Estimation of a counterfactual wage distribution using survey data

Mihaela-Cătălina Anastasiade

University of Neuchâtel

PhD Thesis submitted to the Faculty of Science
Institute of Statistics
University of Neuchâtel
For the degree of PhD in Science.

Accepted by the dissertation committee:
Prof. Jacques Savoy, president of the jury, Université de Neuchâtel
Prof. Yves Tillé, thesis director, Université de Neuchâtel
Prof. Camelia Goga, Université de Bourgogne Franche-Comté
Prof. Maria Giovanna Ranalli, Université de Pérouse
Dr. Alina Matei, Université de Neuchâtel

Thesis defended on July 3rd, 2018

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
E-mail : secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Mihaela ANASTASIADE

Titre:

"Estimation of a counterfactual wage distribution using survey data"

sur le rapport des membres du jury composé comme suit:

- Prof. Yves Tillé, directeur de thèse, Université de Neuchâtel, Suisse
- Prof. Jacques Savoy, Université de Neuchâtel, Suisse
- Prof M. Giovanna Ranalli, Université de Pérouse, Italie
- Prof. Camelia Goga, Université de Bourgogne Franche-Comté, France
- Dr Alina Matei, Université de Neuchâtel, Suisse

Neuchâtel, le 12 juillet 2018

Le Doyen, Prof. R. Bshary

To the loving memory of my uncle, Jordan.

Acknowledgements

First of all, I want to express my gratitude to my thesis supervisor, Prof. Yves Tillé, for the huge opportunity that he gave me when he accepted me as a PhD student. I am honoured to have worked with you and I am grateful for the patience and the trust that you have given me. Thank you for your encouragements in times of doubt and for your constant availability to answer questions and to give good advice!

I want to thank Prof. Jacques Savoy for accepting to be the president of the jury committee, as well as all the members of the committee for taking the time to review this manuscript: Dr. Alina Matei, Prof. Camelia Goga and Prof. Maria Giovanna Ranalli.

This manuscript could not have been done without the contribution of the Swiss Federal Statistical Office, which provided me with data from a survey that they conduct. I hereby express my gratitude for their contribution.

I spent an amazing time at the Institute of Statistics, thanks to my colleagues. You filled these years with many good memories! Panda bears, summer schools, stories about beavers and trains and many other things that I will always remember with a lot of pleasure. I want to thank Monique Graf for sharing with me the secrets of the GB2 distribution and for her suggestions, she helped me to have a better view of my work. I would also like to thank Katia, Katharina and Rosa, three very talented persons, who helped me learn German and Spanish. I spent some a really nice time with them.

I want to say a special thank you to Dr. Alina Matei, who is the driving force behind it all. I would not have considered coming back to Switzerland, nor would I have made it to the end without your help and her encouragements. Thank you for your patience and your kindness and for believing more than me that this was possible.

There are not enough words to express my gratitude to my family for all that they have done for me. First and foremost, to my parents, who have always put me first, encouraged and supported me in every way. Their sacrifices made this achievement possible. To my wonderful sister, who is my role model and my definition of intelli-

gence and kindness. To my two grandmothers, for all the things they taught me and for making my childhood so beautiful. To my aunt, Michaela, who has has always been there for me with good advice and with time to listen. To my uncle, Jordan, to whom I dedicate this manuscript, a truly unique person with an infinite strengh of will. His gentleness and his generosity are forever engraved in my heart. You, my dear family, are a blessing in my life.

Family is not only about blood connections. Therefore, I would also like to thank Irina, for our neverending, amazing bond that never lost its power, in spite of the distance. I am really greatful for all that we have shared together! Thank you to Adina, a wonderful friend whom I admire for her courage, although I have never told her that. I am sure that we share a friendship for life and it is a true joy to have you. To Mami and Heinz, for taking care of me and for always being so nice and kind.

Life here in Switzerland would not have been the same without the great people I met here. First of all, I want to thank Alain's family for being so warm and welcoming with me. You gave me a family here and I am very grateful to you for this. Thank you, Vittoria and Isabella, for letting me share such beautiful moments with you and your wonderful children! Sybille, Sophie and Ruth, I am lucky to have you as my friends for so many years now and to know that I can always count on you!

I also have a special thought for the people who gave me the opportunity of a lifetime: Anne and Stéphane, thank you for a great internship at the Swiss Federal Statistical Office. Rhyem and Rick,you fulfilled a great dream of mine on November 23rd, 2017 and I do not have enough words to thank you. I also want to thank my former and my current colleagues from the Swiss Federal Statistical Office, for the pleasure of working with them.

One chapter ends and another one starts. I am very lucky to go through these chapters in life next to the most wonderful and incredible person in the world. Thank you, Alain, for making my life so unbelievably happy. You mean the world to me.

Les Charbonnières, July 16th, 2018.

Abstract

Wage discrimination based on gender is a debated topic nowadays. While wage discrimination is prohibited by the Universal Declaration of Human Rights, studies show that women are systematically paid less than men. Wage discrimination is defined as the fact that two groups sharing the same characteristics and performing the same tasks are paid differently. However, this definition is not easy to translate in a statistical context. On the one hand, there is the problem of equal characteristics, and on the other hand, that of equal tasks. In this thesis, we focus on the first problem and propose three methods to estimate gender wage discrimination. Using all of them, we reweigh the distribution of women's characteristics such that it is the same to that of men. In this way, we compare the pay that two groups with the same characteristics obtain. We propose a nonparametric and two parametric methods that take into account the presence of survey weights. These methods are illustrated using real data obtained from the Swiss Federal Statistical Office and their results are compared to those of well-established methods from the statistical literature.

Keywords: decomposition methods, income distribution, reweighting.

De nos jours, la discrimination salariale basée sur le sexe est un sujet débattu. Bien que la discrimination salariale soit interdite par la Déclaration universelle des droits de l'homme, des études montrent que les femmes sont systématiquement payées moins que les hommes. La discrimination salariale se produit quand deux groupes partageant les mêmes caractéristiques et remplissant les mêmes tâches sont payés différemment. Cette définition pose des difficultés à partir du moment où elle est présentée dans un contexte statistique. D'un côté, il y a le problème du partage des mêmes caractéristiques et de l'autre côté, celui des tâches similaires. Dans cette thèse, nous nous concentrons sur le premier et nous proposons trois méthodes d'estimation de la discrimination salariale à l'encontre des femmes. A travers ces méthodes,

nous repondérons la distribution des caractéristiques des femmes de telle manière qu'elle soit la même que celle des hommes. De cette manière, nous aboutissons à la comparaison des salaires de deux groupes à caractéristiques égales. Nous proposons deux approches: l'une est basée sur une méthode non-paramétrique, l'autre sur deux méthodes paramétriques qui prennent en compte les poids de sondage. Ces méthodes sont illustrées en utilisant une base de données obtenue de l'Office fédéral de la statistique, Suisse. Les résultats sont comparés à ceux obtenus avec d'autres méthodes établies dans la littérature statistique.

Mots-clé: méthodes de décomposition, distribution du revenu, repondération.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation

In 1918, Millicent G. Fawcett narrated the story of Mr. Jones, who braided military tunics for a living (Fawcett, 1918). When he fell ill, he had the permission to work from home. As the illness progressed, he taught his wife, Mrs. Jones, how to braid these tunics, so that the household would still have a revenue. However, the employer was not aware that the job was being done by Mrs. Jones until her husband died. When it became clear that Mrs. Jones had been braiding the tunics, this led to her wage being reduced by one-third. As the end of World War I was approaching, the idea of "equal pay for equal work" gained more attention. The prejudiced belief that women were unable to perform skilled work was refuted during the war (Fawcett, 1918), which broadened their horizons and made that for them to earn their living was no longer inconceivable. However Rathbone (1917) doubted that a "fair competition between men workers and women workers" was possible, because of the "customary difference in the wage level of the two sexes and the causes of that customary difference". One hundred years later, this question may still be valid. Overall, the fairness of pay is a question that researchers have addressed thoroughly, but differently: if at the beginning, the idea of "equal pay for equal work" was controversial, now it is widely accepted. However, accepted does not always mean put into practice.

In spite of the enormous progress done by society since War World I, in our days, gender discrimination still represents an issue present in various aspects of our daily lives. Indeed, the traditional roles of individuals are challenged, since it is no longer expected for men to be the breadwinners of a household and for women to look after children (see, for instance, Cancian et al., 1992; Schwartz, 2010). Lips (2017)

overviews the progress recorded in the last years in our society. The author notes that whereas at the beginning of the 20^{th} century, women were not allowed to vote, today not only do they have this right, but in more than 30 countries, they have held the highest position in state. However, this does not necessarily mean that the way has opened up for women, giving them access to any kind of job. In the same book, Lips (2017) also deplores the fact that we, as individuals are still biased by stereotypes, which are strongly related to gender. One environment where gender discrimination may be quite obvious is the workplace. One still classifies workplaces into male-dominated and female-dominated and as Lips (2013) notes, hearing about a man who occupies a job usually held by a woman or the other way around "is a reminder of the way work has been gendered". An obvious consequence of the unequal treatment in the workplace leads to a wage gap between men and women.

In Switzerland, gender wage discrimination in the workplace is prohibited by the Swiss Federal Act on Gender Equality (The Federal Assembly of the Swiss Confederation, 1995). Gender wage equality is also covered by the Swiss Constitution. Moreover, the Federal Office for Gender Equality is mandated to ensure that discrimination is not present in either aspect of the daily life and that gender equality is attained (Federal Office for Gender Equality, 2015).

However, inequality does not necessarily mean discrimination. The Cambridge dictionary defines discrimination as "treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin color, sex, sexuality etc.". Inequality does not necessarily imply unfairness, whereas discrimination does. In one of the earliest papers written about Switzerland and devoted to gender wage discrimination, Kugler (1988) advises not to take inequality for discrimination. This is because the reported discrimination level might in fact be the result of one of the two phenomena: women may discriminated against before they enter the labor market, which may affect their choices in their later life. Perpetuating stereotypes related to the traditional roles of men and women may lead to lower educational levels of women, which finally result in lower wages for women. On the other hand, women may simply choose to work in a certain field, associated to lower pay, or to take care of the household, without external influences. Both these situations lead to distorted discrimination levels. However, the author agrees that "the significance of these potential sources of distortion is difficult to empirically estimate" [1] (Kugler, 1988).

[1] The original text is: *Die Bedeutung dieser potentiellen Verzerrungsquellen ist empirisch fundiert schwer abzuschätzen*

The research on gender wage discrimination in Switzerland is limited to a small number of studies. All of them agree upon the fact that wage equality between men and women is not yet attained. Kugler (1988) finds that a small proportion of the wage difference, namely around 7%, is attributable to discrimination. All the studies thereafter find larger values attributed to discrimination. Diekmann et al. (1995) for instance, report values between 20% and 35%, depending on the model used. Bonjour and Gerfin (2001) find that women face a higher discrimination in lower-paying jobs and credit low educational levels as the main cause. Schmid (2016) analyses cohorts of employees and shows that overall, women who work part-time (occupational degree of less that 50%) are those who are mostly discriminated against.

There is a number of possible causes behind wage inequality, and some of these causes are found in the studies cited above. These may stem from the labor demand or supply sides (Schmid, 2016). The starting point may be the educational attainment of individuals. Finzi (2007) shows that in the period 1970-2000, the proportion of women who had only completed compulsory education was larger than that of men, whereas for university degrees, the inverse held. Secondly, there are the life plans of the individual: while women may devote more time to children and the household, men tend to concentrate on their professional careers (Schmid, 2016). This tendency is confirmed in Switzerland. In Schön and Liechti (2013), it is shown that in a typical family with children under 15 years old, the man works fulltime and the woman either works part-time or is unemployed. Fthenakis and Minsel (2002) note that in families without children, there is a link between the desire to have a child and the status in education of the woman. If the woman is pursuing studies, then a child does not fit into the plans, as opposed to when she has already graduated. For men, the status in education with respect to the choice of having a child is irrelevant. Second, the returns to endowments are different for men and for women. Following Schmid (2016), flexibility pays for men, but not for women. Men are more willing to travel long distances if the wage is higher, but for women it is not the case. This can be linked with the hypothesis that women plan their professional life in function of their personal life. These factors may result in segregation: if women choose their jobs in such a way to accommodate their family plans, they may end up in lower-paying jobs. This concentration of women in lower-paying jobs and of men in higher-paying jobs is called occupational segregation. Schmid (2016) finds that occupational segregation is an influential factor of inequality and quantifies this segregation. According to the author, “40% of the male or female employees would need to change jobs”. Diekmann et al. (1995) however find that even if there were

no segregation in the economic activities, the wage differences would remain the same. These characteristics on the demand side result in a reaction from the supply side. Job segregation leads to lowering wages in a sector where the majority of the positions is held by women (see, for instance, Hausmann et al., 2015; England et al., 2007; England, 1992).

Wage inequality may also be the result of behavior, both that of employees and of employers. Rathbone (1917) notes that "I have not yet met the feminist whose principles compel her to pay her waitress the wages that would be demanded by a butler." Jann (2003) shows that both women and men tend to give less value to the same job if a woman fills it. Lips (2013) cites the work of Alice Eagly, a social psychologist who summarized the workplace bias as a circle in which certain jobs are associated with traits that are more likely to pertain to one gender. The more this association occurs, the more members of the respective gender fill it. This is how biases arise and can give rise to barriers for people who want to do a job associated to the other gender. These biases become "virtually automatic in their activation", leading us to associate for instance an engineer to a man and a nurse to a woman. However, assigning a person to a job should be done solely on the basis of that person's ability to fill in the job, and not on that of their gender. For instance, Goldin and Rouse (1997) show that women were more likely to be hired in a symphony orchestra when blind auditions are done so that the selection process was done objectively and talent was evaluated unbiasedly. Biases linking gender to jobs give rise to gender segregation in the workplace. This has led to differences in wages (Blau et al., 2013). Schmid (2016) reports that " occupational segregation accounts for a quarter of the gender wage gap." Moreover, when the proportion of women is higher in a given occupation, the wages in that occupation are lower (Levanon et al., 2009; Siltanen, 1994). Following the theory of devaluation, "decisions of employers about the relative pay of "male" and "female" occupations are affected by gender bias" (Levanon et al., 2009). Occupations in which women represent a high proportion are seen as having less value than those in which men are more involved. The assortative theory states that individuals choose partners who are similar to themselves. This results in homogamy, a state in which individuals display "the preference to choose a similar partner over other potential partners" (Kuhn and Ravazzini, 2017). Homogamy may lead to a homogeneous couple in terms of income, but to a heterogeneous society. Using data from the United States, Schwartz and Mare (2005) show in an analysis covering the period from 1940 to 2000, that "college graduates, in particular, were increasingly likely to marry each other rather than

persons with less education". Kuhn and Ravazzini (2017) investigate the impact of homogamy on wage inequality in Switzerland and do not find any significant effect.

In his book, Becker (2010) defined wage discrimination as having different wage rates in two groups when the members of one group are identical to the members of the second group. Following this definition, gender wage discrimination occurs when a man and a woman receive different remuneration for a job that requires the same qualifications or which implies identical productivity (see, for instance, Neumark, 1988; Gardeazabal and Ugidos, 2005; Schmid, 2016).

In this thesis, we use this definition in the attempt to develop statistical methods to estimate gender wage discrimination. We analyze statistically the fairness of the Swiss labor market through the lens of wage equality. Selection bias, employment processes or other mechanisms on the labor market are not covered. This is because we assume that an individual is free to choose their job, that the labor market offers equal opportunities to men and women and that an employer is objective when choosing a candidate to fill in the job, judging them by their qualifications and not by subjective criteria.

In reality however, it is extremely rare to have employees with identical qualifications, which makes the estimation of the discrimination level challenging. Nevertheless, in the statisticals literature, the estimation of discrimination levels is mostly linked to differences in characteristics. Quantifying other factors, such as the ones presented above, that may lead to differences in wages is perhaps impossible.

In this thesis, we propose two approaches that link the characteristics of employees to their wages. In the two proposed approaches, we focus on matching the characteristics of men with those of women in order to have two comparable groups. When having two groups with similar characteristics, we can estimate the wage differences between them and thus estimate the discrimination level against one of these two groups.

1.2 Outline

This thesis is structured as follows: in Chapter 2, two wage decomposition methods are described. There are other well-established methods in the literature, however, we only address these two because our proposed approaches are directly related to them. Both methods share the hypothesis of an existing relationship between an individual's wage and their characteristics (for instance age or work experience). This relationship is assumed differently in each method. The first method, proposed

by Blinder (1973) and Oaxaca (1973), assumes a linear relationship between the logarithm of the wage and the characteristics of an employee. Using it results in a decomposition of the average wage differences. It represents the ground of numerous methods that follow in the statistical literature. The second method, developed by DiNardo et al. (1996) resorts to logistic regression to estimate the probability of being a man and a woman, respectively, given the observed characteristics. The second step is the computation of a reweighting factor that is used to render women's distribution of characteristics similar to that of men. This method allows the decomposition of the difference in wages at other points than the mean, namely at quantiles. Chapter 2 serves as the literature review and we will often make reference to it. In Chapter 3, we present the first proposed approach, which is a nonparametric method. It is a generalization the methods described in Chapter 2 and it addresses their potential drawbacks. Two applications to real survey data are also included in Chapter 3. The data are provided from the Swiss Federal Statistical Office. In Chapter 4, the second approach is introduced. It is a parametric approach, where we assume that the wage of each individual follows a distribution that has one of the parameters expressed as a function of their characteristics. We do not assume a global distribution for the wage of one group, but a distribution fitted for each individual. This will allow us to estimate a wage distribution for an individual given their attributes. Finally, Chapter 5 contains a discussion and the conclusions. In Appendix A, a description of the categorical variables used in the applications is done and in Appendix B, first and second-order derivatives useful for the application in Chapter 4 are detailed.

Chapter 2

An introduction to wage decomposition methods

Abstract

This chapter is dedicated to two existing methods in the literature and serves as the literature review that we will turn to later when motivating the proposed approaches. In Section 2.1, we briefly explain what decomposition methods are and discuss their aim. In Section 2.2 the general framework is presented, that will be used throughout the entire thesis. In Sections 2.3 and 2.4, we review the two decomposition methods. There are many other techniques in the literature, but our proposed approaches will be directly related to these two. The first technique, developed by Blinder (1973) and Oaxaca (1973) decomposes the difference in average wages. The second one, the semi-parametric approach of DiNardo et al. (1996) extends it and allows the decomposition of wage differences at other points. A part of this chapter is found in Anastasiade and Tillé (2017a).

2.1 Introduction

In labor economics, decomposition methods are used to answer questions related to changes in the labor market. For example, one wants to estimate how the affiliation to unions impacts employees' wages (Doiron and Riddell, 1994; Lewis, 1986) or which factors account for the increase in inequality from one time period to another (see, for instance, Machado and Mata, 2005; Pereira and Martins, 2002; Melly, 2005). Fortin et al. (2011) provide a comprehensive summary of these methods. Generally speaking, decomposition methods allow to divide a difference in the values of an outcome measured on two groups (or at two moments in time) into two parts: a part that can be explained by some factors that characterize the two groups and the other one that can not. In the context of decompositions of wage differences of two groups, these explaining factors can be the different characteristics (or attributes) of the two groups. For instance, the members in one group earn more because they have more work experience or more years of schooling. A more detailed discussion is given in Section 2.3. Decomposition methods in wage differences are done through

so-called counterfactual exercises. There are two such exercises: either estimating a counterfactual average or a counterfactual distribution. This counterfactual attribute represents what one group would earn if they had the characteristics of the other group. A counterfactual object is never observed in reality, however, it is a tool that enables the estimation of the two parts that make up the difference in outcomes of the two groups in question. The counterfactual average wage is discussed in Subsection 2.3.2 and the counterfactual wage distribution in Subsection 2.3.5.

There are two types of decompositions: aggregate and detailed (Fortin et al., 2011). In an aggregate decomposition, only the two parts that make up the difference are estimated. In the detailed decomposition, the contribution of each characteristic to these two parts is estimated. Fortin et al. (2011) summarize the limitations of these methods and note that they do not identify the exact relationships between the characteristics and the outcome, but only "provide useful indications of particular hypotheses or explanations to be explored in more detail" (Fortin et al., 2011). We only examine aggregate decompositions, because the focus is to estimate the overall differences, and not to investigate the impact of a particular characteristic on them.

2.2 The general setup

The setup presented in this section will be used throughout this thesis. Consider a finite population of employees with the labels $U = \{1, 2, \ldots, N\}$. From this population, we randomly select a sample S of size n, without replacement. The sample is selected through a sampling design $p(s) = \Pr(S = s), \forall s \subseteq U$. To each unit $k \in S$, a survey weight w_k is associated. These weights can be equal to the inverse of the inclusion probabilities or can be more complicated weights, like calibration weights. The inclusion probability in the sample is defined as

$$\pi_k = \sum_{s \subset U | k \ni s} p(s), \ \ \forall k \in U. \tag{2.1}$$

The set U is divided in two subpopulations of labels corresponding to men and women, denoted by U_M and U_F respectively, such that $U_M \cup U_F = U$ and $U_M \cap U_F = \emptyset$. There are N_M and N_F individuals in U_M and U_F, respectively. Similarly, the sample S is divided into two random subsamples of men and women, denoted by $S_M = S \cap U_M$ and $S_F = S \cap U_F$ respectively. We denote these subsamples as $S_g \subseteq U_g, g \in \{M, F\}$, with n_M and n_F being the number of employees in the subsamples, respectively, such that $n_M + n_F = n$. The variable of interest, denoted by y, is in this case the logarithm

of the wage. The totals of the variable of interest in the two subpopulations are given by

$$Y_g = \sum_{k \in U_g} y_k, h \in \{F, M\},$$

where y_k is the logarithm of the wage of the kth individual. Since not all units in the subpopulations are observed, the totals can be estimated by

$$\widehat{Y}_g = \sum_{k \in S_g} d_k y_k, g \in \{F, M\}, \tag{2.2}$$

where d_k is a sampling weight allotted to the kth unit of the sample. Basically, the sampling weights are the inverse of the inclusion probabilities given by

$$d_k = \frac{1}{\pi_k}, k \in S. \tag{2.3}$$

If the weights are defined as in Expression (2.3), the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is obtained in (2.2), that is unbiased provided that $\pi_k > 0$, for all k in S. However, in real survey data applications, the weights w_k are obtained after several statistical treatments. Auxiliary totals are totals of variables such as age, gender or geographical variables. The auxiliary totals are known at the population level from sources such as censuses. The sampling weights are adjusted to compensate for the questionnaire nonresponse. This is generally done using the information provided by known auxiliary totals. The adjustment is done for example by means of the Deville-Särndal calibration procedure (Deville and Särndal, 1992). So, in what follows, suppose that for the sample units, there is a weighting system that was previously computed.

The population means of the logarithms of the wages are given by

$$\overline{Y}_g = \frac{1}{N_g} \sum_{k \in U_g} y_k, g \in \{F, M\},$$

and can be estimated by

$$\widehat{\overline{Y}}_g = \frac{\sum_{k \in S_g} w_k y_k}{\sum_{k \in S_g} w_k}, g \in \{F, M\}.$$

Moreover, assume that for each kth individual in either of the two subsamples, there is a vector of J auxiliary variables denoted by

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top \in \mathbb{R}^J.$$

This vector is supposed to be known for each unit selected in the sample. The auxiliary variables contain some characteristics of the individual, for instance the age, the education level or the seniority level. They can be quantitative or qualitative variables, thus x_{kj} can be a categorical variable or a quantity. Also assume that the first auxiliary variable is a constant, i.e. $x_{k1} = 1$, for all $k \in U$.

The totals of these auxiliary variables at the subpopulation level are given by

$$\mathbf{X}_g = \sum_{k \in U_g} x_k, g \in \{F, M\}.$$

Using the weights w_k defined above, these two totals can be estimated by

$$\widehat{\mathbf{X}}_g = \sum_{k \in S_g} w_k \mathbf{x}_k, g \in \{F, M\}.$$

Vectors of average values can be analogously estimated. The average values at the subpopulation levels are given by

$$\overline{\mathbf{X}}_g = \frac{1}{N_g} \sum_{k \in U_g} \mathbf{x}_k, g \in \{F, M\},$$

and estimated by

$$\widehat{\overline{\mathbf{X}}}_g = \frac{\sum_{k \in S_g} w_k \mathbf{x}_k}{\sum_{k \in S_g} w_k}, g \in \{F, M\}. \tag{2.4}$$

2.3 The adapted Blinder-Oaxaca method

2.3.1 Introduction

Using the setup in Section 2.2, the findings of Blinder (1973) and Oaxaca (1973) (hereafter, BO) are summarized in the context of sampling theory, namely by using sampling weights. Assume that in each sample, a linear relationship is suitable between the J characteristics that are available and the logarithm of the wage. This can be expressed as

$$Y_{k,g} = \mathbf{X}_{k,g}^\top \beta_g + \varepsilon_{k,g}, k \in U_g, \tag{2.5}$$

where ε_g is the error term in group g which follows a normal distribution $\varepsilon_g \sim N(0, \sigma_g^2)$, where σ_g^2 represents the variance of the error term in group $g \in \{F, M\}$ and β_g denotes the vector of regression coefficients in group $g \in \{F, M\}$.

By using Model (2.5) and assuming that Y_g is a random variable, $\mathbf{X}_g$ is a vector of random covariates and U is a random sample from an infinite population, one obtains the conditional expectation $E(Y_g \mid \mathbf{X}_g = \mathbf{x}_g) = \mathbf{x}_g^\top \beta_g$ and the unconditional expectation

$$E(Y_g) = E\left(E(Y_g \mid \mathbf{X}_g)\right) = E(\mathbf{X}_g)\beta_g + E(\varepsilon_g) = E(\mathbf{X}_g)\beta_g,$$

where $\mathbf{X}_g$ and ε_g are independent.

The difference between the conditional expectations of log of wages of two groups can be written as

$$\begin{aligned} \Delta &= E(Y_M) - E(Y_F) \\ &= E\left(E(Y_M \mid \mathbf{X}_M)\right) - E\left(E(Y_F \mid \mathbf{X}_F)\right). \\ &= \left(E\left(\mathbf{X}_M\right) - E(\mathbf{X}_F)\right)\beta_F + E(\mathbf{X}_M)\left(\beta_M - \beta_F\right). \end{aligned} \tag{2.6}$$

The difference between the average of the log of wages of the groups in Expression (2.6) contains two elements: an explained part, also called the *composition effect* $(E(\mathbf{X}_M) - E(\mathbf{X}_F))\beta_F$ and an unexplained part, or the *structure effect* $E(\mathbf{X}_M)(\beta_M - \beta_F)$. The former encompasses differences in characteristics between the two groups. The latter is the difference in the returns on characteristics between the two groups, the part that is not attributable to objective factors (Oaxaca, 1973; Blinder, 1973).

At the U_g level, $E(\mathbf{X}_g)$ is reduced to a finite mean $\overline{\mathbf{X}}_g = \sum_{k \in U_g} \mathbf{X}_{k,g}/N_g$, and the regression coefficients are given by

$$\beta_g = \left(\sum_{k \in U_g} \mathbf{X}_{k,g}\mathbf{X}_{k,g}^\top\right)^{-1} \sum_{k \in U_g} \mathbf{X}_{k,g} Y_{k,g},\ g \in \{M, F\}.$$

The vector β_g can be consistently estimated from the subsamples S_g by

$$\widehat{\beta}_g = \left(\sum_{k \in S_g} w_k \mathbf{x}_{k,g}\mathbf{x}_{k,g}^\top\right)^{-1} \sum_{k \in S_g} w_k \mathbf{x}_{k,g} y_{k,g},\ g \in \{M, F\}, \tag{2.7}$$

where $y_{k,g}$ is the realization of $Y_{k,g}, k \in S_g$.

The difference Δ can be estimated at sample level by

$$\widehat{\Delta} = (\widehat{\overline{\mathbf{X}}}_M - \widehat{\overline{\mathbf{X}}}_F)^\top \widehat{\beta}_F + \widehat{\overline{\mathbf{X}}}_M^\top (\widehat{\beta}_M - \widehat{\beta}_F), \tag{2.8}$$

where $\widehat{\overline{\mathbf{X}}}_g = \sum_{k \in S_g} w_k \mathbf{x}_{k,g} / \sum_{k \in S_g} w_k$ represents the estimator of $\overline{\mathbf{X}}_g$.

The regression coefficients $\widehat{\beta}_g$ are called the group wage structure or the returns on characteristics and they represent the contribution of each characteristic to the wage. If the returns on characteristics are identical in both groups, there will be no difference in the average wages of men and those of women.

Result 1 *A sufficient condition to obtain the following equalities*

$$\overline{Y}_g = \overline{\mathbf{X}}_g^\top \beta_g \text{ and } \widehat{\overline{Y}}_g = \widehat{\overline{\mathbf{X}}}_g^\top \widehat{\beta}_g$$

is that there exists a vector $\zeta \in \mathbb{R}^p$, such that $\zeta^\top \mathbf{x}_k = 1$, for all $k \in U_g$.

Proof

We only give the proof for $\widehat{\overline{Y}}_F = \widehat{\overline{\mathbf{X}}}_F \widehat{\beta}_F$, the other equalities are obtained in a similar manner.

$$\begin{aligned}
\widehat{\mathbf{X}}_F \widehat{\beta}_F &= \widehat{\mathbf{X}}_F \left(\sum_{k \in S_F} w_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{\ell \in S_F} w_\ell \mathbf{x}_\ell y_\ell \\
&= \sum_{j \in S_F} w_j \mathbf{x}_j^\top \left(\sum_{k \in S_F} w_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{\ell \in S_F} w_\ell \mathbf{x}_\ell y_\ell \\
&= \left(\sum_{j \in S_F} \zeta^\top w_j \mathbf{x}_j \mathbf{x}_j^\top \right) \left(\sum_{k \in S_F} w_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{\ell \in S_F} w_\ell \mathbf{x}_\ell y_\ell \\
&= \sum_{\ell \in S_F} \zeta^\top w_\ell \mathbf{x}_\ell y_\ell = \sum_{\ell \in S_F} w_\ell y_\ell = \widehat{Y}_F.
\end{aligned}$$

By dividing this equation by $\sum_{k \in S_F} w_k$, we get Result 1. □

Since it is assumed that $x_{k1} = 1$ for all $k \in U$, with $\zeta^\top = (1\, 0 \ldots 0)$, the equality is always fulfilled.

