Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Decomposition of gender wage inequalities through calibration: Application to the Swiss structure of earnings survey

by Mihaela-Catalina Anastasiade and Yves Tillé

Release date: December 21, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 211-234
Statistics Canada, Catalogue No. 12-001-X

Decomposition of gender wage inequalities through calibration: Application to the Swiss structure of earnings survey

Mihaela-Catalina Anastasiade and Yves Tillé[1]

Abstract

This paper proposes a new approach to decompose the wage difference between men and women that is based on a calibration procedure. This approach generalizes two current decomposition methods that are re-expressed using survey weights. The first one is the Blinder-Oaxaca method and the second one is a reweighting method proposed by DiNardo, Fortin and Lemieux. The new approach provides a weighting system that enables us to estimate such parameters of interest like quantiles. An application to data from the Swiss Structure of Earnings Survey shows the interest of this method.

Key Words: Blinder-Oaxaca; Gender wage discrimination; Quantiles; Reweighting; Wages.

1 Introduction

Wage discrimination can be based on different criteria, such as gender, race or religion. 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). Since a quantification of discrimination is required in order to assess its magnitude, the topic has awakened the interest of statisticians. The original technique proposed by Blinder (1973) and Oaxaca (1973) estimates how much of the difference between the average wages of men and the average wages of women is due to discrimination. However, in general, there is an uneven allocation of women and men among jobs (see, for instance Bielby and Baron, 1986). If the members of one of these two groups, usually women, are concentrated in lower paying jobs, the difference in average wages might not be of great relevance. So instead of analysing the discrimination level in average wages, it might be interesting to see if discrimination occurs uniformly in all types of jobs. A detailed reference of the different statistical papers devoted to the estimation of discrimination can be found in Fortin, Lemieux, and Firpo (2011).

While there are many decomposition methods available in the literature, only two of them will be discussed in this paper. These two methods are not presented in their original forms, but by taking into account survey weights. They are the Blinder-Oaxaca method (hereafter, BO) and the semi-parametric method developed by DiNardo, Fortin, and Lemieux (1996) (hereafter, DFL). Originally, the BO method analysed the difference between the average wages of men and the average wages of women. However, it does not allow for an analysis of the wage differences for other parameters, such as quantiles. The original DFL method addresses this issue. Its starting point is a logistic model where, for each observation, the probability of being a man or a woman is modelled as a function of the observed characteristics. The ratio of these probabilities is used to construct a reweighting factor. Its aim is to approach the distribution of the characteristics of women to the distribution of characteristics of men. By having similar distributions of the

1. Mihaela-Catalina Anastasiade and Yves Tillé, Institute of Statistics, University of Neuchâtel, 51 Avenue de Bellevaux, 2000 Neuchâtel, Switzerland. E-mail: mihaela.anastasiade@unine.ch; yves.tille@unine.ch.

characteristics, an estimation of the discrimination level at parameters other than the mean is achievable. However, the reweighting factor may have a large variance in cases where one or more characteristics are good predictors of the gender. Moreover, the reweighted distribution of characteristics of women may not match the distribution of characteristics of men. We address the problems related to the two methods through a calibration approach. The idea behind calibration is the same as that of the DFL method. It consists of approaching the distribution of characteristics of women to that of men, in order to estimate the discrimination level along the entire wage distributions.

The paper is structured as follows: after the definition of the notation in Section 2, the BO decomposition is re-expressed with the use of survey data in Section 3. Sampling weights are taken into account in order to correct for the difference between the sample and the population of interest. Therefore, the decomposition will be termed "weighted BO". The key concept of women's counterfactual wage distribution is also presented. It is defined as the wage distribution of women if they had the same characteristics as men. Next, we discuss the use of the counterfactual wage distribution in the wage difference decomposition. In Section 4, the DFL method is developed, again using survey weights. Since the original method does not include survey weights, it will be termed "weighted DFL". Next in Section 5, a new approach to compute the counterfactual wage distribution is proposed, using the calibration method (Deville and Särndal, 1992). The use of two particular cases of calibration are discussed. These are the linear calibration and the raking-ratio calibration. The first case yields the same result as the weighted BO method for average wages. The second case has a similar approach to the weighted DFL method, but without assuming a logistic model. In other words, the proposed technique can be regarded as a generalization of the two methods discussed above. Section 6 includes an overview of the dataset used as well as descriptive statistics on the observed wages. A brief description of the model used and the results obtained using the discussed methods are presented. Finally, Section 7 summarizes the conclusions and in Appendix B, the computation of the variance of the counterfactual wage is shown.

