\left[\mathrm{v}_A(X)|D_B = 0\right]\right\} + \beta_A' \left\{Var\left[X|D_B = 1\right] - Var\left[X|D_B = 0\right]\right\}\beta_A$$

and

$$\Delta_S^V = \left\{\mathbb{E}\left[\mathrm{v}_B(X)|D_B = 1\right] - \mathbb{E}\left[\mathrm{v}_A(X)|D_B = 1\right]\right\} + (\beta_B - \beta_A)' Var\left[X|D_B = 1\right](\beta_B - \beta_A).$$

While it is straightforward to estimate the regression coefficients (β_A and β_B) and the covariance matrices of the covariates ($Var\left[X|D_B = 0\right]$ and $Var\left[X|D_B = 1\right]$), the within-group (or residual) variance terms $\mathrm{v}_A(X)$ and $\mathrm{v}_B(X)$ also have to be estimated to compute the decomposition.

Several approaches have been used in the literature to estimate $\mathrm{v}_A(X)$ and $\mathrm{v}_B(X)$. The simplest possible approach is to assume that the error term is homoscedastic, in which case $\mathrm{v}_A(X) = \sigma_A^2$ and $\mathrm{v}_B(X) = \sigma_B^2$, and the two relevant variance parameters can be estimated from the sampling variance of the error terms in the regressions. The

[40] See for example, Theorem B.4 in Greene (2003).

homoscedasticity assumption is very strong, however. When errors are heteroscedastic, differences between σ_A^2 and σ_B^2 can reflect spurious composition effects, in which case the decomposition will attribute to the wage structure effect (Δ_S^V) what should really be a composition effect (Δ_X^V). Lemieux (2006b) has shown this was a major problem when looking at changes in residual wage inequality in the United States since the late 1980s.

A simple way of capturing at least some of the relationship between the covariates and the conditional variance is to compute the variance of residuals for a limited number of subgroups of "cells". For instance, Lemieux (2006b) shows estimates for 20 different subgroups of workers (based on education and experience), while Card (1996) divides the sample in five quintiles based on predicted wages $X\widehat{\beta}$.

Finally, one could attempt to estimate a more general specification for the conditional variance by running a "second step" model for squared regression residual $\widehat{v}^2 = \left(Y - X\widehat{\beta}\right)^2$ on some specification of the covariates. For example, assuming that $\mathrm{v}_A(X) = X\delta$, we can estimate estimate $\widehat{\delta}$ by running a regression of $\widehat{v}^2$ on X.[41] We can then write the two aggregate components of the variance decomposition as:

$$\Delta_X^V = \left\{\left(\mathbb{E}\left[X|D_B=1\right] - \mathbb{E}\left[X|D_B=0\right]\right)\delta_A\right\} + \beta_A' \left\{Var\left[X|D_B=1\right] - Var\left[X|D_B=0\right]\right\}\beta_A \tag{25}$$

and

$$\Delta_S^V = \left\{\mathbb{E}\left[X|D_B=1\right](\delta_B - \delta_A)\right\} + (\beta_B - \beta_A)' Var\left[X|D_B=1\right](\beta_B - \beta_A). \tag{26}$$

Compared to the standard OB decomposition for the mean, which only requires estimating a (regression) model for the conditional mean, in the case of the variance, we also need to estimate a model for the conditional variance. While this is quite feasible in practice, we can already see a number of challenges involved when decomposing distributional parameters beyond the mean:

- The estimation is more involved since we need to estimate models for two, instead of just one, conditional moment. Furthermore, little guidance is typically available on "reasonable" specifications for the conditional variance. For instance, in the case of wages, the Mincer equation provides a reasonably accurate and widely accepted specification for the conditional mean, while no such standard model is available for the conditional variance.

[41] Estimating these simple models of the conditional cross-sectional variance is a special case of the large time-series literature on the estimation of auto-regressive conditional heteroskedasticity models (ARCH, GARCH, etc.).

- Computing the detailed decomposition is more complicated since the between-group component is a quadratic form in the β's. This yields a number of interaction terms that are difficult to interpret.

Since the complexity of decomposition methods already increases for a distributional measure as simple and convenient as the variance, this suggests these problems will be compounded in the case of other distributional measures such as quantiles. Indeed, we show in the next subsection that for quantiles, attempts at generalizing the approach suggested here require estimating the entire conditional distribution of Y given X. This is a more daunting estimation challenge, and we now discuss solutions that have been suggested in the literature.

4.2 Going beyond the variance: general framework

An important limitation of summary measures of dispersion such as the variance, the Gini coefficient or the Theil coefficient is that they provide little information regarding what happens where in the distribution. This is an important shortcoming in the literature on changes in wage inequality where many important explanations of the observed changes have specific implications for specific points of the distribution. For instance, the minimum wage explanation suggested by DiNardo, Fortin, and Lemieux (1996) should only affects the bottom end of the distribution. At the other extreme, explanations based on how top executives are compensated should only affect the top of the distribution. Other explanations based on de-unionization (Freeman, 1993, Card, 1992, and DiNardo, Fortin, and Lemieux, 1996) and the computerization of "routine" jobs (Autor, Levy and Murnane, 2003) tend to affect the middle (or "lower middle") of the distribution. As a result, it is imperative to go beyond summary measures such as the variance to better understand the sources of growing wage inequality.

Going beyond summary measures is also important in many other interesting economic problems such the sources of the gender wage gap and the impact of social programs on labor supply.[42] The most common approach for achieving this goal is to perform a decomposition for various quantiles (or differences between quantiles like the 90-10 gap) of the distribution. Unfortunately, as we point out in the introduction, it is much more difficult to decompose quantiles than the mean or even the variance. The basic problem is that the law of iterated expectations does not hold in the case of quantiles, i.e.

[42] See Albrecht, Björklund, and Vroman (2003) who look at whether there is a glass ceiling in female earnings, and Bitler, Gelbach and Hoynes (2006) who study the distributional effects of work incentive programs on labor supply.

$Q_{g,\tau} \neq \mathbb{E}_X[Q_{g,\tau}(X)]$, where $Q_{g,\tau}$, is the τ^{th} quantile of the (unconditional) distribution of Y_g, and $Q_{g,\tau}(X)$ is the corresponding conditional quantile.

As it turns out, one (implicitly) needs to know the entire conditional distribution of Y_g given X given to compute $Q_{g,\tau}$. To see this, note that

$$\tau = F_{Y_g}(Q_{g,\tau}) = \mathbb{E}[F_{Y_g|X_g}(Q_{g,\tau}|X)] = \int F_{Y_g|X_g}(Q_{g,\tau}|X)dF_{X_g}(X), \quad g = A, B,$$

where $F_{Y_g|X_g}(\cdot)$ is the cumulative distribution of Y conditional on X in group g. Given τ, it is possible to implicitly use this equation to solve out for $Q_{g,\tau}$. It is also clear that in order to do so we need to know the conditional distribution function $F_{Y_g|X_g}(\cdot)$, as opposed to just the conditional mean and variance, as was the case for the variance. Estimating an entire conditional distribution function for each value of $(Y_g|X)$ is a difficult problem. Various decomposition methods that we discuss in detail below suggest different ways of handling this challenge.

But before covering them in detail, we recall the basic principles underlying these methods. As in Section 2, we focus on cumulative distributions since any standard distribution statistic, such as a quantile, can be directly computed from the cumulative distribution. For instance, quantiles of the counterfactual distribution can be obtained by inverting $F_{Y_A^C}$: $Q_{A,\tau}^C = F_{Y_A^C}^{-1}(\tau)$.

For the sake of presentational simplicity, we introduce a simplified notation relative to Section 2. We use F_{X_g} instead of $F_{X|D_g}$ to represent the marginal distribution of X, and $F_{Y_g|X_g}$ to represent $F_{Y_g|X,D_g}$ the conditional distributions, for $g = A, B$ introduced in equation (4). We use the shorthand $F_{Y_A^C}$ instead of $F_{Y_A^C:X=X|D_B}$ to represent the key counterfactual distribution of interest introduced in equation (5), which mixes the distribution of characteristics of group B with the wage structure from group A:

$$F_{Y_A^C}(y) = \int F_{Y_A|X_A}(y|X)dF_{X_B}(X). \tag{27}$$

Three general approaches have been suggested in the decomposition literature for estimating the counterfactual distribution $F_{Y_A^C}(y)$. A first general approach, initially suggested by Juhn, Murphy and Pierce (1993), replaces each value of Y_B for group B with a counterfactual value of $Y_A^C = g(Y_B, X)$, where $g(\cdot, \cdot)$ is an imputation function. The idea is to replace Y_B from group B with a counterfactual value of Y_A^C that holds the same rank in the conditional distribution $F_{Y_A|X_A}(\cdot|\cdot)$ as it did in the original distribution of Y_B. As we discussed in Section 2.2.3, this is done in practice using a residual imputation

procedure. Machado and Mata (2005) and Autor, Katz, and Kearney (2005) have later suggested other approaches, based on conditional quantile regressions, to transform a wage observation Y_B into a counterfactual observation Y_A^C.

A second approach proposed by DiNardo, Fortin, and Lemieux (1996) [DFL] is based on the following manipulation of equation (27):

$$F_{Y_A^C}(y) = \int F_{Y_A|X_A}(y|X)\Psi(X)dF_{X_A}(X), \tag{28}$$

where $\Psi(X) = dF_{X_B}(X)/dF_{X_A}(X)$ is a reweighting factor. This makes it clear that the counterfactual distribution $F_{Y_A^C}(\cdot)$ is simply a reweighted version of the distribution $F_{Y_A}(\cdot)$. The reweighting factor is a simple function of X that can be easily estimated using standard methods such as a logit or probit. The basic idea of the DFL approach is to start with group A, and then replace the distribution of X of group A ($F_{X_A}(\cdot)$) with the distribution of X of group B ($F_{X_B}(\cdot)$) using the reweighting factor $\Psi(\cdot)$.

The third set of approaches also works with equation (27) starting with group B, and then replacing the conditional distribution $F_{Y_B|X_B}(Y|X)$ with $F_{Y_A|X_A}(Y|X)$. Doing so is more involved, from an estimation point of view, than following the DFL approach. The problem is that the conditional distributions depend on both X and y, while the reweighting factor $\Psi(X)$ only depends on X.

Under this third set of approaches, one needs to directly estimate the conditional distribution $F_{Y|X}(y|X)$. Parametric approaches for doing so were suggested by Donald, Green, and Paarsch (2000) who used a hazard model approach, and Fortin and Lemieux (1998) who suggested estimating an ordered probit. More recently, Chernozhukov, Fernandez-Val, and Melly (2009) suggest estimating distributional regressions (e.g. a logit, for each value of y). In all cases, the idea is to replace the conditional distribution for group B, $F_{Y_B|X_B}(y|X)$, with an estimate of the conditional distribution $F_{Y_A|X_A}(y|X)$ obtained using one of these methods.

In the next subsections, we discuss how these various approaches can be implemented. We also present some results regarding their statistical properties, and address computational issues linked to their implementation.

4.3 Residual Imputation Approach: JMP

Procedure

As we explain above, Juhn, Murphy, and Pierce (1993) propose an imputation approach where the wage Y_B from group B is replaced by a counterfactual wage Y_A^C where

both the returns to observables and unobservables are set to be as in group A. The implementation of this procedure is divided in two steps. First, unobservables are replaced by counterfactual unobservables, as in equation (9). Second, counterfactual returns to observables are also imputed, as in equation (12).[43]

Under the assumption of additive linearity (Assumption 10), the original wage equation for individual i from group B,

$$Y_{Bi} = X_i\beta_B + \upsilon_{Bi} \quad \text{where} \quad \upsilon_{Bi} = h_B(\varepsilon_i)$$

allows the returns to unobservables to be group-specific. Under the assumption of rank preservation (14), the first counterfactual is computed as

$$Y_{Ai}^{C,2} = X_i\beta_B + \upsilon_{Ai}^{C,2}, \tag{29}$$

where

$$\upsilon_{Ai}^{C,2} = F_{\upsilon_A|X}^{-1}(\tau_{Bi}(x_i), x_i),$$

and $\tau_{Bi}(x_i)$ is the conditional rank of υ_{Bi} in the distribution of residuals for group B ($\tau_{Bi}(x_i) = F_{\upsilon_B|X}(\upsilon_{Bi}|X = x_i)$). A second counterfactual is then obtained by also replacing the returns to observable characteristics β_B with β_A

$$Y_{Ai}^{C,3} = X_i\beta_A + \upsilon_{Ai}^{C,2}.$$

Under the assumptions of linearity and rank preservation, this counterfactual wage should the same as Y_{Ai}^{C}, the counterfactual wage obtained by replacing the wage structure $m_B(\cdot)$ with $m_A(\cdot)$.

In practice, it is straightforward to estimate β_A and β_B using OLS under the assumptions of linearity and zero conditional mean. It is much less clear, however, how to perform the residual imputation procedure described above. Under the strong assumption that the regression residuals υ_g are independent of X, it follows that

$$\upsilon_{Ai}^{C,2} = F_{\upsilon_A}^{-1}(\tau_{Bi}).$$

Under this independence assumption, one simply needs to compute the rank of the resid-

[43] Juhn, Murphy, and Pierce (1993) actually consider multiple time periods and proposed an additional counterfactual where the returns to observables are set to their mean across time periods, a complex counterfactual treatment.

ual υ_{Bi} in the marginal distribution (distribution over the whole sample) of residuals for group B, and then pick the corresponding residuals in the marginal distribution of residuals for group A. If υ_{Bi} is at the 70^{th} percentile of the distribution of residuals of group B ($\tau_{Bi} = .7$), then $\upsilon_{Ai}^{C,2}$ will simply be the 70^{th} percentile of the distribution of residuals for group A. In practice, most applications of the JMP procedure use this strong assumption of independence because there is little guidance as to how a conditional imputation procedure could be used instead.

Limitations

Since independence of regression residuals is unrealistic, a more accurate implementation of JMP would require deciding how to condition on X when performing the imputation procedure. If X consists of a limited number of groups or "cells", then one could perform the imputation within each of these groups. In general, however, it is difficult to know how to implement this ranking/imputation procedure in more general cases. As a result, other procedures such as the quantile method of Machado and Mata (2005) are increasingly being used as an alternative to JMP.

Another limitation of the JMP procedure is that there is no natural way of extending it to the case of the detailed decomposition for the composition effect.

Advantages

One advantage of the two-step procedure is that it provides a way of separating the between- and within-group components, as in a variance decomposition. This plays an important role in the inequality literature, since JMP concluded that most of the inequality growth from the 1960s to the 1980s was linked to the residual inequality component.

It is not clear, however, what is meant by between- and within-group components in the case of distributional measures like the 90-10 gap that are not decomposable. A better way of justifying JMP is that $Y = X\beta + \upsilon$ represents a structural model where X are observed skills, while υ represents unobserved skills. One can then perform simulation exercises asking what happens to the distribution when one either replaces returns to observed or unobserved skills (see also Section 2.2.3).

This economic interpretation also requires, however, some fairly strong assumptions. The two most important assumptions are the linearity of the model (assumption 10, $m_g(X_i, \varepsilon_i) = X_i\beta_g + \upsilon_{gi}$) and rank preservation (assumption 14). While linearity can be viewed as a useful approximation, rank preservation is much stronger since it means that someone with the same unobserved skills would be in the exact same position, conditional on X, in either group A or B. Just adding measurement error to the model would result

in a violation of rank preservation.

Finally, if one is willing to interpret a simple regression as a decomposition between observed and unobserved skills, this can be combined with methods other than JMP. For instance, DFL perform regression adjustments to illustrate the effects of supply and demand factors on wages.[44]

4.4 Methods based on conditional quantiles

Procedure

Like JMP, Machado and Mata (2005, MM from hereinafter) propose a procedure based on transforming a wage observation Y_B into a counterfactual observation Y_A^C. The main advantage relative to JMP is that their estimation procedure based on quantile regressions (Koenker and Bassett, 1978) provides an explicit way of estimating the (inverse) conditional distribution function $F^{-1}_{Y_A|X_A}(\cdot,\cdot)$ in the transformation $g(Y,X) = F^{-1}_{Y_A|X_A}(F_{Y_B|X_B}(Y|X),X)$. One important difference, however, is that instead of transforming each actual observation of Y_{Bi} into a counterfactual Y^C_{Ai}, MM use a simulation approach where quantiles are drawn at random.

More specifically, since

$$Y_A^C = F^{-1}_{Y_A|X_A}(F_{Y_B|X_B}(Y|X),X),$$

and $\tau_B(Y|X) = F_{Y_B|X_B}(Y|X)$ follows a uniform distribution, one can think of doing the following:

1. Draw a simulated value τ_s from a uniform distribution $s = 1,\ldots,S$.

2. Estimate a linear quantile regression for the τ_s^{th} quantile, and use the estimated result to predict simulated values of both Y_{Bs} and Y^C_{As}.[45] The reason for using quantile regressions is that:

$$Y^C_{As} = F^{-1}_{Y_A|X_A}(\tau_s,X) \quad \text{and} \quad Y_{Bs} = F^{-1}_{Y_B|X_B}(\tau_s,X),$$

where $F^{-1}_{Y_A|X_A}(\cdot,\cdot)$ and $F^{-1}_{Y_B|X_B}(\cdot,\cdot)$ are the conditional quantile functions for the τ_s^{th} quantile in group A and B, respectively.

3. Compare the simulated distributions of Y_{Bs} and Y^C_{As} to obtain measures of the wage

[44] See also Lemieux (2002).

[45] For each random draw s, MM also draw a vector of covariates X_s from the observed data and perform the prediction for this value only. Melly (2005) discusses more efficient ways of computing dsitributions using this conditional quantile regression approach.

structure effect. The composition effect is computed as the complement to the overall difference.