2.3.2 The counterfactual average wage

The term $\widehat{\overline{\mathbf{X}}}_M \widehat{\beta}_F$ that appears in Equation (2.8) is called the women's counterfactual average wage. We interpret it as the estimated average wage of women if they had the

same average characteristics as men and if their return on characteristics remained unchanged. This counterfactual exercise is also found in Fortin and Lemieux (2013). Women's counterfactual wage distribution is obtained by using the characteristics of men ($\mathbf{X}_M$) and the wage structure of women (β_F).

Women's counterfactual mean of log of wage equals

$$E(E(Y_F \mid \mathbf{X}_M)) = E(\mathbf{X}_M)\beta_F.$$

Using Result 1 from the previous section, at the U level, women's counterfactual average wage $\overline{Y}_{F|M}$ equals

$$\overline{Y}_{F|M} = \overline{\mathbf{X}}_M^\top \beta_F,$$

and is estimated from the sample by

$$\widehat{\overline{Y}}_{F|M} = \widehat{\overline{\mathbf{X}}}_M^\top \widehat{\beta}_F,$$

where $\widehat{\overline{\mathbf{X}}}_M$ are estimated in Equation (2.4) and $\widehat{\beta}_F$ are the coefficients estimated by means of Equation (2.7). With this notation, the BO decomposition given in (2.6) is re-expressed at the sample level as

$$\widehat{\Delta} = \widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F = (\widehat{\overline{\mathbf{X}}}_M - \widehat{\overline{\mathbf{X}}}_F)^\top \widehat{\beta}_F + \widehat{\overline{\mathbf{X}}}_M^\top (\widehat{\beta}_M - \widehat{\beta}_F) = (\widehat{\overline{Y}}_{F|M} - \widehat{\overline{Y}}_F) + (\widehat{\overline{Y}}_M - \widehat{\overline{Y}}_{F|M}). \quad (2.9)$$

2.3.3 The composition and the structure effects

As already mentioned, the estimated difference between the average wages of the groups contains two elements: an explained part, also called the estimated *composition effect* $(\widehat{\overline{\mathbf{X}}}_M - \widehat{\overline{\mathbf{X}}}_F)^\top \widehat{\beta}_F$ and an unexplained part, or the estimated *structure effect* $\widehat{\overline{\mathbf{X}}}_M^\top (\widehat{\beta}_M - \widehat{\beta}_F)$. The former encompasses differences in characteristics between the two groups. The latter is the difference in the returns on characteristics between the two groups, the part that is not attributable to objective factors (Oaxaca, 1973; Blinder, 1973). The estimation of the *structure effect* is the central element of this paper. Equation (2.6) has the same elements as the one proposed by Oaxaca (1973) and Blinder (1973). The methodology applied to obtain the estimated average values and coefficients differs from the traditional regression technique. The BO method uses the estimated regression coefficients obtained through ordinary least squares (OLS) and the vectors of average values of the observed explanatory variables. The proposed approach takes into account the survey weights and a finite population. However, the weighted

BO method is the same as the original BO method if the sampling weights are all equal to 1.

The two elements in Equation (2.6) have different names across the literature. The first one, whose denomination we retained as *composition effect* is also termed *endowments effect*. The second one, which we call *structure effect* is also found in the literature as *unexplained residual*, *price effect*, *sex effect*, *calculated effect* or *unequal treatment* (Weichselbaumer and Winter-Ebmer, 2006). Using the BO method, the structure effect is interpreted as discrimination. However, discrimination is an intricate phenomenon that might not be always fully observed. Inclusion of unobserved variables or some good interpretation of the mechanisms on the labour market can help to increase the explained part of the wage difference. Moreover, Weichselbaumer and Winter-Ebmer (2005) note two potential issues regarding the chosen regression model. First, if the characteristics chosen in the linear model are themselves subject to discrimination, then the resulting structure effect will be over-estimated. Second, if the characteristics are not a proper measure of the productivity, then again, the structure effect might be under- or over-estimated. Weichselbaumer and Winter-Ebmer (2006) warn about the legitimacy of the characteristics as productivity indicators, since "wages may also be determined by bargaining power, compensating differentials or efficiency wages". However, for simplicity, in what follows, we will assume that there are no such issues and that the estimated structure effect is the result of discrimination on the labor market. Moreover, we do not examine sample selection bias or other mechanisms underlying the gender distribution in certain jobs.

2.3.4 The non-discriminatory wage structure

In decomposition methods, a partial equilibrium approach is assumed (Fortin et al., 2011; Cappellari et al., 2016). This approach implies that the wage structure of one group is the wage structure that would prevail in a non-discriminatory world. When we build women's counterfactual average wage as $\widehat{\overline{\mathbf{X}}}_M \widehat{\beta}_F$, we assume that women's wage structure is the one that would prevail in a non-discriminatory world. Therefore, on average, this is what they would earn if they had the same characteristics as men. This assumption means that women earn a fair wage and that discrimination takes place in men earning more than they should (Oaxaca, 1973).

In the general equilibrium approach, the assumption is that the wage structure that would prevail in a non-discriminatory world is different from the ones pertaining to the two groups. For instance, Cotton (1988) defines a non-discriminatory wage

structure as

$$\widehat{\beta}^{*} = p_M \widehat{\beta}_M + p_F \widehat{\beta}_F, \tag{2.10}$$

where p_M and p_F are the proportions of men and women, respectively on the labour market. From Equation (2.10), it is clear that the non-discriminatory wage structure will be closer to the wage structure that belongs to the group that makes up the larger proportion of the sample. Cotton (1988) argues that discrimination is actually the sum of two components: the amount by which men are overpaid and the amount by which women are underpaid. The difference between average wages, denoted as Δ_C, is thus given by

$$\begin{aligned} \Delta_C &= \widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F \\ &= (\widehat{\overline{\mathbf{X}}}_M^{\top} - \widehat{\overline{\mathbf{X}}}_F^{\top})\widehat{\beta}^{*} + \widehat{\overline{\mathbf{X}}}_M^{\top}(\widehat{\beta}_M - \widehat{\beta}^{*}) + \widehat{\overline{\mathbf{X}}}_F^{\top}(\widehat{\beta}^{*} - \widehat{\beta}_F) \end{aligned} \tag{2.11}$$

Heckman (1977) argues that before writing the wage equation, it is important to estimate the probability of being included in the observed sample using a probit model. From this probit model, the inverse of Mill's ratio is estimated. Reimers (1983) follows this argument and writes the average of the logarithm of the wage as

$$\widehat{\overline{Y}}_g = \widehat{\overline{\mathbf{X}}}_g^{\top} \widehat{\beta}_g + \widehat{c}_g \widehat{\overline{\xi}}_g, \tag{2.12}$$

where $\widehat{c}_g$ is the estimated covariance between the residuals in the probit model and those in the wage equation in Equation (2.5) and $\widehat{\overline{\xi}}_g$ is Mill's ratio average from the probit model. We obtain

$$\begin{aligned} \Delta_R &= \widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F \\ &= (\widehat{\overline{\mathbf{X}}}_M^{\top} - \widehat{\overline{\mathbf{X}}}_F^{\top})[W\widehat{\beta}_M + (\mathbf{I} - \mathbf{W})\widehat{\beta}_F] + [\widehat{\overline{\mathbf{X}}}_M^{\top}(\mathbf{I} - \mathbf{W}) + \widehat{\overline{\mathbf{X}}}_F^{\top}\mathbf{W}](\widehat{\beta}_M - \widehat{\beta}_F) + (\widehat{c}_M \widehat{\overline{\xi}}_M - \widehat{c}_F \widehat{\overline{\xi}}_F), \end{aligned} \tag{2.13}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{W}$ is a diagonal matrix of weights. Reimers (1983) selects $\mathbf{W} = 0.5\mathbf{I}$. However, Oaxaca and Ransom (1994) claim that neither one of these alternatives is "completely satisfactory since each chooses the weight in an arbitrary manner" (Oaxaca and Ransom, 1994). The choice of a non-discriminatory wage structure is always debatable, but necessary to make a decomposition exercise.

2.3.5 The counterfactual wage distribution

While the method of Blinder (1973) and Oaxaca (1973) enables the decomposition of the differences in average wages, sometimes it is of higher interest to decompose the differences in wages at other points. For instance, one may be interested in estimating the composition and the structure effects in lower-paying jobs, since women tend to be concentrated in such jobs. In order to do so, one should be able to estimate a counterfactual wage distribution.

A counterfactual distribution is defined "as the result of either a change in the distribution of a set of covariates X that determine the outcome variable of interest Y, or as a change in the relationship of the covariates with the outcome, i.e. a change in the conditional distribution of Y given X" (Chernozhukov et al., 2013).

In what follows, we will construct the counterfactual wage distribution as the distribution resulting from the change in the distribution of covariates. We will then compare the observed and the counterfactual wage distributions to measure the effects of the change. More specifically, we will build the counterfactual wage distribution of women using the characteristics of men. We will interpret it as the estimated wage distribution of women if they had the characteristics of men. Therefore, the differences that we will estimate between the estimated wage distribution of men and the countefactual wage distribution will represent the estimation of the structure effect. This is because we assume that the two distributions are estimated using the same set of characteristics. On the other hand, the differences between the estimated counterfactual wage distribution and the estimated wage distribution of women will represent the estimation of the composition effect. Since the two distributions are estimated using different characteristics, we assume that the resulting differences are accounted for by the difference in characteristics.

The counterfactual distribution serves as a tool that enables the estimation of the discrimination level between men and women. However, such a distribution is *never* observed in reality. Its estimation is backed-up by hypotheses that are stated by the researcher.

2.4 The adapted DiNardo-Fortin-Lemieux method

2.4.1 Introduction

The method proposed by DiNardo et al. (1996) (heareafter, DFL) uses a reweighting function by which women's distribution of characteristics is rendered similar to

men's distribution of characteristics. The reweighted distribution is the women's counterfactual distribution of characteristics. The DFL method is presented through the use of survey weights in order to take the sampling design into account.

The DFL method is presented through the use of sampling weights in order to take the sampling design into account.

The reweighting factor is equal to

$$\psi(\mathbf{x}_k) = \frac{\Pr(D_{Mk} = 1 \mid \mathbf{x}_k)/\Pr(D_{Mk} = 1)}{\Pr(D_{Mk} = 0 \mid \mathbf{x}_k)/\Pr(D_{Mk} = 0)},$$

where $D_{Mk} = 1$ if individual k is a man and $D_{Mk} = 0$ otherwise and $\mathbf{x}_k$ is the vector of observed characteristics for individual $k, k \in U$. Obviously, $\Pr(D_{Mk} = 1 \mid \mathbf{x}_k)$ and $\Pr(D_{Mk} = 0 \mid \mathbf{x}_k)$ must be estimated. For this type of estimation, DiNardo et al. (1996) suggested the use of a logit or a probit model. Using the information from the sample, the reweighting factor $\widehat{\psi}(\mathbf{x}_k)$ is estimated by

$$\widehat{\psi}(\mathbf{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 \mid \mathbf{x}_k)/\widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 \mid \mathbf{x}_k)/\widehat{\Pr}(D_{Mk} = 0)}. \tag{2.14}$$

The computation of $\widehat{\psi}(\mathbf{x}_k)$ is discussed in the next section. Using the estimated reweighting factor, women's counterfactual wage mean is estimated by

$$\widehat{\overline{Y}}_{F|M}^{DFL} = \frac{\sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k y_k}{\sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k}, \tag{2.15}$$

and women's counterfactual means of characteristics by

$$\widehat{\overline{\mathbf{X}}}_{F|M}^{DFL} = \frac{\sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k \mathbf{x}_k}{\sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k}. \tag{2.16}$$

Using the reweighting factor, women's counterfactual coefficients can also be computed. They are given at the U_F level by

$$\beta_F^{DFL} = \left(\sum_{k \in U_F} \psi(\mathbf{x}_k) \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U_F} \psi(\mathbf{x}_k) \mathbf{x}_k y_k,$$

and estimated by

$$\widehat{\beta}_F^{DFL} = \left(\sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_F} \widehat{\psi}(\mathbf{x}_k) w_k \mathbf{x}_k y_k. \tag{2.17}$$

The coefficients above have to be computed, because under the same condition as in Result 1, women's counterfactual wage mean defined in (2.15) is given by

$$\widehat{\overline{Y}}_{F|M}^{DFL} = \widehat{\overline{\mathbf{X}}}_{F|M}^{DFL^\top} \widehat{\beta}_F^{DFL}.$$

Using the reweighting function, the BO decomposition formula can be written as

$$\begin{aligned} \widehat{\Delta} = \widehat{\overline{Y}}_M - \widehat{\overline{Y}}_F &= (\widehat{\overline{Y}}_{F|M}^{DFL} - \widehat{\overline{Y}}_F) + (\widehat{\overline{Y}}_M - \widehat{\overline{Y}}_{F|M}^{DFL}) \\ &= \left(\widehat{\overline{\mathbf{X}}}_{F|M}^{DFL^\top} \widehat{\beta}_F^{DFL} - \widehat{\overline{\mathbf{X}}}_F^\top \widehat{\beta}_F\right) + \left(\widehat{\overline{\mathbf{X}}}_M^\top \widehat{\beta}_M - \widehat{\overline{\mathbf{X}}}_{F|M}^{DFL^\top} \widehat{\beta}_F^{DFL}\right), \quad (2.18) \end{aligned}$$

where $\widehat{\beta}_M$ and $\widehat{\beta}_F$ are defined in (2.7). The first term of Equation (2.18) is the composition effect and the second one the structure effect.

2.4.2 Estimation of the reweighting factor

Conditional to $\mathbf{x}_k$, D_{Mk} is an independent Bernoulli random variable having the conditional probability

$$\Pr(D_{Mk} = 1 \mid \mathbf{x}_k) = \frac{1}{1 + \exp(-\mathbf{x}_k^\top \gamma)},$$

where γ is a vector of parameters. This implies that

$$\Pr(D_{Mk} = 0 \mid \mathbf{x}_k) = 1 - \frac{1}{1 + \exp(-\mathbf{x}_k^\top \gamma)} = \frac{\exp(-\mathbf{x}_k^\top \gamma)}{1 + \exp(-\mathbf{x}_k^\top \gamma)}.$$

The ratio of these two probabilities equals

$$\frac{\Pr(D_{Mk} = 1 \mid \mathbf{x}_k)}{\Pr(D_{Mk} = 0 \mid \mathbf{x}_k)} = \exp(\mathbf{x}_k^\top \gamma).$$

Thus the estimated reweighting factor defined in Equation (2.14) will be equal to

$$\widehat{\psi}(\mathbf{x}_k) = \widehat{a} \exp(\mathbf{x}_k^\top \widehat{\gamma}),$$

where $\widehat{\gamma}$ is the estimation of γ from the sample and $\widehat{a}$ is the ratio of estimated proportions of women and men. It is given by:

$$\widehat{a} = \frac{\widehat{\Pr}(D_{Mk}=0)}{\widehat{\Pr}(D_{Mk}=1)} = \frac{\sum_{k\in S_F} w_k}{\sum_{k\in S_M} w_k}.$$

Vector γ can be estimated by using the maximum likelihood method, namely

$$\mathscr{L}(\gamma) = \prod_{k\in U} \Big[\Pr(D_{Mk}=1 \mid \mathbf{x}_k)^{D_{Mk}} \Pr(D_{Mk}=0 \mid \mathbf{x}_k)^{1-D_{Mk}} \Big].$$

By taking the logarithm, the log-likelihood function is obtained

$$l(\gamma) = -\sum_{k\in U} \log[1+\exp(-\mathbf{x}_k^\top \gamma)] - \sum_{k\in U} (1-D_{Mk})\mathbf{x}_k^\top \gamma.$$

Since the entire population is not observed, but only a sample, the empirical likelihood is used. Thus,

$$\widehat{l}(\gamma) = -\sum_{k\in S} w_k \log[1+\exp(-\mathbf{x}_k^\top \gamma)] - \sum_{k\in S} (1-D_{Mk}) w_k \mathbf{x}_k^\top \gamma.$$

By setting to zero the derivative of $\widehat{l}(\gamma)$ with respect to γ, it results that

$$\frac{\partial \widehat{l}(\gamma)}{\partial \gamma} = \sum_{k\in S} w_k \mathbf{x}_k \frac{\exp(-\mathbf{x}_k^\top \gamma)}{1+\exp(-\mathbf{x}_k^\top \gamma)} - \sum_{k\in S} (1-D_{Mk}) w_k \mathbf{x}_k = \mathbf{0},$$

which can be written as

$$\sum_{k\in S} w_k \mathbf{x}_k \frac{1}{1+\exp(-\mathbf{x}_k^\top \gamma)} = \sum_{k\in S_M} w_k \mathbf{x}_k.$$

The vector γ is estimated by $\widehat{\gamma}$ from the sample using empirical likelihood.

Women's counterfactual wage mean can be re-expressed as

$$\widehat{\overline{Y}}_{F|M}^{DFL} = \frac{\sum_{k\in S_F} w_k \widehat{\psi}(\mathbf{x}_k) y_k}{\sum_{k\in S_F} w_k \widehat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k\in S_F} w_k \exp(\mathbf{x}^\top \widehat{\gamma}) y_k}{\sum_{k\in S_F} w_k \exp(\mathbf{x}^\top \widehat{\gamma})}. \tag{2.19}$$

Women's counterfactual means of characteristics are given by

$$\widehat{\overline{\mathbf{X}}}_{F|M}^{DFL} = \frac{\sum_{k\in S_F} w_k \widehat{\psi}(\mathbf{x}_k) \mathbf{x}_k}{\sum_{k\in S_F} w_k \widehat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k\in S_F} w_k \exp(\mathbf{x}^\top \widehat{\gamma}) \mathbf{x}_k}{\sum_{k\in S_F} w_k \exp(\mathbf{x}^\top \widehat{\gamma})}.$$

This method allows for the construction of a counterfactual wage distribution. This in turn allows for the comparison between this new distribution and the observed wage distributions of women and men. The drawback of the method is that it may happen that at least one characteristic is a good predictor of the gender (for instance, the economic sector). This implies that $\Pr(D_{Mk} = 1 \mid \mathbf{x}_k)$ will get close to 1 and that the reweighting factor will take take on a large value (Fortin et al., 2011). This obviously leads to a large variance of the estimated factor.

2.4.3 Further decomposition of the structure effect

As Fortin et al. (2011) note, the purpose of the DFL reweighting factor is to render the distribution of women's characteristics identical to that of men. This implies that the means of the auxiliary variables in the two groups should be equal. However, with the DFL method, it is not the case. Indeed,

$$\widehat{\overline{\mathbf{X}}}_{F|M}^{DFL} \neq \widehat{\overline{\mathbf{X}}}_{M} \tag{2.20}$$

(see, for instance, Fortin et al., 2011; Donzé, 2013). The reweighting factor thus fails to match the two distributions perfectly, which means that the residual part of the estimated structure effect will not be equal to 0 (see, for instance, Fortin et al., 2011; Donzé, 2013).

The structure effect in Equation (2.18) can be further divided in the following elements

$$\left(\widehat{\overline{\mathbf{X}}}_{M}^{\top}\widehat{\beta}_{M} - \widehat{\overline{\mathbf{X}}}_{F|M}^{DFL^{\top}}\widehat{\beta}_{F}^{DFL}\right) = \widehat{\overline{\mathbf{X}}}_{M}^{\top}\left(\widehat{\beta}_{M} - \widehat{\beta}_{F}^{DFL}\right) + \left(\widehat{\overline{\mathbf{X}}}_{M} - \widehat{\overline{\mathbf{X}}}_{F|M}^{DFL}\right)\widehat{\beta}_{F}^{DFL}, \tag{2.21}$$

where $\widehat{\overline{\mathbf{X}}}_{F|M}^{DFL}$ and $\widehat{\beta}_{F}^{DFL}$ are defined in Equations (2.16) and (2.17), respectively (Fortin et al., 2011). The first element of the right-hand side of Equation (2.21) is the *pure effect* and the second the *residual effect* or the *total reweighting error* (Fortin et al., 2011). The pure effect is the actual unexplained part of the wage difference. The residual effect contains the misfit of the model, in other words, what the reweighting factor fails to match between men's and women's distribution of characteristics. This method allows for the construction of a counterfactual wage distribution. This in turn allows for the comparison between this new distribution and the observed wage distributions of women and men.

2.5 Conclusions

In this chapter, we presented two methods related to the decomposition of wage differences. The seminal work of Blinder (1973) and Oaxaca (1973) opened the floor to this topic and numerous researchers have suggested other methods which extend it. The BO method is straightforward and relies on linear regression, however it can only be used for differences in average wages. The reweighting method of DiNardo et al. (1996) allows the decomposition of wage differences at other points than the mean, namely quantiles in our application. Their method is based on logistic regression for the estimation of the reweighting factor. The idea of the method is to reweigh the distribution of characteristics in one group, such that they are similar to those in the other group. The idea is intuitive, because it leads to the comparison of the wages of two groups with the same characteristics. Moreover, questions such as whether the structure effect is the same for low- and high-paying jobs can be explored. This is what will be done in Chapter 3.

Chapter 3

The calibration approach

Abstract

In this chapter, we propose an approach to estimate the structure and the composition effects at different quantiles. This approach is based on the calibration method of Deville and Särndal (1992). The idea is similar to that of DiNardo et al. (1996) in that the estimation of wage discrimination can be done by comparing the wages of two groups with similar characteristics. Therefore, we calibrate women's characteristics on men's by computing calibration weights. These calibration weights are then used to estimate different parameters of women's counterfactual wage distribution. The calibration approach is a generalization of the method of Blinder (1973) and Oaxaca (1973) and of the reweighting method of DiNardo et al. (1996), as it will be shown in this chapter.[1]

3.1 The calibration method

The calibration method was introduced by Deville and Särndal (1992). The idea behind the technique is to make use of the information known at the population level on some auxiliary variables to estimate a function of a variable of interest. Usually, the auxiliary variables and the variable of interest are correlated.

Assuming that the sampling weights d_k are available and that the totals of auxiliary information at the population level given by

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k,$$

are known, new weights $w_k, k \in S$ should be constructed, such that the following constraint (or calibration equation) is respected

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{3.1}$$

[1] This chapter is a reprint of Anastasiade and Tillé (2017a) and Anastasiade and Tillé (2017b)

Obviously, there is an infinite number of solutions for the constraint in (3.1), therefore weights w_k as close as possible to the sampling weights d_k should be found. This involves minimizing a pseudo-distance criterion defined as

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k}, \tag{3.2}$$

where q_k are coefficients that indicate the importance of each unit in the sample. If all units have the same importance, $q_k = 1$, for each $k \in S$. In what follows, we will assume that $q_k = 1$ for all $k \in S$. The function $G_k(.,.)$ should be convex and positive, such that

$$G_k(d_k, d_k) = 0.$$

By minimizing (3.2) subject to (3.1), using a Lagrangian function yields

$$w_k = d_k F_k(\mathbf{x}_k^\top \lambda), \tag{3.3}$$

where $F_k(\mathbf{x}_k^\top \lambda)$ is called the calibration function. The function $d_k F_k(\mathbf{x}_k^\top \lambda)$ is the inverse of $g_k(\mathbf{x}_k^\top \lambda, w_k)/q_k, k \in S$, where

$$g_k(w_k, d_k) = \frac{\partial G_k(w_k, d_k)}{\partial w_k},$$

and the vector λ contains the Lagrange multipliers.

The weights are determined by solving in λ the calibration equations that become

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in S} d_k F_k(\mathbf{x}_k^\top \lambda) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k,$$

where $F_k(\mathbf{x}_k^\top \lambda)$ is the calibration function. The resulting calibration estimation of $Y = \sum_{k \in U} y_k$ is

$$\widehat{Y} = \sum_{k \in S} d_k y_k F_k(\mathbf{x}_k^\top \lambda). \tag{3.4}$$

The calibration method yields different weights, depending on the pseudo-distance $G_k(w_k, d_k)$. Deville and Särndal (1992) propose a list of pseudo-distances that can be used to calibrate the weights. Table 1 shows the two pseudo-distances that will be used in the following sections.

Table 3.1 Pseudo-distances for calibration

Pseudo-distance function	$G_k(w_k, d_k)$	$g_k(w_k, d_k)$	$F(\mathbf{x}_k^\top \lambda)$
Chi-squared	$\frac{(w_k - d_k)^2}{2d_k}$	$\frac{w_k}{d_k} - 1$	$1 + q_k \mathbf{x}_k^\top \lambda$
Entropy	$w_k \log \frac{w_k}{d_k} + d_k - w_k$	$\log \frac{w_k}{d_k}$	$\exp(q_k \mathbf{x}_k^\top \lambda)$

3.2 The proposed approach

In the current context, the auxiliary variables that are used in the calibration process are some selected characteristics measured for every individual. The aim is to 'divert' the calibration technique in order to compute a weighting system that adjusts the totals of the auxiliary variables of women on the totals of men. The variable of interest is the logarithm of the wage.

In the women sample, new weights w_k close to d_k are computed, such that $\sum_{k \in S_F} G(w_k, d_k)$ is minimized. The following calibration equation is satisfied

$$\sum_{k \in S_F} w_k \mathbf{x}_k = \widehat{\widehat{\mathbf{X}}}_M, \tag{3.5}$$

where the vector $\widehat{\widehat{\mathbf{X}}}_M$ stores the totals of men's characteristics adjusted on the total of the weights of the women over the total of the weights of the men, that is

$$\widehat{\widehat{\mathbf{X}}}_M = \frac{\sum_{k \in S_F} d_k}{\sum_{k \in S_M} d_k} \sum_{k \in S_M} d_k \mathbf{x}_k.$$

Dividing the calibration Equation (3.5) by $\sum_{k \in S_F} d_k$ yields

$$\frac{\sum_{k \in S_F} w_k \mathbf{x}_k}{\sum_{k \in S_F} d_k} = \frac{\widehat{\widehat{\mathbf{X}}}_M}{\sum_{k \in S_F} d_k} = \widehat{\overline{\mathbf{X}}}_M. \tag{3.6}$$

So with the new weights w_k, the new women's means of characteristics are equal to those of men. Another interesting equality is

$$\sum_{k \in S_F} w_k = \sum_{k \in S_F} d_k, \tag{3.7}$$

which holds because $x_{k1} = 1, k \in S_M$ and calibration is performed on it. If

$$\widehat{\widetilde{\mathbf{X}}}_M = \frac{\sum_{k \in S_F} w_k \mathbf{x}_k}{\sum_{k \in S_F} w_k},$$

by putting together Equations (3.6) and (3.7), this means that

$$\widehat{\widetilde{\mathbf{X}}}_M = \widehat{\overline{\mathbf{X}}}_M. \tag{3.8}$$

After choosing a pseudo-distance $G_k(u_k, w_k)$, the new weights can be determined by solving in λ the calibration equations

$$\sum_{k \in S_F} w_k \mathbf{x}_k F_k(\lambda^\top \mathbf{x}_k) = \widehat{\widetilde{\mathbf{X}}}_M, \tag{3.9}$$

and next by computing the weights by

$$u_k = w_k F_k(\lambda^\top \mathbf{x}_k).$$

Women's counterfactual wage mean estimator is thus given by

$$\widehat{\overline{Y}}_{F|M} = \frac{\sum_{k \in S_F} w_k y_k}{\sum_{k \in S_F} d_k} = \frac{\sum_{k \in S_F} w_k y_k}{\sum_{k \in S_F} w_k}.$$

3.2.1 Linear calibration

Result 2 *Women's counterfactual wage mean obtained using linear calibration is equal to the counterfactual wage mean obtained using the weighted BO method, i.e.* $\widehat{\overline{Y}}_{F|M} = \widehat{\widetilde{\mathbf{X}}}^\top \widehat{\beta}_F$.

Proof
In order to determine the vector λ in the case when the chi-squared pseudo-distance is used, the following equation must be solved

$$\begin{aligned} \widehat{\widetilde{\mathbf{X}}}_M &= \sum_{k \in S_F} d_k \mathbf{x}_k F(\mathbf{x}_k^\top \lambda) = \sum_{k \in S_F} d_k \mathbf{x}_k (1 + \mathbf{x}_k^\top \lambda) \\ &= \sum_{k \in S_F} d_k \mathbf{x}_k + \left(\sum_{k \in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top \right) \lambda. \end{aligned}$$

Thus,

$$\lambda = \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\widehat{\widehat{\mathbf{X}}}_M - \sum_{k\in S_F} d_k \mathbf{x}_k\right) = \mathbf{T}^{-1}\left(\widehat{\widehat{\mathbf{X}}}_M - \widehat{\mathbf{X}}_F\right),$$

where

$$\mathbf{T} = \sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top.$$

Thus

$$w_k = d_k F(\mathbf{x}_k^\top \lambda) = d_k \left\{1 + \mathbf{x}_k^\top \mathbf{T}^{-1}\left(\widehat{\widehat{\mathbf{X}}}_M - \widehat{\mathbf{X}}_F\right)\right\}.$$

Using the result from the previous equation, we obtain

$$\begin{aligned} \widehat{Y}_{F|M}^{LC} &= \sum_{k\in S_F} d_k F(\mathbf{x}_k^\top \lambda) y_k \\ &= \sum_{k\in S_F} d_k y_k + \left(\widehat{\widehat{\mathbf{X}}}_M - \widehat{\mathbf{X}}_F\right)^\top \mathbf{T}^{-1} \sum_{k\in S_F} d_k \mathbf{x}_k y_k, \end{aligned} \tag{3.10}$$

where $\widehat{Y}_{F|M}^{LC}$ denotes the total of the logarithm of the wage in the women sample, when the total is constructed using the chi-squared pseudo-distance. Let

$$\widehat{\beta}_F = \mathbf{T}^{-1} \sum_{k\in S_F} d_k \mathbf{x}_k y_k.$$

Vector $\widehat{\beta}_F$ has already been defined in the same way in Equation (2.7) for the weighted BO method. Equation (3.10) is rewritten as

$$\begin{aligned} \widehat{Y}_{F|M}^{LC} &= \sum_{k\in S_F} d_k y_k + \left(\widehat{\widehat{\mathbf{X}}}_M - \widehat{\mathbf{X}}_F\right)^\top \widehat{\beta}_F \\ &= \widehat{Y}_F + \left(\widehat{\widehat{\mathbf{X}}}_M - \widehat{\mathbf{X}}_F\right)^\top \widehat{\beta}_F \\ &= \widehat{\widehat{\mathbf{X}}}_M^\top \widehat{\beta}_F, \end{aligned} \tag{3.11}$$

because under the condition of Result 1, $\widehat{\mathbf{X}}_F^\top \widehat{\beta}_F = \widehat{Y}_F$. By dividing (3.11) by $\sum_{k\in S_F} w_k$, Result 2 is obtained. □

Using the chi-squared pseudo-distance, the resulting weights have no bounds. This means that the calibration weights might be negative. Even though this cali-

bration instance yields the same results as the BO method for average wages, we advocate for the use of an instance that gives nonnegative weights.