2 Problem and notation

The question of interest is the estimation of the wage differences between women and men, more specifically, how much of this difference is attributable to discrimination. Assume there is a finite population U of size N that can be divided into two subpopulations, women and men, denoted by $U_h, h \in \{F, M\}$, of size N_h. Additionally, a random sample S is drawn from U, which contains both women and men. Sample S is selected by means of a sampling design $p(s) = \Pr(S = s)$ for any $s \subset U$, where

$$p(s) \geq 0 \quad \text{and} \quad \sum_{s \subset U} p(s) = 1.$$

Sample S can be split into two subsamples, $S_h, h \in \{F, M\}$, women and men, such that $S = \cup S_h$. 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

$$Y_h = \sum_{k \in U_h} y_k, h \in \{F, M\},$$

where y_k is the logarithm of the wage of the k^{th} individual. Since not all units in the subpopulations are observed, the totals can be estimated by

$$\hat{Y}_h = \sum_{k \in S_h} d_k y_k, h \in \{F, M\},$$

where d_k is a sampling weight assigned to the k^{th} unit of the sample. Sampling weights are obtained after several statistical treatments (for example, adjustment for non-response).

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

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, h \in \{F, M\},$$

and can be estimated by

$$\hat{\bar{Y}}_h = \frac{\sum_{k \in S_h} d_k y_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}.$$

Moreover, assume that for each k^{th} individual in either of the two subsamples, there is a vector of p auxiliary variables denoted by

$$\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kj}, \ldots, x_{kp}\right)^\top \in \mathbb{R}^p.$$

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

$$\mathbf{X}_h = \sum_{k \in U_h} x_k, h \in \{F, M\}.$$

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

$$\hat{\mathbf{X}}_h = \sum_{k \in S_h} d_k \mathbf{x}_k, h \in \{F, M\}.$$

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

$$\bar{\mathbf{X}}_h = \frac{1}{N_h} \sum_{k \in U_h} \mathbf{x}_k, h \in \{F, M\},$$

and estimated by

$$\hat{\bar{\mathbf{X}}}_h = \frac{\sum_{k \in S_h} d_k \mathbf{x}_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}. \quad (2.1)$$

3 The weighted BO decomposition

3.1 The decomposition

Using the setup in Section 2, the findings of Blinder (1973) and Oaxaca (1973) 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 p characteristics that are available and the logarithm of the wage. A regression is done separately in each subpopulation U_h, $h = \{M, F\}$. At the subpopulation level, the values of the regression coefficients are given by

$$\boldsymbol{\beta}_h = \left(\sum_{k \in U_h} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U_h} \mathbf{x}_k y_k .$$

They can be estimated from the sample by

$$\hat{\boldsymbol{\beta}}_h = \left(\sum_{k \in S_h} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_h} d_k \mathbf{x}_k y_k, \tag{3.1}$$

where d_k are the sampling weights. The regression coefficients $\hat{\boldsymbol{\beta}}_h$ are called the group wage structure or the returns on characteristics and they represent the contribution of each characteristic to the wage.

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

$$\bar{Y}_h = \bar{\mathbf{X}}_h^\top \boldsymbol{\beta}_h \quad \textit{and} \quad \hat{\bar{Y}}_h = \hat{\bar{\mathbf{X}}}_h^\top \hat{\boldsymbol{\beta}}_h$$

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

Since it is assumed that $x_{k1} = 1$ for all $k \in U$, with $\boldsymbol{\varsigma}^\top = (1\ 0 \ldots 0)$, the equality is always fulfilled. The proof of Result 1 can be found in Appendix A. Putting together the result above, equations (2.1) and (3.1), the average difference between the wages of two groups can be written as

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right). \tag{3.2}$$

The difference between average wages of the groups contains two elements: an explained part, also called the *composition effect* $\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F$ and an unexplained part, or the *structure effect* $\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right)$. 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). It is obtained using characteristics as a proxy for productivity. The estimation of the *structure effect* is the central element of this paper. Equation (3.2) 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. However, the weighted BO method is the same as the original BO method if the sampling weights are all equal to 1.

3.2 A note on the structure effect

The two elements in equation 3.2 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 an estimation of the discrimination level. However, discrimination is an intricate phenomenon that might not be always fully observed. Unobserved variables, selection bias or some 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 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 distribution of men and women in certain jobs.