A key implementation question is how to specify the functional forms for the conditional quantile functions. MM suggest a linear specification in the X that can be estimated using quantile regression methods. The conditional quantile regression models can be written as:

$$Q_{g,\tau}(Y|X)=F^{-1}_{Y_g|X_g}(\tau, X) = X\beta_{g,\tau}, \quad g = A, B$$

Table 4 reports in the top panel the results of the Machado-Mata procedure applied to our gender gap example using the male wage structure as reference.[46] It shows that the median gender log wage gap in the central column gives almost the same results for the aggregate decomposition as the OB mean gender log wage gap decomposition displayed in column (1) of Table 3. Going across the columns to compare quantile effects shows that gender differences in characteristics are much more important at the bottom (10th centile) than at the top (90th centile) of the wage distribution. Indeed, some significant wage structure effects emerge at the 90th percentile.

Limitations

This decomposition method is computationally demanding, and becomes quite cumbersome for data sets numbering more than a few thousand observations. Bootstrapping quantile regressions for sizeable number of quantiles τ (100 would be a minimum) is computationally tedious with large data sets. The implementation of the procedure can be simplified by estimating a large number of quantile regressions (say 99, one for each percentile from 1 to 99) instead of drawing values of τ_s at random.[47]

Another limitation is that the linear specification is restrictive and finding the correct functional for the conditional quantile regressions can be very tedious. For instance, if there is a spike at the minimum wage in the wage distribution, this will result in flat spots in quantile regressions that would have to be captured with spline functions with knots that depend on X. Accurately describing simple distribution with mass points (as is commonly observed in wage data) can, therefore, be quite difficult to do using quantile regressions.

As pointed out by Chernozhukov, Fernandez-Val, and Melly (2009), it is not very

[46] The estimates were computed with Melly's implementation "rqdeco" in Stata.

[47] See Melly (2005) for a detailed description of this alternative procedure. Gosling, Machin, and Meghir (2000) and Autor, Katz, and Kearney (2005) also use a similar idea in their empirical applications to changes in the distribution of wages over time.

natural to estimate *inverse* conditional distribution functions (quantile regressions) when the main goal of counterfactual exercises is to replace the conditional distribution function $F_{Y_B|X_B}$ with $F_{Y_A|X_A}$ to obtain equation (27). Chernozhukov, Fernandez-Val, and Melly (2009) suggest instead to estimate directly distributional regression models for $F_{Y|X}(\cdot,\cdot)$, which is a more direct way of approaching the problem.

Advantages

One advantage of the MM approach is that it provides a natural way of performing a detailed decomposition for the wage structure component. The idea is to successively replace the elements of $\beta_{B,\tau}$ by those of $\beta_{A,\tau}$ when performing the simulations, keeping in mind that this type of detailed decomposition is path dependent. Unfortunately, the MM does not provide a way of performing the detailed decomposition for the composition effect.[48] This is a major drawback since the detailed decomposition of the composition effects is always clearly interpretable, while the detailed decomposition of the wage structure effect arbitrarily depends on the choice of the omitted group.

4.5 Reweighting methods

Procedure

As we mention in Section 4.2, another way of estimating the counterfactual distribution $F_{Y_A^C}(\cdot)$ is to replace the marginal distribution of X for group A with the marginal distribution of X for group B using a reweighting factor $\Psi(X)$. This idea was first introduced in the decomposition literature by DiNardo, Fortin and Lemieux [DFL] (1996). While DFL focus on the estimation of counterfactual densities in their empirical application, the method is easily applicable to any distributional statistic.

In practice, the DFL reweighting method is similar to the propensity score reweighting method commonly used in the program evaluation literature (see Hirano, Imbens, and Ridder, 2003). For instance, in DFL's application to changes in wage inequality in the United States, time is viewed as a state variable, or in the context of the treatment effects literature as a treatment.[49] The impact of a particular factor or set of factors on changes in the wage distribution over time is constructed by considering the counterfactual state

[48] Machado and Mata (2005) suggest computing the detailed decomposition for the composition effect using an unconditional reweighting procedure. This is invalid as a way of performing the decomposition for the same reason that a OB decomposition would be invalid if the β coefficient used for one covariate was estimated without controlling for the other covariates. We propose a conditional reweighting procedure in the next section that deals adequately with this issue.

[49] This view of course makes more sense when some policy or other change has taken place over time (see Biewen (2001).

of the world where the distribution of this factor remained fixed in time, maintaining the assumption 6 of invariance of the conditional distribution. Note that by contrast with the notation of this chapter, in DFL, time period 1 is used as reference group.[50] The choice of period 0 or period 1 as the reference group is analogous to the choice of whether the female or the male wage structure should be the reference wage structure in the analysis of the gender wage gap and is expected to yield different results in most cases.

In DFL, manipulations of the wage distributions, computed through reweighting, are applied to non-parametric estimates of the wage density, which can be particularly useful when local distortions, from minimum wage effects for example, are at play. To be consistent with the rest of this section, however, we focus our discussion on the cumulative distribution instead of the density. The key counterfactual distribution of interest, shown in equation (27) (distribution of wages that would prevail for workers in group A if they had the distribution of characteristics of group B) is constructed, as shown in equation (28), using the reweighting factor

$$\Psi(X) = \frac{dF_{X_B}(X)}{dF_{X_A}(X)}.$$

Although the reweighting factor is the ratio of two multivariate marginal distribution functions (of the covariates X), this expression can be simplified using Bayes' rule. Remembering that Bayes' rule states that

$$P(B_i|A) = \ P(A|B_i) \cdot P(B_i) / \sum_j P(A|B_j) \cdot P(B_j).$$

We have

$$\Pr(X|D_B = 1) = \frac{\Pr(D_B = 1|X) \cdot dF(X)}{\int_x \Pr(D_B = 1|X) \cdot dF(X)} = \frac{\Pr(D_B = 1|X)}{\Pr(D_B = 1)}$$

and a similar expression for $D_B = 0$. Since $dF_{X_A}(X) = \Pr(X|D_B = 0)$ and $dF_{X_B}(X) = \Pr(X|D_B = 1)$, the reweighting factor that keeps all conditioning variables as in period 0 becomes

$$\Psi(X) = \frac{\Pr(X|D_B = 1)}{\Pr(X|D_B = 0)} = \frac{\Pr(D_B = 1|X)/\Pr(D_B = 1)}{\Pr(D_B = 0|X)/\Pr(D_B = 0)}.$$

The reweighting factor can be easily computed by estimating a probability model for $\Pr(D_B = 1|X)$, and using the predicted probabilities to compute a value $\widehat{\Psi}(X)$ for each observation. DFL suggest estimating a flexible logit or probit model, while Hirano,

[50] On the other hand, by analogy with the treatment effects literature, Firpo, Fortin, and Lemieux (2007) use time period 0 as the reference group.

Imbens, and Ridder propose to use a non-parametric logit model.[51]

The reweighting decomposition procedure can be implemented in practice as follows:

1. Pool the data for group A and B and run a logit or probit model for the probability of belonging to group B:

$$\Pr(D_B{=}1|X) = 1 - \Pr(D_B{=}0|X) = 1 - \Pr(\varepsilon > -h(X)\beta) = \Lambda(-h(X)\alpha) \tag{30}$$

where $\Lambda()$ is either a normal or logit link function, and $h(X)$ is a polynomial in X.

2. Estimate the reweighting factor $\widehat{\Psi}(X)$ for observations in group A using the predicted probability of belonging to group B ($\widehat{\Pr}(D_B = 1|X)$) and A ($\widehat{\Pr}(D_B = 0|X) = 1 - \widehat{\Pr}(D_B = 1|X)$), and the sample proportions in group B ($\widehat{\Pr}(D_B = 1)$) and A ($\widehat{\Pr}(D_B = 0)$):

$$\widehat{\Psi}(X) = \frac{\widehat{\Pr}(D_B = 1|X)/\widehat{\Pr}(D_B = 1)}{\widehat{\Pr}(D_B = 0|X)/\widehat{\Pr}(D_B = 0)}.$$

3. Compute the counterfactual statistic of interest using observations from the group A sample reweighted using $\widehat{\Psi}(X)$.

In DFL, the main object of interest is the probability density function, which is estimated using kernel density methods. The density for group A and the counterfactual density can be estimated as follows using kernel density methods, where $K(.)$ is the kernel function:[52]

$$\begin{aligned}
\widehat{f}_{Y_A}(y) &= \frac{1}{h \cdot N_A} \sum_{i \in A} K\left(\frac{Y_i - y}{h}\right), \\
\widehat{f}_{Y_A^C}(y) &= \frac{1}{h \cdot N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot K\left(\frac{Y_i - y}{h}\right).
\end{aligned}$$

Consider the density function for group A, $f_{Y_A}(y)$, and the counterfactual density $f_{Y_A^C}(y)$. The composition effect in a decomposition of densities is:

$$\Delta_X^{f(y)} = f_{Y_A^C}(y) - f_{Y_A}(y). \tag{31}$$

Various statistics from the wage distribution, such as the 10th, 50th, and 90th percentile, or the variance, Gini, or Theil coefficients can be computed either from the

[51] The estimator suggested by Hirano, Imbens, and Ridder (2003) is a series estimator applied to the case of a logit model. The idea is to add increasingly higher order polynomial terms in the covariates as the size of the sample increases. Importantly, they also show that this approach yields an efficient estimate of the treatment effect.

[52] The two most popular kernel functions are the Gaussian and the Epanechnikov kernel.

counterfactual density or the counterfactual distribution using the reweighting factor. The latter procedure is easier to use as it simply involves computing (weighted) statistics using standard computer packages. For example, the counterfactual variance can be computed as:

$$\widehat{Var}_{Y_A^C} = \frac{1}{N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot \left(Y_i - \widehat{\mu}_{Y_A^C} \right)^2,$$

where the counterfactual mean $\widehat{\mu}_{Y_A^C}$ is:

$$\widehat{\mu}_{Y_A^C} = \frac{1}{N_A} \sum_{i \in A} \widehat{\Psi}(X_i) \cdot Y_i.$$

For the 90-10, 90-50, and 50-10 wage differentials, the sought-after contributions to changes in inequality are computed as differences in the composition effects, for example,

$$\Delta_X^{90-10} = [Q_{A,.9}^C - Q_{A,.9}] - [Q_{A,.1}^C - Q_{A,.1}]. \tag{32}$$

Table 5 presents, in panel A, the results of a DFL decomposition of changes over time in male wage inequality using large samples from combined MORG-CPS data as in Firpo, Fortin, and Lemieux (2007). In this decomposition, the counterfactual distribution of wages in 1983/85 is constructed by reweighting the characteristics of workers in 1983/85 (time period 0) so that they look like those of 2003/05 (time period 1) workers, holding the conditional distribution of wages in 1983/05 fixed.[53] The results of the aggregate decomposition, reported in the first three rows of Table 5, show that composition effects play a large role in changes in overall wage inequality, as measured by the 90-10 log wage differential or the variance of log wages. But the wage structure effects are more important when looking for increases at the top of the wage distribution, as measured by the 90-50 log wage differential, or decreases in the bottom, as measured by the 50-10 log wage differential.

Advantages

The main advantage of the reweighting approach is its simplicity. The aggregate decomposition for any distributional statistics is easily computed by running a single probability model (logit or probit) and using standard packages to compute distributional statistics with $\widehat{\Psi}(X_i)$ as weight.

[53] By contrast, in the original DiNardo, Fortin, and Lemieux (1996) decomposition, workers in 1988 (time period 1) were reweighed to look like workers in 1979 (time period 0). The counterfactual distribution of wages was asking what would the distribution of wages look like if the workers' characteristics had remained at 1979 levels.

Another more methodological advantage is that formal results from Hirano, Imbens, and Ridder (2003) and Firpo (2007, 2010) establish the efficiency of this estimation method. Note that although it is possible to compute analytically the standard errors of the different elements of the decomposition obtained by reweighting, it is simpler in most cases to conduct inference by bootstrapping.[54]

For these two reasons, we recommend the reweighting approach as the method of choice for computing the aggregate decomposition. This recommendation even applies in the simple case of the mean decomposition. As pointed out by Barsky et al. (2002), a standard OB decomposition based on a linear regression model will yield biased estimates of the decomposition terms when the underlying conditional expectation of Y given X is non-linear (see Section 3.4). They suggest using a reweighting approach as an alternative, and the results of Hirano, Imbens, and Ridder (2003) can be used to show that the resulting decomposition is efficient.

Limitations

A first limitation of the reweighting method is that it is not straightforwardly extended to the case of the detailed decomposition. One exception is the case of binary covariates where it is relatively easily to compute the corresponding element of the decomposition. For instance, in the case of the union status (a binary covariate), DFL show how to compute the component of the composition corresponding to this particular covariate. It also relatively easy to compute the corresponding element of the wage structure effect. We discuss in Section 5 other options that can be used in the case of non-binary covariates.

As in the program evaluation literature, reweighting can have some undesirable properties in small samples when there is a problem of common support. The problem is that the estimated value of $\Psi(X)$ becomes very large when $\Pr(D_B = 1|X)$ gets close to 1. While lack of common support is a problem for any decomposition procedure, Frolich (2004) finds that reweighting estimators perform particularly poorly in this context, though Busso, DiNardo, and McCrary (2009) reach the opposite conclusion using a different simulation experiment.[55]

Finally, even in cases where a pure reweighting approach has some limitations, there may be gains in combining reweighting with other approaches. For instance, we discuss

[54] The analytical standard errors have to take account of the fact that the logit or probit model used to construct the reweighting factor is estimated. Firpo, Fortin and Lemieux (2007) show how to perform this adjustment. In practice, however, it is generally simpler to bootstrap the whole estimation procedure (both the estimation of the logit/probit to construct the weights and the computation of the various elements of the decomposition).

[55] In principle, other popular methods in the program evaluation literature such as matching could be used instead of reweighting.

in the next section how reweighting can be used to improve a decomposition based on the RIF-regression approach of Fortin, Firpo, and Lemieux (2009). Lemieux (2002) also discusses how an hybrid approach based on DFL reweighting and the JMP decomposition procedure can be used to compute both the between- and within-group components of the composition and wage structure effects.

4.6 Methods based on estimating the conditional distribution

Procedure(s)

As mentioned above, when we first introduced the key counterfactual distribution of interest in equation (5), an alternative approach to the construction of this counterfactual is based on the estimation of the conditional distribution of the outcome variable, $F_{Y_A|X_A}(y|X)$. The counterfactual distribution is then estimated by integrating this conditional distribution over the distribution of X in group B.

Two early parametric methods based on this idea were suggested by Donald, Green, and Paarsch (2000), and Fortin and Lemieux (1998).[56] Donald, Green and Paarsch propose estimating the conditional distribution using a hazard model. The (conditional) hazard function is defined as

$$h(y|X) = \frac{f(y|X)}{S(y|X)},$$

where $S(y|X) = 1 - F(y|X)$ is the survivor function. Therefore, the conditional distribution of the outcome variable, $F(y|X)$, or its density, $f(y|X)$, is easily recovered from the estimates of the hazard model. For instance, in the standard proportional hazard model[57]

$$h(y|X) = \exp(X\alpha)h_0(y),$$

estimates of α and of the baseline hazard $h_0(y)$ can be used to recover the conditional distribution

$$F(y|X) = 1 - \exp(-\Lambda_0(y)\exp(X\alpha)),$$

where $\Lambda_0(y) = \int h_0(u)du$ is the integrated baseline hazard.

[56] Foresi and Perrachi (1995) proposed to use a sequence of logit models to estimate the conditional distribution of excess returns.

[57] Donald, Green and Paarsch (2000) use a more general specification of the proportional hazard model where α and $h_0(y)$ are allowed to vary for different values (segments) of y.

Fortin and Lemieux (1998) suggest estimating an ordered probit model instead of a hazard model. They consider the following model for the outcome variable Y:

$$Y = \Lambda^{-1}(Y^*),$$

where $\Lambda(\cdot)$ is a monotonically increasing transformation function. The latent variable Y^*, interpreted as a latent "skill index" by Fortin and Lemieux, is defined as

$$Y^* = X\alpha + \varepsilon,$$

where ε is assumed to follow a standard normal distribution. It follows that the conditional distribution of Y is given by

$$F(y|X) = \Phi(-X\alpha + \Lambda(y)).$$

Fortin and Lemieux implement this in practice by discretizing the outcome variable into a large number of small bins. Each bin j corresponds to values of Y between the two thresholds c_{j-1} and c_j. The conditional probability of Y being in bin j is

$$Prob(c_{j-1} \leq Y \leq c_j|X) = \Phi(-X\alpha + \Lambda(c_j)) - \Phi(-X\alpha + \Lambda(c_{j-1})).$$

This corresponds to an ordered probit model where the $\Lambda(c_j)$ parameters (for $j = 1, ...J$) are the usual latent variable thresholds. The estimated values of α and of the thresholds can then be used to construct the counterfactual distribution, just as in Donald, Green, and Paarsch (2000).

To be more concrete, the following steps could be used to estimate the counterfactual distribution $F_{Y_A^C}(y)$ at the point $y = c_j$:

1. Estimate the ordered probit for group A. This yields estimates $\widehat{\alpha}_A$ and $\widehat{\Lambda}_A(c_j)$, the ordered probit parameters.

2. Compute the predicted probability $\widehat{F}_{Y_A|X_A}(c_j|X_i) = \Phi(-X_i\widehat{\alpha}_A + \widehat{\Lambda}_A(c_j))$ for each individual i in group B.

3. For each threshold c_j, compute the sample average of $\widehat{F}_{Y_A|X_A}(c_j|X_i)$ over all observations in group B:

$$\widehat{F}_{Y_A^C}(c_j) = \frac{1}{N_B}\sum_{i\in B}\Phi(-X_i\widehat{\alpha}_A + \widehat{\Lambda}_A(c_j)).$$

Repeating this for a large number of values of $y = c_j$ will provide an estimate of the

counterfactual distribution $F_{Y_A^C}(y)$.