3.2.2 Raking-ratio calibration

The second instance of calibration uses the entropy pseudo-distance. It is also known as "raking-ratio" calibration. Using the entropy pseudo-distance, Equation (3.5) becomes

$$\widehat{\widehat{\mathbf{X}}}_M = \sum_{k \in S_F} d_k \mathbf{x}_k F(\mathbf{x}_k^\top \lambda) = \sum_{k \in S_F} d_k \mathbf{x}_k \exp(\mathbf{x}_k^\top \lambda). \tag{3.12}$$

This resulting system of equations cannot be solved analytically. However, the value of λ can be found through the Newton-Raphson algorithm.

The equation (3.4) can be now written as

$$\widehat{Y}_{F|M}^{RRC} = \sum_{k \in S_F} d_k \exp(\mathbf{x}_k^\top \lambda) y_k,$$

where $\widehat{Y}_{F|M}^{RRC}$ denotes the total of the logarithm of the wage in the women sample, when the total is constructed using the raking-ratio calibration. The counterfactual wage mean of women is written as

$$\widehat{\overline{Y}}_{F|M}^{RRC} = \frac{\sum_{k \in S_F} d_k \exp(\mathbf{x}_k^\top \lambda) y_k}{\sum_{k \in S_F} d_k \exp(\mathbf{x}_k^\top \lambda)}.$$

The equation above is very similar to Equation (2.19). The only difference lies in the estimation of the parameters λ and γ. The vector λ contains the Lagrangian multipliers solving Equation (3.12) under constraint (3.1), while the vector γ is found through maximum likelihood.

After computing the calibration weights w_k defined in (3.5) and by using the information in Equation (3.8), it results that

$$\widehat{\overline{\mathbf{X}}}_M = \widehat{\overline{\mathbf{X}}}_{F|M}^{RRC} = \frac{\sum_{k \in S_F} \mathbf{x}_k w_k}{\sum_{k \in S_F} w_k},$$

which ensures that the residual part of the structure effect defined in Equation (2.21) will equal 0. This is a solution to the problem shown in Section 2.43. This instance of calibration also remedies the issue of the negative weights that may arise when using the chi-squared pseudo-distance.

In Section 3.4 and 3.5, we will present two applications to real data. In the first one, presented in Anastasiade and Tillé (2017a), we compare the two calibration instances with the reweighted DFL approach. In the second one, presented in Anastasiade and Tillé (2017b), we compare the wages in the private sector with those in the public sector and we estimate the discrimination levels at different quantiles in both sectors.

3.3 Variance of women's counterfactual wage mean through linearization

To estimate the variance of women's counterfactual wage mean, we use the linearization method. Following Graf (2011), we compute the partial derivative of an estimator with respect to the sample indicator. This derivative provides the linearized variable that can be plugged in the variance estimator. We compute the partial derivatives of the estimators of the wage averages of men and women, and that of the estimator of the counterfactual wage average.

3.3.1 Linearization of the estimator of men's and women's wage means

We have the estimators of the wage means:

$$\widehat{\overline{Y}}_M = \frac{\sum_{k\in S_M} d_k y_k}{\sum_{k\in S_M} d_k} \quad \text{and} \quad \widehat{\overline{Y}}_F = \frac{\sum_{k\in S_F} d_k y_k}{\sum_{k\in S_F} d_k}.$$

When we linearize these estimators, we take the derivative of the indicator function. Therefore, we have

$$\frac{\partial \widehat{\overline{Y}}_M}{\partial I_k} \quad \text{and} \quad \frac{\partial \widehat{\overline{Y}}_F}{\partial I_k}.$$

We can also write $\widehat{\overline{Y}}_M$ as

$$\widehat{\overline{Y}}_M = \frac{\sum_{k\in U} d_k y_k I_k}{\sum_{k\in U} d_k I_k},$$

where

$$I_k = \begin{cases} 1, & \text{if } k \in S_M \\ 0, & \text{otherwise} \end{cases}$$

We compute the derivative and we have the quotient rule which implies that if

$$f(x) = \frac{g(x)}{h(x)},$$

then

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{[h(x)]^2}.$$

Therefore, in our case, $g(x) = \sum_{k\in U} d_k y_k I_k$ and $h(x) = \sum_{k\in U} d_k I_k$. This means that

$$g'(x) = d_k y_k$$

and

$$h'(x) = d_k.$$

Putting it all together, we have that

$$\begin{aligned}
f'(x) &= \frac{d_k y_k \sum_{k\in U} d_k I_k - d_k \sum_{k\in U} d_k y_k I_k}{[\sum_{k\in U} d_k I_k]^2} \\
&= \frac{d_k y_k \sum_{k\in S_M} d_k - d_k \sum_{k\in S_M} d_k y_k}{[\sum_{k\in S_M} d_k]^2} \\
&= \frac{d_k y_k}{\sum_{k\in S_M} d_k} - \frac{d_k}{\sum_{k\in S_M} d_k} \frac{\sum_{k\in S_M} d_k y_k}{\sum_{k\in S_M} d_k} \\
&= \frac{d_k y_k}{\sum_{k\in S_M} d_k} - \frac{d_k}{\sum_{k\in S_M} d_k} \widehat{\overline{Y}}_M \\
&= \frac{d_k (y_k - \widehat{\overline{Y}}_M)}{\sum_{k\in S_M} d_k}.
\end{aligned}$$

We next look for a set of weights v_k such that

$$w_k = v_k d_k, k \in S_F$$

and

$$\frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} w_k} = \frac{\sum_{k\in S_F} d_k}{\sum_{k\in S_M} d_k} \widehat{\overline{\mathbf{X}}}_M.$$

In other words,

$$\mathbf{A} = \sum_{k\in S_F} v_k d_k \mathbf{x}_k = \widehat{\overline{\mathbf{X}}}_M \sum_{k\in S_F} d_k.$$

In terms of dimension, we have that

- $\mathbf{x}_k$ is of size $p \times 1$,

- $\mathbf{x}_k^\top$ is of size $1 \times p$,
- λ is of size $p \times 1$,
- $\widehat{\overline{\mathbf{X}}}_M$ is of size $p \times 1$,
- $F(\mathbf{x}_k^\top \lambda)$ is a scalar,
- $\mathbf{A}$ is of size $p \times 1$.

Now when it comes to the derivation of $\mathbf{A}$,

$$\frac{\partial \mathbf{A}}{\partial I_k} = \frac{d_k \partial F(\mathbf{x}_k^\top \lambda)\mathbf{x}_k}{\partial I_k}$$

It is important to note that λ depends on I_k.

We have that $\dfrac{\partial \mathbf{A}}{\partial I_k} = \dfrac{\partial \sum_{k \in S_F} v_k d_k \mathbf{x}_k}{\partial I_k} = \dfrac{\partial \sum_{k \in S_F} F(\mathbf{x}_k^\top \lambda) d_k \mathbf{x}_k}{\partial I_k}$.
Therefore,

$$\frac{\partial \sum_{k \in S_F} F(\mathbf{x}_k^\top \lambda) d_k \mathbf{x}_k}{\partial I_k} = \frac{\partial \sum_{k \in U} F(\mathbf{x}_k^\top \lambda) d_k \mathbf{x}_k I_k}{\partial I_k}$$

Here, we have two functions that depend on I_k. They are

$$f(I) = d_k \mathbf{x}_k I_k \quad \text{and} \quad g(x) = F(\mathbf{x}_k^\top \lambda).$$

Using the product rule that says:

$$(f \cdot g)' = f' \cdot g + f \cdot g',$$

we can write $\mathbf{A}$ as

$$\mathbf{A} = d_1 F(\mathbf{x}_1^\top \lambda)\mathbf{x}_1 + \cdots + d_n F(\mathbf{x}_n^\top \lambda)\mathbf{x}_n.$$

When we derive with respect to I_1, we have

$$\begin{aligned}\frac{\partial \mathbf{A}}{\partial I_1} &= \frac{\partial [d_1\mathbf{x}_1 I_1 F(\mathbf{x}_1^\top \lambda) + \cdots + d_n F(\mathbf{x}_n^\top \lambda)\mathbf{x}_n I_n]}{\partial I_1} \\ &= d_1\mathbf{x}_1 F(\mathbf{x}_1^\top \lambda) + d_1 x_1 F'(\mathbf{x}_1^\top \lambda)\mathbf{x}_1^\top \frac{\partial \lambda}{\partial I_1} \\ &+ 0 + d_2\mathbf{x}_2 F'(\mathbf{x}_2^\top \lambda)\mathbf{x}_2^\top \frac{\partial \lambda}{\partial I_1} \\ &+ \dots \\ &+ 0 + d_n\mathbf{x}_n F'(\mathbf{x}_n^\top \lambda)\mathbf{x}_n^\top \frac{\partial \lambda}{\partial I_1} \\ &= d_1 v_1 \mathbf{x}_1 + \frac{\partial \lambda}{\partial I_1} \sum_{j=1}^{n} d_j \mathbf{x}_j F'(\mathbf{x}_j^\top \lambda)\mathbf{x}_j^\top\end{aligned}$$

This is valid for all $j \in S_F$.
We also know that $\mathbf{A} = \widehat{\overline{\mathbf{X}}}_M \sum_{k\in S_F} d_k = \widehat{\overline{\mathbf{X}}}_M \sum_{k\in U} d_k I_k$.
The derivative with respect to $I_j, j \in S_F$ is

$$\frac{\partial \mathbf{A}}{\partial I_j} = d_j v_j \mathbf{x}_j + \left[\sum_{k=1}^{n} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda)\mathbf{x}_k^\top\right] \frac{\partial \lambda}{\partial I_j} = d_j \widehat{\overline{\mathbf{X}}}_M.$$

Solving for $\dfrac{\partial \lambda}{\partial I_j}$, we have

$$d_j v_j \mathbf{x}_j - d_j \widehat{\overline{\mathbf{X}}}_M = -\left[\sum_{k=1}^{n} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda)\mathbf{x}_k'\right] \frac{\partial \lambda}{\partial I_j}.$$

Therefore,

$$\frac{\partial \lambda}{\partial I_j} = -\Big[\sum_{k=1}^{n} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda)\mathbf{x}_k^\top)^{-1}\Big] d_j (v_j \mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M).$$

When $j \in S_M$, we compute the derivative of

$$\mathbf{A} = \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l} \sum_{l\in S_M} d_l \mathbf{x}_l.$$

$\mathbf{A}$ can be written as

$$\frac{\sum_{k\in U} d_k I_k}{\sum_{l\in U} d_l I_l} \sum_{l\in U} d_l \mathbf{x}_l I_l.$$

This means that there are three functions that depend on I_j:

- $f(I) = \sum_{k \in U} d_k I_k$,
- $g(I) = \sum_{l \in U} d_l \mathbf{x}_l I_l$,
- $h(I) = \sum_{l \in U} d_l I_l$.

The derivative of the expression $\frac{f(I)g(I)}{h(I)}$ will be computed. Applying the product and the quotient rule, this means that the derivative will be

$$\begin{aligned} \left[\frac{f(I)g(I)}{h(I)}\right]' &= \frac{[f(I)g(I)]'h(I) - f(I)g(I)h'(I)}{[h(I)]^2} \\ &= \frac{[f'(I)g(I) + f(I)g'(I)]h(I) - f(I)g(I)h'(I)}{[h(I)]^2}. \end{aligned} \tag{3.13}$$

Taking each function separately, we compute the derivatives with respect to the indicator I_j, where $j \in S_M$. We obtain

$$\begin{aligned} \frac{\partial f(I)}{\partial I_j} &= \frac{\partial \sum_{k \in U} d_k I_k}{\partial I_j} = 0, j \neq k. \\ \frac{\partial g(I)}{\partial I_j} &= \frac{\partial \sum_{l \in U} d_l \mathbf{x}_l I_l}{\partial I_j} = d_l \mathbf{x}_l, j = l. \\ \frac{\partial h(I)}{\partial I_j} &= \frac{\partial \sum_{l \in U} d_l I_l}{\partial I_j} = d_l, j = l. \end{aligned}$$

Equation (3.13) then becomes

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial I_j} &= \frac{[f'(I)g(I) + f(I)g'(I)]h(I) - f(I)g(I)h'(I)}{[h(I)]^2} \\ &= \frac{[d_j \mathbf{x}_j \sum_{k \in U} d_k I_k] \sum_{l \in U} d_l I_l - d_l \sum_{k \in U} d_k I_k \sum_{l \in U} d_l \mathbf{x}_l I_l}{[\sum_{l \in U} d_l I_l]^2} \\ &= \frac{d_j \mathbf{x}_j \sum_{k \in S_F} d_k] \sum_{k \in S_M} d_l - d_l \sum_{k \in S_F} d_k \sum_{l \in S_M} d_l \mathbf{x}_l}{[\sum_{l \in S_M} d_l]^2} \\ &= \frac{d_j \mathbf{x}_j \sum_{k \in S_F} d_k \sum_{k \in S_M} d_l}{[\sum_{l \in S_M} d_l]^2} - \frac{d_l \sum_{k \in S_F} d_k \sum_{l \in S_M} d_l \mathbf{x}_l}{[\sum_{l \in S_M} d_l]^2} \\ &= d_j \mathbf{x}_j \frac{\sum_{k \in S_F}}{\sum_{l \in S_M} d_l} - d_l \widehat{\overline{\mathbf{X}}}_M \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \\ &= d_j \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} [\mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M] \end{aligned}$$

We have to find $\frac{\partial \lambda}{\partial I_j}, j \in S_M$. If we write that

$$\mathbf{A} = \sum_{k \in U} F(\mathbf{x}_k^\top \lambda) d_k \mathbf{x}_k I_k,$$

again, we have two functions that depend on I_j. These are $f = d_j \mathbf{x}_j I_j$ and $g(I) = F(\mathbf{x}_k^\top \lambda)$. Thus

$$\begin{aligned}
\frac{\partial \mathbf{A}}{\partial I_j} &= f'g + g'f \\
&= 0 + \sum_{k \in S_F} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k^\top \frac{\partial \lambda}{\partial I_j} \\
&= \frac{\partial \lambda}{\partial I_j} \sum_{k \in S_F} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k^\top .
\end{aligned}$$

Therefore, we have that

$$d_l \frac{\sum_{k \in S_F}}{\sum_{l \in S_M} d_l} [\mathbf{x}_l - \widehat{\widehat{\mathbf{X}}}_M] = \frac{\partial \lambda}{\partial I_j} \sum_{k \in S_F} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k^\top ,$$

meaning that

$$\frac{\partial \lambda}{\partial I_j} = [\sum_{k \in S_F} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k^\top]^{-1} d_l \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} [\mathbf{x}_l - \widehat{\widehat{\mathbf{X}}}_M]$$

3.3.2 Linearization of the estimator of women's counterfactual wage mean

Now we compute the derivative of $\widehat{\widehat{Y}}_{F|M}$ with respect to the indicator function. We write $\widehat{\widehat{Y}}_{F|M}$ as

$$\widehat{\widehat{Y}}_{F|M} = \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} d_k}.$$

We look at two cases: when $j \in S_F$ and when $j \in S_M$.

In the first case, when $j \in S_F$, we have that

$$\begin{aligned}
\frac{\partial \widehat{\overline{Y}}_{F|M}}{\partial I_j} &= \frac{\left\{ d_j y_j F(\mathbf{x}_j^\top \lambda) + \left[\sum_{k \in S_F} \mathbf{x}_k^\top d_k y_k F'(\mathbf{x}_k^\top \lambda) \right] \frac{\partial \lambda}{\partial I_j} \right\} \sum_{k \in S_F} d_k - d_j \sum_{k \in S_F} d_k F(\mathbf{x}_k^\top \lambda) y_k}{[\sum_{k \in S_F} d_k]^2} \\
&= \frac{d_j y_j F(\mathbf{x}_j^\top \lambda) \sum_{k \in S_F} d_k}{[\sum_{k \in S_F} d_k]^2} \\
&+ \frac{\left\{ \sum_{k \in S_F} \mathbf{x}_k^\top d_k y_k F'(\mathbf{x}_k^\top \lambda) \sum_{k \in S_F} d_k \right\} \frac{\partial \lambda}{\partial I_j}}{[\sum_{k \in S_F} d_k]^2} \\
&= -\frac{d_j \sum_{k \in S_F} d_k F(\mathbf{x}_k^\top \lambda) y_k}{[\sum_{k \in S_F} d_k]^2} \\
&= \frac{d_j y_j F(\mathbf{x}_j^\top \lambda)}{\sum_{k \in S_F} d_k} + \frac{d_j (v_j \mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M)(-[\sum_{k=1}^n d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k'])^{-1}] \sum_{k \in S_F} \mathbf{x}_k^\top d_k y_k F'(\mathbf{x}_k^\top \lambda)}{\sum_{k \in S_F} d_k} \\
&- \frac{d_j \widehat{\overline{Y}}_{F|M}}{\sum_{k \in S_F} d_k} \\
&= \frac{d_j}{\sum_{k \in S_F} d_k} [v_j y_j + (v_j \mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M) \widehat{\mathbf{B}}_F - \widehat{\overline{Y}}_{F|M}],
\end{aligned}$$

where $\widehat{\mathbf{B}}_F$ are the regression coefficients in group F.

In the second case, when $j \in S_M$, we have

$$\begin{aligned}
\frac{\partial \widehat{\overline{Y}}_{F|M}}{\partial I_j} &= \frac{\partial \frac{\sum_{k \in S_F} v_k d_k y_k}{\sum_{k \in S_F} d_k}}{\partial I_j} \\
&= \frac{\partial \frac{\sum_{k \in U} F(\mathbf{x}_k^\top \lambda) d_k y_k I_k}{\sum_{k \in U} d_k I_k}}{\partial I_j}.
\end{aligned}$$

Again, we have three functions that depend on I_j:

- $f(I) = F(\mathbf{x}_k^\top \lambda)$,
- $g(I) = \sum_{k \in U} d_k y_k I_k$,
- $h(I) = \sum_{k \in U} d_k I_k$.

When we derive them, we obtain

- $f'(I) = \dfrac{\partial F(\mathbf{x}_k^\top \lambda)}{\partial \lambda} \dfrac{\partial \lambda}{\partial I_j}$
- $g'(I) = 0, k \neq j$
- $h'(I) = 0, k \neq j$

It must be noted that in $f(I)$, there is λ which depends on I_j, when $j \in S_M$. Again, we compute the derivative of a function of the form

$$t = \frac{f \cdot g}{h}.$$

This is

$$t' = \frac{(f \cdot g)' h - f \cdot g \cdot h'}{h^2}.$$

When computing the derivative of the denominator, we do the same as when computing the derivative $\dfrac{\partial \mathbf{A}}{\partial I_j}$. Therefore, we have that

$$\begin{aligned}
t' &= \frac{\sum_{k \in S_F} \dfrac{\partial F(\mathbf{x}_k^\top \lambda)}{\partial \lambda} \dfrac{\partial \lambda}{\partial I_j} \mathbf{x}_k^\top y_k I_k \sum_{k \in S_F} d_k}{[\sum_{k \in S_F} d_k]^2} \\
&= \frac{\partial \lambda}{\partial I_j} \frac{\sum_{k \in S_F} F'(\mathbf{x}_k^\top \lambda) \mathbf{x}_k^\top y_k}{\sum_{k \in S_F} d_k} \\
&= [\sum_{k \in S_F} d_k \mathbf{x}_k F'(\mathbf{x}_k \lambda) \mathbf{x}_k^\top]^{-1} d_l \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} [\mathbf{x}_l - \widehat{\overline{\mathbf{X}}}_M] \frac{\sum_{k \in S_F} F'(\mathbf{x}_k^\top \lambda) \mathbf{x}_k^\top y_k}{\sum_{k \in S_F} d_k} \\
&= \frac{1}{\sum_{l \in S_M d_l}} d_j [\mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M] \widehat{\mathbf{B}}_F.
\end{aligned}$$

Thus, the linearized variable is:

$$t_j = \begin{cases} \dfrac{d_j \left[v_j y_j - \widehat{\overline{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M \right)^\top \mathbf{B}_F \right]}{\sum_{k \in S_F} d_k} & \text{if } j \in S_F \\ \dfrac{d_j \left(\mathbf{x}_j - \widehat{\overline{\mathbf{X}}}_M \right)^\top \mathbf{B}_F}{\sum_{\ell \in S_M} d_\ell} & \text{if } j \in S_M. \end{cases}$$

The linearized variable must only be plugged in the variance estimator corresponding to the sampling design. That is we have:

$$\widehat{\text{var}}(\widehat{\overline{Y}}_{F|M}) \approx \widehat{\text{var}}(\sum_{k \in S} t_k w_k) = \widehat{\text{var}}(\sum_{k \in S_F} t_k w_k + \sum_{l \in S_M} t_l w_l).$$

Note that the variance of the counterfactual depends on the variance computed for the sample of men for the part that is explained by the regression and for the sample of women for the part that remains unexplained.

3.4 First application to real data

3.4.1 The dataset

The dataset used contains information collected in 2008 by the Swiss Federal Statistical Office from a survey called Survey on Earnings Structure. A questionnaire was sent to public and private organizations from the secondary and tertiary sectors to collect information on particular aspects. These aspects include the size of the organization, employment contract types and employee remuneration within the organization. The questionnaire was filled in by an authorized member of the organization and not by employees. This enhances data reliability and makes it less prone to approximations. The analyses that follow were restricted to the private sector. The valid observations that were included were the individuals with no missing values, who worked more than one hour per week and whose difference between the age and the work experience was greater than or equal to 15 (according to the Swiss employment laws, this represents the legal minimum age to be eligible to work). Thus, 29,048 cases were excluded from the original dataset. The final dataset contains 647,139 men and 435,507 women. The sampling weights are also provided in the dataset by the Swiss Federal Statistical Office, therefore no treatment or computation of these weights were done in this application.

In the next tables, the values expressed in Swiss francs are given in parentheses. However, the figures are plotted using the logarithms of the wages. The values are obtained taking the survey weights into consideration.

Table 3.2 contains the median and wage averages for the entire sample and for women and men.

Table 3.2 Wage mean and median computed for the entire dataset, women and men, in Swiss francs

	Mean	Median
Entire dataset	6977	5905
Women	5843	5220
Men	7725	6346

Both the wage mean and the median values of men are above the values in the entire dataset, while those of women are below. Table 3.3 shows the distribution of women and men in low and high paying jobs. The weighted quantiles of the wage of the entire dataset are computed on the first row. The following two lines show the cumulative proportions of women and men who earn less than the value of the quantile.

Table 3.3 Weighted quantiles of the logarithm of the wage and proportions of women and men who earn less than the value that represents a particular quantile of the wage computed for the entire dataset (values in Swiss francs are given in parantheses)

	Quantile										
	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Logarithm of wage	7.89	8.27	8.39	8.50	8.59	8.68	8.78	8.89	9.03	9.27	10.09
	(2683)	(3897)	(4412)	(4905)	(5400)	(5905)	(6488)	(7233)	(8380)	(10667)	(24202)
Cumulative proportion of women	0.02	0.17	0.32	0.43	0.53	0.63	0.72	0.81	0.89	0.96	1
Cumulative proportion of men	0.006	0.06	0.12	0.21	0.31	0.42	0.52	0.63	0.74	0.86	0.99

While 43% of women have a wage smaller than CHF 4905 (as opposed to only 21% of men), there are only 11% of women who earn between CHF 8380 and CHF 24202 (compared to 25% of men). Moreover, 63% of women earn below the median value of the wage of the entire dataset, compared to only 42% of men. The potential generating mechanisms of this allocation should be investigated. Nevertheless, it is not the purpose of this paper. For a closer insight into the distribution of the wages in each sample, Table 3.4 displays the weighted quantiles of the logarithms of the wages of women and men, as well as the difference between them. A surprising value of the

Table 3.4 Wages of women and men and the difference between wages of men and women, in terms of logarithms (values in Swiss francs are given in parantheses)

	Quantile										
	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Women	7.80	8.19	8.30	8.38	8.47	8.56	8.66	8.76	8.88	9.06	9.67
	(2432)	(3602)	(4005)	(4344)	(4756)	(5220)	(5743)	(6353)	(7154)	(8577)	(15761)
Men	8.01	8.36	8.49	8.58	8.67	8.76	8.86	8.98	9.14	9.38	10.26
	(3000)	(4259)	(4850)	(5344)	(5820)	(6346)	(7012)	(7908)	(9291)	(11905)	(28571)
Difference	0.21	0.17	0.19	0.21	0.20	0.20	0.20	0.22	0.26	0.33	0.59
	(568)	(657)	(845)	(1000)	(1064)	(1126)	(1269)	(1555)	(2137)	(3328)	(12810)

difference between the wages is observed at the quantile of order 1%. It is expected that these jobs fall into the type of jobs that do not require extensive qualifications or high education levels. While only 0.6% of men occupy such positions (see Table 3.3), they earn more than the 2% of women who have similar jobs. Figure 3.1 shows the data presented in Table 3.4 above in a graphical form.

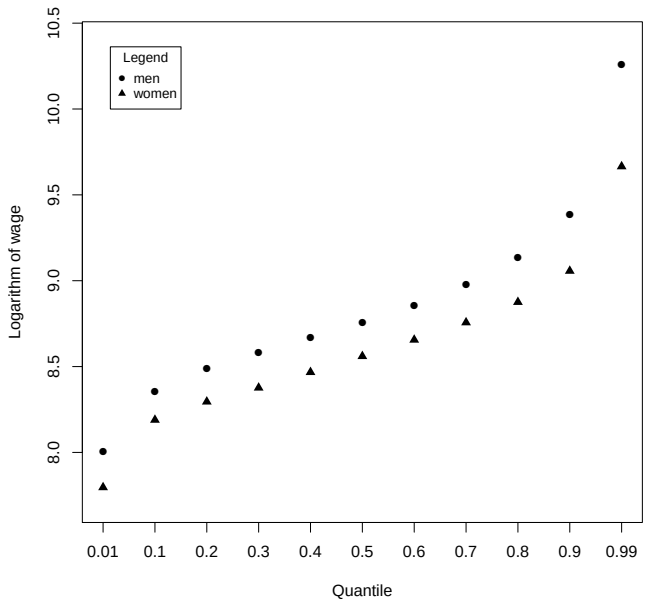


Fig. 3.1 Weighted quantiles of the logarithm of the wages of women and men.

The distance between the two sets of points increases toward the higher-level quantiles, which means that the differences between the wages become higher. It has to be established how much of these differences are not attributable to differing characteristics of women and men. As a final graphical evidence of wage inequalities, Figure 3.2 shows the distributions of the logarithm of the wages of women and men.

3.4.2 The model

The regression model used includes eight explanatory variables:

- education level : nominal variable with 9 categories indicating the highest educational degree attained;
- number of years of service in the current position (proxy for work experience);
- qualification requirements : ordinal variable with 4 levels indicating the level of qualification required for the position;
- region of the institution: nominal variable with 7 categories;
- economic sector: nominal variable with 10 categories;

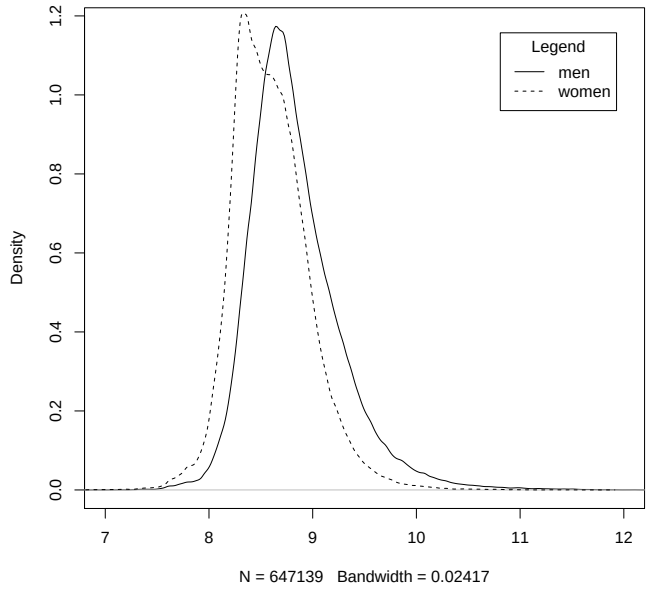


Fig. 3.2 Estimated densities of the logarithm of the wages of women and men.

- degree of occupation - the occupation rate of the employee (if the value is 1, then the employee works full-time);
- age : the actual age;
- the square of the age: the square of the age is also included, because it has been observed that the wage increases until a certain age and decreases afterwards (see, for instance Williams, 2010).

A description of the categorical variables can be found in Appendix A. The model was selected from a number of models with several variables using the AIC criterion. The dependent variable is the logarithm of the standardized wage. By standardized wage of an individual, we mean the wage computed for that individual if they worked full-time. This variable is provided by the Swiss Federal Statistical Office in the dataset.

3.4.3 Weights and counterfactual distributions

This section only includes results in terms of logarithms. When using the BO method and the previous regression model, the difference between average wages of men and women is 0.23, out of which only 0.09 represent the explained part and 0.14

the unexplained part. The results obtained through the methods presented above are compared. The counterfactual wage distribution using the weighted DFL and the calibration methods was estimated by reweighting women's wage by the factor ψ and by the calibration weights, respectively. The calibration method through the chi-squared pseudo-distance is denoted as "linear", the calibration through the entropy pseudo-distance as "raking-ratio" and the method proposed by DiNardo et al. (1996) adjusted to take the survey weights into consideration as "weighted DFL". First, Table 3.5 shows the minimum and the maximum values, as well as the standard deviations of the weights, obtained using the linear calibration, the raking-ratio calibration and the weighted DFL method.