3.3 The counterfactual wage distribution

In general, the counterfactual wage distribution is an artificial distribution obtained by using the characteristics of a group to estimate the wages of another group (see, for instance Bourguignon, Ferreira, and Leite, 2002). Examples of counterfactual distributions are found in DiNardo et al. (1996) or DiNardo (2002). The term $\hat{\bar{\mathbf{X}}}_M\hat{\boldsymbol{\beta}}_F$ that appears in equation (3.2) is called the women's counterfactual average wage. It is interpreted as the estimated average wage of women if they had the same average characteristics as men and if their return on characteristics remained unchanged. Women's counterfactual wage distribution is obtained by using the characteristics of men $(\mathbf{X}_M)$ and the wage structure of women $(\boldsymbol{\beta}_F)$. In terms of interpretation, it is the wage distribution of women, if they had the same characteristics as men.

Using Result 1 from the previous section, women's counterfactual mean wage equals

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

and is estimated from the sample by

$$\hat{\bar{Y}}_{F|M} = \hat{\bar{\mathbf{X}}}_M^\top\hat{\boldsymbol{\beta}}_F,$$

where $\hat{\bar{\mathbf{X}}}_M$ are estimated in equation (2.1) and $\hat{\boldsymbol{\beta}}_F$ are the coefficients estimated by means of equation (3.1). With this notation, the BO decomposition given in (3.2) is re-expressed as

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right). \quad (3.3)$$

3.4 Using the counterfactual distribution to estimate the composition and the structure effects

Building the counterfactual average wage allows for the estimation of the two effects that make up the wage difference at the average levels. From equation (3.3), the composition effect is equal to

$$\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right).$$

The composition effect can be interpreted as the difference between what women would earn on average if they had the characteristics of men and what they actually earn. Thus, it reflects the inequality due to the differences in characteristics. The structure effect in equation (3.3) is equal to

$$\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right).$$

The structure effect is the difference between the actual average wage of men and what women would earn if they had the average characteristics of men and their wage own structure. The equations above express the composition and structure effects at the average levels, since this is the limitation of the BO method. The next section presents a method that allows for the construction of the entire counterfactual distribution. This in turn results in the ability of estimating the composition and structure effects along the entire wage distribution.

4 The weighted DFL method

4.1 The method

The method proposed by DiNardo et al. (1996) 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 reweighting function is equal to

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

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. Obviously, $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ and $\Pr(D_{Mk} = 0 | \mathbf{x}_k)$ 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,

$$\hat{\psi}(\mathbf{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 | \mathbf{x}_k) / \widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 | \mathbf{x}_k) / \widehat{\Pr}(D_{Mk} = 0)}. \tag{4.1}$$

Using the reweighting factor, women's counterfactual wage mean is estimated by

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}, \tag{4.2}$$

and women's counterfactual means of characteristics by

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}. \tag{4.3}$$

The estimated reweighting factor defined in equation (4.1) will be equal to

$$\hat{\psi}(\mathbf{x}_k) = \hat{a} \ \exp\left(\mathbf{x}_k^\top \hat{\boldsymbol{\gamma}}\right),$$

where $\hat{\boldsymbol{\gamma}}$ is the estimation of $\boldsymbol{\gamma}$ from the sample using empirical likelihood and $\hat{a}$ is the ratio of estimated proportions of women and men. It is given by:

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

Since the DFL method is presented taking the survey weights into account, the reweighting factor ψ_k will be multiplied by $d_k, k \in S_F$. This resulting factor will be termed "weighted DFL factor". Women's estimated counterfactual wage mean can be re-expressed as

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right)}. \tag{4.4}$$

Women's counterfactual means of characteristics are estimated as

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right)}.$$

Through the use of the reweighting factor, the counterfactual coefficients in the women's sample are given by

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

and estimated by

$$\hat{\boldsymbol{\beta}}_F^{\text{DFL}} = \left(\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k y_k. \tag{4.5}$$

The coefficients above have to be computed, because under the same condition as in Result 1, women's counterfactual wage mean defined in (4.2) is given by

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}.$$

The BO decomposition formula can now be expressed as

$$\begin{aligned} \hat{\bar{Y}}_M - \hat{\bar{Y}}_F &= \left(\hat{\bar{Y}}_{F|M}^{\text{DFL}} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}^{\text{DFL}}\right) \\ &= \left(\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}} - \hat{\bar{\mathbf{X}}}_F^{\top} \hat{\boldsymbol{\beta}}_F\right) + \left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right), \end{aligned} \tag{4.6}$$

where $\hat{\boldsymbol{\beta}}_M$ and $\hat{\boldsymbol{\beta}}_F$ are defined in (3.1). The first term of equation (4.6) is the composition effect and the second one the structure effect.