In a similar spirit, Chernozhukov, Fernandez-Val, and Melly (2009) suggest a more flexible distribution regression approach for estimating the conditional distribution $F(y|X)$. The idea is to estimate a separate regression model for each value of y. They consider the model $F(y|X) = \Lambda(X\alpha(y))$, where $\Lambda(\cdot)$ is a known link function. For example, if $\Lambda(\cdot)$ is a logistic function, $\alpha(y)$ can be estimated by creating a dummy variable $\mathbb{1}\{Y_i \leq y\}$ indicating whether the value of Y_i is below y, where $\mathbb{1}\{\cdot\}$ is the indicator function, and running a logit regression of $\mathbb{1}\{Y_i \leq y\}$ on X_i to estimate $\alpha(y)$.

Similarly, if the link function is the identity function ($\Lambda(z) = z$) the probability model is a linear probability model. If the link function is the normal CDF ($\Lambda(z) = \Phi(z)$) the probability model is a probit. Compared to Fortin and Lemieux (1998), Chernozhukov, Fernandez-Val, and Melly (2009) suggest estimating a separate probit for each value of y, while Fortin and Lemieux use a more restrictive model where only the intercept (the threshold in the ordered probit) is allowed to change for different values of y.

As above, the counterfactual distribution can be obtained by first estimating the regression model (probit, logit, or LPM) for group A to obtain the parameter estimates $\widehat{\alpha}_A(y)$, computing the predicted probabilities $\Lambda(X_i\widehat{\alpha}_A(y))$, and averaging over these predicted probabilities to get the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$:

$$\widehat{F}_{Y_A^C}(y) = \frac{1}{N_B}\sum_{i\in B}\Lambda(X_i\widehat{\alpha}_A(y)).$$

Once the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$ has been estimated, counterfactual quantiles can be obtained by inverting the estimated distribution function. Consider $Q^C_{\tau,A}$, the τ^{th} quantile of the counterfactual distribution $F_{Y_A^C}(\cdot)$. The estimated counterfactual quantile is:

$$\widehat{Q}^C_{A,\tau}=\widehat{F}^{-1}_{Y_A^C}(\tau).$$

It is useful to illustrate graphically how the estimation of the counterfactual distribution $\widehat{F}_{Y_A^C}(y)$ and the inversion into quantiles can be performed in practice. Figure 1 first shows the actual CDF's for group A, $F_{Y_A}(\cdot)$, and B, $F_{Y_B}(\cdot)$, respectively. The squares in between the two cumulative distributions illustrate examples of counterfactuals computed using the one of the method discussed above.

For example, consider the case of the median wage for group B, $Q_{B,.5}$. Using the distribution regression approach of Chernozhukov, Fernandez-Val, and Melly (2009), one can estimate, for example, a LPM by running a regression of $\mathbb{1}\{(Y_i \leq Q_{B,.5}\}$ on X_i for

group A. This yields an estimate of $\widehat{\alpha}_A(y = Q_{B,.5})$ that can then be used to compute $\widehat{F}_{Y_A^C}(y = Q_{B,.5})$. This counterfactual proportion is represented by the square on the vertical line over $y = Q_{B,.5}$ in Figure 1.

Figure 2 then illustrates what happens when a similar exercise is performed for a larger number of values of y (100 in this particular figure). It now becomes clear from the figure how to numerically perform the inversion. In the case of the median, the total gap between group A and B is $Q_{B,.5}-Q_{A,.5}$. The counterfactual median can then be estimated by picking the corresponding point $Q^C_{A,.5}$ on the counterfactual function defined by the set of points estimated by running a set of LPM at different values of y. In practice, one could compute the precise value of $Q^C_{A,.5}$ by estimating the LPMs (or a logit or probit) for a large number of values of y, and then "connecting the dots" (i.e. using linear interpolations) between these different values.

Figure 2 also illustrates one of the key messages of the chapter listed in the introduction, namely that is it easier to estimate models for proportions than quantiles. In Figure 2, the difference in the proportion of observations under a given value of y is simply the vertical distance between the two cumulative distributions, $F_{Y_B}(y) - F_{Y_A}(y)$. Decomposing this particular gap in proportion is not a very difficult problem. As discussed in Section 3.5, one can simply run a LPM and perform a standard OB decomposition. An alternative also discussed in Section 3.5 is to perform a nonlinear decomposition using a logit or probit model. The conditional distribution methods of Fortin and Lemieux (1998) and Chernozhukov, Fernandez-Val, and Melly (2009) essentially amount to computing this decomposition in the vertical dimension.

By contrast, it is not clear at first glance how to decompose the **horizontal** distance, or quantile gap, between the two curves. But since the vertical and horizontal are just two different ways of describing the same difference between the two cumulative distributions $F_{Y_B}(y)$ and $F_{Y_A}(y)$, one can perform a first decomposition either vertically or horizontally, and then invert back to get the decomposition in the other dimension. Since decomposing proportions (the vertical distance) is relatively easy, this suggests first performing the decomposition on proportions at many points of the distribution, and then inverting back to get the decomposition in the quantile dimension (the horizontal distance).

Table 5 reports, in panels B and C, the results of the aggregate decomposition results for male wages using the method of Chernozhukov, Fernandez-Val, and Melly (2009). The counterfactual wage distribution is constructed by asking what would be the distribution of wages in 1983/85 if the conditional distribution was as in 2003/05. Panel B uses the

LPM to estimate $\Lambda(X_i\widehat{\alpha}_A(y))$ while the logit model is used in Panel C.[58] The first rows of Panel B and C show the changes in the wage differentials based on the fitted distributions, so that any discrepancies between these rows in the first row of Panel A shows the estimation errors. The second rows report the composition effects computed as the difference between the fitted distribution in 1983/85 and the counterfactual distribution. Given our relatively large sample, the differences across estimators in the different panels are at times statistically different. However, the results from the logit estimation in Panel C give results that are qualitatively similar to the DFL results shown in Panel A, with composition effects being relatively more important in accounting for overall wage inequality, as measured by the 90-10 log wage differential, and wage structure effects playing a relatively more important role in increasing wage inequality at the top and reducing wage inequality at the bottom.

Limitations

If one is just interested in performing an aggregate distribution, it is preferable to simply use the reweighting methods discussed above. Like the conditional quantile methods discussed in Section 4.4, conditional distribution methods require some parametric assumptions on the distribution regressions that may or may not be valid. Chernozhukov, Fernandez-Val, and Melly's distribution regression approach is more flexible than earlier suggestions by Donald, Green and Paarsch (2000) and Fortin and Lemieux (1998), but it potentially involves estimating a large number of regressions.

Running unconstrained regressions for a large number of values of y may result, however, in non-monotonicities in the estimated counterfactual distribution $\widehat{F}_{Y_A^C}(y)$. Smoothing or related methods then have to be used to make sure that the counterfactual distribution is monotonic and, thus, invertible into quantiles.[59] By contrast, reweighting methods require estimating just one flexible logit or probit regression, which is very easy to implement in practice.

Advantages

An important advantage of distribution regression methods over reweighting is that they can be readily generalized to the case of the detailed decomposition, although these decomposition will be path dependent. We show in the next section how Chernozhukov, Fernandez-Val, and Melly's distribution regression approach, and the related RIF regression method of Firpo, Fortin and Lemieux (2009) can be used to perform a detailed

[58] The estimation was performed using Melly's "counterfactual" Stata procedure. The computation of the variance and gini were based on the estimation of 100 centiles.

[59] Chernozhukov, Fernandez-Val, and Melly (2009) use the method of Chernozhukov, Fernandez-Val, and Galichon (2010) to ensure that the function is monotonic.

decomposition very much in the spirit of the traditional OB decomposition for the mean.

4.7 Summary

In this section we discuss most of the existing methods that have been proposed to perform an aggregate decomposition for general distributional statistics. While all these methods could, in principle, yield similar results, we argue that DFL reweighting is the method of choice in this context for two main reasons. First, it is simple to implement as it simply involves estimating a single logit or probit model for computing the reweighting factors. Counterfactual values of any distributional statistical can then be readily computed from the reweighted sample. By contrast, methods that yield counterfactual estimates of quantiles or the whole CDF require estimating a separate model at a large number of points in the distribution.

The second advantage of reweighting is that there are well established results in the program evaluation that show that the method is asymptotically efficient (Hirano, Imbens, and Ridder, 2003, and Firpo, 2007).

5 Detailed decompositions for general distributional statistics

In this Section, we extend the methods introduced above for the aggregate decomposition to the case of the detailed decomposition. We first show that conditional distribution methods based on distribution regressions can be used to compute both the composition and wage structure subcomponents of the detailed decomposition. We then discuss a related method based the RIF-regressions introduced in Firpo, Fortin, and Lemieux (2009). The main advantage of this last procedure is that it is regression based and, thus, as easy to use in practice as the traditional OB method.

The other methods proposed in Section 4 are not as easy to extend to the case of the detailed decomposition. We discuss, nonetheless, which elements of the detailed decomposition can be estimated using these various methods, and under which circumstances it is advantageous to use these methods instead of others.

5.1 Methods based on the conditional distribution

Procedure

In the case where the specification used for the distribution regression is the LPM, the aggregate decomposition of Section 4.6 can be generalized to the detailed decomposition as follows. Since the link function for the LPM is $\Lambda(z) = z$, the counterfactual distribution used earlier becomes:

$$\widehat{F}_{Y_A^C}(y) = \frac{1}{N_B} \sum_{i \in B} X_i \widehat{\alpha}_A(y) = \overline{X}_B \widehat{\alpha}_A(y).$$

We can also write:

$$\begin{aligned} \widehat{F}_{Y_B}(y) - \widehat{F}_{Y_A}(y) &= \left[\widehat{F}_{Y_B}(y) - \widehat{F}_{Y_A^C}(y)\right] + \left[\widehat{F}_{Y_A^C}(y) - \widehat{F}_{Y_A}(y)\right] \\ &= \overline{X}_B \left(\widehat{\alpha}_B(y) - \widehat{\alpha}_A(y)\right) + \left(\overline{X}_B - \overline{X}_A\right) \widehat{\alpha}_A(y), \end{aligned}$$

where the first term is the familiar wage structure effect, while the second term is the composition effect. The above equation can, therefore, be used to compute a detailed decomposition of the difference in the proportion of workers below wage y between groups A and B. We obtain the detailed distribution of quantiles by *i)* computing the different counterfactuals for each element of X and α sequentially, for a large number of values of y, and *ii)* inverting to get the corresponding quantiles for each detailed counterfactual. A similar approach could also be used when the link function is a probit or a logit by using the procedure suggested in Section 3.5.

Advantages

The main advantage of this method based on distribution regressions and the global inversion of counterfactual CDF into counterfactual quantiles (as in Figure 2) is that it yields a detailed decomposition comparable to the OB decomposition of the mean.

Limitations

One limitation of this method is that it involves computing a large number of counterfactuals CDFs and quantiles, as the procedure has to be repeated for a sizable number of values of y. This can become cumbersome because of the potential non-monotonicity problems discussed earlier. Furthermore, the procedure suffers from the problem of path dependence since the different counterfactual elements of the detailed decomposition have to be computed sequentially. For these reasons, we next turn to a simpler approach based on a local, as opposed to a global, inversion of the CDF.

5.2 RIF-regression methods

Procedure

RIF-regression methods provide a simple way of performing detailed decompositions for any distributional statistic for which an influence function can be computed. Although we focus below on the case of quantiles of the unconditional distribution of the outcome variable, our empirical example includes the case of the variance and Gini. The procedure can be readily used to address glass ceiling issues in the context of the gender wage gap, or changes in the interquartile range in the context of changes in wage inequality. It can be used to either perform OB- type detailed decompositions, or a slightly modified "hybrid" version of the decomposition suggested by Firpo, Fortin, and Lemieux (2007) (reweighting combined with RIF regressions, as in Section 3.4 for the mean).

A RIF-regression (Firpo, Fortin, and Lemieux, 2009) is similar to a standard regression, except that the dependent variable, Y, is replaced by the (recentered) influence function of the statistic of interest. Consider $\text{IF}(y;\nu)$, the influence function corresponding to an observed wage y for the distributional statistic of interest, $\nu(F_Y)$. The recentered influence function (RIF) is defined as $\text{RIF}(y;\nu) = \nu(F_Y) + \text{IF}(y;\nu)$, so that it aggregates back to the statistics of interest ($\int \text{RIF}(y;\nu) \cdot dF(y) = \nu(F_Y)$). In its simplest form, the approach assumes that the conditional expectation of the $\text{RIF}(Y;\nu)$ can be modelled as a linear function of the explanatory variables,

$$\mathbb{E}\left[\text{RIF}(Y;\nu)\,|X\right] = X\gamma + \varepsilon,$$

where the parameters γ can be estimated by OLS.[60]

In the case of quantiles, the influence function $\text{IF}(Y, Q_\tau)$ is given by $(\tau - \mathbb{1}\{Y \leq Q_\tau\}) / f_Y(Q_\tau)$, where $\mathbb{1}\{\cdot\}$ is an indicator function, $f_Y(\cdot)$ is the density of the marginal distribution of Y, and Q_τ is the population τ-quantile of the unconditional distribution of Y. As a result, $\text{RIF}(Y;Q_\tau)$ is equal to $Q_\tau + \text{IF}(Y, Q_\tau)$, and can be rewritten as

$$\text{RIF}(y;Q_\tau) = Q_\tau + \frac{\tau - \mathbb{1}\{y \leq Q_\tau\}}{f_Y(Q_\tau)} = c_{1,\tau} \cdot \mathbb{1}\{y > Q_\tau\} + c_{2,\tau}, \tag{33}$$

where $c_{1,\tau} = 1/f_Y(Q_\tau)$ and $c_{2,\tau} = Q_\tau - c_{1,\tau} \cdot (1-\tau)$. Except for the constants $c_{1,\tau}$ and $c_{2,\tau}$, the RIF for a quantile is simply an indicator variable $\mathbb{1}\{Y \leq Q_\tau\}$ for whether the outcome variable is smaller or equal to the quantile Q_τ. Using the terminology introduced above, running a linear regression of $\mathbb{1}\{Y \leq Q_\tau\}$ on X is a distributional

[60] Firpo, Fortin, and Lemieux (2009) also propose other more flexible estimation procedures.

regression estimated at $y = Q_\tau$, using the link function of the linear probability model ($\Lambda(z) = z$).

There is, thus, a close connection between RIF regressions and the distributional regression approach of Chernozhukov, Fernandez-Val, and Melly (2009). In both cases, regression models are estimated for explaining the determinants of the proportion of workers earning less than a certain wage. As we saw in Figure 2, in Chernozhukov, Fernandez-Val and Melly (2009) estimates of models for proportions are then *globally* inverted back into the space of quantiles. This provides a way of decomposing quantiles using a series of simple regression models for proportions.

Figure 3 shows that RIF-regressions for quantiles are based on a similar idea, except that the inversion is only performed *locally*. Suppose that after estimating a model for proportions, we compute a counterfactual proportion based on changing either the mean value of a covariate, or the return to the covariate estimated with the LPM regression. Under the assumption that the relationship between counterfactual proportions and counterfactual quantiles is locally linear, one can then go from the counterfactual proportion to the counterfactual quantile (both illustrated in Figure 3) by moving along a line with a slope given by the slope of the counterfactual distribution function. Since the slope of a cumulative distribution is the just the probability density function, one can easily go from proportions to quantiles by dividing the elements of the decomposition for proportions by the density.

While the argument presented in Figure 3 is a bit heuristic, it provides the basic intuition for how we can get a decomposition model for quantiles by simply dividing a model for proportions by the density. As we see in equation (33), in the RIF for quantiles, the indicator variable $\mathbb{1}\{y \leq Q_\tau\}$ is indeed divided by $f_Y(Q_\tau)$ (i.e. multiplying by the constant $c_{1,\tau}$).

Firpo, Fortin, and Lemieux (2009) explain how to first compute the RIF, and then run regressions of the RIF on the vector of covariates. In the case of quantiles, the RIF is first estimated by computing the sample quantile $\widehat{Q}_\tau$, and estimating the density at that point using kernel methods. An estimate of the RIF of each observation, $\widehat{\text{RIF}}(Y_i; Q_\tau)$, is then obtained by plugging in the estimates $\widehat{Q}_\tau$ and $\widehat{f}(\widehat{Q}_\tau)$ into equation (33).

Letting the coefficients of the unconditional quantile regressions for each group be

$$\widehat{\gamma}_{g,\tau} = \left(\sum_{i \in G} X_i \cdot X_i^\intercal \right)^{-1} \cdot \sum_{i \in G} \widehat{\text{RIF}}(Y_{gi}; Q_{g,\tau}) \cdot X_i, \quad g = A, B \tag{34}$$

we can write the equivalent of the OB decomposition for any unconditional quantile as

$$\begin{aligned} \widehat{\Delta}_O^\tau &= \overline{X}_B \left(\widehat{\gamma}_{B,\tau} - \widehat{\gamma}_{A,\tau}\right) + \left(\overline{X}_B - \overline{X}_A\right) \widehat{\gamma}_{A,\tau} && (35) \\ &= \widehat{\Delta}_S^\tau + \widehat{\Delta}_X^\tau. && (36) \end{aligned}$$

The second term in equation (36) can be rewritten in terms of the sum of the contribution of each covariate as

$$\widehat{\Delta}_X^\tau = \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \widehat{\gamma}_{Ak,\tau}.$$

That is, the detailed elements of the composition effect can be computed in the same way as for the mean. Similarly, the detailed elements of the wage structure effects can be computed, but as in the case of the mean, these will also be subject to the problem of the omitted group.

Table 4 presents in its bottom panel such OB like gender wage gap decomposition of the 10th, 50th, and 90th percentiles of the unconditional distribution of wages corresponding to Tables 2 and 3 using the male coefficients as reference group and without reweighting. As with the MM decomposition presented in the top panel, the composition effects from the decomposition of the median gender pay gap reported in the central column of Table 4 are very close to those of the decomposition of the mean gender pay gap reported in column (1) of Table 3. As before, the wage structure effects in the relatively small NLSY sample are generally not statistically significant, with the exception of the industrial sectors which are, however, subject to the categorical variables problem. The comparison of the composition effects at the 10th and 90th percentiles shows that the impact of differences in life-time work experience is much larger at the bottom of the distribution than at the top where it is not statistically significant. Note that the aggregate decomposition results obtained using either the MM method or the RIF regressions do not exhibit statistically significant differences.