Table 3.5 Minimum, maximum and standard deviation of the weights.

Method	Minimum	Maximum	Standard deviation
Linear	-39.06	319.8	4.97
Raking-ratio	0.0011	904.7	6.79
Weighted DFL	0.0022	804.4	6.16

The linear case yields the same results as the weighted BO method. However, as seen in Table 3.19, this particular case yields negative weights. There were 69,553 such weights (14.59%). The raking-ratio alternative always yields positive weights, however, the standard deviation of the weights is higher. The weighted DFL factor has a smaller standard deviation than the weights obtained by the raking-ratio calibration method. There are 1319 cases where the conditional probability of being a man is larger than 0.98. Originally, the DFL factor is multiplied by the ratio between the sum of sampling weights of women and the sum of sampling weights of men. Since $\widehat{a}$ is smaller than one, the reweighting factor will shrink. If on the other hand, $\widehat{a}$ is larger than one (for instance, for sectors such as the public sector), the reweighting factor might be larger. Table 3.6 shows the structure effect estimated at the average levels of the wages. The two calibration approaches yield equal structure and composition effects. Using the DFL reweighting factor, results in a slightly lower structure effect and a higher composition effect than the other two methods.

Given that negative weights are obtained in the first case of calibration, the corresponding estimated density can not be graphically represented. Only women's counterfactual wage distributions constructed using the raking-ratio and the DFL reweighting factor are constructed. They are presented in Figure 3.3.

Figure 3.3 shows that the two counterfactual wage distributions are very close to each other around the tails. However, toward the middle, the two methods do

Table 3.6 Estimated composition and structure effects in the difference in mean averages.

Method	Total	=	Composition effect	+	Structure effect
Linear	0.23		0.09		0.14
Raking-ratio	0.23		0.09		0.14
Weighted DFL	0.23		0.10		0.13

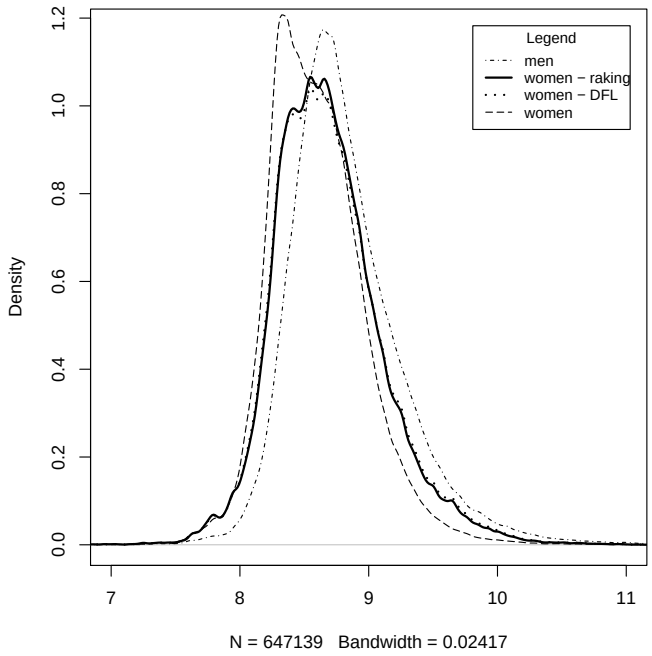


Fig. 3.3 Estimated densities of the logarithm of the wages of women and men and the counterfactual distributions of the logarithm of the wage of women constructed using the raking-ratio and the weighted DFL factor, respectively.

not yield the same results. As previously mentioned, using DFL reweighting and calibration methods allow the estimation the composition and structure effects not only at the average levels, but also along the entire distribution. Table 3.7 displays the estimated structure and composition effects of the wage differences between men and women computed using the three methods at some selected quantiles.

The proportion of the structure effect of the entire wage difference between men and women decreases as the order of the quantile increases. This means that for jobs with higher salaries, more of the wage differences can be explained by differences in group characteristics than for jobs with lower salaries. The raking-ratio and the

Table 3.7 Estimated composition and structure effects of the wage difference at selected quantiles.

Quantile	Method	Total	=	Composition effect (%)	+	Structure effect (%)
1%	Linear	0.21		0.01		0.20
				(3%)		(97%)
	Raking	0.21		-0.01		0.22
				(-3.5%)		(103.5%)
	Weighted DFL	0.21		-0.01		0.22
				(-3.4%)		(103.4%)
10%	Linear	0.17		0.05		0.12
				(28.8%)		(71.2)%
	Raking	0.17		0.04		0.13
				(22.4%)		(77.6%)
	Weighted DFL	0.17		0.03		0.14
				(19.4%)		(80.6%)
20%	Linear	0.20		0.07		0.13
				(34.2%)		(65.8%)
	Raking	0.19		0.06		0.13
				(29.7%)		(70.3%)
	Weighted DFL	0.19		0.05		0.14
				(28.2%)		(71.8%)
50%	Linear	0.19		0.09		0.10
				(46.3%)		(53.7%)
	Raking	0.20		0.09		0.11
				(44.7%)		(55.3%)
	Weighted DFL	0.20		0.09		0.11
				(45.7%)		(54.3%)
80%	Linear	0.26		0.11		0.15
				(43.9%)		(56.1%)
	Raking	0.26		0.12		0.14
				(46.5%)		(53.5%)
	Weighted DFL	0.26		0.13		0.13
				(50.8%)		(49.2%)
90%	Linear	0.33		0.15		0.18
				(46.0%)		(54.0%)
	Raking	0.33		0.17		0.16
				(51.6%)		(48.4%)
	Weighted DFL	0.33		0.19		0.14
				(58.0%)		(42.0%)
99%	Linear	0.60		0.24		0.36
				(40.0%)		(60.0%)
	Raking	0.60		0.27		0.33
				(45.3%)		(54.7%)
	Weighted DFL	0.59		0.29		0.30
				(49.4%)		(50.6%)

DFL reweighting factor yield similar results up to the quantile of order 90%. The composition effect at the first percentile is estimated to be negative, meaning that at this point, the differences in wages are due solely to discrimination.

Figure 3.4 shows the weighted quantiles of the logarithms of the wage of men, those of women and contrast the counterfactual distributions obtained through the raking-ratio calibration and the DFL reweighting factor. Because the linear calibration yielded negative weights, the same graph is not reproduced for it.

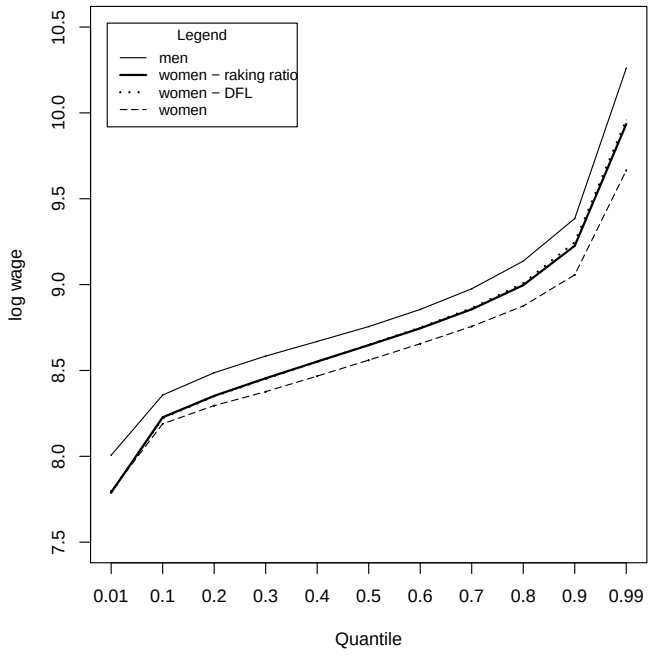


Fig. 3.4 Weighted quantiles of the logarithms of the wage of women and men and the weighted quantiles of the counterfactual distribution of the logarithm of the wage of women constructed using the raking-ratio calibration and the weighted DFL factor.

3.4.4 Further decomposition of the structure effect

A logistic model using the same covariates from Section 3.4.2 for the probability of being a man yields estimated values between 0.002 and 0.99. For the variables "years in the current position", "age" and "square of the age" the difference between the average values of men and the reweighted averages of women computed using the reweighting factor are the largest. In Equation (2.21), the structure effect is composed of the pure effect and the residual effect. Using the DFL reweighting factor, the residual effect equals -0.00474. In contrast, by using either one of the calibration techniques, in both cases, it equals 0. Moreover, the calibration approach allows overriding the computation of the counterfactual regression coefficients. This

is because the technique ensures the equality between the means $\widehat{\overline{\mathbf{X}}}_M$ and $\widehat{\overline{\mathbf{X}}}_{F|M}$. Calibration thus represents a generalization of the DFL reweighting factor technique, because it allows for a more precise estimation of the structure effect, since the resulting value only includes the pure part.

3.5 Second application to real data

3.5.1 The dataset

The dataset used emerged from the Survey on the Structure of Earnings, a survey that is sent to public and private institutions. The variables collected on employees include for instance the number of years in service in the current position, the education level or the qualification requirements. There are 307480 observations (128387 men and 179093 women) in the public sector and 667987 observations (391325 men and 276662 women) in the private sector. The observations taken into account are individuals between the ages of 15 and 65, who work more than 1 hour per week and who have a fixed monthly wage. The dataset includes observations for which all the variables have been observed, except for the education level. Following the guidelines of Strub and Stocker (2010), missing values of this variable were kept in the final dataset, to avoid the loss of a significant proportion of observations (7.6% in the private sector and 13.52% in the public sector). The variable of interest is the logarithm of the standardized full-time wages. This gives a common base for comparison. The sampling weights adjusted for non-response were provided by the Swiss Federal Statistical Office.

3.5.2 The calibration variables

The variables retained for calibration are

- education level - ordinal variable with 9 categories;
- number of years of service in the current position;
- region - nominal variable with 7 categories;
- professional position - ordinal variable with 5 categories (from senior management to no management position);
- age;
- the square of the age;
- qualification requirements - ordinal variable with 4 categories (the degree of task complexity required for the position);

• occupancy rate - binary variable (0 if the individual works less than 80% and 1 otherwise).

A description of the categorical variables used in this application can be found in Appendix A. All the variables are correlated with the variable of interest. They were selected using a stepwise regression using the AIC criterion. The choice of selecting the variables to calibrate on through a regression method is justified by the idea of rendering the results comparable with those obtained through the original decomposition technique and through the reweighting method of DiNardo et al. (1996). In order to draw comparisons between sectors, the calibration variables had to be the same both in the private and in the public sector. As such, the economic sector is not taken into account because there are some sectors where there are observations only in one sector, and not in the other one. For instance, there are no observations in the private sector for the activity "public administration" and no observations in the public sector for manufacturing related activities (for instance, tobacco or textiles). For the nominal and ordinal variables selected, the categories are the same in both sectors. A binary variable was created for each category in the context of the calibration approach. The first category of each qualitative variable was removed, in order to avoid multicollinearity with the constant. In the linear and the logistic models used in the Blinder-Oaxaca and the reweighting method of DiNardo et al. (1996), the first category of the qualitative variables were the reference categories. Calibration was done such that in each category, the proportion of women computed using the calibration weights that was in that category was matched to the proportion of men. The continuous variables were left unchanged in all three methods.

3.5.3 Descriptive results

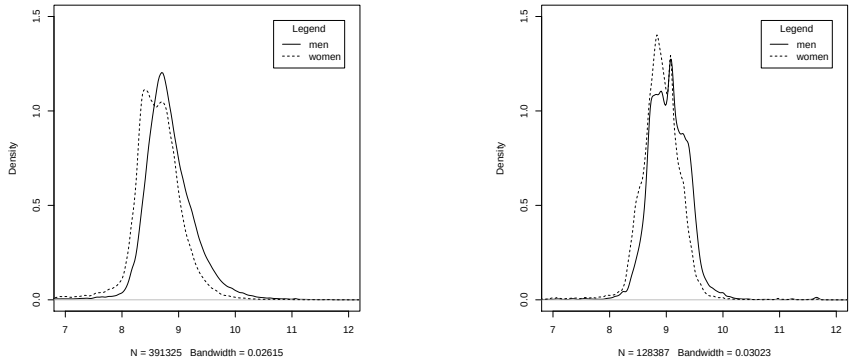
Table 3.8 displays the average and median wages in both sectors, for the entire sample, women and men. All wages are standardized for a full-time occupancy rate. This means that every individual has a reported wage as if they worked full-time. This provides a common base for comparison. The values in the public sector are higher than those in the private sector. In the private sector, the median in the women sample is smaller than the median in the entire dataset by around CHF 1000. In both sectors, the values in the sample of women are below those in the entire dataset, and the values in the sample of men above.

Figure 3.5 shows the wage densities (in logarithms) for women and men. In the public sector, the wage density function is higher for women than for men, while the

Table 3.8 Wage averages and medians for women and men in the private and public sectors in terms of Swiss francs in 2012.

	Private sector		Public sector	
	Median	Mean	Median	Mean
Entire dataset	6143	7194.92	7710	8411.60
Women	5534	6183.96	7318	7846.09
Men	6540	7886.82	8411	9155.15

opposite holds true for the private sector. In terms of dispersion, in both sectors, the densities of women's wages and men's wages are comparable.



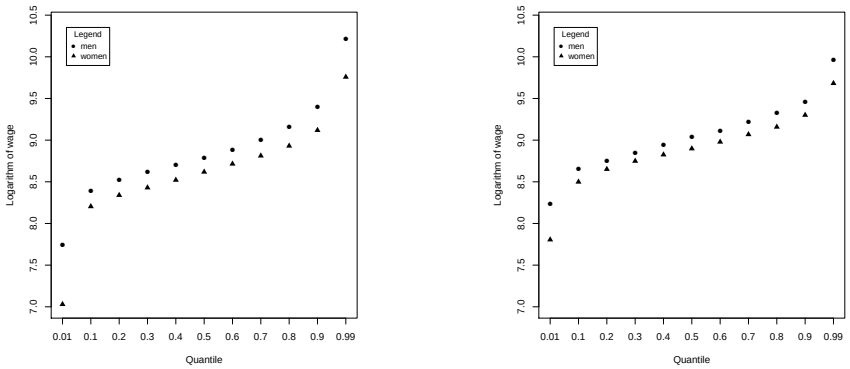
(a) private sector (b) public sector

Fig. 3.5 Wage (in logarithms) densities of women and men in the private (left panel) and public (right panel) sectors in 2012.

Figure 3.6 shows the wage quantiles in both sectors. Both figures have the same values on the vertical axis. Wages in the public sector tend to be higher than in the private sector, except for the last quantile. This might occur because of differences in education. 22% of individuals in the public sector have a university degree, compared to little under 8% in the private sector. It might also occur because on average, workers in the public sector have more work experience in the current position (9.26 years) than those in the private sector (7.81 years). In the public sector, 48.57% of employees work in jobs where the task complexities are at the highest level, compared to 26.42% in the private sector. These results are summarized in Table 3.9. At all quantiles, women earn less than men, with greater differences at quantiles 1% and 99%. The wage differences between men and women are smaller in the public sector than in the private sector all along the wage distribution. Another aspect is that the wages at the lower and upper values of the quantiles are more extreme in the private sector.

Table 3.9 Selected employee characteristics in the private and public sectors.

Characteristic	Private sector	Public sector
University degree	7.73%	22.38%
Average work experience in current position (in years)	9.26	7.81
Proportion of employees with the highest task complexity	26.42%	48.57



(a) private sector (b) public sector

Fig. 3.6 Wage (in logarithms) quantiles of women and men in the private and public sectors in 2012.

In the private sector, looking at employees who earn above the value of quantile 99% (in each group, the quantile 99% is the value such that 99% of the employees of the group in the private sector earn less than this value), we find that 63.69% of men have a position in senior management, compared to 27.13% of women. There is also a difference in education level, with 50% of men having a university degree and only 36.41% of women who have such a degree. The experience level in the current position stands at an average value of 10.45 years for men and 8.59 years for women. On the other side, 12.18% of men and 6.62% of women who earn below the value of quantile 1% have a position in senior management (in each group, the quantile 1% is the value such that 1% of the employees of the group in the private sector earn less than it). While it may seem interesting that employees who earn low wages might be in senior management, we found that they work in activities related to accommodation and commerce. 6.31% of men and 1.90% of women have a university degree and the average work experience in the current position of women is 5.03 years for women and 4.83 years for men. This means that women spend on average more time in low-paying jobs than men. These findings are summarized in Tables 3.10 and 3.11.

Table 3.10 Selected employee characteristics of men and women who earn above the value of quantile 99% in the private sector.

Characteristic	Men	Women
Proportion of employees in senior management	63.69%	27.13%
Proportion of employees with a university degree	49.99%	36.41%
Average work experience in current position (in years)	10.45	8.59

Table 3.11 Selected employee characteristics of men and women who earn below the value of quantile 1% in the private sector.

Characteristic	Men	Women
Proportion of employees in senior management	12.18%	6.62%
Proportion of employees with a university degree	6.31%	1.90%
Average work experience in current position (in years)	4.83	5.03

In the public sector, for employees who earn above the value of quantile 99% in their group, 30.49% of men and 14.90% of women are present in senior management positions. 63.16% of men and 43.23% of women have a university degree, while the average work experience in the current position is 13.57 years for men and 10.34 years for women. For employees who earn less than the value of quantile 1%, 1.96% of men and 0.49% of women have a position in senior management. In terms of education, 21.44% of men and 6.50% of women have a university degree. The proportion of men is quite high, with the majority working in teaching-related activities. However, there was no information available with respect to the exact activities. The average work experience is 5.38 years for women and 4.80 years for men. Therefore, similarly to the private sector, in the public sector, women also tend to spend more time on average in low-paying jobs. These findings are summarized in Tables 3.12 and 3.13 .

Table 3.12 Selected employee characteristics of men and women who earn above the value of quantile 99% in the public sector.

Characteristic	Men	Women
Proportion of employees in senior management	30.49%	14.90%
Proportion of employees with a university degree	63.16%	43.23%
Average work experience in current position (in years)	13.57	10.34

Tables 3.14 and 3.15 show the following information for each sector: on the first two lines the wage quantiles for the entire dataset expressed in logarithms and Swiss francs; the last two lines show the cumulative proportions of men and women who earn below the particular value of the quantile. One quantity of interest in Tables 3.14

Table 3.13 Selected employee characteristics of men and women who earn below the value of quantile 1% in the public sector.

Characteristic	Men	Women
Proportion of employees in senior management	1.96%	0.49%
Proportion of employees with a university degree	21.44%	6.50%
Average work experience in current position (in years)	4.87	5.38

and 3.15 is the median, or the quantile 50%. It represents the value such that 50% of the population earns less than that value. The median is of interest because it is less sensitive to extreme values, compared to the mean. In both sectors, the cumulative proportion of women increases faster than that of men for lower-paying jobs. Thus, in the private sector, 61% of women earn less than the value of the median, whereas only 43% of men earn below that amount. In the public sector, 57% of women earn less than the value of the median computed for the entire dataset, compared to 41% of men. For higher-paying jobs, in the private sector, 13% of men and 5% of women earn above the value of quantile 90%. The situation is similar in the public sector, with 16% of men and 6% of women earning above the value of quantile 90%.

Table 3.14 Proportions of men and women who earn less than the value of a particular quantile of the wage computed for the entire dataset in the private sector (values in Swiss francs are given in parantheses).

	Quantile										
	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Logarithm of wage	7.39	8.30	8.43	8.54	8.63	8.72	8.81	8.92	9.07	9.30	10.10
	(1617)	(4018)	(4597)	(5111)	(5625)	(6143)	(6724)	(7494)	(8667)	(10921)	(24297)
Cumulative proportion of men	0.0056	0.06	0.13	0.22	0.32	0.43	0.53	0.64	0.75	0.87	0.9859
Cumulative proportion of women	0.0164	0.16	0.30	0.42	0.52	0.61	0.70	0.79	0.88	0.95	0.9960

Table 3.15 Proportions of men and women who earn less than the value of a particular quantile of the wage computed for the entire dataset in the public sector (values in Swiss francs are given in parantheses).

	Quantile										
	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Logarithm of wage	7.98	8.55	8.70	8.79	8.87	8.95	9.04	9.12	9.24	9.38	9.85
	(2921)	(5190)	(6007)	(6559)	(7099)	(7710)	(8474)	(9155)	(10292)	(11900)	(18890)
Cumulative proportion of men	0.0057	0.06	0.14	0.24	0.32	0.41	0.51	0.61	0.72	0.84	0.9849
Cumulative proportion of women	0.0133	0.13	0.25	0.35	0.46	0.57	0.67	0.77	0.86	0.94	0.9939

3.5.4 Weights

Table 3.16 shows the sampling, the reweighting factor and the calibration weights' minimum, maximum, standard deviations and coefficients of variation.

Table 3.16 Weights' minimum, maximum, standard deviation and coefficient of variation

	Private sector				Public sector			
	Minimum	Maximum	Standard deviation	Coefficient of variation	Minimum	Maximum	Standard deviation	Coefficient of variation
Sampling weights w_k	0.83	282.20	3.28	1.22	1.00	118.10	3.67	2.04
Reweighting factor ψ_k	0.07	1313.00	7.86	2.01	0.03	197.90	4.58	3.33
Calibration weights u_k	0.04	926.80	5.46	2.03	0.02	255.60	5.61	3.11

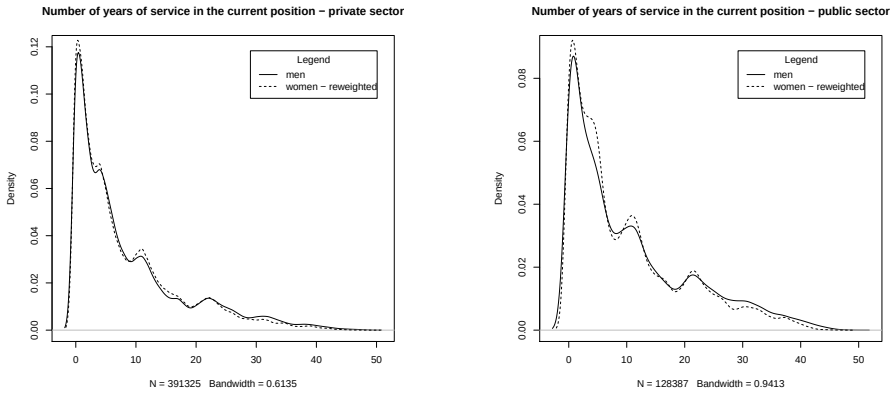
In the public sector, both sampling and calibration weights have a higher standard deviation than in the private sector. The calibration weights for both sectors have close variances, just like the sampling weights.

In each sector, the mean of the sampling weights equals the mean of the calibration weights. Since the standard deviations of the two sets of calibration weights are higher than those of the sampling weights, the coefficient of variation is higher for the former than for the latter.

The reweighting factor did not perfectly match the distribution of characteristics of men to the reweighted distribution of characteristics of women for none of the chosen variables. The largest differences were found for the variables "number of years of service in the current position" and "the square of the age". Figures 3.7 and 3.8 display the densities of these variables.

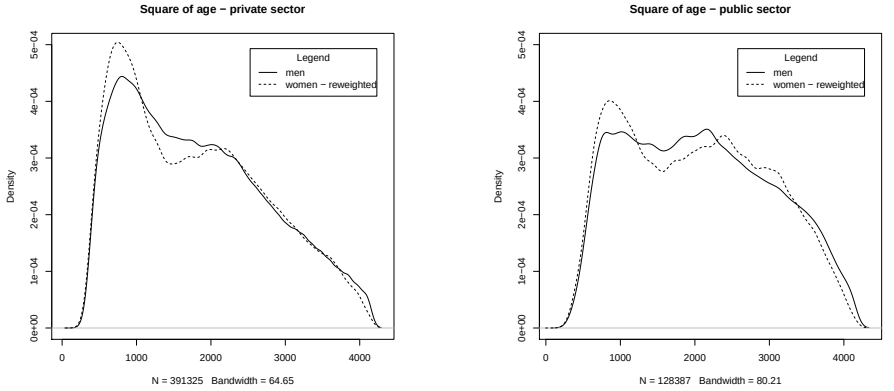
3.5.5 Estimated structure and composition effects for the difference in average values

Tables 3.17 and 3.18 summarize the results obtained through the Blinder-Oaxaca decomposition method, the reweighting factor of DiNardo et al. (1996) and the calibration approach proposed in this paper. The decomposition is only done for the difference in average wages. Tables 3.17 refers to the private sector, while Table 3.18 refers to the public sector. In the private sector, the three methods yield similar results. The residual part in the structure effect computed through the reweightig factor equals 0.002. In the public sector, the reweighting factor yields a structure value that is smaller than the one obtained through calibration. This is because the residual part is negative and equals -0.017.



(a) private sector (b) public sector

Fig. 3.7 Estimated densities in the men's sample and the reweighted distribution in the women's sample of the variable "number of years of service in the current position" in the private and public sectors in 2012.



(a) private sector (b) public sector

Fig. 3.8 Estimated densities in the men's sample and the reweighted distribution in the women's sample of the variable "the square of age" in the private and public sectors in 2012.

Tables 3.17 and 3.18 show that for differences in average wages, the results obtained with the three methods are identical. However, the method of Blinder (1973) and Oaxaca (1973) is only restricted to the decomposition of the differences in average values. The reweighting method of DiNardo et al. (1996) and the calibration approach allow for the decomposition of the wage differences at points other than the mean. Such an analysis might be informative in order to assess whether the estimated composition and structure effects are constant all along the wage distributions.

Table 3.17 Estimated composition and structure effects of the difference in average wages in the private sector (the proportion of each effect of the difference in parantheses).

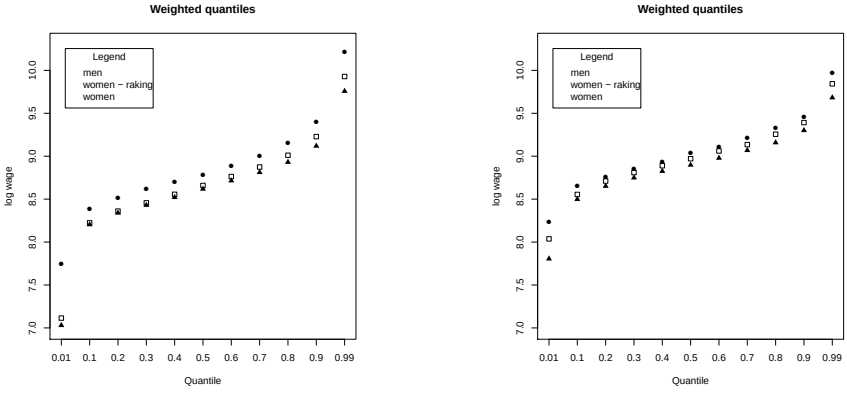
Method	Total	=	Composition effect	+	Structure effect
Blinder-Oaxaca	0.220		0.054 (24.55%)		0.166 (75.45%)
Reweighting factor	0.220		0.053 (24.09%)		0.167 (75.91%)
Calibration approach	0.220		0.055 (25.00%)		0.165 (75.00%)

Table 3.18 Estimated composition and structure effects of the difference in average wages in the public sector (the proportion of each effect of the difference in parantheses).

Method	Total	=	Composition effect (%)	+	Structure effect (%)
Blinder-Oaxaca	0.143		0.075 (52.45%)		0.068 (47.55%)
Reweighting factor	0.143		0.094 (65.73%)		0.049 (34.27%)
Calibration approach	0.143		0.076 (53.15%)		0.067 (46.85%)

3.5.6 Women's counterfactual wage distribution

In what follows, only results obtained through the calibration approach will be discussed. Figure 3.9 shows the quantiles of women's counterfactual wage distributions in both sectors.



(a) private sector (b) public sector

Fig. 3.9 Women's counterfactual wage distributions (in logarithms) in the private and public sectors in 2012.

The differences between men's observed wages (the dots) and women's counterfactual wages (the squares) are the structure effect. The differences between women's counterfactual wages (the squares) and women's observed wages (the triangles) are

the composition effect. These differences can be computed at every quantile in Figure 3.9. Tables 3.19 and 3.20 show how much of the wage differences is explained (the composition effect) and how much is not (the structure effect) at selected quantiles.

Table 3.19 Estimated composition and structure effects of the wage difference at selected quantiles in the private sector.

Quantile	Total	=	Composition effect (%)	+	Structure effect (%)
1%	0.71		0.08 (12%)		0.63 (88%)
10%	0.18		0.02 (11%)		0.16 (89%)
20%	0.18		0.02 (12%)		0.16 (88%)
50%	0.17		0.04 (24%)		0.13 (76%)
80%	0.23		0.08 (35%)		0.15 (65%)
90%	0.28		0.11 (39%)		0.17 (61%)
99%	0.46		0.17 (37%)		0.29 (63%)

Table 3.20 Estimated composition and structure effects of the wage difference at selected quantiles in the public sector.