4.2 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,

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} \neq \hat{\bar{\mathbf{X}}}_M \tag{4.7}$$

(see, for instance, Fortin et al. 2011; Donzé, 2013). The reweighting factor thus fails to match the two distributions perfectly.

The structure effect in equation (4.6) can be further divided in the following elements

$$\left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right) = \hat{\bar{\mathbf{X}}}_M^{\top} \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right) + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}}\right) \hat{\boldsymbol{\beta}}_F^{\text{DFL}}, \tag{4.8}$$

where $\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}}$ and $\hat{\boldsymbol{\beta}}_F^{\text{DFL}}$ are defined in equations (4.3) and (4.5), respectively (Fortin et al. 2011). The first element of the right-hand side of equation (4.8) 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. 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 $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ may 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 factor. This will be shown in Section 8.

5 The calibration approach

5.1 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. The resulting estimates are consistent and efficient.

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

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k . \tag{5.1}$$

The weights are determined by solving in $\boldsymbol{\lambda}$ the calibration equations that become

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

where $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$ is the calibration function. The resulting calibration estimation of Y is

$$\hat{Y} = \sum_{k \in S} d_k y_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right). \tag{5.2}$$

In what follows, we will use the linear case, where the pseudo-distance function is the chi-square distance and the calibration function is given by $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = 1 + \mathbf{x}_k^\top \boldsymbol{\lambda}.$ In the second case, we will use the raking-ratio, which uses the Entropy pseudo-distance and where the calibration function is given by $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = \exp \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right).$

5.2 Calibration of women's characteristics on the men's characteristics

Suppose that for all the units of the sample, there is a given sampling weight d_k. 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 $\sum_{k \in S_F} G \left(w_k , d_k \right)$ is minimized. The following calibration equation is satisfied

$$\sum_{k \in S_F} w_k \mathbf{x}_k = \hat{\tilde{\mathbf{X}}}_M , \tag{5.3}$$

where the vector $\hat{\tilde{\mathbf{X}}}_M$ 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.

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

Dividing the calibration equation (5.3) by $\sum_{k\in S_F} d_k$ yields

$$\frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} d_k} = \frac{\hat{\hat{\mathbf{X}}}_M}{\sum_{k\in S_F} d_k} = \hat{\hat{\mathbf{X}}}_M . \tag{5.4}$$

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

$$\sum_{k\in S_F} w_k = \sum_{k\in S_F} d_k , \tag{5.5}$$

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

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

by putting together equations (5.4) and (5.5), this means that

$$\hat{\hat{\tilde{\mathbf{X}}}}_M = \hat{\hat{\mathbf{X}}}_M . \tag{5.6}$$

Women's counterfactual wage mean estimator is thus

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

5.3 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.,* $\hat{\hat{\bar{Y}}}_{F|M} = \hat{\hat{\mathbf{X}}}^\top \hat{\boldsymbol{\beta}}_F$.

Proof

In order to determine the vector $\boldsymbol{\lambda}$ in the case when the chi-squared pseudo-distance is used, the following equation must be solved

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

Thus,

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

where

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

Thus

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

Using the result from the previous equation, the numerator of expression (5.2) becomes

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k \\ &= \sum_{k \in S_F} d_k y_k + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k , \end{aligned} \tag{5.7}$$

where $\hat{Y}_{F|M}^{\mathrm{LC}}$ 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

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

Vector $\hat{\boldsymbol{\beta}}_F$ has already been defined in the same way in equation (3.1) for the weighted BO method. Equation (5.7) is rewritten as

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k y_k + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{Y}_F + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{\tilde{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F , \end{aligned} \tag{5.8}$$

because under the condition of Result 1, $\hat{\mathbf{X}}_F^\top \hat{\boldsymbol{\beta}}_F = \hat{Y}_F$. By dividing (5.8) 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 calibration instance yields the same results as the BO method for average wages, we advocate for the use of an instance that gives nonnegative weights.

5.4 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 (5.3) becomes

$$\hat{\tilde{\mathbf{X}}}_M = \sum_{k \in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k \in S_F} d_k \mathbf{x}_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right). \tag{5.9}$$

This resulting system of equations cannot be solved analytically. However, the value of $\boldsymbol{\lambda}$ can be found through the Newton-Raphson algorithm.