Table 5 presents in Panel D the results of the aggregate decomposition using RIF-regressions without reweighting. The results are qualitatively similar to those of Panels A and C. Table 6 extends the analysis of the decomposition of male wage inequality presented in Table 5 to the detailed decomposition. For each inequality measures, the detailed decomposition are presented both for the extension of the classic OB decomposition in equation (36), and for the reweighted-regression decomposition, described in the

case of the mean in section 3.4. [61] For the reweighted-regression decomposition, Table 6 reports the detailed elements of the main composition effect $\widehat{\Delta}^{\tau}_{X,p}$ and the detailed elements of the main wage structure effect $\widehat{\Delta}^{\tau}_{S,p}$, where

$$\widehat{\Delta}^{\tau}_{X,p} = \left(\overline{X}^{C}_{A} - \overline{X}_{A}\right)\widehat{\gamma}_{A,\tau} \quad \text{and} \quad \widehat{\Delta}^{\tau}_{S,p} = \overline{X}_{B}\left(\widehat{\gamma}_{B,\tau} - \widehat{\gamma}^{C}_{A,\tau}\right),$$

and where the group A sample is reweighted to mimic the group B sample, which means we should have $plim(\overline{X}^{C}_{A}) = plim(\overline{X}_{B})$. The total reweighting error $\widehat{\Delta}^{\tau}_{S,e} = \left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\gamma}^{C}_{A,\tau}$ corresponds to the difference between the "Total explained" across the classic OB and the reweighted-regression decomposition. For example, for the 90-10 log wage differential, it is equal to 0.0617-0.0619=0.0002. [62] The total specification error, $\widehat{\Delta}^{\tau}_{X,e} = \overline{X}^{C}_{A}\left(\widehat{\gamma}^{C}_{A,\tau} - \widehat{\gamma}_{A,\tau}\right)$, corresponds to the difference between the "Total wage structure" across the classic OB and the reweighted-regression decomposition and is found to be more important. In terms of composition effects, de-unionization is found to be an important factor accounting for the polarization of male wage inequality. It is also found to reduce inequality at the bottom, as measured by the 50-10 log wage differential, and to increase inequality at the top, as measured by the 90-50 log wage differential. In terms of wage structure effects, increases in the returns to education are found, as in Lemieux (2006a), to be the dominant factor accounting for overall increases in male wage inequality.

Advantages

Linearization provides several advantages. It is straightforward to invert the proportion of interest by dividing by the density. Since the inversion can be performed locally, another advantage is that we don't need to evaluate the global impact at all points of the distribution and worry about monotonicity. One gets a simple regression which is easy to interpret. As a result, the resulting decomposition is path independent.

Limitations

Like many other methods, RIF regressions assume the invariance of the conditional distribution (i.e., no general equilibrium effects). Also, a legitimate practical issue is how good the approximation is. For relatively smooth dependent variables, such as test scores, it may be a mute point. But in the presence of considerable heaping (usually displayed

[61] Using a reweighted regression approach can be particulalry important in the cases of RIF-regressions that are unlikely to be linear for distributional statistics besides the mean.

[62] The reweighting error reflects the fact that the composition effect in the reweighted-regression decomposition, $\left(\overline{X}_{B} - \overline{X}^{C}_{A}\right)\widehat{\gamma}^{C}_{A,\tau}$, is not exactly equal to the standard composition effect $\left(\overline{X}_{B} - \overline{X}_{B}\right)\widehat{\gamma}^{C}_{A,\tau}$ when the reweighted mean $\overline{X}^{C}_{A}$ is not exactly equal to $\overline{X}_{B}$.

in wage distribution), it may advisable to oversmooth to density estimates and compare its values around the quantile of interest. This can be formally looked at by comparing reweighting estimates to the OB-type composition effect based on RIF regressions (the specification error discussed earlier).

5.3 A reweighting approach

Procedure(s)

As we mention in Section 4, it is relatively straightforward to extend the DFL reweighting method to perform a detailed decomposition in the case of binary covariates. DFL show how to compute the composition effect corresponding to a binary covariate (union status in their application). Likewise, DiNardo and Lemieux (1997) use yet another reweighting technique to compute the wage structure component. We first discuss the case where a covariate is a binary variable, and then discuss the case of categorical (with more than 2 categories) and continuous variables.

Binary covariate

Consider the case of one binary covariate, X_1, and a vector of other covariates, X_2. For instance, DiNardo, Fortin, and Lemieux (1996) look at the case of unionization. They are interested in isolating the contribution of de-unionization to the composition effect by estimating what would have happened to the wage distribution if the distribution of unionization, but of none of the other covariates, had changed over time.

Letting A index the base period and B the end period, consider the counterfactual distribution $F_{Y_A^{C,X_1}}$, which represents the period A distribution that would prevail if the conditional distribution of unionization (but of none of the other covariates X_2) was as in period B.[63] Note that we are performing a counterfactual experiment by changing the conditional, as opposed to the marginal, distribution of unionization. Unless unionization is independent of other covariates ($X_1 \perp X_2$), the marginal distribution of unionization, $F_X(X_1)$, will depend on the distribution of X_2, $F_X(X_2)$. For instance, if unionization is higher in the manufacturing sector, but the share of workers in manufacturing declines over time, the overall unionization rate will decline even if, conditional on industrial composition, the unionization rate remains the same.

Using the language of program evaluation, we want to make sure that secular changes in the rate of unionization are not confounded by other factors such as industrial change. This is achieved by looking at changes in the conditional, as opposed to the marginal

[63] Note that in DFL, it is the opposite; group B is the 1988 time period and group A is the 1979 time period.

distribution of unionization. Note that the main problem with the procedure suggested by MM to compute the elements of the composition effect corresponding to each covariates is that it fails to control for this problem. MM suggest using an unconditional reweighting procedure based on change in the marginal, as opposed to the conditional distribution of covariates. Unless the covariates are independent, this will yield biased estimates of the composition effect elements of the detailed decomposition.

The counterfactual distribution $F_{Y_A^{C,X_1}}$ is formally defined as

$$\begin{aligned} F_{Y_A^{C,X_1}}(y) &= \int\int F_{Y_A|X_A}(y|x_1,x_2)\, dF_{X_B}(x_1|x_2)\, dF_{X_A}(x_2) \\ &= \int\int F_{Y_A|X_A}(y|x_1,x_2)\ \Psi_1(x_1,x_2)\, dF_{X_A}(x_1|x_2)\, dF_{X_A}(x_2) \\ &= \int\int F_{Y_A|X_A}(y|x_1,x_2)\ \Psi_1(x_1,x_2)\, dF_{X_A}(x_1,x_2), \end{aligned}$$

where the reweighting function is

$$\Psi_{X_1}(x_1,x_2) \equiv \frac{dF_{X_B}(x_1|x_2)}{dF_{X_A}(x_1|x_2))} \tag{37}$$

$$= x_1 \cdot \frac{\Pr_B(x_1{=}1|x_2)}{\Pr_A(x_1{=}1|x_2)} + (1-x_1) \cdot \frac{\Pr_B(x_1{=}0|x_2)}{\Pr_A(x_1{=}0|x_2)}, \tag{38}$$

Note that the conditional distribution $F_{Y_A}(Y_A|X_1,X_2)$ is assumed to be unaffected by the change in the conditional distribution of unionization (assumption of invariance of conditional distribution in Section 2). This amounts to assuming away selection into union status based on unobservables (after controlling for the other covariates X_2).

The reweighting factor $\Psi_1(x_1,x_2)$ can be computed in practice by estimating two probit or logit models for the probability that a worker is unionized in period A and B, respectively. The resulting estimates can then be used to compute the predicted probability of being unionized ($\Pr_A(x_1{=}1|x_2)$ and $\Pr_B(x_1{=}1|x_2)$) or not unionized ($\Pr_A(x_1{=}0|x_2)$ and $\Pr_B[x_1{=}0|x_2]$), and then plugging these estimates into the above formula.

DiNardo and Lemieux (1997) use a closely related reweighting procedure to compute the wage structure component of the effect of unions on the wage distribution. Consider the question of what would happen to the wage distribution if no workers were unionized. The distribution of wages among non-union workers:

$$F_{Y_g}(y|X_1=0) = \int\int F_{Y_g|X_g}(y|X_1=0,X_2)\, dF_{X_A}(X_2|X_1=0),$$

is not a proper counterfactual since the distribution of other covariates, X_2, may not be the same of union and non-union workers. DiNardo and Lemieux (1997) suggest solving this problem by reweighting non-union workers so that their distribution of X_2 is the same as for the entire workforce. The reweighting factor that accomplishes this at time A and B are $\Psi_{A,S_1}(X_2)$ and $\Psi_{B,S_1}(X_2)$, respectively, where:

$$\Psi_{g,S_1}(X_2) = \frac{\Pr_g(X_1{=}\,0)}{\Pr_g(X_1{=}\,0|X_2)}, \quad g = A, B.$$

Using these reweighting terms, we can write the counterfactual distribution of wages that would have prevailed in absence of unions as:

$$F_{Y_g^{C,S_1}}(y) = \int\int F_{Y_g|X_g}(y|X_1 = 0, X_2)\ \Psi_{g,S_1}(X_2)\, dF_{X_g}(X_2, X_1 = 0), \quad g = A, B.$$

These various counterfactual distributions can then be used to compute the contribution of unions (or another binary variable x_1) to the composition effect, $\Delta_{X_1}^{F(y)}$, and to the wage structure effect, $\Delta_{S_1}^{F(y)}$:

$$\Delta_{X_1}^{F(y)} = F_{Y_A}(y) - F_{Y_A^{C,X_1}}(y), \tag{39}$$

and

$$\Delta_{S_1}^{F(y)} = \left[F_{Y_A}(y) - F_{Y_A^{C,S_1}}(y)\right] - \left[F_{Y_B}(y) - F_{Y_B^{C,S_1}}(y)\right]. \tag{40}$$

Although we need three different reweighting factors ($\Psi_{X_1}(x_1, x_2)$, $\Psi_{A,S_1}(x_2)$, and $\Psi_{B,S_1}(x_2)$) to compute the elements of the detailed wage decomposition corresponding to x_1, these three reweighting factors can be constructed from the estimates of the two probability models $\Pr_A(x_1{=}\,1|x_2)$ and $\Pr_B(x_1{=}\,1|x_2)$. As before, once these reweighting factors have been computed, the different counterfactual statistics are easily obtained using standard statistical packages.

General covariates

It is difficult to generalize the approach suggested above to the case of covariates that are not binary. In the case of the composition effect, one approach that has been followed in the applied literature consists of sequentially adding covariates in the probability model $\Pr(D_B{=}\,1|X)$ used to compute $\Psi(X)$.[64] For instance, start with $\Pr(D_B{=}\,1|X_1)$, compute $\Psi_1(X_1)$ and the counterfactual statistics of interest by reweighting. Then do the same

[64] See, for example, Butcher and DiNardo (2002) and Altonji, Bharadwaj, and Lange (2008).

thing with $\Pr(D_B{=}1|X_1, X_2)$, etc.

One shortcoming of this approach is that the results depend of the order in which the covariates are sequentially introduced, just like results from a sequential decomposition for the mean also depend on the order in which the covariates are introduced in the regression. For instance, estimates the effect of unions that fail to control for any other covariates may be overstated if union workers tend to be concentrated in industries that would pay high wages even in the absence of unions. As pointed out by Gelbach (2009), the problem with sequentially introducing covariates can be thought of as an omitted variable problem. Unless there are compelling economic reasons for first looking at the effect of some covariates without controlling for the other covariates, sequential decompositions will have the undesirable property of depending (strongly in some cases) on the order of the decomposition (path dependence).[65]

Fortunately, there is a way around the problem of path dependence when performing detailed decompositions using reweighting methods. The approach however still suffers from the adding-up problem and is more appropriate when only the effect of a particular factor is of interest. To illustrate this approach, consider a case with three covariates X_1, X_2, and X_3. In a sequential decomposition, one would first control for X_1 only, then for X_1 and X_2, and finally for X_1, X_2, and X_3. On the one hand, the regression coefficient on X_1 and/or X_2 in regressions that fail to control for X_3 are biased because of the omitted variable problem. The corresponding elements of a detailed OB decomposition for the mean based on these estimated coefficients would, therefore, be biased too.

On the other hand, the coefficient on the last covariate to be introduced in the regression (X_3) is not biased since the other covariates (X_1 and X_2) are also controlled for. So although order matters in a sequential regression approach, the effect of the last covariate to be introduced is not affected by the omitted variable bias.

The same logic applies in the case of detailed decompositions based on a reweighting approach. Intuitively, the difference in the counterfactual distribution one gets by reweighting with X_1 and X_2 only, comparing to reweighting with X_1, X_2, **and** X_3 should yield the appropriate contribution of X_3 to the composition effect.

To see this more formally, consider the group A counterfactual distribution that would

[65] Both Butcher and DiNardo (2002) and Altonji, Bharadwaj, and Lange (2008) consider cases where there is indeed a good reason for following a particular order in the decomposition. For instance, Altonji, Bharadwaj, and Lange (2008) argue that, when looking at various youth outcomes, one should first control for predetermined factors like gender and race before controlling for other factors determined later in life (AFQT score, educational achievement, etc.). In such a situation, the decomposition is econometrically interpretable even if gender and race are introduced first without controlling for the other factors.

prevail if the distribution of X_3, conditional on X_1, X_2, was as in group B:

$$\begin{aligned} F_{Y_A^{C,X_3}}(y) &= \int F_{Y_A|X_A}(y|X)dF_{X_B}(X_3|X_1,X_2)dF_{X_A}(X_1,X_2), \\ &= \int F_{Y_A|X_A}(y|X)\Psi_{X_3|X_1,X_2}(X_1,X_2)dF_{X_A}(X_3|X_1,X_2)dF_{X_A}(X_1,X_2), \\ &= \int F_{Y_A|X_A}(y|X)\Psi_{X_3|X_1,X_2}(X_1,X_2)dF_{X_A}(X_1,X_2,X_3), \end{aligned}$$

where the reweighting factor $\Psi_{X_3|X_1,X_2}(X_1,X_2)$ can be written as:

$$\begin{aligned} \Psi_{X_3|X_1,X_2}(X_1,X_2) &\equiv \frac{dF_{X_B}(X_3|X_1,X_2)}{dF_{X_A}(X_3|X_1,X_2)} \\ &= \frac{dF_{X_B}(X_1,X_2,X_3)/dF_{X_B}(X_1,X_2)}{dF_{X_A}(X_1,X_2,X_3)/dF_{X_A}(X_1,X_2)} \\ &= \Psi(X_1,X_2,X_3)/\Psi_{X_1,X_2}(X_1,X_2). \end{aligned}$$

$\Psi(X_1,X_2,X_3)$ is the reweighting factor used to compute the aggregate decomposition in Section 4.5. $\Psi_{X_1,X_2}(X_1,X_1)$ is a reweighting factor based on all the covariates except the one considered for the detailed decomposition (X_3). As before, Bayes' rule can be used to show that:

$$\Psi_{X_3|X_1,X_2}(X_1,X_2) = \frac{\Pr(X_1,X_2|D_B=1)}{\Pr(X_1,X_2|D_B=0)} = \frac{\Pr(D_B=1|X_1,X_2)/\Pr(D_B=1)}{\Pr(D_B=0|X_1,X_2)/\Pr(D_B=0)}.$$

Once again, this new reweighting factor is easily computed by running a probit or logit regression (with X_1 and X_2 as covariates) and using predicted probability to estimate $\Psi_{X_3|X_1,X_2}(X_1,X_2)$.

This reweighting procedure for the detailed decomposition is summarized as follows:

1. Compute the reweighting factor using all covariates, $\Psi(X)$.

2. For each individual covariate k, compute the reweighting factor using all covariates but X_k, $\Psi_{X_{-k}}(X_{-k})$.

3. For each covariate k, compute the counterfactual statistic of interest using the ratio of reweighting factors $\Psi(X)/\Psi_{X_{-k}}(X_{-k})$ as weight, and compare it to the counterfactual statistic obtained using only $\Psi(X)$ as weight. The difference is the estimated contribution of covariate k to the composition effect.

Note that while this procedure does not suffer from path dependence, the contribution of each covariates does not sum up to the total contribution of covariates (aggregate composition effect). The difference is an interaction effect between the different covariates

which is harder to interpret.

Advantages

This reweighting procedure shares most of the advantages of the other reweighting procedures we proposed for the aggregate decomposition. First, it is generally easy to implement in practice. Second, by using a flexible specification for the logit/probit, it is possible to get estimates of the various components of the decomposition that depend minimally on functional form assumptions. Third, the procedure yields efficient estimates.

Limitations

With a large number of covariates, one needs to compute a sizable number of reweighting factors to compute the various elements of the detailed decomposition. This can be tedious, although it does not require that much in terms of computations since each probit/logit is easy to estimate. Another disadvantage of the suggested decomposition is that although it does not suffer from the problem of path dependence, we are still left with an interaction term which is difficult to interpret. For these reasons, we suggest to first use a regression-based approach like the RIF-regression approach discussed above, which is essentially as easy to compute as a standard OB decomposition. The reweighting procedure suggested here can then be used to probe these results, and make sure they are robust to the functional-form assumptions implicit in the RIF-regression approach.

5.4 Detailed decomposition based on conditional quantiles

As we mentioned earlier, the method of Machado and Mata (2005) can be used to compute the wage structure sub-components of the detailed decomposition. These components are computed by sequentially switching the coefficients of the quantile regressions for each covariate from their estimated valued for group B to their estimated values for group A. This sequential switching cannot be used, however, to compute the sub-components of the composition effect of the detailed decomposition. Rather, Machado and Mata (2005) suggest an unconditional reweighting approach to do so. This does not provide a consistent effect since the effect of the reweighted covariate of interest gets confounded by other covariates correlated with that same covariate. For instance, if union workers are more concentrated in manufacturing, doing an unconditional reweighting on unions will also change the fraction of workers in manufacturing. In this sense the effect of unions is getting confounded by the effect of manufacturing.