Quantile	Total	=	Composition effect (%)	+	Structure effect (%)
1%	0.42		0.23 (55%)		0.19 (45%)
10%	0.15		0.06 (37%)		0.10 (63%)
20%	0.10		0.06 (57%)		0.04 (43%)
50%	0.14		0.07 (52%)		0.07 (48%)
80%	0.17		0.10 (57%)		0.07 (43%)
90%	0.15		0.09 (59%)		0.06 (41%)
99%	0.28		0.16 (57%)		0.12 (43%)

In the public sector, the wage differences are smaller than in the private sector. Moreover, in the public sector, discrimination occurs more uniformly than in the private sector. In other words, in the public sector, at all the quantiles where the wage differences were computed, the structure effect represented around 50% of the differ-

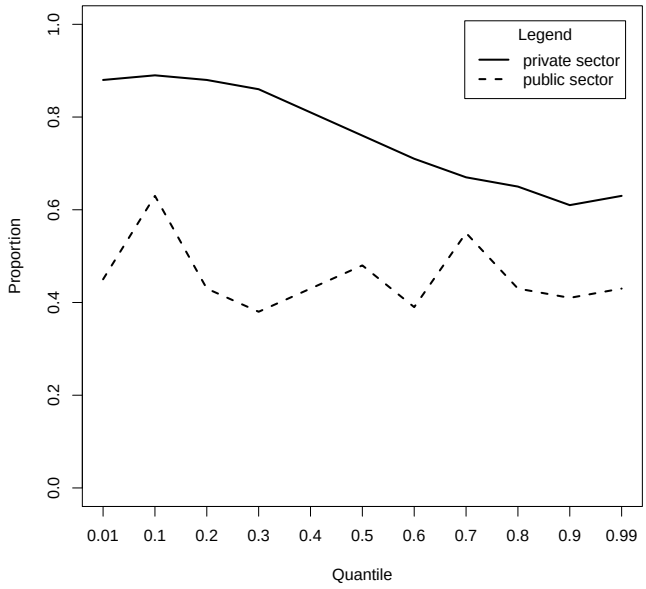
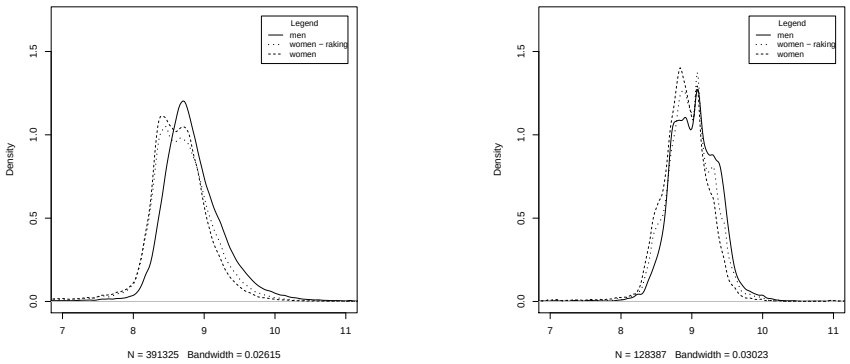


Fig. 3.10 Proportion of the structure effect from the wage differences in the private and public sectors in 2012.

ence. In the private sector, lower-paying jobs display a high level of discrimination. Figure 3.10 displays the proportion of the structure effect from the wage differences. It can be seen that in the private sector, the structure effect displays a downward trend, as shown in Figure 3.10. In the public sector, this trend does not exist. Indeed, at high-order quantiles, the two lines become closer. This means that the difference between the discrimination levels in the private and public sectors diminishes in higher-paying jobs. For the highest-paying jobs, the discrimination level is higher in the private sector than in the public one (63% of the difference is unexplained in the private sector and 43% in the public one). It should be expected that when the job requirements are stricter, as is the case for the highest-paying jobs, much of the difference in wages should be explained by differences in characteristics, especially in the public sector. However, in both sectors, in jobs requiring more qualifications, the discrimination levels are still quite close and represent a fairly large proportion of the total wage difference. As mentioned in Section 2.3.3, the structure effect may include some other factors, such as labour market discrimination or some omitted variables. However, in this paper, it is considered that any non-zero value of the structure effect is only due to discrimination. Figure 3.11 displays women's counterfactual wage

densities. In both sectors, the right tail approaches that of men. Moreover, in the public sector, the counterfactual density follows the shape of that of men.



(a) private sector (b) public sector

Fig. 3.11 Estimated densities of women's counterfactual wage (in logarithms) in the private and public sectors in 2012.

3.6 Conclusions

Using the calibration approach, wage differences can be measured at points other than the mean. While Blinder (1973) and Oaxaca (1973) have laid the groundwork for the estimation of gender wage discrimination, their technique is only limited to analysis of the difference in average wages. The question of interest was whether discrimination occurs constantly, regardless of the type of job (in terms of remuneration). The answer was provided using the calibration approach. First, the proposed approach was compared to the reweighting technique developed by DiNardo et al. (1996), since both methods allow for the examination of wage differences all along wage distributions. The reweighting technique aims at matching the distribution of characteristics of women to that of men, wheareas the calibration approach matches the average characteristics of women to those of men. Using the data, it was shown that the reweighting factor in the DFL method does not match the reweighted distribution of characteristics of women to the distribution of characteristics of men. This misfit is stored in a residual part that is included in the structure effect estimated through the reweighting technique. In the calibration approach, the residual part is always equal to 0.

The pseudo-distance measure in the calibration approach was chosen to generalize the Blinder-Oaxaca method and the reweighting technique of DiNardo et al. (1996). The distance measure implicitly indicates the model that governs the data. Other measures can be chosen. For instance, a Euclidean distance will also yield the same results as the Blinder-Oaxaca method, however, the resulting calibration weights might be negative. This calibration approach extends the Blinder-Oaxaca method beyond the mean and it addresses the issue of the model misfit encountered by the reweighting technique of DiNardo et al. (1996).

Previous research on earnings in the public and private sectors have shown that, on average, employees are better paid in the public sector than in the private sector (see, for instance Lucifora and Meurs, 2006; Heitmueller, 2006; Popli, 2013). Low-skilled workers are better paid in the public sector, whereas the opposite holds true for the private one. These conclusions were confirmed here for the Swiss public sector. Section 3.4.3 includes some descriptive results on the wage distributions in the two sectors. It was shown through graphical tools that the wages are less dispersed in the public sector and that overall, there are smaller wage differences between men and women than in the private sector. Except for the quantile 99%, wages in the public sector are higher than in the private sector. Using the calibration approach, women's counterfactual wage distribution was built. This artificial distribution enables the measurement of the unexplained part of the wage differences at different quantiles, namely the part that was not attributable to objective factors (such as previous work experience or education). While overall the wage differences between men and women in the public sector are smaller than in the private sector, the structure effect amounts to around 50% of these differences all along the wage distribution. This means that in the public sector, employees are treated more fairly that in the private sector.

The wage differences are not constant at all quantiles in either sector. In both sectors, the differences are high at quantile 1%, meaning for jobs that imply lower skills and lower pay. While in the public sector, for these jobs, 55% of the difference is explained, in the private sector only 12% of it is due to objective factors. The private sector displays overall higher differences between men's and women's wages, with more than 50% these differences being unexplained. In the private sector, the proportion of the structure effect of the wage difference has a downward trend. In the public sector, this proportion tends to be constant. While the structure effect represents a fairly large proportion from the wage difference in both sectors, it might not reflect "the effect of the labour market discrimination" (Popli, 2013) or other

factors that influence the dynamics of the labour market, such as sample selection bias. This paper did not investigate such factors and the results obtained are globally termed as discrimination.

Our study confirms some conclusions on the wage differences between the public and private sector in some countries. Not all conclusions are comparable, since other studies focused on wage differences between sectors by gender, whereas ours compares the wage differences between men and women by sectors. We found that the public sector treats its employees more fairly in all types of jobs (lower and higher-paying jobs). Estimating the discrimination levels for lower and higher-paying jobs is useful, because, as Popli (2013) states, "Knowing which segment of the distribution has the largest wage gap can be important for policy" (Popli, 2013).

Chapter 4

A parametric approach to estimate parameters of the counterfactual wage distribution using survey data

Abstract

In this chapter, our main interest is to model wages by using heavy-tailed distributions. Since, in general, men are more likely to earn higher wages, their estimated wage distribution is more likely to be more heavy-tailed. Conditional to some characteristics, we assume that the conditional wage distribution of each woman follows a given theoretical distribution with unknown parameters. First, we estimate the parameters of the distribution of each woman given their characteristics. Next, we estimate what women would earn, if they had the characteristics of men. The comparison between this hypothetical wage and the observed wages in the two groups will lead to the estimation of the wage discrimination against women.

Our goal is to capture the shape of the wage distributions and to go beyond the simple interpretation of mean differences, by determining the estimator of gender wage discrimination at different *quantiles*. Following for instance Melly (2006) and Chernozhukov et al. (2013), we extend to quantiles the classical decomposition method of Blinder (1973) and Oaxaca (1973) for the mean. We provide two parametric methods to estimate quantiles by assuming a given theoretical distribution of conditional wages of men and women given their characteristics.[1]

4.1 Introduction

The wage of an employee is hypothetically a reflection of their characteristics, such as the education level or the previous work experience. As shown in Chapter 2, the regression approach of Blinder (1973) and Oaxaca (1973) consists of modelling the mean of the wage of each individual conditionally on their characteristics. However, on the labor market, sharing the same attributes does not necessarily mean obtaining the same wage. The aim of decomposition methods is to estimate the part of the wage difference that is explained by the differing characteristics and the part which is not.

[1] This chapter is a reprint of Anastasiade et al. (2018)

In this chapter, we focus on modelling wages by using heavy-tailed distributions. Since, in general, men are more likely to earn higher wages, their estimated wage distribution is more likely to be more heavy-tailed. Conditional to some characteristics, we assume that the conditional wage distribution of each woman follows a given theoretical distribution with unknown parameters. First, we estimate the parameters of the distribution of each woman given their characteristics. Next, the marginal wage distribution of women is fitted based on the individual woman wage distributions.

We use again the concept of counterfactual wage distribution (for an overview see Fortin et al., 2011). In Chapters 2 and 3, the counterfactual wage distribution was estimated by putting together the parameters of one group and the characteristics of the other group. This is done in order to estimate what the former group would earn, had they had the characteristics of the other group. We followed this guideline to estimate the wages of women if they had the same characteristics of men. This leads to the estimation of gender wage discrimination conditional to fixed covariates. The literature covers inference issues for counterfactual distributions (see for instance Chernozhukov et al., 2013).

Our goal is to capture the shape of the wage distributions and to go beyond the simple interpretation of mean differences, by determining the estimator of gender wage discrimination at different *quantiles*. Following for instance Melly (2006) and Chernozhukov et al. (2013) we extend to quantiles the classical decomposition method of Blinder (1973) and Oaxaca (1973) for the mean. We provide two parametric methods to estimate quantiles by assuming a given theoretical distribution of conditional wages of men and women given their characteristics.

The unexplained part of the wage difference is usually associated with discrimination. However, as already mentioned in Section 3.5, there are other mechanisms that result in a pay difference between men and women on the labor market. When the combination of the characteristics of men's group and of the parameters of the women's group is used to estimate this unexplained part, its interpretation can be done in two ways: it is either the bonus in pay that men currently have over women, or it is the penalty that women face on the labor market. What leads to this difference is not necessarily discrimination, but other mechanisms may be hidden behind it. For simplicity, we will refer at the unexplained part as discrimination and we use the second interpretation.

Motivated by a flexible way to model wage distributions, we illustrate the proposed methods in this chapter by fitting a generalized beta distribution of the second

kind (hereafter, GB2) distribution to conditional wages in our examples. Following the work of Thurow (1970), who considered that "the beta distribution seems the most flexible" distribution to capture income changes, McDonald (1984) introduced the GB2 distribution to model the income distribution. McDonald (1984); Bandourian et al. (2002); McDonald and Ransom (2008) among others, showed that the GB2 distribution provides a good fit for income. The link between the beta distribution and the GB2 distribution is that if the random variable X follows a beta distribution of parameters p, q, and $Y = X/(1-X)$, then the variable $Z = bY^{1/a}$, follows a GB2 distribution with parameters a, b, p, q (see, for instance, Graf et al., 2011). Thus, a random variable that follows a GB2 distribution is the result of the transformation of another random variable that follows a beta distribution.

The GB2 distribution can used to fit either positively or negatively skewed distributions and is a generalization of several distributions, such as the log-normal, the exponential or the Fisk distributions (Kleiber and Kotz, 2003; McDonald, 1984; McDonald and Xu, 1995; McDonald and Butler, 1990). This distribution is already well-covered in the literature (see, for instance, Kleiber and Kotz, 2003; Graf et al., 2011). Our novelty consists in estimating the GB2 parameters through pseudo-maximum likelihood, when survey weights and characteristics are associated to sampled employees, by expressing the scale parameter of a GB2 distribution as a function of the characteristics of the employees. We also show how to estimate the standard errors of the estimated parameters in a GB2 regression model, using a sandwich estimator and a parametric bootstrap approach. These are useful for the construction of confidence intervals and for inference for the GB2 regression parameters.

In this chapter, we propose two parametric methods to estimate the structure and the composition effects at the quantile level. The chapter is structured as follows: in Section 4.2, we revisit the setup presented in Chapter 2. We also explain how the design-based approach and the model-based approach are combined through the use of survey weights. In Section 4.3 we re-express the counterfactual wage distribution in the context of the setup and in Section 4.4, we discuss how to estimate the composition and the structure effects at the quantile levels. We take the parametric approach, in assuming that wages follow a certain distribution with several parameters, one of which can be expressed as a function of the individual characteristics. This method is flexible in that several distributions can be assumed. In Section 4.5, we introduce the two proposed methods to estimate the composition and the structure effects at the quantile level. As a reminder, these effects are estimated from the differences

between the observed wages and the counterfactual wages. In the first method, we compute the inverse of the cumulative distribution function to estimate the difference in quantiles of the aforementioned wages. In the second method, when the inverse of the cumulative distribution function cannot be computed, we show a simulation-based approach that results in the estimation of the quantiles. In Section 4.7, we develop the estimation of the counterfactual wage distribution for the two cases covered by the two methods, as well as the estimation of the quantiles. Section 4.7 includes a discussion about variance estimation and confidence intervals of the estimated quantiles. Next, in Section 4.8, results of Monte-Carlo simulation are shown, using again the GB2 distribution. We discuss the algorithm to estimate the parameters of the distribution, as well as their standard errors through the sandwich estimator or a parametric bootstrap approach. Finally, in Section 4.9, we apply the proposed methods to real data delivered by the Swiss Federal Statistical Office and in Section 4.10, we draw the conclusions.

4.2 Revisiting the setup

Consider a finite population of employees with the labels $U = \{1, 2, \ldots, N\}$. From this population, we randomly select a sample S of size n, without replacement. The sample is selected through a sampling design $p(s) = \Pr(S = s), \forall s \subseteq U$. To each unit $k \in S$, a survey weight w_k is associated. These weights can be equal to the inverse of the inclusion probabilities or can be more complicated weights, like calibration weights.

The set U is divided in two subsets of labels corresponding to men and women, denoted by U_M and U_F respectively, such that $U_M \cup U_F = U$ and $U_M \cap U_F = \emptyset$. Similarly, the sample S is divided into two random subsamples of men and women, denoted by $S_M = S \cap U_M$ and $S_F = S \cap U_F$ respectively. We denote these subsamples as $S_g \subseteq U_g, g \in \{M, F\}$, with n_M and n_F being the number of employees in the subsamples, respectively, such that $n_M + n_F = n$.

Let Y be the variable wage. We work in a superpopulation framework and assume that the finite population is a random sample drawn from an infinite population. First, we consider that Y is a random variable generated by a distribution model ξ in the infinite population. Next, the finite population $\{Y_1, Y_2, \ldots, Y_N\}$ is randomly generated from the model ξ, where Y_k is the variable wage associated to each $k \in U$. We assume that the estimation process refers to a finite population parameter, and is executed in the design-based approach, while still assuming that Y_k associated

with unit $k \in U$ is random. In this way, we insert survey weights into a model-based analysis.

We also assume the existence of a general regression model that relates the random variable Y to some covariates $X_1, X_2, \ldots, X_J$. The covariates are the same in each $U_g, g \in \{M, F\}$, but for coherence with the subset notation we denote by $X_{1,g}, X_{2,g}, \ldots, X_{J,g}$ the covariates in group $g \in \{M, F\}$. For each unit $k \in U_g$, $g \in \{M, F\}$, the wage is denoted by $Y_{k,g}$ and the J covariates are stored in the vector $\mathbf{X}_{k,g}$

$$\mathbf{X}_{k,g} = (1, X_{1k,g}, X_{2k,g}, \ldots, X_{Jk,g})^\top. \tag{4.1}$$

One realization of $\mathbf{X}_{k,g}$ is denoted by $\mathbf{x}_{k,g} = (1, x_{1k,g}, x_{2k,g}, \ldots, x_{Jk,g})^\top$. The last $J-1$ elements of the vector $\mathbf{x}_{k,g}$ represent realizations of variables $X_{1,g}, X_{2,g}, \ldots, X_{J,g}$, respectively, $g \in \{M, F\}$. In what follows, we also denote by y_k a realization of $Y_k, k \in U$ and use $\mathbf{X}_g = (X_{1,g}, X_{2,g}, \ldots, X_{J,g})$, $g \in \{M, F\}$.

4.3 Revisiting the counterfactual wage distribution

The counterfactual distribution is an artificial distribution, defined "as the result of either a change in the distribution of a set of covariates X that determine the outcome variable of interest Y, or as a change in the relationship of the covariates with the outcome, i.e. a change in the conditional distribution of Y given X" (Chernozhukov et al., 2013). We construct a counterfactual wage distribution as the distribution resulting from the change in the distribution of covariates.

Let $F^{(Y_F|\mathbf{X}_F)}(.)$ and $F^{(Y_M|\mathbf{X}_M)}(.)$ be the cumulative distribution functions (CDFs) of the conditional wage distributions of women and men, with respect to the characteristics $\mathbf{X}_F$ and $\mathbf{X}_M$, respectively. We also denote by $F^{\mathbf{X}_F}(.)$ and $F^{\mathbf{X}_M}(.)$ the CDFs of distributions corresponding to $\mathbf{X}_F$ and $\mathbf{X}_M$, respectively. Let also $F^C(.)$ be the CDF of the counterfactual distribution of women. Following Chernozhukov et al. (2013), the CDF in the point $y \in \mathscr{Y}_\mathscr{F}$, where $\mathscr{Y}_\mathscr{F}$ is women's wage support is defined as

$$F^C(y) = \int_{\mathscr{X}_\mathscr{M}} F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}), \tag{4.2}$$

where $\mathscr{X}_\mathscr{M}$ is the support of $\mathbf{X}_M$. The counterfactual wage distribution is well defined if the support of $\mathbf{X}_F$ ($\mathscr{X}_\mathscr{F}$) includes the support of $\mathbf{X}_M$: $\mathscr{X}_\mathscr{M} \subseteq \mathscr{X}_\mathscr{F}$. Note that the counterfactual wage distribution represents a marginal distribution, and not a conditional one.

4.4 Estimation of structure and composition effects at the quantile level

4.4.1 Quantile decomposition

We express the change at quantile α between the wage of men and the counterfactual wage of women (corresponding to the structure effect at quantile α) as

$$\Delta^{ST}_{(\alpha)} = Q^{M}_{(\alpha)} - Q^{C}_{(\alpha)}, \tag{4.3}$$

where $Q^{M}_{(\alpha)}$ and $Q^{C}_{(\alpha)}$ represent the quantile of order α of the men's wage distribution and of the counterfactual distribution, respectively. The change at quantile α between the counterfactual wage distribution and women's wage distribution (corresponding to the composition effect) is expressed as

$$\Delta^{CO}_{(\alpha)} = Q^{C}_{(\alpha)} - Q^{F}_{(\alpha)}. \tag{4.4}$$

Section 4.5 provides methods to estimate $Q^{M}_{(\alpha)}$, $Q^{C}_{(\alpha)}$, and $Q^{F}_{(\alpha)}$. It follows that afterwards $\Delta^{ST}_{(\alpha)}$ and $\Delta^{CO}_{(\alpha)}$ can be estimated by plugging in the corresponding estimators of $Q^{M}_{(\alpha)}$, $Q^{C}_{(\alpha)}$, and $Q^{F}_{(\alpha)}$ in Expressions (4.3) and (4.4), respectively.

4.4.2 The conditional wage distributions

Consider that the conditional wage distribution of $Y_{k,g} \mid \mathbf{X}_g = \mathbf{x}_{k,g}$, $g \in \{M,F\}$ is a continuous distribution denoted by $A(\gamma_{k,g}, \delta_g)$, $k \in U_g$, and is characterized by a number of parameters, where $\gamma_{k,g}$ is the one of interest and δ_g denote the remaining parameters. We assume that the distribution A is known, but its parameters are unknown. We also consider that $\gamma_{k,g} = h(\mathbf{x}^{\top}_{k,g}\beta_g)$, where h is a known continuous function, $\mathbf{x}_{k,g}$ represents the vector of the characteristics of unit $k \in U_g$ and β_g is a set of regression parameters corresponding to group $g \in \{M,F\}$. Note that $\gamma_{k,g}$ can be different for each unit k. Also note that the corresponding regression model used here may be very general: linear or nonlinear, and with the error terms following a very general distribution. For example, in Sections 4.8 and 4.9 we use a GB2 regression model.

We assume that $Y_{k,g} \mid \mathbf{X}_g = \mathbf{x}_g \sim A(h(\mathbf{x}^{\top}_{k,g}\beta_g), \delta_g)$, $k \in U_g$ are independent. The overall distribution of all $Y_{k,g}, k \in U_g$ is a mixture distribution with N_g components,

having the density function

$$f^{Y_g}(y) = \sum_{k \in U_g} \lambda_k f_{A(\gamma_{k,g}, \delta_g)}(y \mid \mathbf{x}_{k,g}), \tag{4.5}$$

where $\lambda_k = 1/N_g$, and $f_{A(\gamma_{k,g},\delta_g)}(. \mid \mathbf{x}_{k,g})$ is the density function of the distribution $A(\gamma_{k,g} = h(\mathbf{x}_{k,g}^\top \beta_g), \delta_k), k \in U_g$. The parameter $\gamma_{k,g}$ is estimated by $\widehat{\gamma}_{k,g} = h(\mathbf{x}_{k,g}^\top \widehat{\beta}_g)$, where $\widehat{\beta}_g$ are the estimated regression parameters computed on the sample S_g, using a weighted approach with weights $w_k, k \in S_g$. The remaining the parameters δ_g are estimated by $\widehat{\delta}_g$, the corresponding sample weighted estimates computed on S_g.

4.5 Two methods to estimate parameters of the wage distributions of men and women

We propose to estimate the quantiles of the previous mixture distribution using two methods. In the first method, we express the estimated quantiles of a wage distribution as the inverse of its cumulative distribution function. In case that this inverse cannot be computed, we present in the second method a simulation-based approach that bypasses this problem and makes the estimation of the quantiles possible.

4.5.1 First method (Method 1)

We integrate $f^{Y_g}(.)$ on $(-\infty, y]$ in Expression (4.5) to obtain the corresponding CDF in the point $y \in \mathscr{Y}_g$, where $\mathscr{Y}_g$ is the wage support in group g

$$\begin{aligned} F^{Y_g}(y) &= \int_{-\infty}^{y} \sum_{k \in U_g} \lambda_k f_{A(\gamma_{k,g},\delta_g)}(z \mid \mathbf{x}_{k,g}) dz \\ &= \sum_{k \in U_g} \lambda_k \int_{-\infty}^{y} f_{A(\gamma_{k,g},\delta_g)}(z \mid \mathbf{x}_{k,g}) dz \\ &= \sum_{k \in U_g} \lambda_k F_{A(\gamma_{k,g},\delta_g)}(y \mid \mathbf{x}_{k,g}) \\ &= \frac{1}{N_g} \sum_{k \in U_g} F_{A(\gamma_{k,g},\delta_g)}(y \mid \mathbf{x}_{k,g}), \end{aligned} \tag{4.6}$$

where $F_{A(\gamma_{k,g},\delta_g)}(. \mid \mathbf{x}_{k,g})$ is the CDF of the distribution $A(\gamma_{k,g} = h(\mathbf{x}_{k,g}^\top \beta_g), \delta_g)$. Next, $F^{Y_g}(.)$ is estimated by

$$\widehat{F}^{Y_g}(y) = \sum_{k \in S_g} w_k F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(y \mid \mathbf{x}_{k,g}) / \sum_{k \in S_g} w_k. \tag{4.7}$$

The quantile $Q^g_{(\alpha)}$ is estimated using the following relationship

$$\widehat{Q}^g_{(\alpha)} = \inf\{y \mid \widehat{F}^{Y_g}(y) \geq \alpha\}, \tag{4.8}$$

where $\widehat{Q}^g_{(\alpha)}$ is the estimator of $Q^g_{(\alpha)}, g \in \{M, F\}$.

Remark 1 *1. At the superpopulation level we have*

$$E_{\mathbf{X}_g}(E_{Y_g}(I(Y_g \leq y) \mid \mathbf{X}_g)) = E_{Y_g}(I(Y_g \leq y)) = F(y),$$

and

$$E_{Y_g}(I(Y_g \leq y) \mid \mathbf{X}_g) = F^{Y_g|\mathbf{X}_g}(y \mid \mathbf{X}_g),$$

where $E_{\mathbf{X}_g}(.)$ denotes the expectation with respect to $\mathbf{X}_g$, $E_{Y_g}(.)$ with respect to Y_g, and $F(.)$ is the model CDF of Y_g. It follows that

$$E_{\mathbf{X}_g}(F^{Y_g|\mathbf{X}_g}(y \mid \mathbf{X}_g)) = F(y).$$

At the finite population level, $E_{\mathbf{X}_g}(F^{Y_g|\mathbf{X}_g}(y \mid \mathbf{X}_g))$ is written as $\frac{1}{N_g}\sum_{k \in U_g} F_{A(\gamma_{k,g},\delta_g)}(y \mid \mathbf{x}_{k,g})$, while $F(y)$ is estimated by $F^{Y_g}(y)$. This shows the correctness of the first proposed method.

2. *The proposed method is based on a parametric approach. Since we use auxiliary information $\mathbf{x}_{k,g}$ in estimating $F^{Y_g}(y)$, we expect to reduce the variance of the estimator given by Expression (4.7) compared to that of the estimator*

$$\widetilde{F}^{Y_g}(y) = \sum_{k \in S_g} w_k I(y_k \leq y) / \sum_{k \in S_g} w_k. \tag{4.9}$$

4.5.2 Second method (Method 2)

If the inverse function of $\widehat{F}^{Y_g}(y)$ cannot be computed, we propose to use the following Monte Carlo method based on parametric bootstrap:

1. Generate a large number m of n_g independent draws from the distribution $A(h(\mathbf{x}_{k,g}^\top \widehat{\beta}_g), \widehat{\delta}_g)$, $k \in S_g$, respectively. A matrix of dimension $m \times n_g$ of such

draws is obtained. Each element $(i,k), i = 1,\ldots,m, k = 1,\ldots n_g$ in this matrix is the realization $y_{i,k}$ of a random variable $Y_{i,k}$ with $Y_{i,k} \mid \mathbf{x}_{k,g} \sim A(h(\mathbf{x}_{k,g}^{\top}\widehat{\beta}_g), \widehat{\delta}_g)$; conditional to $\mathbf{x}_g$, all the random variables $Y_{i,k} \mid \mathbf{x}_{k,g}$ are independent.

2. Associate to each element $(i,k), i = 1,\ldots,m, k = 1,\ldots,n_g$ the weight $w_k, k \in S_g$ and estimate the empirical weighted quantiles of order $\alpha \in [0,1]$, $\widehat{Q}_{\alpha}^{(i)}(Y_g)$, by using
$$\widehat{Q}_{\alpha}^{(i)}(Y_g) = \inf\{y \mid \widehat{F}_i^{(Y_g)}(y) \geq \alpha\}, \tag{4.10}$$
where $\widehat{F}_i^{(Y_g)}(y) = \sum_{k=1}^{n_g} w_k I(y_{i,k} \leq y) / \sum_{k=1}^{n_g} w_k$.

3. For each $\alpha \in [0,1]$, compute the mean of the $\widehat{Q}_{\alpha}^{(i)}(Y_g), i = 1,\ldots,m$; this mean denoted by $\widehat{\overline{Q}}_{\alpha}(Y_g)$ represents an estimate of the quantile of order α of the mixture distribution with the density given in Expression (4.5).

Remark 2 *1. The method provides a reliable estimator of the quantile of order α of the wage distribution in the group g if the conditional distribution of $Y_g \mid \mathbf{X}_g$ is correctly specified. The correctness of the method is assured by the fact that we generate a random variable $Y_{i,k}$ corresponding to each $\mathbf{x}_k, k = 1,\ldots n_g$. Since we consider the entire set of covariates $\mathbf{x}_g$ the resulting quantile $\widehat{Q}_{\alpha}^{(i)}(Y_g)$ (computed in each run i of the algorithm) is an estimator of the unconditional quantile of order α of the wage distribution in group g.*

2. Monte Carlo simulation results indicate that both methods used to estimate the quantiles of $Y_g, g \in \{M,F\}$ give similar performances in terms of Monte-Carlo variance. These results are not shown here.

4.6 Two methods to estimate parameters of the counterfactual wage distribution

We are interested in estimating the quantiles of the counterfactual distribution. This is necessary for a comparison between them and the estimated quantiles of the unconditional distribution of women's wages and those of men, respectively.

The counterfactual wage is a potential wage of a woman if she has the characteristics of a man. Under the assumption that $\mathscr{X}_M = \mathscr{X}_F$, DiNardo et al. (1996) reexpressed the counterfactual distribution given in Expression (4.2) as

$$F^C(y) = \int_{\mathscr{X}_M} F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}) = \int_{\mathscr{X}_F} F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}) \psi(\mathbf{x}) dF^{\mathbf{X}_F}(\mathbf{x}), \tag{4.11}$$

where $\psi(\mathbf{x}) = dF^{\mathbf{X}_M}(\mathbf{x})/dF^{\mathbf{X}_F}(\mathbf{x})$. DiNardo et al. (1996) rewrite the $\psi(.)$ factor as

$$\psi(\mathbf{x}_k) = \psi_k = \frac{P(D_k = 1 \mid \mathbf{x}_k)/P(D_k = 1)}{P(D_k = 0 \mid \mathbf{x}_k)/P(D_k = 0)}, \tag{4.12}$$

where $D_k = 1$ if individual k is a man and $D_k = 0$ otherwise and $\mathbf{x}_k$ is the vector of observed characteristics for individual k. The parameter $\psi(\mathbf{x}_k)$ can be estimated by using a probit or a logistic regression model (DiNardo et al., 1996) or by calibration (Anastasiade and Tillé, 2017a). The difference between the two methods is discussed in Anastasiade and Tillé (2017a).