The equation (5.2) can be now written as

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

where $\hat{Y}_{F|M}^{\text{RRC}}$ 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

$$\hat{\bar{Y}}_{F|M}^{\text{RRC}} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right)}.$$

The equation above is very similar to equation (4.4). The only difference lies in the estimation of the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$. The vector $\boldsymbol{\lambda}$ contains the Lagrangian multipliers solving equation 5.9 under constraint (5.1), while the vector $\boldsymbol{\gamma}$ is found through maximum likelihood.

After computing the calibration weights w_k defined in (5.3) and by using the information in equation (5.6), it results that

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

which ensures that the residual part of the structure effect defined in equation (4.8) will equal 0. This is a solution to the problem shown in Section 4.3. This instance of calibration also remedies the issue of the negative weights that may arise when using the chi-squared pseudo-distance.

6 Application to the Swiss Structure of Earnings Survey

6.1 Data description

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

Table 6.1
Wage mean and median computed for the entire dataset, women and men, in Swiss francs

	Mean	Median
Entire dataset	6,977	5,905
Women	5,843	5,220
Men	7,725	6,346

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 6.2 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 6.2
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 (2,683)	8.27 (3,897)	8.39 (4,412)	8.50 (4,905)	8.59 (5,400)	8.68 (5,905)	8.78 (6,488)	8.89 (7,233)	9.03 (8,380)	9.27 (10,667)	10.09 (24,202)
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 of under CHF 4,905 (as opposed to only 21% of men), there are only 11% of women who earn between CHF 8,380 and CHF 24,202 (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 6.3 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 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 6.3), they earn more than the 2% of women who have similar jobs. Figure 6.1 shows the data presented in Table 6.3 below in a graphical form.

Table 6.3
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
	(2,432)	(3,602)	(4,005)	(4,344)	(4,756)	(5,220)	(5,743)	(6,353)	(7,154)	(8,577)	(15,761)
Men	8.01	8.36	8.49	8.58	8.67	8.76	8.86	8.98	9.14	9.38	10.26
	(3,000)	(4,259)	(4,850)	(5,344)	(5,820)	(6,346)	(7,012)	(7,908)	(9,291)	(11,905)	(28,571)
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)	(1,000)	(1,064)	(1,126)	(1,269)	(1,555)	(2,137)	(3,328)	(12,810)

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 6.2 shows the distributions of the logarithm of the wages of women and men.