This is a significant drawback since it is arguably more important to conduct a detailed

decomposition for the composition effect than for the wage structure effect. As discussed earlier, there are always some interpretation problems with the detailed components of the wage structure effect because of the omitted group problem.

One solution is to use the conditional reweighting procedure described above instead. But once this type of reweighting approach is used, there is no need to estimate (conditional) quantile regressions. Unless the quantile regressions are of interest on their own, it is preferable to use a more consistent approach, such as the one based on the estimation of RIF-regressions, for estimating the detailed components of both the wage structure and composition effects.

6 Extensions

In this section, we present three extensions to the decomposition methods discussed earlier. We first consider the case where either the ignorability or the zero conditional mean assumptions are violated because of self-selection or endogeneity of the covariates. We next discuss the situation where some of these problems can be addressed when panel data are available. We conclude the section by discussing the connection between conventional decomposition methods and structural modelling.

6.1 Dealing with self-selection and endogeneity

The various decomposition procedures discussed up to this point provide consistent estimates of the aggregate composition and wage structure effects under the ignorability assumption. Stronger assumptions, such as conditional mean independence (for decompositions of the mean) or straight independence, have to be invoked to perform the detailed decomposition. In this section we discuss some alternatives for estimating the decomposition when these assumptions fail. We mostly focus on the case of the OB decomposition of the mean, though some of the results we present could be extended to more general distributional statistics.

We consider three scenarios, first introduced in Section 2.1.6, under which the OB decomposition is inconsistent because of a failure of the ignorability or conditional independence assumption. In the first case, the problem is that individuals from groups A and B may self-select differently into the labor market. For instance, participation decisions of men (group B) may be different from participation decisions of women (group A) in ways that are not captured by observable characteristics. In the second case, we

consider what happens when individuals can self-select into group A or B (for instance union and non-union jobs) on the basis of unobservables. The third case is a standard endogeneity problem where the covariates are correlated with the error term. For example, education (one of the covariate) may be correlated with the error term because more able individuals tend to get more schooling.

1. Differential self-selection within groups A and B.

One major concern when decomposing differences in wages between two groups with very different labor force participation rates is that the probability of participation depends on unobservables ε in different ways for groups A and B. This is a well known problem in the gender wage gap literature (Blau and Kahn, 2006, Olivetti and Petrongolo, 2008, Mulligan and Rubinstein, 2008, etc.) and in the black-white wage gap literature (Neal and Johnson, 1996).

Our estimates of decomposition terms may be directly affected when workers of groups A and B self-select into the labor market differently. Thus, controlling for selection based on observables and unobservables is necessary to guarantee point identification of the decomposition terms. If no convincing models for self-selection is available a more agnostic approach based on bounds has also been recently proposed. Therefore, following Machado (2009), we distinguish three branches in the literature of self-selection:i) selection on observables; ii) selection based on unobservables; iii) bounds.

Selection based on observables and, when panel data are available, on time-invariant unobserved components can be used to impute values for the missing data on wages of non-participants. Representative papers of this approach are Neal and Johnson (1996), Johnson et al. (2000), Neal (2004), Blau and Kahn (2006) and Olivetti and Petrongolo (2008). These papers are typically concerned with mean or median wages. However, extensions to cumulative distribution functions or general ν-wage gaps could also be considered.

When labor market participation is based on unobservables, correction procedures for the mean wages are also available. In these procedures, a control variate is added as a regressor in the conditional expectation function. The exclusion restriction that an available instrument Z does not belong to the conditional expectation function also needs to be imposed.[66] Leading parametric and nonparametric examples are Heckman (1974, 1976), Duncan and Leigh (1980), Dolton, Makepeace, and Van Der Klaauw (1989), Vella

[66] As is well known, selection models can be identified on the basis of functional restrictions even when an excluded instrumental variable is not available. This is no longer viewed, however, as a credible identification strategy. We, therefore, only focus on the case where an instrumental variable is available.

(1998), Mulligan and Rubinstein (2008).

In this setting, the decomposition can be performed by adding a control variate $\lambda_g(X_i, Z_i)$ to the regression. In most applications, $\lambda_g(X_i, Z_i)$ is the usual inverse Mills' ratio term obtain by fitting a probit model of the participation decision. Note that the addition of this control variate slightly changes the interpretation of the decomposition. The full decomposition for the mean is now

$$\begin{aligned}\Delta^{\mu} &= (\beta_{B0} - \beta_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} (\beta_{Bk} - \beta_{Ak}) + \overline{\lambda}_B(\sigma_B - \sigma_A) \\ &\quad + \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \beta_{Ak} + \left(\overline{\lambda}_B - \overline{\lambda}_A\right) \sigma_A.\end{aligned}$$

where σ_A and σ_B are the estimated coefficients on the control variates. The decomposition provides a full accounting for the wage gap that also includes differences in both the composition of unobservables ($\left(\overline{\lambda}_B - \overline{\lambda}_A\right) \sigma_A$) and in the return to unobservables ($\overline{\lambda}_B(\sigma_B - \sigma_A)$). This treats symmetrically the contribution of observables (the X's) and unobservables in the decomposition.

A third approach uses bounds for the conditional expectation function of wages for groups A and B. With those bounds one can come up with bounds for the wage structure effect, Δ^{μ}_S, and the composition effect, Δ^{μ}_X. Let $\Delta^{\mu}_S = \mathbb{E}\big[\big(\mathbb{E}\left[Y_B|X, D_B = 1\right] - \mathbb{E}\left[Y_A|X, D_B = 1\right]\big)|D_B = 1\big]$. Then, letting D_S be a dummy indicating labor force participation, we can write the conditional expected wage as,

$$\begin{aligned}\mathbb{E}\left[Y_g|X, D_g\right] &= \mathbb{E}\left[Y_g|X, D_g, D_S = 0\right] \\ &\quad + \Pr\left(D_S = 1|X, D_g\right) \left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - \mathbb{E}\left[Y_g|X, D_g, D_S = 0\right]\right)\end{aligned}$$

and therefore

$$\begin{aligned}L_g &+ \Pr\left(D_S = 1|X, D_g\right) \left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - L_g\right) \\ &\leq \mathbb{E}\left[Y_g|X, D_g\right] \\ &\leq U_g + \Pr\left(D_S = 1|X, D_g\right) \left(\mathbb{E}\left[Y_g|X, D_g, D_S = 1\right] - U_g\right)\end{aligned}$$

where L_g and U_g are lower and upper bounds of the distribution of Y_g, for $g = A, B$.

Therefore,

$$\begin{aligned}
&(\mathbb{E}[Y_B|X, D_B = 1, D_S = 1] - \mathbb{E}[Y_A|X, D_B = 1, D_S = 1]) \Pr(D_S = 1|X, D_B = 1) \\
&\quad + (L_B - U_A) \Pr(D_S = 0|X, D_B = 1) \\
\leq\ & \mathbb{E}[Y_B|X, D_B = 1] - \mathbb{E}[Y_A|X, D_B = 1] \\
\leq\ & (\mathbb{E}[Y_B|X, D_B = 1, D_S = 1] - \mathbb{E}[Y_A|X, D_B = 1, D_S = 1]) \Pr(D_S = 1|X, D_B = 1) \\
&\quad + (U_B - L_A) \Pr(D_S = 0|X, D_B = 1).
\end{aligned}$$

This bounding approach to the selection problem may also use restrictions motivated by econometric or economic theory to narrow the bounds, as in Manski (1990) and Blundell et al. (2010).

2. Self-Selection into groups A and B

In the next case we consider, individuals have the choice to belong to either group A or B. The leading example is the choice of the union status of workers. The traditional way of dealing with the problem is to model the choice decision and correct for selection biases using control function methods.[67]

As discussed in Section 2.1.6, it is also possible to apply instrumental variable methods more directly without explicitly modelling the selection process into groups A and B. Angrist and Imbens (1994) show that this will identify the wage gap for the subpopulation of compliers who are induced by the instrument to switch from one group to the other.

3. Endogeneity of the covariates

The standard assumption used in the OB decomposition is that the outcome variable Y is linearly related to the covariates, X, and that the error term υ is conditionally independent of X, as in equation (1). Now consider the case where the conditional independence assumption fails because one or several of the covariates are correlated with the error term. Note that while the ignorability assumption may hold even if conditional independence fails, we consider a general case here where neither assumption holds.

As is well known, the conventional solution to the endogeneity problem is to use instrumental variable methods. For example, if we suspect years of education (one of the covariate) to be correlated with the error term in the wage equation, we can still estimate the model consistently provided that we have a valid instrument for years of

[67] See for instance, the survey of Lewis (1986) who concludes that these methods yield unreliable estimates of the union wage gap. Given these negative results and the lack of credible instruments for unionization, not much progress has been made in this literature over the last two decades. One exception is DiNardo and Lee (2004) who use a regression discontinuity design.

education. The decomposition can then be performed by replacing the OLS estimates of the β coefficients by their IV counterparts.

Of course, in most cases it is difficult to come up with credible instrumentation strategies. It is important to remember, however, that even when the zero conditional mean assumption $\mathbb{E}(\upsilon|X) = 0$ fails, the aggregate decomposition may remain valid, provided that ignorability holds. This would be the case, for example, when unobserved ability is correlated with education, but the correlation (more generally the conditional distribution of ability given education) is the same in group A and B. While we are not able to identify the contribution of education vs. ability in this context (unless we have an instrument), we know that they are no systematic ability differences between groups A and B once we have controlled for education. As a result, the aggregate decomposition remains valid.

6.2 Panel data

An arguably better way of dealing with the selection and endogeneity problems mentioned above is to use panel data. Generally speaking, panel data methods can be used to compute consistent estimates of the β's in each of the three cases discussed earlier. For example, if the zero conditional mean assumption holds once we also control for a person-specific fixed effects θ_i in a panel of length T ($\mathbb{E}(\upsilon_{it}|X_{i1}, .., X_{iT}, \theta_i)$), we can consistently estimate β using standard panel data methods (fixed effects, first differences, etc.). This provides an alternative way of dealing with endogeneity problems when no instrumental variables are available.

As we also discussed earlier, panel data can be used to impute wages for years where an individual is not participating in the labor market (e.g. Olivetti and Petrongolo (2008)). Note that in cases where groups are mutually exclusive (e.g. men vs. women), it may still be possible to estimate fixed effect models if the basic unit used is the firm (or related concepts) instead, or in addition to the individual (Woodstock, 2008). Care has to be exercised in those circumstances to ensure that the firm fixed effect is the same for both female and male employees of the same firm. Another important issue with these models is the difficulty of interpretation of the differences in male and female intercepts which may capture the unobserved or omitted individual and firm effects.

Panel data methods have also been used to adjust for the selection into groups in cases where the same individual is observed in group A and B. For example, Freeman (1984) and Card (1996) estimate the union wage gap with panel data to control for the selection

of workers into union status. Lemieux (1998) uses a more general approach where the return to the fixed effect may be different in the union and non-union sector. He also shows how to generalize the approach to the case of a decomposition of the variance.

Without loss of generality, assume that the return to the fixed effect for non-union workers (group A) is 1, while it is equal to σ_B for union workers. The mean decomposition adjusted for fixed effects yields:

$$\begin{aligned}\Delta^{\mu} &= (\beta_{B0} - \beta_{A0}) + \sum_{k=1}^{K} \overline{X}_{Bk} (\beta_{Bk} - \beta_{Ak}) + \overline{\theta}_B(\sigma_B - 1) \\ &\quad + \sum_{k=1}^{K} \left(\overline{X}_{Bk} - \overline{X}_{Ak}\right) \beta_{Ak} + \left(\overline{\theta}_B - \overline{\theta}_A\right).\end{aligned}$$

The interpretation of the decomposition is the same as in a standard OB setting except that $\left(\overline{\theta}_B - \overline{\theta}_A\right)$ now represents the composition effect term linked to non-random selection into the union sector, while the wage structure term $\overline{\theta}_B(\sigma_B - 1)$ captures a corresponding wage structure effect.

More sophisticated model with several levels of fixed effects have also been used in practice. For instance, Abowd et al. (2008) decompose inter-industry wage differentials into various components that include both individual- and firm-specific fixed effects.

6.3 Decomposition in structural models

In Section 2, we pointed out that decomposition methods were closely related to methods used in the program evaluation literature where it is not necessary to estimate a fully specified structural model to estimate the main parameter of interest (the ATT). Provided that the ignorability assumption is satisfied, we can perform an aggregate decomposition without estimating an underlying structural model.

There are some limits, however, to what can achieved without specifying any structure to the underlying economic problem. As we just discussed in Section 6.1, one problem is that the ignorability assumption may not hold. Under this scenario, more explicit modelling may be useful for correcting biases in the decomposition due to endogeneity, self-selection, etc.

Another problem that we now address concerns the interpretation of the wage structure components of the detailed decomposition. Throughout this chapter, we have proposed a number of ways of estimating these components for both the mean and more general distributional statistics. In the case of the mean, the interpretation of the de-

tailed decomposition for the wage structure effect is relatively straightforward. Under the assumption (implicit in the OB decomposition) that the wage equations are truly linear and the errors have a zero conditional mean, we can think of the wage setting model as a fully specify structural model. The β coefficients are the "deep" structural parameters of the model, and these structural parameters are used directly to perform the decomposition.

Things become more complicated once we go beyond the mean. For instance, in the case of the variance (section 4.1), recall that the wage structure effect from equation (26) which depends on the parameters of both the models for the conditional mean (β) and for the variance (δ).

Take, for example, the case where one of the covariates is the union status of workers. The parameter δ captures the "compression", or within-group, effect, while the parameter β captures the "wage gap", or between-group, effect. These two terms have a distinct economic interpretation as they reflect different channels through which union wage policies tend to impact the wage distribution.

In the case of more general distributional statistics, the wage structure effect depends on an even larger number of underlying parameters capturing the relationship between the covariates and higher order moments of the distribution. As a result, the wage structure part of the detailed decomposition becomes even harder to interpret, as it potentially depends on a large number of underlying parameters.

In some cases, this may not pose a problem from an interpretation point of view. For instance, we may only care about the overall effect of unions, irrespective of whether it is coming from a between- or within-group effect (or corresponding components for higher order moments). But in other cases this type of interpretation may be unsatisfactory. Consider, for example, the effect of education on the wage structure. Like unions, education may influence wage dispersion through a between- or within-group channel. The between-group component is linked to the traditional return to education (effect on conditional means), but education also has a substantial effect on within-group dispersion (see, e.g., Lemieux, 2006b). All these effects are combined together in the decomposition methods proposed in Section 5, which is problematic if we want to know, for instance, the specific contribution of changes in the return to education to the growth in wage inequality.

In these circumstances, we need to use a more structural approach to get a more economically interpretable decomposition of the wage structure effect. The decomposition method of Juhn, Murphy and Pierce (1993) is, in fact, an early example of a more

structurally-based decomposition. In their setting, the model for the conditional mean is interpreted as an underlying human capital pricing equation. Likewise, changes in residual wage dispersion (given X) are interpreted as reflecting an increase in the return to unobservable skills.

As we discussed in Section 4.3, the fact that Juhn, Murphy and Pierce (1993)'s provides a richer interpretation of the wage structure effect by separating the within- and between-group components is an important advantage of the method. We also mentioned, however, that the interpretation of the decomposition was not that clear for distributional statistics going beyond the variance, and that the procedure typically imposes substantial restrictions on the data that may or may not hold. By contrast, a method like DFL imposes very little restrictions (provided that the probit/logit model used for reweighting is reasonably flexible), though it is more limited in terms of the economic interpretation of the wage structure effect.

In light of this, the challenge is to find a way of imposing more explicit structure on the economic problem while making sure the underlying model "fits" the data reasonably well. One possible way of achieving this goal is to go back to the structural form introduced in Section 2 ($Y_{gi} = m_g(X_i, \varepsilon_i)$), and use recent results from the literature on nonparametric identification of structural functions to identify the functions $m_g(\cdot)$. As discussed in Section 2.2.1, this can be done by invoking results obtained by Matzkin (2003), Blundell and Powell (2007) and Imbens and Newey (2009). Generally speaking, it is possible to identify the functions $m_g(\cdot)$ nonparametrically under the assumptions of independence of ε (Assumption 8), and strict monotonicity of $m_g(\cdot)$ in ε (Assumption 9).

But while it is possible, in principle, to nonparametrically identify the functions $m_g(\cdot)$, there is no guarantee that the resulting estimates will be economically interpretable. As a result, a more common approach used in the empirical literature is to write down a more explicit (and parametric) structural model, but carefully look at whether the model adequately fits the data. Once the model has been estimated, simulation methods can then be used to compute a variety of counterfactual exercises. The counterfactuals then form the basis of a more economically interpretable decomposition of the wage structure effect.

To take a specific example, consider Keane and Wolpin (1997) model of career progression of young men where educational and occupational choices are explicitly modeled using a dynamic programming approach. After carefully looking at whether the estimated model is rich enough to adequately fit the distribution of wages, occupational choices, and educational achievement, Keane and Wolpin use the estimated model to decompose

the distribution of lifetime utility (itself computed using the model). They conclude that 90 percent of the variance of lifetime utility is due to skill endowment heterogeneity (schooling at age 16 and unobserved type). By contrast, choices and other developments happening after age 16 have a relatively modest impact on the variance of lifetime utility.[68] The general idea here is to combine structural estimation and simulation methods to quantify the contribution of the different parameters of interest to some decompositions of interest. These issues are discussed in more detail in the chapter on structural methods by Keane, Wolpin, and Todd (2010).

One last point is that the interpretation problem linked to the wage structure effect does not apply to the detailed decomposition for the composition effect. In that case, each component is based on a clear counterfactual exercise that does not require an underlying structure to be interpretable. The aggregate decomposition is based on the following counterfactual exercise: what would be the distribution of outcomes for group A if the distribution of the covariates for group A were the same as for group B? Similarly, the detailed decomposition is based on a conditional version of the counterfactual. For example, one may want to ask what would be the distribution of outcomes for group A if the distribution of unionization (or another covariate) for group A was the same as for group B, *conditional* on the distribution of the other covariates remaining the same.