The empirical counterfactual CDF defined at the U_F level can be written as

$$F^C_{emp}(y) = \frac{\sum_{k \in U_F} \psi_k I(y_k \leq y)}{\sum_{k \in U_F} \psi_k}. \tag{4.13}$$

The weighted method of DiNardo et al. (1996) and that of Anastasiade and Tillé (2017a) use the estimated empirical counterfactual CDF defined by

$$\widehat{F}^C_{emp}(y) = \frac{\sum_{k \in S_F} \widehat{\psi}_k w_k I(y_k \leq y)}{\sum_{k \in S_F} \widehat{\psi}_k w_k}, \tag{4.14}$$

where $\widehat{\psi}_k$ is an estimator of ψ_k. Next, they estimate the α-quantile $Q^C_{(\alpha)}$ of the counterfactual distribution by using the following relationship

$$\widehat{Q}^C_{(\alpha),emp} = \inf\{y \mid \widehat{F}^C_{emp}(y) \geq \alpha\}, \tag{4.15}$$

where $\widehat{Q}^C_{(\alpha),emp}$ is the empirical estimator of $Q^C_{(\alpha)}$.

We write the counterfactual CDF at the U_F level as

$$F^C_{U_F}(y) = \frac{1}{N_C} \sum_{k \in U_F} \psi_k F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}_F), \tag{4.16}$$

where $N_C = \sum_{k \in U_F} \psi_k$, and propose to estimate $F^C_{U_F}(y)$ by

$$\widehat{F}^C(y) = \frac{\sum_{k \in S_F} \widehat{\psi}_k w_k \widehat{F}^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}_F)}{\sum_{k \in S_F} \widehat{\psi}_k w_k}, \tag{4.17}$$

where $\widehat{F}^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}_{k,F}) = F_{A(h(\mathbf{x}^\top_{k,F}\widehat{\beta}_F),\widehat{\delta}_F)}(y \mid \mathbf{x}_{k,F})$, and $\widehat{\psi}_k$ is estimated by calibration using the raking method (Anastasiade and Tillé, 2017a). To estimate $Q^C_{(\alpha)}$, the following

relationship is used

$$\widehat{Q}^C_{(\alpha)} = \inf\{y \mid \widehat{F}^C(y) \geq \alpha\}, \tag{4.18}$$

where $\widehat{Q}^C_{(\alpha)}$ is the estimator of $Q^C_{(\alpha)}$.

If the inverse function of $\widehat{F}^C(.)$ can be computed, the quantile of order α of the counterfactual wage distribution is estimated by solving in y the equation $\widehat{F}^C(y) = \alpha$. The solution of this equation gives an estimate of $Q^C_{(\alpha)}$. If the inverse function of $\widehat{F}^C(.)$ cannot be computed, the following Monte-Carlo method based on parametric bootstrap is used:

1. Generate a large number m of n_F independent draws from the distribution $A(h(\mathbf{x}^\top_{k,F}\widehat{\beta}_F), \widehat{\delta}_F)$, $k \in S_F$, respectively. A matrix of dimension $m \times n_F$ of such draws is obtained. Each element $(i,k), i = 1,\ldots,m, k = 1,\ldots n_F$ in this matrix is the realization $y_{i,k}$ of a random variable $Y_{i,k}$ with $Y_{i,k} \mid \mathbf{x}_{k,F} \sim A(h(\mathbf{x}^\top_{k,F}\widehat{\beta}_F), \widehat{\delta}_F)$; conditional to $\mathbf{x}_F$, all the random variables $Y_{i,k} \mid \mathbf{x}_{k,g}$ are independent.

2. Associate to each element $(i,k), i = 1,\ldots,m, k = 1,\ldots,n_F$ the weight $\widehat{\psi}_k w_k, k \in S_F$ and estimate the empirical weighted quantile of order $\alpha \in [0,1]$ of $Y_{i,1},\ldots,Y_{i,k},\ldots Y_{i,n_F}$ by

$$\widehat{Q}^{(i)}_\alpha(Y_C) = \inf\{y \mid \widehat{F}_i(y) \geq \alpha\},$$

   where Y_C denotes the counterfactual wage distribution and $\widehat{F}_i(y) = \sum_{k=1}^{n_F} \widehat{\psi}_k w_k I(y_{i,k} \leq y) / \sum_{k=1}^{n_F} \widehat{\psi}_k w_k$.

3. For each $\alpha \in [0,1]$, compute the mean of the $\widehat{Q}^{(i)}_\alpha(Y_C), i = 1,\ldots,m$; this mean denoted by $\widehat{\overline{Q}}_\alpha(Y_C)$ represents an estimate of the quantile of order α of the counterfactual wage distribution.

Remark 3 *The method for estimating $\widehat{Q}^{(i)}_\alpha(Y_C)$ uses random weights $\widehat{\psi}_k w_k, k \in S_F$. The estimation of $\widehat{Q}^{(i)}_\alpha(Y_C)$ is still reliable in this case because $\widehat{\psi}_k w_k, k \in S_F$ are fixed in each run of the algorithm.*

4.7 Variance estimation

Variance estimation of the quantile estimators can be derived by linearization (Deville, 1999). For a generic finite population U of size N, the CDF of wage distribution y_k at the U level is written as

$$F_U(y) = \frac{1}{N}\sum_{k \in U} I(y_k \leq y), \tag{4.19}$$

and is estimated by $\widehat{F}_U(y) = \sum_{k\in S} w_k I(y_k \leq y)/\sum_{k\in S} w_k$. Following Deville (1999), the linearized variable of the α-order quantile $Q_{(\alpha)} = F_U^{-1}(\alpha)$ is given by

$$z_k^{Q_{(\alpha)}} = \frac{1}{f(Q_{(\alpha)})N}\left(\alpha - I(y_k \leq Q_{(\alpha)})\right), k \in U,$$

where $f(.)$ represents the model wage density function. That leads to the following approximation of the variance estimator of $\widehat{Q}_{(\alpha)} = \widehat{F}_U^{-1}(\alpha)$

$$\widehat{var}(\widehat{Q}_{(\alpha)}) \simeq \widehat{var}(\sum_{k\in S_g} \widehat{z}_k^{\widehat{Q}_{(\alpha)}} w_k),$$

where $\widehat{var}(\sum_{k\in S} \widehat{z}_k^{\widehat{Q}_{(\alpha)}} w_k)$ is computed in function of the used sampling design.

4.7.1 Quantiles of the wage distribution

At the U_g level, the CDF of the wage distribution is given in Expression (4.6). Its influence function at $k \in U_g$ in point y is

$$\frac{1}{N_g}\left(F_{A(\gamma_{k,g},\delta_g)}(y \mid \mathbf{x}_{k,g}) - F^{Y_g}(y)\right).$$

Since $F^{Y_g}(Q^g_{(\alpha)}) = \alpha$, the linearized variable of the α-order quantile of the wage distribution in group g is given by

$$z_k^{Q^g_{(\alpha)}} = \frac{1}{f_g(Q^g_{(\alpha)})N_g}\left(\alpha - F_{A(\gamma_{k,g},\delta_g)}(Q^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right), k \in U_g$$

where $f_g(.)$ represents the model wage density function in the group g. For our first proposed method, it can be estimated using

$$\widehat{z}_k^{\widehat{Q}^g_{(\alpha)}} = \frac{1}{\widehat{f}_g(\widehat{Q}^g_{(\alpha)})\widehat{N}_g}\left(\alpha - F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right),$$

where $\widehat{N}_g = \sum_{k\in S_g} w_k$, and $\widehat{f}_g(.) = \sum_{k\in S_g} w_k f_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(. \mid \mathbf{x}_{k,g})/\sum_{k\in S_g} w_k$.

It follows that $\widehat{z}_k^{\widehat{Q}^g_{(\alpha)}} = \left(\alpha - F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right)/\sum_{k\in S_g} w_k f_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})$.

That leads to the following approximation of the variance estimator of $\widehat{Q}^g_{(\alpha)}$

$$\widehat{var}(\widehat{Q}^g_{(\alpha)}) \simeq \widehat{var}(\sum_{k\in S_g} \widehat{z}_k^{\widehat{Q}^g_{(\alpha)}} w_k) = V_{lin}(\widehat{Q}^g_{(\alpha)}).$$

Note that the term $\widehat{var}(\sum_{k\in S_g} \widehat{z}_k^{\widehat{Q}^g_{(\alpha)}} w_k)$ in the previous expression is the estimated variance of the ratio of two random variables

$$\widehat{var}(\sum_{k\in S_g} \widehat{z}_k^{\widehat{Q}^g_{(\alpha)}} w_k) = \widehat{var}\left(\sum_{k\in S_g} w_k \left(\alpha - F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right) / \sum_{k\in S_g} w_k f_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right).$$

Next, $\widehat{var}\left(\sum_{k\in S_g} w_k \left(\alpha - F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right) / \sum_{k\in S_g} w_k f_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^g_{(\alpha)} \mid \mathbf{x}_{k,g})\right)$ is computed in function of the used sampling design based. This computation can be done, for example, using Taylor series expansion (see also Särndal et al., 1992, p. 177–178).

4.7.2 Quantiles of the counterfactual wage distribution

The counterfactual CDF at the U_F level is given in Expression (4.16). Since ψ_k is a factor that does not depend on the wage, the linearized variable of the α-order quantile of the counterfactual wage distribution can be estimated as before by using

$$\widehat{z}_k^{\widehat{Q}^C_{(\alpha)}} = \frac{1}{\widehat{f}^C(\widehat{Q}^C_{(\alpha)})\widehat{N}_C}\left(\alpha - F_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(\widehat{Q}^C_{(\alpha)} \mid \mathbf{x}_{k,g})\,\widehat{\psi}_k\right),$$

where $\widehat{N}_C = \sum_{k\in S_F} \widehat{\psi}_k w_k$, and $\widehat{f}^C(.) = \sum_{k\in S_F} \widehat{\psi}_k w_k f_{A(\widehat{\gamma}_{k,g},\widehat{\delta}_g)}(. \mid \mathbf{x}_{k,g}) / \sum_{k\in S_F} \widehat{\psi}_k w_k$. That leads as before to the following approximation of the variance estimator of $\widehat{Q}^C_{(\alpha)}$

$$\widehat{var}(\widehat{Q}^C_{(\alpha)}) \simeq \widehat{var}(\sum_{k\in S_F} \widehat{z}_k^{\widehat{Q}^C_{(\alpha)}} w_k) = V_{lin}(\widehat{Q}^C_{(\alpha)}),$$

where $\widehat{var}(\sum_{k\in S_F} \widehat{z}_k^{\widehat{Q}^C_{(\alpha)}} w_k)$ is computed in function of the used sampling design.

4.7.3 Approximate confidence intervals for quantiles

For the first proposed method used to estimate the counterfactual wage quantiles, we have that

$$\widehat{var}(\widehat{Q}^C_{(\alpha)}) \simeq \widehat{var}(\sum_{k\in S_F} \widehat{z}_k^{\widehat{Q}^C_{(\alpha)}} w_k) = V_{lin}(\widehat{Q}^C_{(\alpha)}),$$

where $\widehat{var}(\sum_{k\in S_F} \widehat{z}_k^{\widehat{Q}^C_{(\alpha)}} w_k)$ is computed in function of the used sampling design.

Approximate 95% confidence intervals for quantiles $Q_{(\alpha)}$ of the wage in group g can be constructed using

$$[\widehat{Q}^g_{(\alpha)} - 1.96\sqrt{V_{lin}(\widehat{Q}^g_{(\alpha)})}, \widehat{Q}^g_{(\alpha)} + 1.96\sqrt{V_{lin}(\widehat{Q}^g_{(\alpha)})}],$$

and for $Q^C_{(\alpha)}$

$$[\widehat{Q}^C_{(\alpha)} - 1.96\sqrt{V_{lin}(\widehat{Q}^C_{(\alpha)})}, \widehat{Q}^C_{(\alpha)} + 1.96\sqrt{V_{lin}(\widehat{Q}^C_{(\alpha)})}].$$

Alternatively, approximate 95% confidence intervals for quantiles $Q^g_{(\alpha)}$ of the wage in group g can be computed using the following Woodruff-type interval (Woodruff, 1952)

$$\left[\widehat{F}^{-1}\left\{\alpha - 1.96\sqrt{\widehat{V}(\widehat{F}(Q^g_{(\alpha)}))}\right\}, \widehat{F}^{-1}\left\{\alpha + 1.96\sqrt{\widehat{V}(\widehat{F}(Q^g_{(\alpha)}))}\right\}\right],$$

where $\widehat{F} = \widehat{F}^{Y_g}$ is defined in Expression (4.6) and $\widehat{V}(\widehat{F}(.))$ is an estimator of the variance of $\widehat{F}(.)$. Similarly, approximate 95% confidence intervals for quantiles $Q^C_{(\alpha)}$ are given by

$$\left[(\widehat{F}^C)^{-1}\left\{\alpha - 1.96\sqrt{\widehat{V}(\widehat{F}^C(Q^C_{(\alpha)}))}\right\}, (\widehat{F}^C)^{-1}\left\{\alpha + 1.96\sqrt{\widehat{V}(\widehat{F}^C(Q^C_{(\alpha)}))}\right\}\right],$$

where $\widehat{F}^C(.)$ is defined in Expression (4.17) and $\widehat{V}(\widehat{F}^C(.))$ is an estimator of the variance of $\widehat{F}^C(.)$.

Remark 4 *The variance estimation of the quantile estimator given in the second proposed method is taken to be the Monte-Carlo variance over the columns of the matrix of* $\widehat{Q}^{(i)}_\alpha(Y_g)$, $g \in \{M,F\}, i = 1,\ldots,m : V_{MC}(\widehat{Q}^g_{(\alpha)}) = \frac{1}{m(m-1)}\sum_{i=1}^m \left(\widehat{Q}^{(i)}_\alpha(Y_g) - \widehat{\overline{Q}}_\alpha(Y_g)\right)^2$. *Similarly, for the counterfactual distribution, the variance estimation of the quantile estimator given in the second proposed method is* $V_{MC}(\widehat{Q}^C_{(\alpha)}) = \frac{1}{m(m-1)}\sum_{i=1}^m \left(\widehat{Q}^{(i)}_\alpha(Y^C) - \widehat{\overline{Q}}_\alpha(Y^C)\right)^2$. *Approximate 95% confidence intervals for quantiles* $Q_{(\alpha)}(Y_g)$ *can be constructed by using*

$$[\widehat{\overline{Q}}_\alpha(Y_g) - 1.96\sqrt{V_{MC}(\widehat{Q}^g_{(\alpha)})}, \widehat{\overline{Q}}_\alpha(Y_g) + 1.96\sqrt{V_{MC}(\widehat{Q}^g_{(\alpha)})}],$$

and for $Q^C_{(\alpha)}$ *by*

$$[\widehat{\overline{Q}}_\alpha(Y^C) - 1.96\sqrt{V_{MC}(\widehat{Q}^C_{(\alpha)})}, \widehat{\overline{Q}}_\alpha(Y^C) + 1.96\sqrt{V_{Mc}(\widehat{Q}^C_{(\alpha)})}].$$

4.8 The GB2 distribution

In this section, we present an application to the GB2 distribution. The GB2 distribution is characterized by four parameters, namely a, b, p and q. In Kleiber and Kotz (2003) and McDonald and Xu (1995), the probability density function of a $GB2(a,b,p,q)$ distribution is given by

$$f(y;a,b,p,q) = \frac{\mid a \mid y^{ap-1}}{b^{ap}\mathrm{B(p,q)}\left[1+\left(\frac{\mathrm{y}}{\mathrm{b}}\right)^{\mathrm{a}}\right]^{\mathrm{p+q}}}, \tag{4.20}$$

where $B(p,q)$ represents the function $Beta(p,q)$ with arguments p and q. Using the notation of Graf et al. (2011); Graf and Nedyalkova (2015), Equation (4.20) is rewritten as

$$f(y;a,b,p,q) = \frac{a\left(\frac{y}{b}\right)^{ap-1}}{b\mathrm{B(p,q)}\left[1+\left(\frac{\mathrm{y}}{\mathrm{b}}\right)^{\mathrm{a}}\right]^{\mathrm{p+q}}}, y > 0. \tag{4.21}$$

The parameters a, p and q are shape parameters and b is the scale parameter (Kleiber and Kotz, 2003). All of them are strictly positive. The peak of the distribution is controlled by a, the other two shape parameters control for the left and the right tail respectively. In several studies, the GB2 distribution has proved to provide a good fit for wages. Moreover, it includes several other distributions as special cases or as limiting cases (Kleiber and Kotz, 2003; McDonald and Xu, 1995)

4.8.1 The GB2 regression model

We exemplify the proposed methods to estimate the structure and composition effects at the quantile level by using the GB2 distribution. We consider the following regression model

$$\log(Y_k) = \mathbf{X}_k^\top \boldsymbol{\beta} + \log(\varepsilon_k), \tag{4.22}$$

where $\varepsilon_k \sim GB2(a,1,p,q)$. In the regression setup, conditional to $\mathbf{X}_k$, the distribution of ε_k is the distribution of Y_k, but with some different parameters. We have that $E(Y_k \mid \mathbf{X}_{k,g} = \mathbf{x}_{k,g}) = \exp(\mathbf{x}_k^\top \boldsymbol{\beta}_g)\mathrm{B(p_g + 1/a_g, q_g - 1/a_g)/B(p_g, q_g)}, \mathrm{k} \in \mathrm{U_g}$. We are thus able to fit

a conditional wage distribution for any individual in U_g, given their characteristics. As $\varepsilon_k \sim GB2(a,1,p,q)$ we have that $Y_k \mid \mathbf{X}_k = \mathbf{x}_k \sim GB2(a,\exp[\mathbf{x}_k^\top \beta],p,q)$ (see McDonald and Butler, 1990). Since ε_k follows a GB2 distribution, we refer to the model in Equation (4.22) as a GB2 regression model.

We borrow from McDonald and Butler (1990) the idea of changing the scale parameter, by expressing it as a function of the observed characteristics of the employees. In each group, $g \in \{M,F\}$, we assume that the conditional wage of $k \in U_g$, $Y_k \mid \mathbf{X}_{k,g} = \mathbf{x}_{k,g} \sim GB2(a_g,\exp(\mathbf{x}_k\beta_g),p_g,q_g)$. Thus, for each $k \in U_g$, Expression (4.21) becomes

$$f[y_k;a_g,\exp(\mathbf{x}_k^\top\beta_g),p_g,q_g] = \frac{a\left[\frac{y_k}{\exp(\mathbf{x}_k^\top\beta_g)}\right]^{a_g p_g - 1}}{\exp(\mathbf{x}_k^\top\beta_g)\mathrm{B}(\mathrm{p_g},\mathrm{q_g})\left\{1+\left[\frac{\mathrm{y_k}}{\exp(\mathbf{x}_\mathrm{k}^\top\beta_\mathrm{g})}\right]^{\mathrm{a}}\right\}^{\mathrm{p_g+q_g}}},\ y_k > 0. \tag{4.23}$$

We have that $E(Y_k \mid \mathbf{X}_{k,g} = \mathbf{x}_{k,g}) = \exp(\mathbf{x}_k^\top\beta_g)\mathrm{B}(\mathrm{p_g}+1/\mathrm{a_g},\mathrm{q_g}-1/\mathrm{a_g})/\mathrm{B}(\mathrm{p_g},\mathrm{q_g}),\mathrm{k}\in\mathrm{U_g}$. We are thus able to fit a conditional wage distribution for any individual in U_g, given their characteristics.

Remark 5 *Note that the classical Oaxaca-Blinder type decomposition given in Expression (2.6) changes if we replace Model (2.5) by Model (4.22), because $E(\log(\varepsilon_g))$ in GB2 regression is not equal to zero anymore. Instead we can write Expression (2.6) as*

$$\Delta_{GB2} = (E(\mathbf{X}_M) - E(\mathbf{X}_F))\,\beta_F + E(\mathbf{X}_M)\,(\beta_M - \beta_F) + E(\log(\varepsilon_M)) - E(\log(\varepsilon_F)), \tag{4.24}$$

where $\log(\varepsilon_M) \sim EGB2(\delta_M,\zeta_M,p_M,q_M)$ and $\log(\varepsilon_F) \sim EGB2(\delta_F,\zeta_F,p_F,q_F)$, and $EGB2(.)$ denotes the exponential GB2 distribution and δ_g and ζ_g represent the first two parameters of the $EGB2$ distribution in group g.

4.8.2 Estimation of the parameters

4.8.2.1 The algorithm

From Equation (4.23), the logarithm of the GB2 density for a given value y_k gives

$$\log[f(y_k;a,\exp(\mathbf{x}_k^\top\beta),p,q)] = \log(a) + (ap-1)\log(y_k) - (ap-1)\exp(\mathbf{x}_k^\top\beta) - \mathbf{x}_k^\top\beta - \log(\mathrm{B(p,q)} - (\mathrm{p+q})\log\left\{1+\left[\frac{\mathrm{y_k}}{\exp(\mathbf{x}_\mathrm{k}^\top\beta)}\right]^{\mathrm{a}}\right\} \tag{4.25}$$

We estimate the parameters using pseudo-maximum likelihood and a generic sample S of size n. The pseudo log-likelihood function is

$$l = \frac{\sum_{k=1}^{n} w_k \log f[y_k; a, \exp(\mathbf{x}_k^\top \beta), p, q]}{\sum_{k=1}^{n} w_k}, \tag{4.26}$$

where w_k is the survey weight allotted to individual k and $f(\cdot)$ is the density defined in Expression (4.23). The function l in Expression (4.26) is maximized w.r.t to the parameters.

The number of parameters in the log-likelihood function depends on the number of covariates in the model. If there are J covariates (including the intercept), then there are $J+3$ parameters to be estimated. We assume that a_g, p_g and q_g are the three shape parameters that characterize the conditional GB2 distribution of each wage Y_k, $k \in S_g \in \{M, F\}$, and that the scale parameter is a function of the covariates.

The following iterative algorithm is used to estimate the parameters of a GB2 distribution from data of a generic sample S, when covariates and weights are involved. The maximization of the log-likelihood function is a time-consuming procedure, because of a number local maximum points. This is why the iterative algorithm takes first into account several samples and second, the entire group. The starting points of the optimization step are very important and we explain below how they are chosen. The algorithm contains the following steps:

1. Create a matrix $\widetilde{M}$ of size $t_1 \times (J+3)$, where t_1 is fixed.

2. Draw from the initial sample S a simple random sample without replacement of size $0.90 \times n$. Let $\tilde{s}$ be a realization of this sample.

3. Fit on the wage of $\tilde{s}$ an iid GB2 distribution (without covariates) where the starting values of a, β_0, p and q are given from the Fisk distribution with $p = q = 1$. First initial values of $\widehat{a}, \widehat{\beta}_0, \widehat{p}$ and $\widehat{q}$ are obtained.

4. Use the estimated values from Step 3 as starting values for a, β_0, p, q and take as starting values for the vector $\beta = (log(\widehat{b}), 0, \ldots, 0)'$ for an iterative optimization process to maximize the log-likelihood function l. The process is applied on the data of the initial sample S, when covariates and weights are involved. Repeat t_2 times this iterative optimization.

5. The resulting parameters in Step 4 will represent line i of the matrix $\widetilde{M}$, given as

$$(\widehat{a}_i, \widehat{\beta}_{0,i}, \widehat{\beta}_{1,i}, \ldots, \widehat{\beta}_{p,i}, \widehat{p}_i, \widehat{q}_i).$$

6. Repeat Steps 2 to 5 t_1 times.

7. Select from matrix $\widetilde{M}$ the line with the estimated parameters that maximize the log-likelihood function l given in Expression (4.26).

4.8.2.2 Estimation of the standard errors using parametric bootstrap

We use a parametric bootstrap approach to estimate the standard errors of the GB2 parameters. We run the following algorithm:

1. We estimate the parameters using the algorithm described in Subsection 4.8.2.1.

2. We create a matrix M^* of size $10000 \times (J+3)$. Each line of this matrix will represent a pseudo-population following a GB2 distribution with the estimated parameters from Step 1.

3. We fill the matrix M^* with the 10000 pseudo-populations and for each one of them, we re-estimate the parameters using the same algorithm described in Subsection 4.8.2.1.

4. For each parameter, we will have a vector of 10000 estimated values from Step 3. We compute the standard deviation for each of these vectors and this will be the resulting estimated standard error based on a parametric bootstrap approach. Note that we consider the pseudo-populations to be independently and identically distributed.

4.8.2.3 Estimation of the standard errors using the sandwich estimator

The sandwich estimator is covered in various articles and books, for instance Wolter (1985), Huber (1981) or Freedman (2006). The sandwich estimator for the GB2 distribution with four parameters is discussed in Graf et al. (2011). In this Section, we show how standard errors can be estimated through the sandwich estimator, when the scale parameter is replaced by a vector.

Assume that θ_g is the vector of parameters to be estimated in group g, of the form

$$\theta_g = (a_g \; \beta_{0g} \; \beta_{1g} \; \ldots \; \beta_{Jg} \; p_g \; q_g).$$

We denote by $l(\theta)$ the log-likelihood, which should have a maximum value for a given $\widehat{\theta}$. The value of $\widehat{\theta}$ that results in the maximum value of $l(\theta)$ is obtained when

$$l'(\theta) = 0. \tag{4.27}$$

As found in Graf et al. (2011),

$$l'(\widehat{\theta}) \approx l'(\theta) + l''(\theta)[\widehat{\theta} - \theta], \tag{4.28}$$

where $l'(\widehat{\theta}_g)$ is the vector of size $(J+3) \times 1$ containing the derivatives of the log-likelihood with respect to the $J+3$ parameters, $l''(\theta)$ is the $(J+3) \times (J+3)$ matrix containing the second-order derivatives of the log-likelihood with respect to the parameters and $[\widehat{\theta} - \theta]$ is a vector of size $(J+3) \times 1$.

Putting together Equations (4.27) and (4.28), we obtain that

$$\begin{aligned} l'(\theta) + l''(\theta)[\widehat{\theta} - \theta] &\approx 0 \\ [\widehat{\theta} - \theta] &\approx -\frac{l'(\theta)}{l''(\theta)}. \end{aligned}$$

Next,

$$\begin{aligned} \text{var}(\theta) = \text{E}[\theta - \widehat{\theta}]^2 &\approx [l''(\theta)]^{-1}\{\text{E}[l'(\theta)l'(\theta)^\top] - \text{E}^2[l'(\theta)]\}[-l''(\theta)]^{-1} \\ &\approx [l''(\theta)]^{-1}\{\text{E}[l'(\theta)l'(\theta)^\top]\}[-l''(\theta)]^{-1}. \end{aligned}$$

The variance of θ is estimated as the diagonal of the matrix V of size $(J+3) \times (J+3)$ given by

$$V = [l''(\widehat{\theta})]^{-1}\{\text{E}[l'(\widehat{\theta})l'(\widehat{\theta})]^\top\}[-l''(\widehat{\theta})]^{-1} \tag{4.29}$$

The term $\text{E}[l'(\widehat{\theta}_g)l'(\widehat{\theta}_g)^\top]$ is estimated from the sample as

$$\sum_{k \in S_g}^{n_g} w_k l'(\widehat{\theta}_g)l'(\widehat{\theta}_g)^\top.$$

The first and second-order derivatives are detailed in Appendix B.

4.8.2.4 Examples

Two different settings are used to show the performance of the algorithm used to estimate the parameters of a GB2 distribution given in Subsection 4.8.2.1:

- Example 1, where we generate a conditional wage distribution for women, $Y_{k,F} = \exp[1.44 + 0.15X_{1k,F} - 0.2X_{2k,F} + \log(\varepsilon_{k,F})]$, where $\varepsilon_{k,F} \sim GB2(8, 1, 0.50, 0.90)$ are iid, $X_{1k,F} \sim Student(6)$ and $X_{2k,F} \sim Bernoulli(0.5), k = 1, \ldots n_F$, with $n_F = 8000$. The first variable is standardized, and the covariates are independent.

- Example 2, where we use a phantom independent variable $X_{5,F}$ that has no contribution to the dependent one in the women subsample and generate a conditional wage distribution for women, $Y_{k,F} = \exp[0.88 + 0.07X_{1k,F} - 0.20X_{2k,F} + 0.27X_{3k,F} + 0.25X_{4k,F} + 0X_{5k,F} + \log(\varepsilon_{k,F})]$, where $\varepsilon_{k,F} \sim GB2(6,1,0.58,0.55)$ are iid, $X_{1k,F} \sim$ and $X_{2k,F} \sim ..., k = 1,\ldots n_F$, with $n_F = 8000$. Again, the independent continuous variables are standardized and all covariates are independent.

We estimate the parameters a_F, p_F, q_F, β_F using pseudo maximum likelihood estimation. In order to show the performance of the applied algorithm described in Subsection 4.8.2.1 to estimate the parameters of a GB2 distribution, the true and the estimated parameters are shown in Tables 4.1 and 4.2, respectively, for the two examples. The standard errors of the estimated parameters are given, using respectively the sandwich estimator and the parametric bootstrap method. In both examples, the estimated parameters are close to the true ones. Note that the algorithm used estimated the value of $\beta_{F,5}$ in the second population close to 0, as expected.

Table 4.1 True and estimated parameters for women in Example 1. The standard errors of the estimated parameters are given using the sandwich estimator and the parametric bootstrap method.

		a_F	p_F	q_F	$\beta_{F,0}$	$\beta_{F,1}$	$\beta_{F,2}$
True value		8	0.50	0.90	1.43	0.15	-0.20
Estimated value		7.94	0.50	0.93	1.45	0.15	-0.20
Standard error	Sandwich	0.61	0.05	0.10	0.01	0.003	0.006
	Param. bootstrap	0.42	0.03	0.08	0.01	0.003	0.006

Table 4.2 True and estimated parameters for women in Example 2. The standard errors of the estimated parameters are given using the sandwich estimator and the parametric bootstrap method.

		a_F	p_F	q_F	$\beta_{F,0}$	$\beta_{F,1}$	$\beta_{F,2}$	$\beta_{F,3}$	$\beta_{F,4}$	$\beta_{F,5}$
True value		6	0.58	0.85	0.88	0.07	-0.20	0.27	0.25	0
Estimated value		6.47	0.53	0.77	0.87	0.07	-0.21	0.28	0.25	0
Standard error	Sandwich	0.50	0.05	0.08	0.01	0.01	0.01	0.00	0.00	0.00
	Param. bootstrap	0.47	0.05	0.08	0.02	0.00	0.01	0.00	0.00	0.01

4.8.3 Monte-Carlo study

Monte-Carlo simulation was used to show the performances of the proposed estimators when the quantiles of the counterfactual wage distribution are estimated. Two settings have been employed as follows:

- Setting 1, where we generate a conditional wage distribution for women, $Y_{k,F} = \exp[0.15 + X_{k,F} + \varepsilon_{k,F}]$, where $\varepsilon_{k,F} \sim N(0,1)$ are iid, $X_{k,F} \sim N(5,1)$, $k = 1,\ldots N_F$, with $N_F = 50000$. The covariate for the men is $X_{k,M} \sim N(4,1)$, iid, $k = 1,\ldots N_M$, with $N_M = N_F$. The correlation between $\log(Y_F)$ and X_F is about 0.70.

- Setting 2, where we generate a conditional wage distribution for women, $Y_{k,F} = \exp[1.44 + 0.15X_{k,F} + \log(\varepsilon_{k,F})]$, where $\varepsilon_{k,F} \sim GB2(8,1,0.50,0.90)$ are iid, $X_{k,F} \sim Gamma(9,2)$, $k = 1,\ldots N_F$, with $N_F = 50000$. The covariate for the men is $X_{k,M} \sim Gamma(10,2)$, iid, $k = 1,\ldots N_M$, with $N_M = N_F$. The correlation between $\log(Y_F)$ and X_F is about 0.60.

At the population level, the counterfactual distribution was computed by reweighting Y_F with the factor ψ_k. The factor ψ_k was computed as $dF^{\mathbf{X}_M}(X_F)/dF^{\mathbf{X}_F}(X_F)$. For example, for Setting 1, ψ_k is the ratio between the theoretical density of $N(4,1)$ in the points $X_{F,k}$ and the theoretical density of $N(5,1)$ in the points $X_{F,k}, k \in U_F$. Next, Expression (4.16) was used to construct the counterfactual CDF. For Setting 1, $F^{(Y_F|\mathbf{X}_F)}(y \mid X_{k,F})$ is the CDF of the log-normal distribution with parameters $\mu = \mathbf{X}_F^\top \beta_1$ and $\sigma^2 = 1$, where $\beta_1 = (0.15,1)'$. For Setting 2, $F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}_{k,F})$ is the CDF of the distribution $GB2(8,\exp[X_F^\top \beta_2],0.50,0.90)$, with $\beta_2 = (1.44,0.15)'$. For both settings, at the population level, the quantile Q_α^C was computed using

$$Q_{(\alpha)}^C = \inf\{y \mid F^C(y) \geq \alpha\}, \tag{4.30}$$

where

$$F^C(y) = \frac{\sum_{k \in U_F} \psi_k F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}_F)}{\sum_{k \in U_F} \psi_k}. \tag{4.31}$$

We use m runs and we draw in each run a sample of women and men, respectively. In Setting 1, the number of runs equals 10000 and in Setting 2, due to the time-consuming process of fitting a GB2 distribution, we used only 1000 runs. In Setting 1, we select samples of women and men respectively by simple random sampling without replacement, with sample size $n_F = n_M = 1000$. In Setting 2, we employ systematic sampling with unequal probabilities for both samples with $n_F = n_M = 10000$, where the inclusion probabilities were proportional to X_F and X_M, respectively. In each run of the Monte-Carlo simulation, we compute the quantiles of order 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, and 99%, respectively, of the counterfactual wage distribution using the estimators given by the two proposed methods, the method of Anastasiade and Tillé (2017a) (with raking calibration;

hereafter, the calibration method) and the method of DiNardo et al. (1996) (hereafter, DFL).

For each generic estimator $\widehat{Q}_\alpha^C$ of Q_α^C the following Monte-Carlo measures were used:

- the Monte-Carlo relative bias (in percentages)

$$RB_{MC}(\widehat{Q}_\alpha^C) = 100\left(E_{MC}(\widehat{Q}_\alpha^C) - Q_\alpha^C\right)/Q_\alpha^C,$$

  where $E_{MC}(\widehat{Q}_\alpha) = \sum_{i=1}^m \widehat{Q}_{i,\alpha}^C/m$, and $\widehat{Q}_{i,\alpha}^C$ is the quantile estimator of Q_α^C computed in the ith run;

- the Monte-Carlo variance

$$Var_{MC}(\widehat{Q}_\alpha^C) = \frac{1}{m-1}\sum_{i=1}^m (\widehat{Q}_{i,\alpha}^C - E_{MC}(\widehat{Q}_\alpha^C))^2.$$

- the Monte-Carlo root mean square error

$$RMSE_{MC}(\widehat{Q}_\alpha^C) = \left(Var_{MC}(\widehat{Q}_\alpha^C) + \left(B_{MC}(\widehat{Q}_\alpha^C)\right)^2\right)^{1/2},$$

  where $B_{MC}(\widehat{Q}_\alpha^C) = E_{MC}(\widehat{Q}_\alpha^C) - Q_\alpha^C$.

- the Monte-Carlo coefficient of variation (in percentages)

$$CV_{MC}(\widehat{Q}_\alpha^C) = 100\left(Var_{MC}(\widehat{Q}_\alpha^C)\right)^{1/2}/E_{MC}(\widehat{Q}_\alpha^C).$$

For the proposed methods, we estimated the parameters of the women's wage distribution at each run using the corresponding weights of women selected in the women's sample, as well as the factor ψ_k given by the method of Anastasiade and Tillé (2017a) with raking calibration. The estimated factor ψ_k was also used to compute in each run the calibration estimator for each quantile of the counterfactual distribution. Similarly, the factor ψ_k for the DFL estimator was computed in each run. We used a weighted logistic regression to compute $P(D_k = 1 \mid X_k)$ and $P(D_k = 0 \mid X_k)$, while $P(D_k = 1)$ and $P(D_k = 0)$ were estimated by weighted means $\sum_{k\in S_g} w_k / \sum_{k\in S} w_k, g \in \{M, F\}$; see Expression (4.12). All the results were computed in R using the default definition for an empirical quantile.

All the used estimators are biased. The Monte-Carlo relative biases in percentages are shown in Tables 4.3 and 4.7 for the two settings, while the values of the Monte-Carlo variances and root mean square errors are given in Tables 4.4, 4.5, 4.8, and 4.9, respectively.

The estimator of Anastasiade and Tillé (2017a) using calibration was used in Figures 4.1 and 4.2 as a benchmark in order to visualize the behavior of the other three estimators at different quantiles. The first plot in Figures 4.1 and 4.2 show the ratio between the Monte Carlo variance obtained by using the proposed methods, the method of DiNardo et al. (1996) and those of the calibration method for each of the quantiles of order 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% and 99%. The second plot in the two figures provides the ratio between the Monte Carlo bias of the proposed methods and the DFL method and those of the calibration method. Similarly, the third plot shows the ratio of the Monte-Carlo root mean square errors.

In Setting 1, the calibration and the weighted DFL methods result in estimators that have a higher relative bias for the quantiles of order 1% to 10% and for those of order 70% to 99% than for the other quantiles. In Setting 1, for all quantiles, the calibration and the DFL estimators have a similar behavior in terms of Monte-Carlo variance and root mean square error. The two proposed methods show a larger relative bias than the DFL estimator (in absolute value) at the quantiles of order 30% and 50%. Compared to the calibration estimator, the relative bias is larger at the quantile of order 60% only. However, the two methods provide at each quantile a substantial reduction of the Monte-Carlo variance, and a good behavior with respect to the root mean square error (see Figure 4.1) over the calibration and DFL estimators. The estimators obtained by the two proposed methods also display a smaller coefficient of variation than the calibration and the DFL estimators at all quantiles.

For Setting 2, the Monte-Carlo expectation of the estimated parameters of the GB2 distribution are for a, β_0, β_1, p, and q, respectively : 8.02, 1.44, 0.15, 0.50, 0.91, showing that the proposed algorithm given in Section 4.8.2 provides approximately unbiased estimates under the sampling design, for large sample sizes. In Setting 2, the two proposed methods result in estimators that have a lower Monte-Carlo variance at the extreme quantiles than the calibration and the DFL estimators. We observe the same behavior in terms of the Monte-Carlo root mean square error and, like in Setting 1, the coefficient of variation of the estimators obtained using the two proposed methods are smaller than of those using the last two methods. The

proposed methods sometimes show a larger bias and relative mean square error than the calibration estimator, but provide a reduction of the Monte-Carlo variance at each quantile (except for the quantile of order 20%; see also Figure 4.1).

Table 4.3 Setting 1: Monte Carlo relative bias (in%) of the four estimators of the counterfactual wage quantiles

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.76	0.46	0.39	0.33	0.31	0.29	0.28	0.28	0.27	0.27	0.26	0.26	0.26
Method 2	2.15	0.73	0.53	0.40	0.36	0.33	0.31	0.30	0.29	0.28	0.28	0.27	0.28
Calibration	2.42	1.13	1.68	-0.08	-0.70	-1.18	-1.06	-0.28	1.21	1.75	1.71	1.35	0.65
Weighted DFL	3.03	1.74	2.37	0.58	-0.01	-0.46	-0.27	0.50	1.98	2.57	2.51	2.08	1.38

Table 4.4 Setting 1: Monte-Carlo variance of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.14	0.34	0.58	1.21	2.28	4.22	7.93	15.64	33.73	86.13	330.26	1025.89	8860.94
Method 2	0.15	0.34	0.58	1.21	2.29	4.23	7.96	15.68	33.78	86.27	330.66	1028.16	8898.64
Calibration	0.97	1.04	1.69	2.65	4.57	6.99	13.10	26.03	56.62	161.85	582.61	1678.39	20518.87
Weighted DFL	1.01	1.20	2.08	3.53	6.19	9.82	17.50	30.76	59.07	148.23	467.80	1343.06	18044.12

Table 4.5 Setting 1: Monte-Carlo root mean square error of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.38	0.59	0.77	1.11	1.53	2.08	2.86	4.01	5.89	9.40	18.38	32.35	94.88
Method 2	0.41	0.60	0.78	1.12	1.54	2.09	2.87	4.03	5.91	9.43	18.42	32.43	95.21
Calibration	1.00	1.04	1.39	1.63	2.21	3.01	4.06	5.15	8.71	16.12	30.15	47.41	146.34
Weighted DFL	1.02	1.14	1.60	1.90	2.49	3.18	4.21	5.68	10.52	18.98	34.18	51.83	148.74

Table 4.6 Setting 1: Monte-Carlo coefficient of variation (in %) of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	5.44	3.33	2.62	2.06	1.82	1.69	1.62	1.59	1.60	1.63	1.72	1.81	2.03
Method 2	5.43	3.33	2.62	2.06	1.82	1.69	1.62	1.60	1.60	1.63	1.72	1.81	2.03
Calibration	13.91	5.77	4.43	3.06	2.60	2.21	2.11	2.07	2.05	2.20	2.25	2.29	3.07
Weighted DFL	14.10	6.15	4.87	3.51	3.00	2.60	2.42	2.23	2.08	2.09	2.00	2.03	2.86

Table 4.7 Setting 2: Monte Carlo relative bias (in%) of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	-1.74	-1.31	-1.11	-0.86	-0.64	-0.43	-0.20	0.05	0.36	0.75	1.38	1.98	3.49
Method 2	-1.66	-1.30	-1.10	-0.85	-0.64	-0.43	-0.20	0.05	0.36	0.76	1.39	1.98	3.51
Calibration	-3.54	-1.29	-1.14	-0.64	-0.44	-0.30	-0.19	0.10	0.38	0.55	1.35	2.41	4.43
Weighted DFL	-3.17	-0.86	-0.74	-0.22	-0.03	0.18	0.31	0.60	0.95	1.17	2.09	3.23	5.46

Table 4.8 Setting 2: Monte-Carlo variance of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.004	0.009	0.049
Method 2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.004	0.009	0.049
Calibration	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.004	0.012	0.063
Weighted DFL	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.003	0.003	0.006	0.013	0.067

Table 4.9 Setting 2: Monte-Carlo root mean square error of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.06	0.06	0.06	0.06	0.05	0.04	0.04	0.03	0.05	0.09	0.18	0.30	0.72
Method 2	0.06	0.06	0.06	0.06	0.05	0.04	0.04	0.03	0.05	0.09	0.18	0.30	0.72
Calibration	0.10	0.07	0.06	0.05	0.05	0.04	0.04	0.04	0.06	0.07	0.18	0.37	0.90
Weighted DFL	0.10	0.05	0.05	0.03	0.04	0.04	0.04	0.07	0.10	0.14	0.27	0.48	1.10

Table 4.10 Setting 2: Monte-Carlo coefficient of variation (in %) of the four estimators of the counterfactual wage quantiles.

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	1.45	0.80	0.62	0.52	0.48	0.45	0.43	0.41	0.40	0.41	0.51	0.66	1.10
Method 2	1.45	0.80	0.62	0.52	0.48	0.45	0.43	0.41	0.40	0.41	0.51	0.66	1.10
Calibration	1.87	1.13	0.73	0.51	0.59	0.47	0.46	0.47	0.52	0.44	0.53	0.73	1.23
Weighted DFL	1.92	1.12	0.75	0.53	0.61	0.50	0.49	0.52	0.55	0.53	0.61	0.77	1.26

4.9 Application to real data

4.9.1 The dataset

We use real data from the Survey on Earnings Structure from 2012, proived by the Swiss Federal Statistical Office. We have a sample of 5643 employees working in the economic activity classified as "Manufacture of computer, electronic and optical products". All the cases where there is missing information are removed. We also removed observations where the monthly wage is less than CHF1000 for a full-time job, because we consider them to be data collection errors. The employees have worked at least one hour during the month of October, are between 15 and 64 years

old and perform tasks that require wide knowledge in a precise field. There are 1880 women and 3763 men in the sample.

4.9.2 Descriptive statistics

In Table 4.11, we present some descriptive statistics of the observed hourly wages from the dataset. Next, Table 4.12 and Figure 4.3 represent a descriptive summary of the estimated wage distributions of men and women, respectively.

Table 4.11 Descriptive statistics of the hourly wages of men, women and the entire dataset

Parameter	Men	Women	Entire dataset
Mean	46.61	37.68	43.64
Median	43.97	35.87	41.30

Table 4.12 Characteristics of the wage distributions of men and women

Characteristic	Men	Women
Kurtosis	14.25	11.46
Skewness	2.31	1.86

The estimated wage distribution of men is more assymetric and heavy-tailed than that of women. We expected this, since it is more likely for men to have more extreme wages than women.

4.9.3 The model

We used three explanatory variables in the model. These are:

- the age of the employee
- the education level – an ordinal variable with 9 categories
- the professional position – an ordinal variable with 5 categories.

The variable "age" is standardized. For each category of the ordinal variables, a binary variable is created. The first category is dropped, to avoid multicollinearity. In each group g, there are 17 parameters to be estimated.

4.9.4 Estimated parameters

We fit a GB2 distribution as a conditional distribution of women's wages given the previous characteristics. In Table 4.12, the estimated parameters for the women's sample are given, together with their estimated standard errors. The standard errors are estimated using the sandwich estimator.

Table 4.13 Estimated parameters in the women's sample and their estimated standard errors

Parameter	$\widehat{a}_F$	$\widehat{\beta}_{0F}$	$\widehat{\beta}_{age}$	$\widehat{\beta}_{educ1}$	$\widehat{\beta}_{educ2}$	$\widehat{\beta}_{educ3}$	$\widehat{\beta}_{educ4}$	$\widehat{\beta}_{educ5}$	$\widehat{\beta}_{educ6}$	$\widehat{\beta}_{educ7}$	$\widehat{\beta}_{educ8}$	$\widehat{\beta}_{prof1}$	$\widehat{\beta}_{prof2}$	$\widehat{\beta}_{prof3}$	$\widehat{\beta}_{prof4}$	$\widehat{p}_F$	$\widehat{q}_F$
Estimated value	10.38	3.80	0.07	-0.06	-0.09	-0.29	-0.29	0.29	0.27	0.19	0.05	0.10	0.04 -	0.16	-0.11	0.76	0.72
Standard error	1.86	0.16	0.01	0.15	0.14	0.14	0.14	0.06	0.06	0.06	0.18	0.06	0.06	0.06	0.06	0.18	0.18

In Figures 4.4 and 4.5, we show the QQ-plots of the residuals of the log-linear model (with the log of the wages as dependent variable and the same characteristics as covariates) and the GB2 model, respectively. We observe an improvement of the QQ plot when we fit a GB2 model compared to the normal model. In Figure 4.3, departures of the points from the straight line for some higher and lower can be explained by the fact that used employees characteristics may not have a good explanatory power at the tails.

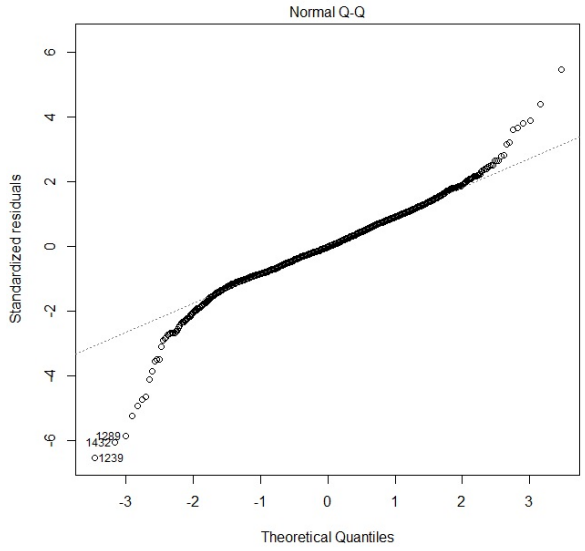


Fig. 4.4 QQ-plot for a log-normal model fitted on real data.

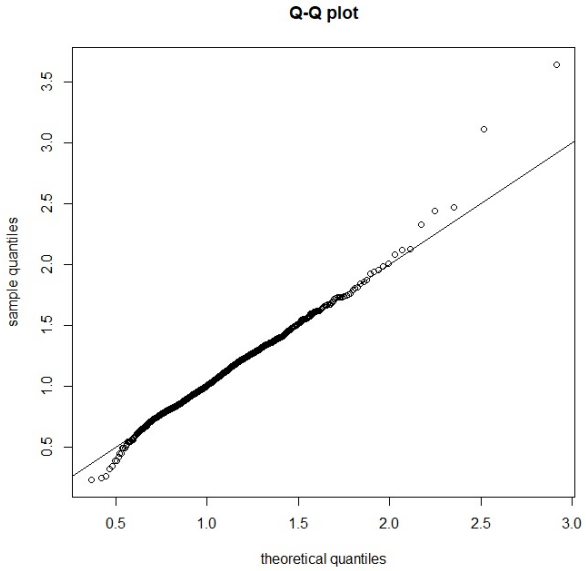


Fig. 4.5 QQ-plot for a GB2 model fitted on real data.

In Table 4.14, we show the estimated quantiles of men's wages, of the counterfactual wage distributions computed using the four methods and finally, those of women's wages. The estimated quantiles of the wage distribution of men and women were estimated using Expression 4.9, those for the calibration and the weighted DFL with Equation 4.15, for Method 1, Equation 4.18 and finally, for Method 2, Equation 4.10.

Table 4.14 Estimated quantiles of the empirical wage distribution of men, of the empirical wage distribution of women and of women's estimated counterfactual wage distribution computed using the four methods. Application to real data.

		1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Men		24.32	29.26	31.70	35.61	38.79	41.41	44.38	47.00	50.68	56.69	64.81	73.99	99.19
Counterfactual	Meth.1	20.73	25.63	28.33	31.77	34.46	36.96	39.52	42.34	45.71	50.14	57.18	63.90	80.23
	Meth.2	20.82	25.64	28.34	31.78	34.47	36.97	39.53	42.36	45.73	50.16	57.21	63.97	80.39
	Calib	21.72	25.81	28.30	31.71	34.47	36.89	39.18	41.90	45.63	49.40	56.60	65.83	85.99
	Weighted DFL	21.51	25.61	28.11	31.53	34.22	36.69	38.75	41.79	45.05	49.16	56.01	64.38	85.99
Women		20.79	23.63	25.61	28.52	30.86	33.04	35.31	37.50	40.24	43.98	49.79	55.40	75.61

As expected, the four estimators provide quantiles estimated of the counterfactual distribution that are between the range of the estimated gender quantiles. At the lowest and the highest order quantiles, the results using the calibration method and the DFL method are higher than for the other two methods.

In Table 4.15, we show the estimated composition and strucure effects of the wage difference at selected quantiles, computed using the three methods proposed in this thesis, as well as the weighted DFL method. Using all four methods, the structure

effects accounts for more than half of the wage difference between men and women at all quantiles. For the quantiles of order 1% and 99%, the estimated proportion is fairly high. However, the results at these quantiles should be interpreted with care. As seen from the QQ-plot shown in Figure 4.5, the points along the tails to not fall on the straight line, which means that there could be a possible misfit for extreme values of wages.

4.10 Conclusions

This chapter had a two-fold purpose: first, to introduce parametric estimators of quantiles of the wage and counterfactual wage distributions. Second, to use the GB2 distribution in the context of modeling wages using survey data. The proposed estimators are based on a strong assumption, namely that wages follow a parametric distribution. If this assumption holds, we expect that the proposed estimators will reduce the variance compared to the estimators given by DiNardo et al. (1996) and Anastasiade and Tillé (2017a), since auxiliary information is introduced in the quantile estimation. In our Monte-Carlo simulation results, the proposed estimators show this variance reduction.

Strictly speaking, the proposed methods are design-based estimators, even though they use an underlying model between the variable of interest and the covariates. As for all design-based estimators, the variance reduction is expected when an important correlation between the variable of interest and the covariates is detected. In Setting 1, the correlation between the logarithm of women's wage and the covariates is larger compared to Setting 2 (0.70 versus 0.60). This difference could explain that the variance reduction of the proposed methods compared to the calibration method is less important in Setting 2 than in Setting 1.

We use in Setting 2 and in Section 4.9 a GB2 distribution for the conditional wage distribution. It is worthy to mention that the parameter estimation is difficult to perform for this distribution. First, when covariates and weights are introduced, the maximization of the pseudo log-likelihood function (see Section 4.8.2) shows multiple local maximum points and the choice of the starting points of the algorithm have a crucial importance. The proposed algorithm is able however to provide approximately unbiased estimators of the parameters for large sample sizes as shown by Setting 2 in Section 4.8.3. Second, as for for iid GB2 fits (already underlined by Graf and Nedyalkova, 2017), the sample sizes should be very large (for example, in Setting 2, it is 10000).

In spite of these two difficulties, the GB2 distribution includes many distributions as special cases, and we expect that its use provides a good fit of wages. In our knowledge, the application shown in Section 4.9 is the first one where the GB2 distribution is fitted on real wages conditional on covariates and also using survey weights.

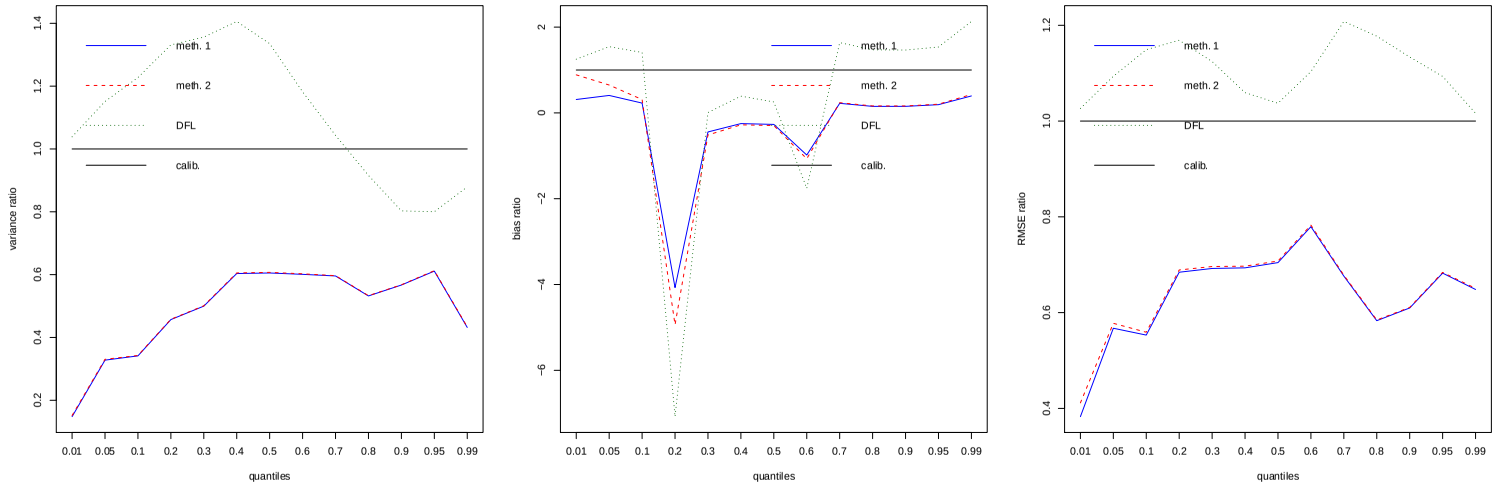


Fig. 4.1 Setting 1, left panel: ratio between the Monte Carlo variance obtained by using the proposed methods, the DFL method and that of the calibration method for each quantile; middle: ratio between the Monte Carlo bias obtained by using the proposed methods and that of the calibration method for each quantile; right panel: ratio between the Monte Carlo RMSE obtained by using the proposed methods and that of the calibration method for each quantile. The horizontal line drawn at level 1 on the y-axis corresponds to the calibration method.

Fig. 4.2 Setting 2, left panel: ratio between the Monte Carlo variance obtained by using the proposed methods, the DFL method and that of the calibration method for each quantile; middle: ratio between the Monte Carlo bias obtained by using the proposed methods and that of the calibration method for each quantile; right panel: ratio between the Monte Carlo RMSE obtained by using the proposed methods and that of the calibration method for each quantile. The horizontal line drawn at level 1 on the y-axis corresponds to the calibration method.

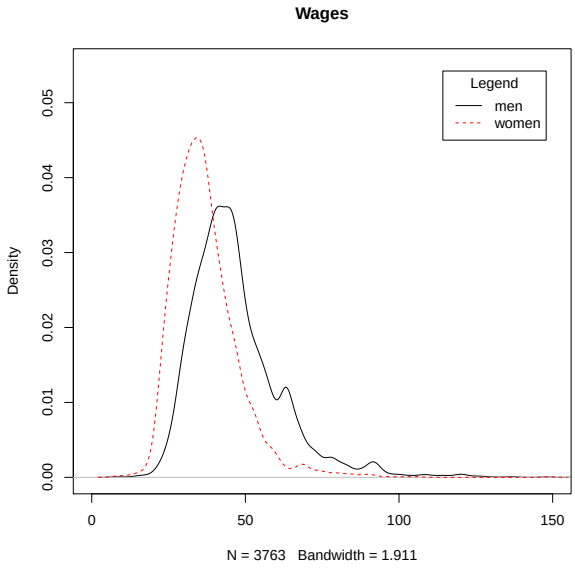


Fig. 4.3 Estimated wage densities of men and women in the dataset.

Table 4.15 Estimated composition and structure effects of the wage difference at selected quantiles, computed using the four methods. Application to real data.

Quantile	Method	Total	= Composition effect (%)	+ Structure effect (%)
1%	Method 1	3.53	-0.06	3.59
			(-0.02%)	(102%)
	Method 2	3.53	0.03	3.50
			(1%)	(99%)
	Calibration	3.53	0.93	2.60
			(28%)	(74%)
	Weighted DFL	3.53	0.72	2.81
			(20%)	(80%)
10%	Method 1	6.09	2.72	3.37
			(45%)	(55%)
	Method 2	6.09	2.73	3.36
			(45%)	(55%)
	Calibration	6.09	2.69	3.40
			(44%)	(56%)
	Weighted DFL	6.09	2.50	3.59
			(41%)	(59%)
20%	Method 1	7.09	3.25	3.84
			(46%)	(54%)
	Method 2	7.09	3.26	3.83
			(-3.5%)	(54%)
	Calibration	7.09	3.19	3.90
			(45%)	(55%)
	Weighted DFL	7.09	3.01	4.08
			(42%)	(58%)
50%	Method 1	9.07	4.21	4.86
			(46%)	(54%)
	Method 2	9.07	4.22	4.85
			(47%)	(53%)
	Calibration	9.07	3.87	5.20
			(43%)	(57%)
	Weighted DFL	9.07	3.44	5.63
			(38%)	(62%)
80%	Method 1	12.71	6.16	6.55
			(48%)	(52%)
	Method 2	12.71	6.18	6.53
			(49%)	(51%)
	Calibration	12.71	5.42	7.29
			(43%)	(57%)
	Weighted DFL	12.71	6.22	7.53
			(41%)	(59%)
90%	Method 1	15.02	7.39	7.63
			(49%)	(51%)
	Method 2	15.02	7.42	7.60
			(49%)	(51%)
	Calibration	15.02	6.81	8.21
			(45%)	(55%)
	Weighted DFL	15.02	6.22	8.80
			(41%)	(59%)
99%	Method 1	23.58	4.62	18.96
			(20%)	(80%)
	Method 2	23.58	4.78	18.80
			(20%)	(80%)
	Calibration	23.58	10.38	13.20
			(44%)	(56%)
	Weighted DFL	23.58	10.38	13.20
			(44%)	(56%)

Chapter 5

Conclusions

In the Universal Declaration of Human Rights, adopted in 1948, it is stated that "Everyone, without any discrimination, has the right to equal pay for equal work." In our days, pay equality may appear, at least hypothetically, as a normal principle. Women's participation in the labor force has increased in the past years and their education levels are similar to those of men. However, there is still evidence that they are paid less then men, even when performing similar jobs.

The wage discrimination is defined as the observed difference in pay that two groups of individuals sharing the same characteristics and who perform the same tasks. This is the basis of the so-called decomposition methods, that isolate the difference in wages due to differing characteristics from that which is not. The latter is attributed to discrimination. This idea was put forward starting with the seminal work of Blinder (1973) and Oaxaca (1973). This work was the departing point of other different decomposition methods.