These interpretation issues aside, it may still be useful to use a more structural approach when we are concerned about the validity of the decomposition because of self-selection, endogeneity, etc. For instance, in Keane and Wolpin (1997), the choice of schooling and occupation is endogenous. Using standard decomposition methods to look, for instance, at the contribution of the changing distribution of occupations to changes in the distribution wages would yield invalid results because occupational choice is endogenous. In such a context, structural modelling, like the IV and selection methods discussed in Section 6.1, can help recover the elements of the decomposition when standard methods fail because of endogeneity or self-selection. But the problem here is quite distinct from issues with the wage structure effect where standard decomposition methods are limited because of an interpretation problem, and where structural modelling provides a natural way of resolving this interpretation problem. By contrast, solutions to the problem of endogeneity or self-selection are only as a good as the instruments (or related assumptions) used to correct for these problems. As a result, the value added of

[68] Note, however, that Hoffman (2009) finds that skill endowments have a sizably smaller impact in a richer model that incorporates comparative advantage (across occupations), search frictions, and exogenous job displacement.

the structural approach is much more limited in the case of the composition effect than in the case of the wage structure effect.

This last point is very clear in the emerging literature where structural modelling is used in conjunction with experimental data. For example, Card and Hyslop (2005) use experimental data from the Self Sufficiency Project (SSP) to look at why individuals offered with a generous work subsidy are less likely to receive social assistance (SA). By definition, there is no composition effect since the treatment and control groups are selected by random assignment. In that context, the average treatment effect precisely corresponds to the wage structure effect (or "SA" structure effect in this context) in a decomposition of the difference between the treatment and control group. It is still useful, however, to go beyond this aggregate decomposition to better understand the mechanisms behind the measured treatment effect. Card and Hyslop (2005) do so by estimating a dynamic search model.

This provides much more insight into the "black box" of the treatment effect than what a traditional decomposition exercise would yield. Remember that the detailed wage structure component in a OB type decomposition is based on the difference between the return to different characteristics in the two groups. In a pure experimental context like the SSP project, this simply reflects some heterogeneity in the treatment effect across different subgroups. Knowing about the importance of heterogeneity in the treatment effect is important from the point of view of the generalizability of the results. But unlike a structural approach, it provides relatively little insight on the mechanisms underlying the treatment effect.

7 Conclusion

The development of new decomposition methods has been a fertile area of research over the last 10-15 years. Building on the seminal work of Oaxaca (1973) and Blinder (1973), a number of procedures that go beyond the mean have been suggested and used extensively in practice. In this chapter, we have reviewed these methods and suggested a number of "best practices" for researchers interested in these issues. We have also illustrated how these methods work in practice by discussing existing applications and working through a set of empirical examples throughout the chapter.

Another important and recent development in this literature has linked decomposition methods to the large and growing literature on program evaluation and treatment effects. This connection is useful for several reasons. First, it helps clarify some interpretation

issues with decompositions. In particular, results from the treatment effect literature can be used to show, for example, that we can give a structural interpretation to an aggregate decomposition under the assumption of ignorability. Another benefit of this connection is that formal results about the statistical properties of treatment effect estimators can also be directly applied to decomposition methods. This helps guide the choice of decomposition methods that have good statistical properties, and conduct inference on these various components of the estimated decomposition.

But this connection with the treatment effect literature also comes at a cost. While no structural modelling is required to perform a decomposition or estimate a treatment effect, these approaches leave open the question of what are the economic mechanisms behind the various elements of the decomposition (or behind the treatment effect). Now that the connection between decomposition methods and the treatment effect literature has been well established, an important direction for future research will be to improve the connection between decomposition methods and structural modelling.

The literature on inequality provides some useful hints on how this connection can be useful and improved upon. In this literature, decomposition methods have helped uncover the most important factors behind the large secular increase in the distribution of wages. Those include the return to education, de-unionization, and the decline in the minimum wage, to include a few examples. These findings have spurred a large number of more conceptual studies trying to provide formal economic explanations for these important phenomena. In principle, these explanations can then be more formally confronted to the data by writing down and estimating a structural model, and using simulation methods to quantify the role of these explanations.

This suggest a two-step research strategy where "over the shelf" decomposition methods, like those discussed in this chapter, can first be used to uncover the main forces underlying an economic phenomenon of interest. More "structural" decomposition methods could then be used to better understand the economics behind the more standard decomposition results. We expect such a research strategy to be a fruitful area of research in the years to come.

References

Abowd, John M., Francis Kramarz, Paul Lengerman, and Sebastien Roux (2008), "Persistent Inter-Industry Wage Differences: Rent Sharing and Opportunity Costs" Working paper.

Albrecht, James, Anders Björklund, and Susan Vroman (2003), "Is There a Glass Ceiling in Sweden?" Journal of Labor Economics 21: 145-178.

Altonji, Joseph G. and Rebecca Blank (1999), "Race and Gender in the Labor Market," in *Handbook of Labor Economics,* Vol. 3C, ed. by O. Ashenfelter, and D. Card. Elsevier Science, Amsterdam.

Altonji, Joseph G. and Rosa L. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Econometrica, 73: 1053-1102.

Altonji, Joseph G., P. Bharadwaj, and Fabian Lange (2008), "Changes in the characteristics of American youth: Implications for adult outcomes. Working paper, Yale University.

Angrist, Joshua and Guido W. Imbens (1994), "Identification and Estimation of Local Average Treatment Effects", Econometrica 62: 467–476.

Athey, Susan and Guido W. Imbens (2006), "Identification and Inference in Nonlinear Difference-In-DifferencesModels," Econometrica 74: 431-497.

Autor, David H., Frank Levy, and Richard Murnane (2003), "The Skill Content of Recent Technological Change: An Empirical Exploration," Quarterly Journal of Economics 118: 1279-1333.

Autor, David H., Lawrence B. Katz and Melissa S. Kearney (2005), "Rising Wage Inequality: The Role of Composition and Prices." NBER Working Paper No. 11628, September.

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

Bauer, Thomas K., Silja Göhlmann, and Mathias Sinning (2007), "Gender differences in smoking behavior", Health Economics 19: 895–909.

Bauer, Thomas K. and Mathias Sinning (2008), "An Extension of the Blinder–Oaxaca Decomposition to Nonlinear Models", Advances in Statistical Analysis 92: 197–206.

Bertrand, Marianne, and Kevin F. Hallock (2001), "The Gender Gap in Top Corporate Jobs," Industrial and Labor Relations Review 55: 3-21.

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

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes (2006), "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments" American Economic Review 96: 988-1012.

Black, Dan, Amelia Haviland, Seth Sanders, and Lowell Taylor (2008), "Gender Wage Disparities among the Highly Educated", Journal of Human Resources 43: 630-659.

Blau, Francine D. and Lawrence M. Kahn (1992), "The Gender Earnings Gap: Learning from International Comparisons", American Economic Review 82: 533–538.

Blau, Francine D. and Lawrence M. Kahn (1997), "Swimming Upstream: Trends in the Gender Wage Differential in the 1980s," Journal of Labor Economics 15: 1-42.

Blau, Francine D. and Lawrence M. Kahn (2003), "Understanding International Differences in the Gender Pay Gap," Journal of Labor Economics 21: 106–144.

Blau, Francine D. and Lawrence M. Kahn (2006), "The US gender pay gap in the 1990s: Slowing convergence," Industrial & Labor Relations Review, 60(1): 45-66.

Blinder, Alan (1973), "Wage Discrimination: Reduced Form and Structural Estimates," Journal of Human Resources 8:436–455.

Blundell, Richard, and James L. Powell (2007), "Censored regression quantiles with endogenous regressors," Journal of Econometrics 141: 65-83.

Blundell, Richard, Martin Browning , and Ian Crawford (2010), "Best Nonparametric Bounds on Demand Responses," Econometrica 76: 1227–1262.

Bourguignon, Francois (1979), "Decomposable Income Inequality Measures", Econometrica 47: 901–920

Bourguignon, F. and Francisco H.G. Ferreira (2005), "Decomposing Changes in the Distribution of Household Incomes: Methodological Aspects," in *The Microeconomics of Income Distribution Dynamics in East Asia and Latin America,"* F. Bourguignon, F.H.G. Ferreira and N. Lustig (eds.) World Bank. pp. 17-46.

Bourguignon,F. , Francisco H.G. Ferreira, and Philippe G. Leite (2008), "Beyond Oaxaca–Blinder: Accounting for differences in household income distributions," Journal of Economic Inequality 6: 117-148.

Busso, Matias, John DiNardo, and Justin McCrary (2009), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators" IZA Discussion Paper No. 3998.

Butcher, Kristin F. and John DiNardo (2002), "The Immigrant and native-born wage distributions: Evidence from United States censuses," Industrial and Labor Relations Review 56: 97-121.

Cain, Glen (1986), "The Economic Analysis of Labor Market Discrimination: a Survey,' in Ashenfelter, O.C. and R. Layard, editors, Handbook of Labor Economics, North-Holland, vol 1, pp.709-730

Card, David (1992), "The Effects of Unions on the Distribution of Wages: Redistribution or Relabelling?" NBER Working Paper 4195, Cambridge: Mass.: National Bureau of Economic Research, 1992.

Card, David (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis" , Econometrica 64: 957-979.

Card, David, and Dean R. Hyslop (2005), "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers" , Econometrica 73:1723–1770.

Chay, Kenneth Y., and David S. Lee (2000), "Changes in Relative Wages in the 1980s: Returns to Observed and Unobserved Skills and Black-White Wage Differentials," Journal of Econometrics, 99(1), 1-38.

Chernozhukov, Victor, Ivan Fernandez-Val, and Blaise Melly (2009), "Inference on Counterfactual Distributions," forthcoming in Econometrica.

Chernozhukov, Victor, Ivan Fernandez-Val, and A. Galichon (2010), "Quantile and Probability Curves without Crossing," forthcoming in Econometrica.

Chiquiar, Daniel and Gordon H. Hanson (2005), "International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States", Journal of Political Economy 113: 239-281.

Cowell, Frank A. (1980), "On the Structure of Additive Inequality Measures", Review of Economic Studies 47: 521–531.

Denison, E.F. (1962), The sources of economic growth in the United States and the alternatives before us, Supplementary Paper No. 13, New York, Committee for Economic Development.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," Econometrica 64: 1001-1044.

DiNardo, John, and David S. Lee (2004), "Economic Impacts of New Unionization On Private Sector Employers: 1984-2001," The Quarterly Journal of Economics 119: 1383-1441.

DiNardo, John and Thomas Lemieux (1997), "Diverging Male Inequality in the United States and Canada, 1981-1988: Do Institutions Explain the Difference," Industrial and Labor Relations Review 50: 629-651.

Dolton, Peter John, Gerald Henry Makepeace and Wilbert Van Der Klaauw (1986), "Sample Selection and Male-Female Earnings Differentials in the Graduate Labour Market". Oxford Economic Papers 38: 317–341.

Doiron, Denise J. and W. Craig Riddell (1994) "The Impact of Unionization on Male-Female Earnings Differences in Canada," Journal of Human Resources 29: 504-534.

Donald, Stephen G., David A. Green, and Harry J. Paarsch (2000), "Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates Source" Review of Economic Studies 67: 609-633.

Duncan, Gregory M. and Duane E. Leigh (1980), "Wage Determination in the Union and Nonunion Sectors: A Sample Selectivity Approach,' Industrial and Labor Relations Review 34: 24-34

Egel, Daniel, Bryan Graham, and Cristine Pinto (2009), "Efficicient Estimation of Data Combination Problems by the Method of Auxiliary-to-Study Tilting." Mimeo.

Even, William E. and David A. Macpherson (1990), Plant size and the decline of unionism", Economics Letters 32: 393–398

Fairlie, Robert W. (1999), The absence of the african-american owned business: an analysis of the dynamics of self–employment", Journal of Labor Economics 17: 80–108.

Fairlie, Robert W. (2005), "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models", Journal of Economic and Social Measurement 30: 305–316.

Fields, Judith, and Edward N.Wolff (1995), "Interindustry Wage Differentials and the Gender Wage Gap." Industrial and Labor Relations Review 49: 105-120.

Firpo, Sergio (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects," Econometrica, 75: 259 - 276.

Firpo, Sergio (2010), "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures" EESP-FGV, mimeo.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2007), "Decomposing Wage Distributions using Recentered Influence Functions Regressions", mimeo, University of British Columbia.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009) "Unconditional Quantile Regressions," Econometrica 77(3): 953-973.

Fitzenberger, Bernd, Karsten Kohn and Qingwei Wang (2009), "The Erosion of Union Membership in Germany: Determinants, Densities, Decompositions", Journal of Population Economics, forthcoming.

Foresi, Silverio and Franco Peracchi (1995), "The Conditional Distribution of Excess Returns: an Empirical Analysis", Journal of the American Statistical Association 90: 451-466.

Fortin, Nicole M., and Thomas Lemieux (1998), "Rank Regressions, Wage Distributions, and the Gender Gap", Journal of Human Resources 33: 610–643.

Fortin, Nicole M. (2008), "The Gender Wage Gap among Young Adults in the United States: The Importance of Money vs. People," Journal of Human Resources, 43: 886-920.

Freeman, Richard B. (1980), "Unionism and the Dispersion of Wages,' Industrial and Labor Relations Review 34: 3-23.

Freeman, Richard B. (1984), "Longitudinal Analysis of the Effect of Trade Unions," Journal of Labor Economics 2: 1-26.

Freeman, Richard B. (1993), "How Much has Deunionization Contributed to the Rise of Male Earnings Inequality?' In Sheldon Danziger and Peter Gottschalk, eds. Uneven Tides: Rising Income Inequality in America. New York: Russell Sage Foundation, 133-63.

Frolich, Markus (2004), "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," Review of Economics and Statistics 86: 77-90.

Gardeazabal, Javier and Arantza Ugidos (2004), "More on the Identification in Detailed Wage Decompositions," Review of Economics and Statistics 86: 1034–57.

Gelbach, Jonah B. (2002), "Identified Heterogeneity in Detailed Wage Decompositions," mimeo, University of Maryland at College Park.

Gelbach, Jonah B. (2009), "When Do Covariates Matter? And Which Ones, and How Much?" mimeo, Eller College of Management, University of Arizona.

Gomulka, Joanna and Nicholas Stern (1990), "The employment of married women in the United Kingdom, 1970–1983", Economica 57: 171–199.

Gosling, Amanda, Stephen Machin, and Costas Meghir (2000), "The Changing Distribution of Male Wages in the U.K," Review of Economic Studies 67: 635-666.

Greene, William H. (2003), Econometric Analysis. 5th ed. Upper Saddle River, NJ: Pearson Education

Heckman, James (1974), "Shadow Prices, Market Wages and Labor Supply," Econometrica 42: 679–694.

Heckman, James (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,' Annals of Economic and Social Measurement 5: 475-492.

Heckman, James (1979), "Sample Selection Bias as a Specification Error,' Econometrica 47: 153-163.

Heckman, James J., Jeffrey Smith and Nancy Clements (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", Review of Economic Studies, 64(4): 487–535.

Heckman, James J., Hidehiko Ichimura, and Petra Todd (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," Review of Economic Studies, 64: 605-654.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998), "Characterizing Selection Bias Using Experimental Data," Econometrica 66: 1017-1098.

Heywood, John S. and Daniel Parent (2008), "The White-Black Wage Gap and the Method of Pay," mimeo, McGill University.

Hirano, Kiesuke, Guido W. Imbens, and Geert Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, 1161–1189.

Holland, Paul W. (1986), "Statistics and Causal Inference," Journal of the American Statistical Association 81(396): 945-960.

Hoffman, Florian (2009), "An Empirical Model of Life-Cycle Earnings and Mobility Dynamics," University of Toronto, Department of Economics, mimeo.

Horrace, William and Ronald L. Oaxaca (2001), "Inter-Industry Wage Differentials and the Gender Wage Gap: An Identification Problem," Industrial and Labor Relations Review, 54: 611-618.

Imbens, Guido W. and Whitney K. Newey (2009) "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," Econometrica 77(5): 1481 - 1512.

Jann, Ben (2005), "Standard Errors for the Blinder-Oaxaca Decomposition", German Stata Users' Group Meetings 2005. Available from http://repec.org/dsug2005/oaxaca se handout.pdf.

Jann, Ben (2008), "The Oaxaca-Blinder Decomposition for Linear Regression Models," Stata Journal 8: 435–479.

Jones, Frank Lancaster (1983), "On Decomposing the Wage Gap: A Critical Comment on Blinder's Method", Journal of Human Resources 18: 126–130.

Johnson, William, Yuichi Kitamura, and Derek Neal (2000), "Evaluating a Simple Method for Estimating Black-White Gaps in Median Wages," American Economic Review 90:339–343.

Jorgenson, D. W. and Z. Griliches (1967), "The Explanation of Productivity Change" Review of Economic Studies 34: 249-283

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce (1991), "Accounting for the Slowdown in Black-White Wage Convergence," in Workers and Their Wages: Changing Patterns in the United States, ed. by M. H. Kosters. American Enterprise Institute, Washington.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce (1993), "Wage Inequality and the Rise in Returns to Skill," Journal of Political Economy 101: 410-442.

Keane, Michael P. and Kenneth I. Wolpin (1997), "The Career Decisions of Young Men," Journal of Political Economy, 105: 473-522.

Keane, Michael P., Kenneth I. Wolpin, and Petra Todd (2010), "Dynamic Structure Modeling," in *Handbook of Labor Economics,* Vol. 4, ed. by O. Ashenfelter, and D. Card. Elsevier Science, Amsterdam.

Kendrick, John W. (1961), Productivity Trends in the United States, Princeton, Princeton University Press.

Kennedy, Peter (1986), "Interpreting Dummy Variables," Review of Economics and Statistics 68): 174-175.

Kline, Pat (2009), "Blinder-Oaxaca as a Reweighting Estimator" UC Berkeley mimeo.

Koenker, Roger and G. Bassett (1978), "Regression Quantiles," Econometrica, 46, 33-50.

Krueger, Alan B. and Lawrence H. Summers (1988), "Efficiency Wages and the Inter-Industry Wage Structure," Econometrica 56(2): 259-293.

Krieg, John M. and Paul Storer (2006), "How Much Do Students Matter? Applying the Oaxaca Decomposition to Explain Determinants of Adequate Yearly Progress," Contemporary Economic Policy, 24: 563–581.

Leibbrandt, Murray, James A. Levinsohn, and Justin McCrary (2005), "Incomes in South Africa Since the Fall of Apartheid" NBER Working Paper 11384

Lemieux, Thomas (2002), "Decomposing Changes in Wage Distributions: a Unified Approach," The Canadian Journal of Economics 35: 646-688.

Lemieux, Thomas (2006a), "Post-secondary Education and Increasing Wage Inequality", American Economic Review 96: 195-199.

Lemieux, Thomas (2006b),"Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?", American Economic Review 96: 461-498.

Lewis, H. Gregg (1963), Unionism and Relative Wages in the United States, Chicago: University of Chicago Press.

Lewis, H. Gregg (1986), Union Relative Wage Effects: A Survey, Chicago: University of Chicago Press.

Neumark, David (1988), "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," Journal of Human Resources, 23: 279-295.

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

Machado, Cecilia (2009), "Selection, Heterogeneity and the Gender Wage Gap", Columbia University, Economics Department, mimeo.

Manski, Charles F. (1990), "Nonparametric Bounds on Treatment Effects," American Economic Review, 80(2): 319-323.

Matzkin, Rosa L. (2003), "Nonparametric estimation of nonadditive random functions," Econometrica 71(5):1339-1375.

McEwan, P.J. and Marshall, J.H. (2004) Why does academic achievement vary across countries? Evidence from Cuba and Mexico," Education Economics 12: 205-217.

Melly, Blaise (2006), "Estimation of counterfactual distributions using quantile regression," University of St. Gallen, Discussion Paper.

Melly, Blaise (2005), "Decomposition of differences in distribution using quantile regression," Labour economics 12: 577-590.

Mulligan, Casey B. and Yona Rubinstein (2008), "Selection, Investment, and Women's Relative Wages Over Time," Quarterly Journal of Economics, 123: 1061-1110.

Neal, Derek A. and W. Johnson (1996), "The Role of Premarket Factors in Black-White Wage Differences," Journal of Political Economy 104: 869-895.

Neal, Derek A. (2004), "The Measured Black-White Wage Gap Among Women Is Too Small," Journal of Political Economy, 112: S1-S28.

Ñopo, Hugo (2008) "Matching as a Tool to Decompose Wage Gaps," Review of Economics and Statistics 90: 290-299.

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

Oaxaca, Ronald L. and Michael R. Ransom(1994), "On discrimination and the decomposition of wage differentials". Journal of Econometrics 61: 5–21.

Oaxaca, Ronald L. and Michael R. Ransom (1998), "Calculation of Approximate Variances for Wage Decomposition Differentials," Journal of Economic and Social Measurement 24: 55–61.

Oaxaca, Ronald L. and Michael R. Ransom (1999), "Identification in Detailed Wage Decompositions," Review of Economics and Statistics 81: 154–157.

Oaxaca, Ronald L. (2007), "The challenge of measuring labor market discrimination against women," Swedish Economic Policy Review, 14: 199–231.

Olivetti, Claudia and Barbara Petrongolo (2008), "Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps," Journal of Labor Economics, 26: 621-654.

O'Neill, June and Dave O'Neill (2006), "What Do Wage Differentials Tell Us about Labor Market Discrimination?' in The Economics of Immigration and Social Policy, edited by Soloman Polachek, Carmel Chiswich, and Hillel Rapoport. Research in Labor Economics 24:293-357.

Reimers, Cornelia W. (1983), "Labor Market Discrimination Against Hispanic and Black Men", Review of Economics and Statistics, 65: 570–579.

Robins, James, Andrea Rotnizky, and Lue Ping Zhao (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." Journal of the American Statistical Association 89: 846-866.

Rosenbaum Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70: 41-55.

Rosenbaum Paul R. and Donald B. Rubin (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," Journal of the American Statistical Association 79: 516-524.

Rothe, Christoph (2009), "Nonparametric Estimation of Distributional Policy Effects, mimeo, University of Mannheim.

Shorrocks, Anthony F. (1980), "The Class of Additively Decomposable Inequality Measures", Econometrica 48: 613–625

Shorrocks, Anthony F. (1984), "Inequality Decomposition by Population Subgroups", Econometrica, 52: 1369–1385.

Shorrocks, Anthony F. (1999), "Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value", University of Essex, Department of Economics, mimeo.

Solow, Robert (1957), "Technical Change and the Aggreagate Production Function" Review of Economics and Statistics, 39: 312-320

Sohn, Ritae (2008), 'The Gender Math Gap: Is It Growing?," mimeo, SUNY Albany.

Vella, Frank (1998), "Estimating Models with Sample Selection Bias: A Survey," Journal of Human Resources 33: 127-169.

Woodcock, Simon D. (2008), "Wage Differentials in the Presence of Unobserved Worker, Firm, and Match Heterogeneity?" Labour Economics 15: 772–794.

Yun, Myeong-Su (2005), "A Simple Solution to the Identification Problem in Detailed Wage Decomposition", Economic Inquiry, 43: 766-772. with "Erratum," Economic Inquiry (2006), 44: 198.

Yun, Myeong-Su (2008), "Identification Problem and Detailed Oaxaca Decomposition: A General Solution and Inference," Journal of Economic and Social Measurement 33: 27-38.

Figure 1: Relationship Between Proportions and Quantiles

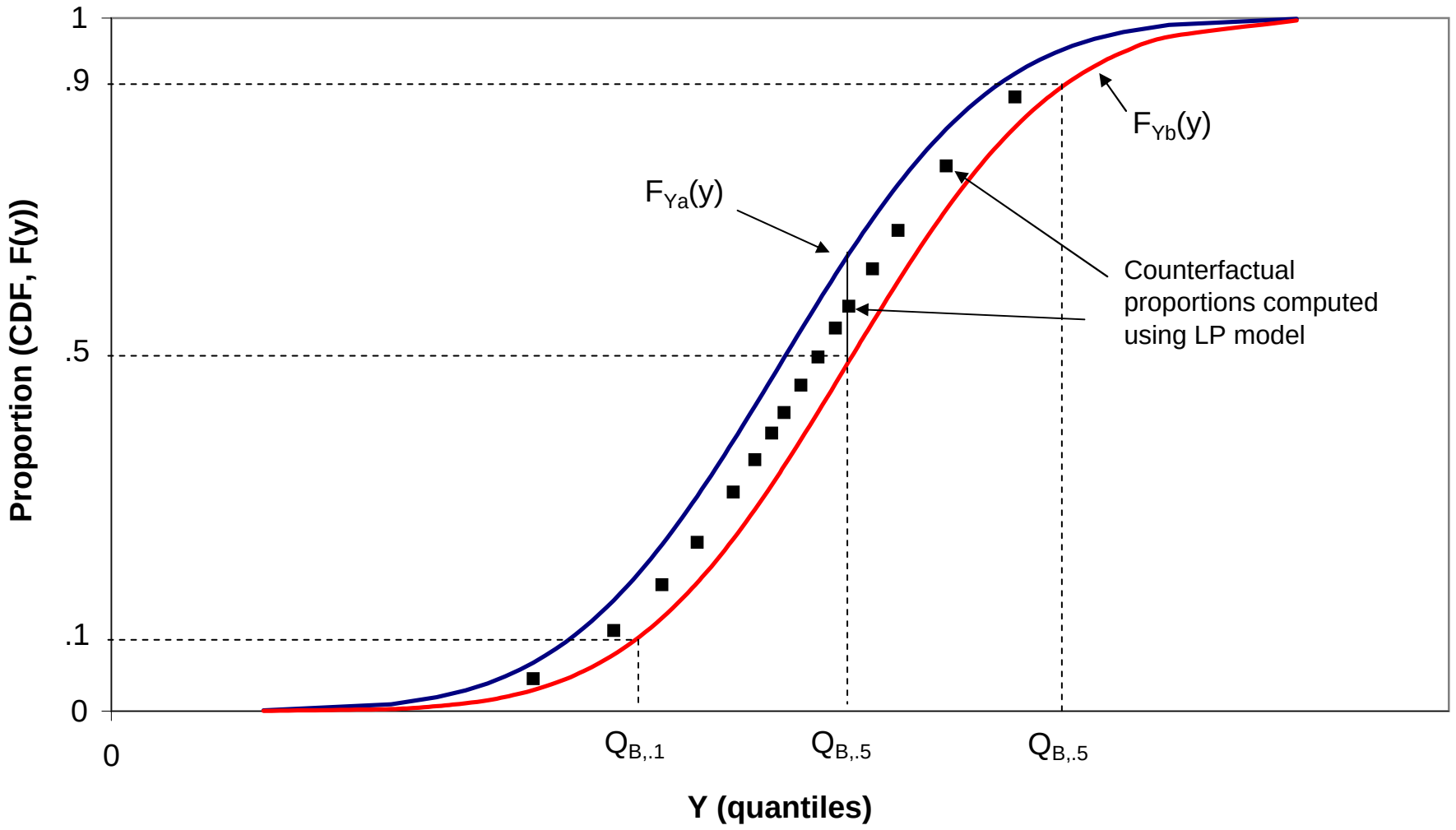

Figure 2: Inverting Globally

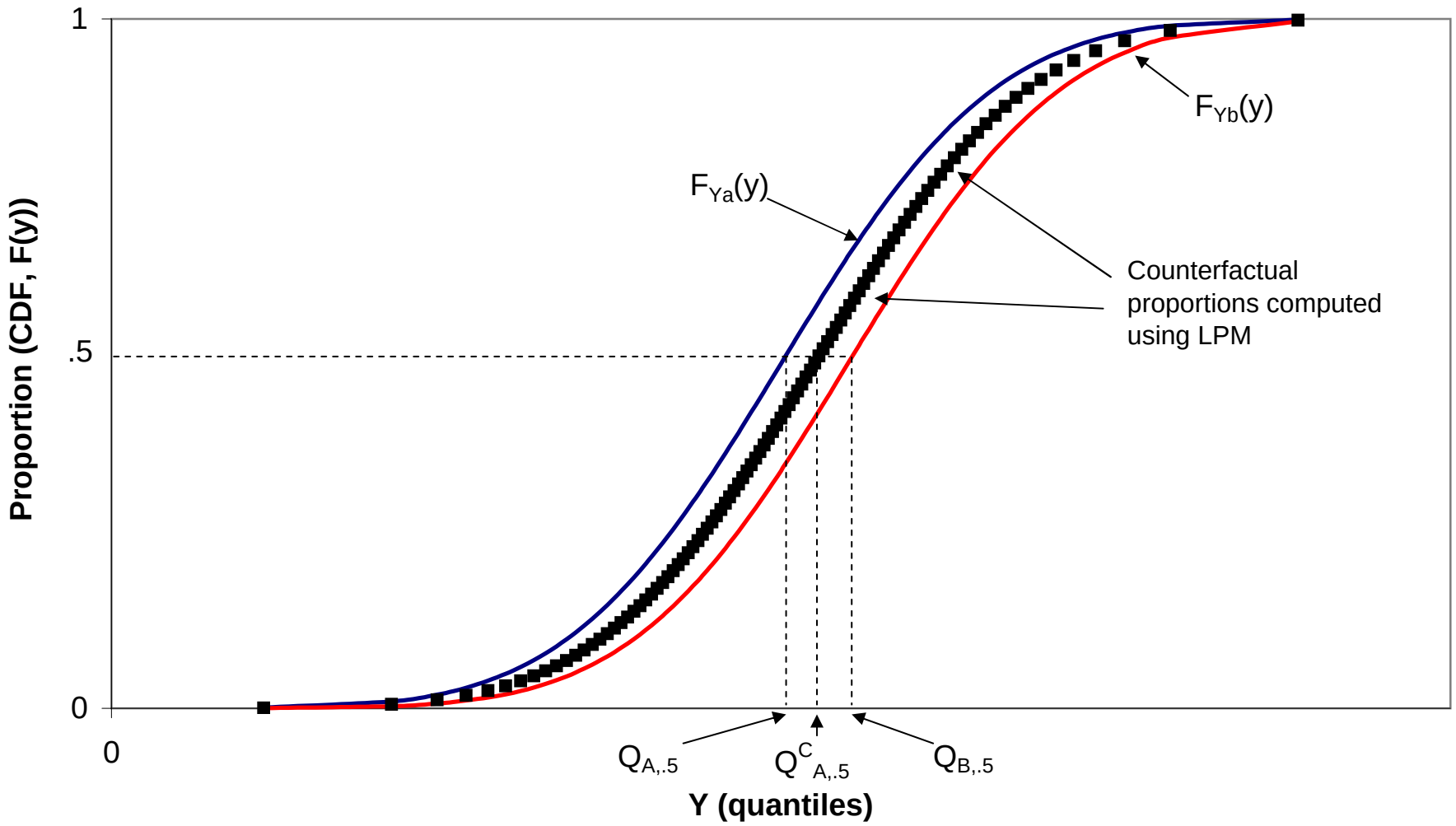

Figure 3: RIF Regressions: Inverting Locally

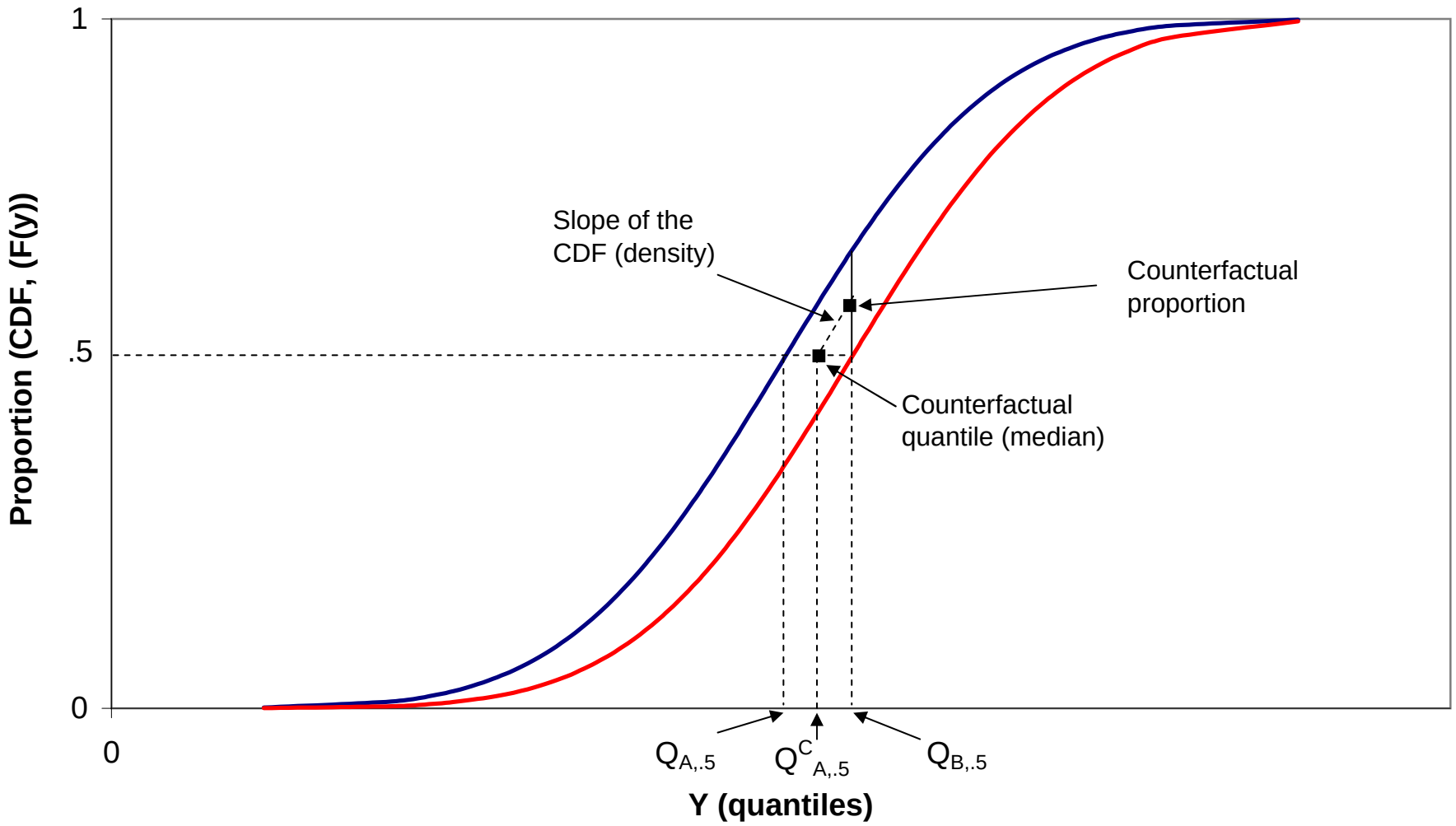

Table 1. Maintained Assumptions and Properties of Major Decomposition Methodologies

Methods	Assumptions[a]	Properties: Limitations and Advantages	Representative Applications
Mean Decomposition:			
3.1. Standard OB: Oaxaca (1973)-Blinder (1973)	Linearity of E[Y|X] Zero Conditional Mean	Path Independent Detailed Decomposition	Gender and Racial Wage Gaps: O'Neill and O'Neill (2006)
3.3. Weighted or Pooled OB: Oaxaca and Ransom (1994), Cotton (1988)	Complex Counterfactual Linearity of E[Y|X] Zero Conditional Mean	Path Independent Detailed Decomposition	Racial/Ethnic Wage Gaps: Reimers (1983)
3.5 Non-linear OB: Fairlie (2005), Bauer and Sinning (2008)	Non-Linearity of E[Y|X]	Path Dependent Detailed Decomposition	Racial Gap in Self-Employment: Fairlie (1999)
Going Beyond the Mean:			
4.1. Variance Decompositions	Linearity of V(Y|X) Invariance of Conditional Variance	No Detailed Decomposition	Union Wage Differentials: Freeman (1980, 1984)
4.3. Residual Imputation Procedure: Juhn, Murphy, and Pierce (1991, 1993)	Linearity of E[Y|X] Conditional Rank Preservation Complex Counterfactual [b]	No Detailed Decomposition	Gender Gap Across Countries: Blau and Kahn (1992)
4.4. Quantile Regressions Methods: Machado and Mata (2005), Chernozhukov, Fernandez-Val, and Melly (2009)	Linearity of $Q_\tau(Y\|X)$ Conditional Rank Preservation	No Detailed Decomposition	Gender Glass Ceiling: Albrecht, Björklund, and Vroman (2003)
4.5 Inverse Propensity Reweighing: DiNardo, Fortin, and Lemieux (1996)	Invariance of Conditional Distribution	Path Dependent	Immigrant/Resident Wage Differentials: Chiquiar and Hanson (2005)
4.6 Estimation of Conditional Distribution: Chernozhukov, Fernandez-Val, and Melly (2009)	Invariance of Conditional Distribution Conditional Rank Preservation	Path Dependent	Racial Wage Gap: Melly (2006)
5.2 RIF Regressions: Firpo, Fortin, and Lemieux (2007, 2009)	Invariance of Conditional Distribution	Path Independent Detailed Decomposition	Racial Wage Gap: Heywood and Parent (2009)

Note: [a]Unless otherwise indicated, the different methodologies appeal to a simple counterfactual treatment.