In this thesis, we revisited the technique proposed by Blinder (1973) and Oaxaca (1973) and that of DiNardo et al. (1996). The first one is based on a linear regression that assumes a relationship between the wage (or the logarithm of the wage) of an individual and their characteristics. The difference in average wages of men and women, respectively, is divided into the part that is explained by the differences in characteristics and into the part that is not. This difference is obtained by estimating *women's counterfactual average wage*, which in this thesis, is interpreted as what women would earn on average if they had the same characteristics of men. However, as mentioned in Chapter 4, it can also be interpreted as the bonus that men have over women. The drawback of this technique is that the decomposition can not be extended to other parameters, such as quantiles. This is precisely what the second method discussed in this thesis achieves. DiNardo et al. (1996) proposed a method

based on logistic regression and reweighting that allows not only a decomposition at the average levels, but also at quantiles. DiNardo et al. (1996) extend the concept of the counterfactual average wage to the *counterfactual wage distribution*. This distribution is interpreted as the wage distribution of women if they had the same characteristics as men. Therefore, questions such as whether the unexplained part in the wage difference is higher for lower-paid jobs can be explored. In Chapter 2, we discussed the original method of Blinder (1973) and of Oaxaca (1973) and that of DiNardo et al. (1996) and re-expressed them by taking into account survey weights. In Chapter 3, we proposed a nonparametric method based on the calibration technique of Deville and Särndal (1992). Two instances of calibration were suggested, the linear and the raking-ratio case. In both instances, the idea is to reweigh women's distributions of characteristics such that their average characteristics are the same with those of men. The three methods were illustrated in two applications on real data supplied by the Swiss Federal Statistical Office. In the first application, a comparison of the three methods was shown, on data from 2008 related to employees in the private sector. In the second application, a comparison between the three methods was shown, for data related to the public and private sectors and the counterfactual wage distribution of women was estimated using the raking-ratio case. It was thus shown that in the public sector, the discrimination level is lower in the public sector and that except for those in the first and last quantiles, employees in the public sector earn higher wages than in the private one.

In Chapter 4, we proposed two parametric methods to estimate a counterfactual wage distribution conditional on some characteristics. We assume that wages follow a certain distribution with a parameter that is a function of the individual's characteristics. Since wage distributions are heavy-tailed and asymmetric, the idea is to capture the shape of these distributions and to achieve and estimation of the discrimination level at the quantile level. Two methods were proposed, one in which we express the quantiles as the inverse of the model cumulative distribution function and one based on simulations, for the case in which the inverse of the cumulative distribution function cannot be computed. An application on real data from the Swiss Federal Statistical Office is shown in the case of a regression where the error term and the wage follow a generalized beta of the second kind distribution (GB2). The distribution in its original form has three shape and a scale parameters. We express the scale parameter as a function of the individuals' characteristics. We show the estimation algorithm of the parameters, as well as of their standard errors. The standard errors are estimated through the sandwich estimator or through a

simulation approach based on parametric bootstrap. Finally, the quantiles of the counterfactual wage distribution of women are computed using the raking-ratio instance, the method of DiNardo et al. (1996) and the two methods proposed in this chapter. The two proposed methods reduce in our examples the Monte-Carlo variance compared to the approach given by calibration and the method of DiNardo et al. (1996) when the assumed model is correct.

It is worth mentioning the fact that the results of the three methods proposed in this thesis highly depend on the chosen covariates. The composition and the structure effect can account for varying proportions of the observed differences if we use different models. This is a drawback of all decomposition methods that take into account employee characteristics and to our knowledge, there is no way that the results can be fixed, regardless of the covariates used.

Decomposition methods share the idea that the discrimination level between two groups is the part of the wage difference that is not explained by the difference in the characteristics of the two groups. However, it is rather difficult to find employees who share the exact same background and who have the same jobs. Therefore, the word "discrimination" should be interpreted with care. The literature is rich in studies that capture potential factors that may affect career paths and consequently, wages. In a recent study done with Australian adolescents, Baxter (2017) showed that 15.8% of boys and 2.6% of girls saw themselves in a job related to technical and trade. In contrast, 5.7% of men and 10.7% of girls considered choosing a job related to services, such as health or protective services. The actual percentages on the labor market differ: there are 23% of men and 5% of women in technical and trade related jobs. In the services related jobs, there are 1.6% men and 9.1% women. The adolescents who were part of this study reported talking to their parents about their future career choices. Therefore, the family can have an influence on the professional paths that children later take on.

A report of the Swiss Federal Statistical Office showed that in 2006, in 54% of couples with no children, both partners worked full-time. When the first child arrives in the family, the percentage dropped to 8%, until the child turned 6, after which it increased to 12%. In 9% of couples without children, the man worked full-time and the woman did not have any job. This percentage increased to 37.5% once a child entered the family (Swiss Federal Statistical Office, 2006).

The expectations related to wage have also been investigated. For instance, Bonnard and Giret (2014) showed that women enrolled in undergraduate studies expect lower wage than their male counterparts. Moreover, women tend to show less self-

confidence when evaluating a job description and choose not to apply if they do not fit in the qualification requirements fully, as opposed to men, who do apply. Mohr (2015) showed that 25% of women and 13% of men cited this reason when choosing not to apply for a job.

Finally, studies have also been done on society's perception on gender roles. In her book, Goodman (2000) claims that "in the social construction of gender, it does not matter what men and women actually do; it does not even matter if they do exactly the same thing. The social institution of gender insists only that what they do is perceived as different". The author gives the example of women enrolled in the US Marine who are required to wear make-up, as a "part of a deliberate policy of making them clearly distinguishable from men Marines" and quotes a drill instructor saying that some women who enroll in the Marine "have the preconceived idea that going into the military means that they can still be a tomboy. They don't realize that you are a *Woman* Marine."

The influence of family, individual expectations and finally, the image that society creates around gender roles can impact the choices of professional careers and consequently, wages. These are all factors that should be kept in mind when portraiting the labor market. This is why the term "discrimination" should be interpreted with care. In this thesis, we followed the guidelines of decomposition methods, that isolate the part of the wage difference that is attributable to differing characteristics from the part that is not and we termed the latter as "discrimination". However, we are aware that our results may be affected by the hidden mechanisms that we could not capture in the data and we do not claim that our results are strictly due to a discriminating labor market.

Appendix A

Datasets

For all the applications, we used data issued from the Survey on Earnings Structure. It is a mandatory survey conducted every two years by the Swiss Federal Statistical Office with public and private institutions. For the application part in Section 3.4, we used the data issued from this survey in 2008 and in Sections 3.5 and 4.9, data from 2012.

A.1 Dataset used in Section 3.4

As mentioned in Section 3.4.2, the regression model used includes eight explanatory variables, out of which three are quantitative, four are qualitative and one is binary. For the qualitative values, we have the following categories (for the regression model, we excluded the first category):

- education level
    1. university (haute école universitaire)
    2. specialized college (haute école spécialisée, haute école pédagogique)
    3. professional education (formation professionnelle supérieure, écoles supérieures)
    4. teaching certificate (brevet d'enseignement)
    5. school leaving examination certificate (maturité)
    6. apprenticeship (apprentissage complet)
    7. practical education acquired in a company (formation acquise en entreprise)

8. without completed professional education (sans formation professionnelle complète)
9. other completed education (autres formations complètes)
10. missing (valeur manquante)

- qualification requirements
  1. difficult tasks
  2. independent and highly qualified tasks
  3. specialized tasks requiring professional knowledge
  4. simple and repetitive tasks
- region of the institution
  1. Cantons of Vaud, Valais, Geneva
  2. Cantons of Berne, Fribourg, Solothurn, Neuchâtel, Jura
  3. Cantons of Basel-Stadt, Basel-Landschaft, Aargau
  4. Canton of Zürich
  5. Cantons of Glaris, Schaffhausen, Appenzell Innerrhoden, Appenzell Ausserrhoden, St. Gallen, Grisons, Thurgau
  6. Cantons of Lucerne, Uri, Schwytz, Obwald, Nidwald, Zug
  7. Canton of Ticino
- economic sector - for the description of the economic activities see Swiss Federal Statistical Office (2008)
  1. Crop and animal production, hunting and related activities; Forestry and logging; Manufacture of food products; Manufacture of leather and related products; Manufacture of wood and of products of wood and cork except furniture. Manufacture of articles of straw and plaiting materials; Manufacture of paper and paper products; Printing and reproduction of recorded media; Manufacture of coke and refined petroleum products; Manufacture of chemicals and chemical products; Manufacture of basic pharmaceutical products and pharmaceutical preparations; Manufacture of rubber and plastic products; Manufacture of other non-metallic mineral products; Manufacture of fabricated metal products, except machinery

and equipment; Manufacture of computer, electronic and optical products; Manufacture of electrical equipment; Manufacture of motor vehicles, trailers and semi-trailers; Manufacture of other transport equipment; Repair and installation of machinery and equipment; Water collection, treatment and supply.

2. Wholesale and retail trade and repair of motor vehicles and motorcycles.
3. Water transport; Air transport; Warehousing and support activities for transportation.
4. Accommodation.
5. Programming and broadcasting activities; Telecommunications; Computer programming, consultancy and related activities.
6. Insurance, reinsurance and pension funding, except compulsory social security; Activities auxiliary to financial services and insurance activities;
7. Activities of head offices, management consultancy activities; Architectural and engineering activities, technical testing and analysis; Scientific research and development; Advertising and market research; Other professional, scientific and technical activities.
8. Security and investigation activities.
9. Education.
10. Creative, arts and entertainment activities; Libraries, archives, museums and other cultural activities; Gambling and betting activities; Sports activities and amusement and recreation activities.

A.2 Dataset used in Section 3.5.1

There are eight variables used for calibration, out of which four are qualitative, two are quantitative and one is binary. Among the qualitative variables, three are the same as in Section 3.4.2 (education level, region and qualification requirements. For the qualification requirements, the order of the categories is inverted). The fourth qualitative variable is the professional position, which has the following categories (the first category was excluded from the model):

1. management – level 1 (cadre supérieur)
2. management – level 2 (cadre moyen)

3. management – level 3 (cadre inférieur)
4. in charge of the execution of works (responsable de l'exécution de travaux)
5. not in management (sans fonction de cadre)

A.3 Dataset used in Section 4.9.1

For Section 4.9.1, we only took into consideration individuals working in the economic activity "Manufacture of computer, electronic and optical products (NOGA code 26). Out of the three variables used, one is quantitative and two (education level and professional position) are qualitative. The categories of the qualitative variables are the same as in Section 3.5.1.

Appendix B

First and second-order derivatives

As already mentioned in Chapter 4, the logarithm of the GB2 density for a given point y_k is

$$\log[f(y_k;a,\exp(\mathbf{x}_k^\top\beta),p,q)] = \log(a)+(ap-1)\log(y_k)-(ap-1)\exp(\mathbf{x}_k^\top\beta)-\mathbf{x}_k^\top\beta- \log(\mathrm{B}(p,q)-(p+q)\log\left\{1+\left[\frac{y_i}{\exp(\mathbf{x}_k^\top\beta)}\right]^a\right\}. \tag{B.1}$$

Using Equation (B.1), the pseudo log-likelihood function is

$$l = \frac{\sum_{i=1}^n w_k \log f[y_k;a,\exp(\mathbf{x}_k^\top\beta),p,q]}{\sum_{i=1}^n w_k}, \tag{B.2}$$

B.1 First-order derivatives

The log-likelihood $\log(f)$ in Equation (B.2) is derived w.r.t. the parameters a,β,p,q. At each point k, the four first-order derivatives will be stored in a $(J+3)\times 1$ vector, where J is the number of covariates.

$$l'[f(y_k)] = \left(\frac{\partial\log(f)}{\partial a},\frac{\partial\log(f)}{\partial\beta_0},\ldots,\frac{\partial\log(f)}{\partial\beta_j},\frac{\partial\log(f)}{\partial p},\frac{\partial\log(f)}{\partial q}\right)^\top.$$

B.1.1 First-order derivative w.r.t. a

$$\begin{aligned}\frac{\partial \log[f(y_k)]}{\partial a} &= \frac{1}{a} + p\log(y_k) - p\mathbf{x}_k^\top\beta - (p+q)\frac{1}{1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a \log(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}) \\ &= \frac{1}{a} + p(\log(y_k) - \mathbf{x}_k^\top\beta) - (p+q)\frac{1}{1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a (\log(y_k) - \mathbf{x}_k^\top\beta) \\ &= \frac{1}{a} + (\log(y_k) - \mathbf{x}_k^\top\beta)\left[p - (p+q)\frac{1}{1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right].\end{aligned}$$

B.1.2 First-order derivatives w.r.t. p

$$\frac{\partial \log[f(y_k)]}{\partial p} = \psi(p+q) - \psi(p) + a[\log(y_k) - \mathbf{x}_k^\top\beta] - \log\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right].$$

B.1.3 First-order derivatives w.r.t. q

$$\frac{\partial \log[f(y_k)]}{\partial q} = \psi(p+q) - \psi(q) - \log\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right].$$

B.1.4 First-order derivatives w.r.t. β

$$\begin{aligned}\frac{\partial \log[f(y_k)]}{\partial \beta} &= \frac{\partial \log(f)}{\partial b}\frac{\partial b}{\partial \beta_k} \\ &= \left[-ap + a(p+q)\frac{\left(\frac{y_k}{b}\right)^a}{1+\left(\frac{y_k}{b}\right)^a}\right]\frac{1}{\exp(\mathbf{x}_k\beta)}\exp(\mathbf{x}_k\beta)x_{ik} \\ &= \left[-ap + a(p+q)\frac{\left(\frac{y_k}{b}\right)^a}{1+\left(\frac{y_k}{b}\right)^a}\right]x_{ik}.\end{aligned}$$

B.2 Second-order derivatives

B.2.1 Second-order derivatives w.r.t a

Putting together the second order derivatives w.r.t. the parameters will result in a matrix M of size $(J+3)\times(J+3)$, where J is the number of covariates in the model that replaces the scale parameter. The elements of the matrix M are given below.

$$\begin{aligned}(*) &= \frac{\partial \log(f)}{\partial a} \\ &= \frac{1}{a} + [\log(y_k) - \mathbf{x}_k^\top\beta]\left[p - (p+q)\frac{1}{1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right].\end{aligned}$$

Element $M[1,1]$ of the matrix is

$$\begin{aligned}
\frac{\partial(*)}{\partial a} &= -\frac{1}{a^2} + [\log(y_k) - \log(\exp(\mathbf{x}_k^\top \beta))] \left[0 - \frac{(p+q)\left(\frac{y_k}{\exp(\mathbf{x}_k^{top}\beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^{top}\beta)}\right)^a]}{\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a\right]^2} \right. \\
&\quad \left. - \frac{(p+q)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)}{\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a\right]^2} \right] \\
&= -\frac{1}{a^2} + \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right) \left[\frac{(p+q)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right) \left[\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a - 1 - \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a \right]}{\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a\right]^2} \right] \\
&= -\frac{1}{a^2} - \frac{(p+q)\log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)}{\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a\right]^2}.
\end{aligned}$$

The elements $M[1,J+2]$ and $M[J+2,1]$ are

$$\begin{aligned}
\frac{\partial(*)}{\partial p} &= 0 + [\log(y_k) - \exp(\mathbf{x}_k^\top \beta)] \left[1 - \frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a} \right] \\
&= \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right) - \frac{\log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a} \\
&= \frac{\log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right) + \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a - \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a} \\
&= \frac{\log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)}\right)^a}.
\end{aligned}$$

The elements $M[1, J+3]$ and $M[J+3, 1]$ are

$$\frac{\partial(*)}{\partial q} = 0 + [\log(y_k) - \exp(\mathbf{x}_k^\top \beta)] \left[0 - \frac{1}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a} \right] \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a$$

$$= -\log\Big(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \Big) \frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a}{1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a}.$$

B.2.2 Second-order derivatives w.r.t. p

$$(**) = \frac{\partial \log(f)}{\partial p} = \psi(p+q) - \psi(p) + a[\log(y_k) - \exp(\mathbf{x}_k^\top \beta)] - \log[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a].$$

The element $M[J+2, J+2]$ is

$$\frac{\partial(**)}{\partial p} = \psi'(p+q) - \psi'(p).$$

The elements $M[J+2, J+3]$ and $M[J+3, J+2]$ are

$$\frac{\partial(**)}{\partial q} = \psi'(p+q).$$

B.2.3 Second-order derivatives w.r.t. q

We have the following

$(***) = \frac{\partial \log(f)}{\partial q} = \psi(p+q) - \psi(q) - \log\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a \right]$ We have to compute $\frac{\partial(***)}{\partial q}$. This will be element $M[J+3, J+3]$.

$$\frac{\partial(***)}{\partial q} = \psi'(p+q) - \psi'(q).$$

B.2.4 Second-order derivatives w.r.t. β

$$
\begin{aligned}
(****) &= \frac{\partial \log(f)}{\partial \beta} \\
&= \frac{\partial \log(f)}{\partial b} \frac{\partial b}{\partial \beta} \\
&= [-ap + a(p+q) \frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a}{\left[1 + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top \beta)} \right)^a \right]}] x_{ij}.
\end{aligned}
$$

The elements $M[1,j]$ and $M[j,1]$ are

$$\frac{\partial(****)}{\partial a} = [-p + -(p+q)\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]} +$$
$$+a(p+q)x_{ij}\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}]$$
$$-a(p+q)x_{ij}\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}]$$
$$= [-p + (p+q)\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^{2a}}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}$$
$$+a(p+q)\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a - \left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}]x_{ij}$$
$$= [-p + (p+q)\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a + \left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^{2a} + a\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a \log\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}]x_{ij}.$$

The elements $M[J+2,J]$ and $M[J,J+2]$ are

$$\begin{aligned}\frac{\partial(****)}{\partial p} &= [-a+a\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}]x_{ij} \\ &= [\frac{-a-a\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a+a\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}]x_{ij} \\ &= [\frac{-a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}]x_{ij}.\end{aligned}$$

The elements $M[J+3,J]$ and $M[J,J+3]$ are

$$\frac{\partial(****)}{\partial q} = [a\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}]x_{ij}.$$

The elements $M[j,j]$ are

$$\begin{aligned}
\frac{\partial(****)}{\partial\beta} &= ax_{ij}(p+q)\frac{y^a(-a)[\exp(\mathbf{x}_k^\top\beta)]^{-a-1}\exp(\mathbf{x}_k^\top\beta)x_{ij}(1+y_k^a[\exp(\mathbf{x}_k^\top\beta)]^{-a})}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2} \\
&\quad - \frac{y^a[\exp(\mathbf{x}_k^\top\beta)]^{-a}y_k^a(-a)[\exp(\mathbf{x}_k^\top\beta)]^{-a-1}\exp(\mathbf{x}_k^\top\beta)x_{ij}}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2} \\
&= ax_{ij}(p+q)\frac{y_k^a(-a)[\exp(\mathbf{x}_k^\top\beta)]^{-a}x_{ij}[1+y^a\exp(\mathbf{x}_k^\top\beta)^{-a}]-y_k^{2a}[\exp(\mathbf{x}_k^\top\beta)]^{-2a}(-a)x_{ij}}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2} \\
&= ax_{ij}(p+q)\frac{-a\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a x_{ij}\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]-\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^{2a}(-a)x_{ij}}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2} \\
&= ax_{ij}(p+q)\frac{-ax_{ij}\left[\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^{2a}-\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^{2a}\right]}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2} \\
&= \frac{-a^2x_{ij}^2(p+q)\left[\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}.
\end{aligned}$$

When deriving with respect to β_j and β_k, $j\neq k$, then

$$\frac{\partial(****)}{\partial\beta_j\beta_k} = -a^2(p+q)x_{ij}x_{ik}\frac{\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a}{\left[1+\left(\frac{y_k}{\exp(\mathbf{x}_k^\top\beta)}\right)^a\right]^2}.$$

Bibliography

Anastasiade, M.-C., Matei, A., and Tillé, Y. (2018). Estimation of a Counterfactual Wage Distribution Using Survey Data. Working paper, Institute of Statistics, University of Neuchâtel.

Anastasiade, M.-C. and Tillé, Y. (2017a). Decomposition of Gender Wage Inequalities through Calibration: Application to the Swiss Structure of Earnings Survey. *Survey Methodology*, 43(2):211–234.

Anastasiade, M.-C. and Tillé, Y. (2017b). Gender Wage Inequalities in Switzerland: the Public versus the Private Sector. *Statistical Methods & Applications*, 26(2):293–316.

Bandourian, R., McDonald, J., and Turley, R. S. (2002). A Comparison of Parametric Models of Income Distribution across Countries and over Time. *Estadistica*, 55:135–152.

Baxter, J. (2017). The Career Aspirations of Young Adolescent Boys and Girls. *Annual Statistical Report 2016*.

Becker, G. S. (2010). *The Economics of Discrimination*. University of Chicago press.

Blau, F. D., Brummund, P., and Liu, A. Y.-H. (2013). Trends in Occupational Segregation by Gender 1970–2009: Adjusting for the Impact of Changes in the Occupational Coding System. *Demography*, 50(2):471–492.

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

Bloch, J. W. (1948). Regional Wage Differentials: 1907-46. *Monthly Lab. Rev.*, 66:371.

Bonjour, D. and Gerfin, M. (2001). The Unequal Distribution of Unequal Pay–an Empirical Analysis of the Gender Wage Gap in Switzerland. *Empirical Economics*, 26(2):407–427.

Bonnard, C. and Giret, J.-F. (2014). Gender Differences in Career Plans and Wage Expectations of French Undergraduates. In *International Conference on Education and Gender*.

Cancian, M., Danziger, S., and Gottschalk, P. (1992). Working Wives and Family Income Inequality among Married Couples. [Unpublished] 1992. Presented at the Annual Meeting of the Population Association of America Denver Colorado April 30-May 2 1992.

Cappellari, L., Tatsiramos, K., and Polachek, S. W. (2016). *Income Inequality Around the World*. Emerald Group Publishing.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on Counterfactual Distributions. *Econometrica*, 81(6):2205–2268.

Cotton, J. (1988). On the Decomposition of Wage Differentials. *The Review of Economics and Statistics*, pages 236–243.

Deville, J.-C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. *Survey Methodology*, 25:193–204.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Diekmann, A., Engelhardt, H., et al. (1995). Einkommensungleichheit zwischen frauen und männern. eine ökonometrische analyse der schweizer arbeitskräfteerhebung. *Swiss Journal of Economics and Statistics (SJES)*, 131(I):57–83.

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

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

Donzé, L. (2013). Erreurs de spécification dans la décomposition de l'inégalité salariale. Presented at Swiss Statistics Meeting 2013, Basel.

Edgeworth, F. Y. (1922). Equal Pay to Men and Women for Equal Work. *The Economic Journal*, 32(128):431–457.

England, P. (1992). *Comparable Worth: Theories and Evidence*. Transaction Publishers.

England, P., Allison, P., and Wu, Y. (2007). Does Bad Pay Cause Occupations to Feminize, Does Feminization Reduce Pay, and How can we Tell with Longitudinal Data? *Social Science Research*, 36(3):1237–1256.

Fawcett, M. G. (1918). Equal Pay for Equal Work. *The Economic Journal*, 28(109):1–6.

Federal Office for Gender Equality (2015). Towards Equality. [Online; accessed October 17, 2017].

Finzi, I. (2007). *Occupational Gender Segregation and Gender Wage Gap in Switzerland*. PhD thesis, Università della Svizzera italiana.

Fortin, N. and Lemieux, T. (2013). Lecture 1: Introduction and General Theory of Decompositions. In *CLSRN Summer School*. Montréal.

Fortin, N., Lemieux, T., and Firpo, S. (2011). *Decomposition Methods in Economics*, volume 4 of *Handbook of Labor Economics*, chapter 1, pages 1–102. Elsevier.

Freedman, D. A. (2006). On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, 60(4):299–302.

Fthenakis, W. E. and Minsel, B. (2002). *Die Rolle des Vaters in der Familie*, volume 213. W. Kohlhammer Stuttgart.

Gardeazabal, J. and Ugidos, A. (2005). Gender Wage Discrimination at Quantiles. *Journal of Population Economics*, 18(1):165–179.

Goldin, C. and Rouse, C. (1997). Orchestrating Impartiality: The Impact of" Blind" Auditions on Female Musicians. Technical report, National bureau of economic research.

Goodman, J. (2000). *Global Perspectives on Gender and Work: Readings and Interpretations*. Rowman & Littlefield Publishers.

Graf, M. (2011). Use of Survey Weights for the Analysis of Compositional Data. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*, pages 114–127. Wiley, Chichester.

Graf, M. and Nedyalkova, D. (2015). *GB2: Generalized Beta Distribution of the Second Kind: Properties, Likelihood, Estimation*. R package version 2.1.

Graf, M. and Nedyalkova, D. (2017). Discretizing a Compound Distribution with Application to Categorical Modelling. *Statistics*, 51(3):685–710.

Graf, M., Nedyalkova, D., Münnich, R., Seger, J., and Zins, S. (2011). Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion. Research Project Report WP2 – D2.1, FP7-SSH-2007-217322 AMELI.

Hausmann, A.-C., Kleinert, C., and Leuze, K. (2015). Entwertung von Frauenberufen oder Entwertung von Frauen im Beruf? *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 67(2):217–242.

Heckman, J. J. (1977). Sample Selection Bias as a Specification Error (with an Application to the Estimation of Labor Supply Functions).

Heitmueller, A. (2006). Public-Private Sector Pay Differentials in a Devolved Scotland. *Journal of Applied Economics*, 9(2):295–323.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Jann, B. (2003). Lohngerechtigkeit und Geschlechterdiskriminierung: Experimentelle Evidenz.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley-Interscience, Hoboken.

Kugler, P. (1988). Lohndiskriminierung in der Schweiz: Evidenz von Mikrodaten. *Swiss Journal of Economics and Statistics (SJES)*, 124(I):23–47.

Kuhn, U. and Ravazzini, L. (2017). The Impact of Assortative Mating on Income Inequality in Switzerland.

Levanon, A., England, P., and Allison, P. (2009). Occupational Feminization and Pay: Assessing Causal Dynamics using 1950–2000 US Census Data. *Social Forces*, 88(2):865–891.

Lewis, H. G. (1986). Union Relative Wage Effects. *Handbook of Labor Economics*, 2:1139–1181.

Lips, H. (2013). *Gender: the Basics*. Routledge.

Lips, H. M. (2017). *Sex and Gender: An Introduction*. Waveland Press.

Lucifora, C. and Meurs, D. (2006). The Public Sector Pay Gap in France, Great Britain and Italy. *Review of Income and Wealth*, 52(1):43–59.

Machado, J. A. and Mata, J. (2005). Counterfactual Decomposition of Changes in Wage Distributions using Quantile Regression. *Journal of applied Econometrics*, 20(4):445–465.

McDonald, J. and Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: Estimation of related measures of inequality. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, volume 5 of *Economic Studies in Equality, Social Exclusion and Well-Being*, pages 147–166. Springer New York.

McDonald, J. B. (1984). Some generalised functions for the size distribution of income. *Econometrica*, 52:647–663.

McDonald, J. B. and Butler, R. J. (1990). Regression models for positive random variables. *Journal of Econometrics*, 43(1-2):227–251.

McDonald, J. B. and Xu, Y. J. (1995). A generalisation of the beta distribution with applications. *Journal of Econometrics*, 66:133–152.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour economics*, 12(4):577–590.

Melly, B. (2006). *Applied quantile regression*. PhD thesis, University of St. Gallen, Switzerland.

Mohr, T. (2015). *Playing Big: Find Your Voice, Your Mission, Your Message*. Avery.

Neumark, D. (1988). Employers' Discriminatory Behavior and the Estimation of Wage Discrimination. *Journal of Human Resources*, 23(3):279–295.

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

Oaxaca, R. L. and Ransom, M. R. (1994). On Discrimination and the Decomposition of Wage Differentials. *Journal of Econometrics*, 61(1):5–21.

Pereira, P. T. and Martins, P. S. (2002). Education and Earnings in Portugal. In *Bank of Portugal Conference Proceedings*.

Popli, G. K. (2013). Gender Wage Differentials in Mexico: A Distributional Approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):295–319.

Rathbone, E. F. (1917). The Remuneration of Women's Services. *The Economic Journal*, 27(105):55–68.

Reimers, C. W. (1983). Labor Market Discrimination against Hispanic and Black Men. *The Review of Economics and Statistics*, pages 570–579.

Rigby, R., Stasinopoulos, D., Heller, G., and De Bastiani, F. (2017). Distributions for Modelling Location, Scale, and Shape: Using GAMLSS in R. *URL www. gamlss. org.(last accessed 5 March 2018)*.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Schmid, F. (2016). The Gender Wage Gap in Switzerland over Time. *Swiss Journal of Sociology*, 42(3):442–467.

Schön, J. and Liechti, C. (2013). Das Engagement der Väter in Haushalt und Familie. Technical report, Swiss Federal Statistical Office.

Schwartz, C. R. (2010). Earnings inequality and the changing association between spouses' earnings. *American Journal of Sociology*, 115(5):1524–1557.

Schwartz, C. R. and Mare, R. D. (2005). Trends in educational assortative marriage from 1940 to 2003. *Demography*, 42(4):621–646.

Siltanen, J. (1994). *Locating Gender: Occupational Segregation, Wages and Domestic Responsibilities. Cambridge Studies in Work and Social Inequality 1.* UCL Press.

Strub, J. and Stocker, D. (2010). Analyse der Lohne von Frauen und Mannern anhand der Lohnstrukturerhebung 2008. Technical report, Buro fur Arbeits- und Sozialpolitische Studien BASS AG.

Swiss Federal Statistical Office (2006). Modèles d'activité dans les couples, partages des tâches et garde des enfants. quelques éléments de la conciliation entre vie familiale et vie professionnelle: la Suisse en comparaison internationale.

Swiss Federal Statistical Office (2008). General Classification of Economic Activities - Explanatory notes. [Online; accessed June 22nd, 2018].

The Federal Assembly of the Swiss Confederation (1995). Federal Act on Gender Equality of 24 March 1995.

Thurow, L. C. (1970). Analyzing the American Income Distribution. *The American Economic Review*, 60(2):261–269.

Weichselbaumer, D. and Winter-Ebmer, R. (2005). A Meta-Analysis of the International Gender Wage Gap. *Journal of Economic Surveys*, 19(3):479–511.

Weichselbaumer, D. and Winter-Ebmer, R. (2006). Rhetoric in Economic Research: the Case of Gender Wage Differentials. *Industrial Relations: A Journal of Economy and Society*, 45(3):416–436.

Williams, C. (2010). Economic Well-Being. Statistical bulletin, Statistics Canada.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer, New York.

Woodruff, R. S. (1952). Confidence Intervals for Medians and Other Position Measures. *Journal of the American Statistical Association*, 47:635–646.