[b] In some applications, the counterfactual is an average over time periods or over countries.

Table 2. Means and OLS Regression Coefficients of Selected Variables from NLSY Log Wage Regressions for Workers Ages 35-43 in 2000

	(1)		(2)	(3)	(4)	(5)
Explanatory Variables	Means		Male Coef.	Female Coef.	Male Coef	Pooled Coef
Female	0	1				-0.092 (0.014)
Education and skill level						
<10 yrs.	0.053	0.032	-0.027 (0.043)	-0.089 (0.05)	-0.027 (0.043)	-0.045 (0.033)
10-12 yrs (no diploma or GED)	0.124	0.104	--- ---	--- ---	--- ---	--- ---
HS grad (diploma)	0.326	0.298	-0.013 (0.028)	-0.002 (0.029)	-0.013 (0.028)	-0.003 (0.02)
HS grad (GED)	0.056	0.045	0.032 (0.042)	-0.012 (0.044)	0.032 (0.042)	0.006 (0.03)
Some college	0.231	0.307	0.164 (0.031)	0.101 (0.03)	0.164 (0.031)	0.131 (0.022)
BA or equiv. degree	0.155	0.153	0.380 (0.037)	0.282 (0.036)	0.380 (0.037)	0.330 (0.026)
MA or equiv. degree	0.041	0.054	0.575 (0.052)	0.399 (0.046)	0.575 (0.052)	0.468 (0.034)
Ph.D or prof. Degree	0.015	0.007	0.862 (0.077)	0.763 (0.1)	0.862 (0.077)	0.807 (0.06)
AFQT percentile score (x.10)	4.231	3.971	0.042 (0.004)	0.041 (0.004)	0.042 (0.004)	0.042 (0.003)
L.F. withdrawal due to family resp.	0.129	0.547	-0.078 (0.025)	-0.083 (0.019)	-0.078 (0.025)	-0.067 (0.015)
Lifetime Work Experience						
Years worked civilian	17.160	15.559	0.038 (0.003)	0.030 (0.002)	0.038 (0.003)	0.033 (0.002)
Years worked military	0.578	0.060	0.024 (0.005)	0.042 (0.013)	0.024 (0.005)	0.021 (0.004)
% worked part-time	0.049	0.135	-0.749 (0.099)	-0.197 (0.049)	-0.749 (0.099)	-0.346 (0.044)
Industrial Sectors						
Primary, Constr. & Utilities	0.186	0.087	--- ---	--- ---	0.059 (0.031)	--- ---
Manufacturing	0.237	0.120	0.034 (0.026)	0.140 (0.035)	0.093 (0.029)	0.072 (0.021)
Education, Health, & Public Adm.	0.130	0.358	-0.059 (0.031)	0.065 (0.03)	--- ---	-0.001 (0.02)
Other Services	0.447	0.436	0.007 (0.024)	0.088 (0.029)	0.066 (0.026)	0.036 (0.018)
Constant			2.993 (0.156)	2.865 (0.144)	2.934 (0.157)	2.949 (0.105)
Dependent Var. (Log Hourly Wage)	2.763	2.529				
Adj. R-Square			0.422	0.407	0.422	0.431
Sample size	2655	2654				

Note: The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). Industrial sectors were added (at a lost of 89 observations) to their analysis to illustrate issues linked to categorical variables. The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. Standard errors are in parentheses.

Table 3. Gender Wage Gap: Oaxaca-Blinder Decomposition Results (NLSY, 2000)

Reference Group:	(1) Using Male Coef. from col. 2, Table 2		(2) Using Male Coef. from col. 4, Table 2		(3) Using Female Coef.		(4) Using Weighted Sum		(5) Using Pooled from col. 5, Table 2	
Unadjusted mean log wage gap :										
$E[\ln(w_m)]-E[\ln(w_f)]$	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)	0.233	(0.015)
Composition effects attributable to										
Age, race, region, etc.	0.012	(0.003)	0.012	(0.003)	0.009	(0.003)	0.011	(0.003)	0.010	(0.003)
Education	-0.012	(0.006)	-0.012	(0.006)	-0.008	(0.004)	-0.010	(0.005)	-0.010	(0.005)
AFQT	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)	0.011	(0.003)
L.T. withdrawal due to family	0.033	(0.011)	0.033	(0.011)	0.035	(0.008)	0.034	(0.007)	0.028	(0.007)
Life-time work experience	0.137	(0.011)	0.137	(0.011)	0.087	(0.01)	0.112	(0.008)	0.092	(0.007)
Industrial sectors	0.017	(0.006)	0.017	(0.006)	0.003	(0.005)	0.010	(0.004)	0.009	(0.004)
Total explained by model	0.197	(0.018)	0.197	(0.018)	0.136	(0.014)	0.167	(0.013)	0.142	(0.012)
Wage structure effects attributable to										
Age, race, region, etc.	-0.098	(0.234)	-0.098	(0.234)	-0.096	(0.232)	-0.097	(0.233)	-0.097	(0.24)
Education	0.045	(0.034)	0.045	(0.034)	0.041	(0.033)	0.043	(0.034)	0.043	(0.031)
AFQT	0.003	(0.023)	0.003	(0.023)	0.003	(0.025)	0.003	(0.024)	0.002	(0.025)
L.T. withdrawal due to family	0.003	(0.017)	0.003	(0.017)	0.001	(0.004)	0.002	(0.011)	0.007	(0.01)
Life-time work experience	0.048	(0.062)	0.048	(0.062)	0.098	(0.067)	0.073	(0.064)	0.092	(0.065)
Industrial sectors	-0.092	(0.033)	0.014	(0.028)	-0.077	(0.029)	-0.085	(0.031)	-0.084	(0.032)
Constant	0.128	(0.213)	0.022	(0.212)	0.193	(0.211)	0.128	(0.213)	0.128	(0.216)
Total wage structure - Unexplained log wage gap	0.036	(0.019)	0.036	(0.019)	0.097	(0.016)	0.066	(0.015)	0.092	(0.014)

Note: The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. In column (1), the omitted industrial sector is "Primary, Construction, and Utilities". In column (2), the omitted industrial sector is "Education, Health and Public Admin". Standard errors are in parentheses. The means of the variables are reported in Table 2.

Table 4. Gender Wage Gap: Quantile Decomposition Results (NLSY, 2000)

Reference Group: Male Coef.	10th percentile		50th percentile		90th percentile	
A: Raw log wage gap :						
$Q_\tau[\ln(w_m)]-Q\tau[\ln(w_f)]$	0.170	(0.023)	0.249	(0.019)	0.258	(0.026)
B: Decomposition Method: Machado-Mata-Melly						
Estimated log wage gap:						
$Q\tau[\ln(wm)]-Q\tau[\ln(w_f)]$	0.192	(0.015)	0.239	(0.016)	0.276	(0.026)
Total explained by characteristics	0.257	(0.028)	0.198	(0.027)	0.143	(0.019)
Total wage structure	-0.065	(0.027)	0.041	(0.024)	0.133	(0.025)
C: Decomposition Method: RIF regressions without reweighing						
Mean RIF gap:						
$E[RIF_\tau(\ln(w_m))]-E[RIF_\tau(\ln(w_f))]$	0.180	(0.023)	0.241	(0.019)	0.260	(0.026)
Composition effects attributable to						
Age, race, region, etc.	0.015	(0.005)	0.013	(0.004)	0.002	(0.004)
Education	-0.011	(0.005)	-0.017	(0.006)	-0.005	(0.01)
AFQT	0.005	(0.02)	0.013	(0.004)	0.013	(0.005)
L.T. withdrawal due to family	0.022	(0.021)	0.042	(0.014)	0.039	(0.017)
Life-time work experience	0.234	(0.026)	0.136	(0.014)	0.039	(0.023)
Industrial Sectors	0.008	(0.012)	0.020	(0.008)	0.047	(0.011)
Total explained by characteristics	0.274	(0.035)	0.208	(0.025)	0.136	(0.028)
Wage structure effects attributable to						
Age, race, region, etc.	-0.342	(0.426)	0.168	(0.357)	0.860	(0.524)
Education	0.023	(0.028)	-0.030	(0.031)	0.023	(0.045)
AFQT	-0.007	(0.03)	0.003	(0.042)	0.008	(0.062)
L.T. withdrawal due to family	-0.075	(0.032)	-0.005	(0.025)	0.018	(0.032)
Life-time work experience	0.084	(0.148)	-0.085	(0.082)	-0.078	(0.119)
Industrial Sectors	0.015	(0.06)	-0.172	(0.046)	-0.054	(0.052)
Constant	0.208	(0.349)	0.154	(0.323)	-0.653	(0.493)
Total wage structure	-0.094	(0.044)	0.033	(0.028)	0.124	(0.036)

Note: The data is an extract from the NLSY79 used in O'Neill and O'Neill (2006). Industrial sectors have been added to their analysis to illustrate issues linked to categorical variables. The other explanatory variables are age, dummies for black, hispanic, region, msa, central city. Bootstrapped standard errors are in parentheses. Means are reported in Table 2.

Table 5. Male Wage Inequality: Aggregate Decomposition Results (CPS, 1983/85-2003/05)

Inequality measure	90-10		90-50		50-10		Variance		Gini	
A.Decomposition Method: DFL - F(X) in 1983/85 reweighted to 2003/05										
Unadjusted change (t_1-t_0):	0.1091	(0.0046)	0.1827	(0.0037)	-0.0736	(0.0033)	0.0617	(0.0015)	0.0112	(0.0004)
Total composition effect	0.0756	(0.0031)	0.0191	(0.0034)	0.0565	(0.0029)	0.0208	(0.0007)	-0.0020	(0.0004)
Total wage effect	0.0336	(0.0048)	0.1637	(0.0043)	-0.1301	(0.004)	0.0408	(0.0017)	0.0132	(0.0003)
B. Decomposition Method: CFVM - LPM - $F^C(Y\|X) = \Lambda$(2003/05 X's with 1983/85 α's)										
Estimated change (t_1-t_0):	0.1100	(0.0055)	0.1921	(0.0057)	-0.0821	(0.0044)	0.0636	(0.0013)	0.0118	(0.0005)
Total composition effect	0.0289	(0.0045)	0.0027	(0.0034)	0.0261	(0.0040)	0.0109	(0.0007)	-0.0046	(0.0003)
Total wage effect	0.0811	(0.0071)	0.1894	(0.0066)	-0.1082	(0.006)	0.0527	(0.0016)	0.0164	(0.0006)
C. Decomposition Method: CFVM - Logit - $F^C(Y\|X) = \Lambda$(2003/05 X's with 1983/85 α's)										
Estimated change (t_1-t_0):	0.1100	(0.0040)	0.1921	(0.0032)	-0.0821	(0.0030)	0.0636	(0.0013)	0.0118	(0.0005)
Total composition effect	0.0872	(0.0044)	0.0392	(0.0040)	0.0480	(0.0018)	0.0212	(0.0008)	-0.0019	(0.0003)
Total wage effect	0.0227	(0.0053)	0.1529	(0.0049)	-0.1301	(0.0030)	0.0424	(0.0016)	0.0137	(0.0005)
D. Decomposition Method: FFL- RIF-OLS - No reweighing										
Estimated change (t_1-t_0):	0.1100	(0.0039)	0.1824	(0.0036)	-0.0724	(0.0031)	0.0617	(0.0013)	0.0112	(0.0004)
Total composition effect	0.0617	(0.0018)	0.0294	(0.0019)	0.0323	(0.0014)	0.0151	(0.0005)	-0.0038	(0.0003)
Total wage effect	0.0483	(0.0043)	0.1530	(0.0043)	-0.1047	(0.0033)	0.0466	(0.0013)	0.0150	(0.0004)

Note: The data is an extract from the Morg CPS 1983/85 (232 784 obs.) and 2003/05 (170 693 obs.) used in Firpo, Fortin and Lemieux (2007). The explanatory variables include union status, 6 education classes (high school omitted), 9 potential experience classes (20-25 years omitted). In Panel B and C, computations were performed using Melly's "counterfactual" procedure. The variance and gini were computed using 100 quantile estimates. In Panel D, the estimated change is computed as $E[RIF\nu(\ln(w_1))]-E[RIF\nu(\ln(w_0))]$. Bootstrapped standard errors (100 reps.) are in parentheses.

Table 6. Male Wage Inequality: FFL Decomposition Results (CPS, 1983/85-2003/05)

Reweighing	No reweighing 1983/85 reference		F(X) in 1983/85 reweighted to 2003/05		No reweighing 1983/85 reference		F(X) in 1983/85 reweighted to 2003/05		No reweighing 1983/85 reference		F(X) in 1983/85 reweighted to 2003/05	
Inequality measure	90-10				90-50				50-10			
Unadjusted change	0.1100	(0.0039)	0.1100	(0.0039)	0.1824	(0.0036)	0.1824	(0.0036)	-0.0724	(0.0031)	-0.0724	(0.0031)
Composition effects attributable to												
Union	0.0353	(0.0008)	0.0356	(0.0008)	0.0568	(0.0009)	0.0572	(0.0009)	-0.0215	(0.0006)	-0.0216	(0.0006)
Education	0.0150	(0.0012)	0.0149	(0.0012)	-0.0155	(0.001)	-0.0156	(0.001)	0.0305	(0.001)	0.0305	(0.001)
Experience	0.0114	(0.0009)	0.0114	(0.0009)	0.0119	(0.0009)	0.0118	(0.0009)	0.0233	(0.0008)	0.0233	(0.0008)
Total explained	0.0617	(0.0018)	0.0619	(0.0018)	0.0294	(0.0019)	0.0298	(0.0019)	0.0323	(0.0014)	0.0322	(0.0013)
Wage structure effects attributable to												
Union	0.0019	(0.0016)	0.0084	(0.0016)	0.0016	(0.0018)	0.0141	(0.0018)	0.0035	(0.0014)	0.0225	(0.0014)
Education	0.1053	(0.0068)	0.1234	(0.0064)	0.0339	(0.007)	0.0754	(0.0067)	0.0714	(0.0053)	0.0480	(0.0059)
Experience	0.0115	(0.0127)	-0.0768	(0.0138)	-0.0120	(0.011)	-0.0451	(0.0116)	0.0235	(0.0081)	0.0318	(0.0092)
Constant	-0.0705	(0.0148)	-0.0211	(0.0158)	0.1326	(0.0129)	0.1477	(0.0134)	-0.2031	(0.0095)	-0.1688	(0.0113)
Total wage structure	0.0483	(0.0043)	0.0339	(0.0042)	0.1530	(0.0043)	0.1639	(0.0043)	-0.1047	(0.0033)	-0.1300	(0.0039)
Inequality measure	Variance				Gini							
Unadjusted change	0.0617	(0.0013)	0.0617	(0.0013)	0.0112	(0.0004)	0.0112	(0.0004)				
Composition effects attributable to												
Union	0.0129	(0.0002)	0.0130	(0.0002)	0.0069	(0.0001)	0.0069	(0.0001)				
Education	0.0013	(0.0003)	0.0013	(0.0003)	-0.0058	(0.0001)	-0.0058	(0.0001)				
Experience	0.0009	(0.0003)	0.0009	(0.0003)	-0.0049	(0.0001)	-0.0049	(0.0001)				
Total explained	0.0151	(0.0005)	0.0152	(0.0005)	-0.0038	(0.0003)	-0.0037	(0.0003)				
Wage structure effects attributable to												
Union	0.0002	(0.0005)	0.0023	(0.0005)	0.0020	(0.0001)	0.0011	(0.0001)				
Education	0.0483	(0.002)	0.0419	(0.002)	0.0070	(0.0007)	0.0064	(0.0007)				
Experience	0.0033	(0.0041)	-0.0177	(0.0041)	-0.0003	(0.0011)	-0.0064	(0.0012)				
Constant	-0.0052	(0.0048)	0.0145	(0.0048)	0.0063	(0.0014)	0.0129	(0.0014)				
Total wage structure	0.0466	(0.0013)	0.041	(0.0013)	0.0150	(0.0004)	0.0132	(0.0004)				

Note: The data is an extract from the Morg CPS 1983/85 (232 784 obs.) and 2003/05 (170 693 obs.) used in Firpo, Fortin and Lemieux (2007). The explanatory variables include union status, 6 education classes (high school omitted), 9 potential experience classes (20-25 years omitted). Bootstrapped standard errors (100 reps.) are in parentheses.