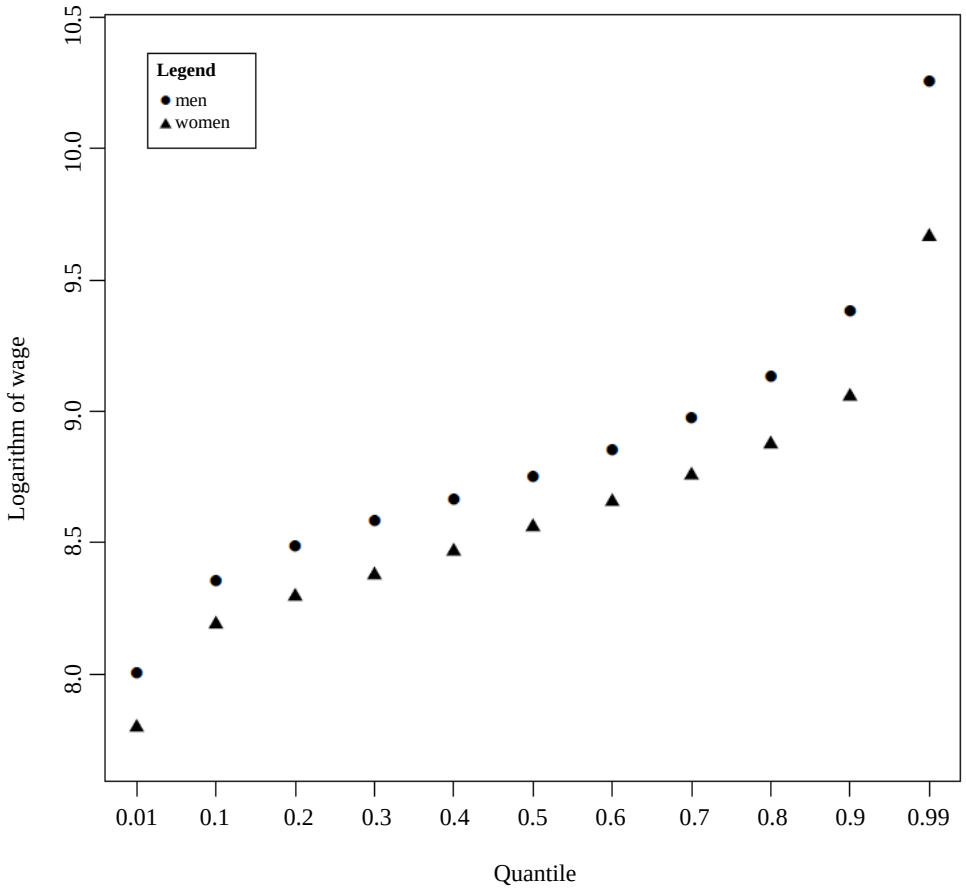

Figure 6.1 Weighted quantiles of the logarithm of the wages of women and men.

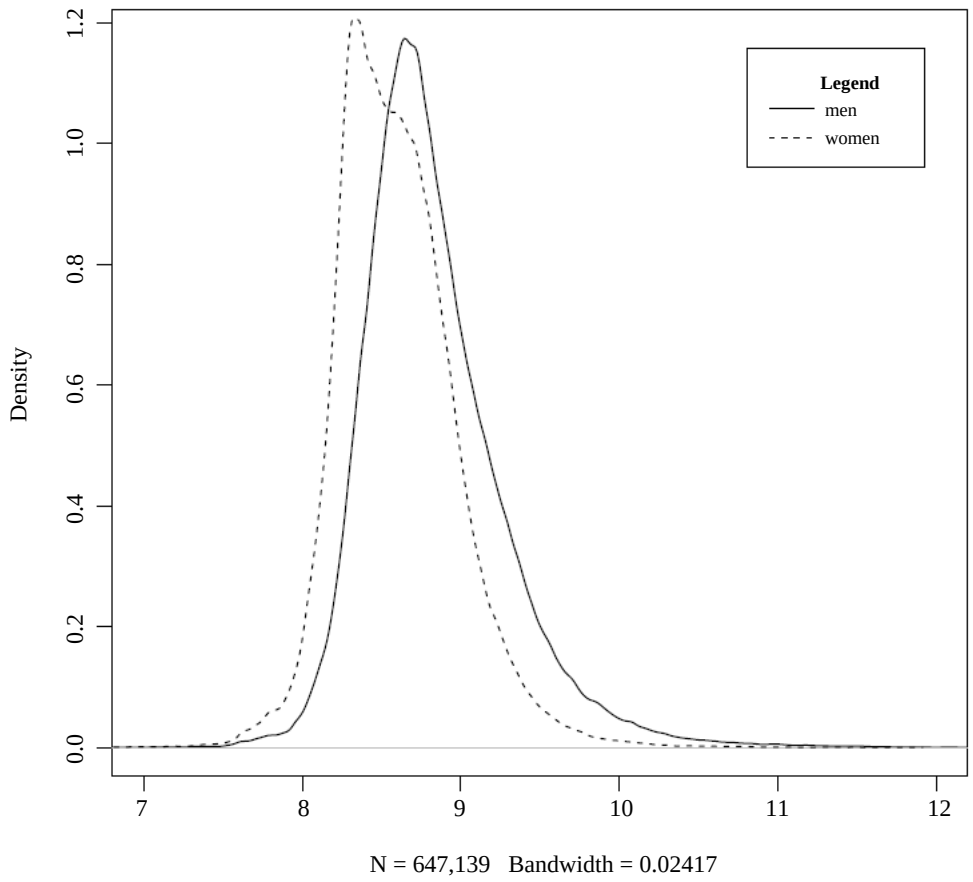

Figure 6.2 Estimated densities of the logarithms of wages of women and men.

6.2 The model

The regression model 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;
- 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).

The model was selected from a number of models with several variables using the minimum 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, therefore no computation was done by the authors.

6.3 Weights and counterfactual distributions

This section only includes results in terms of logarithms. When using the BO method, 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 calibration method through the chi-squared pseudo-distance is denoted as "linear", the calibration through the Kullback-Leibler divergence as "raking-ratio" and the method proposed by DiNardo et al. (1996) adjusted to take the survey weights into consideration as "DFL". First, Table 6.4 shows the minimum and the maximum values of the weights, as well as the standard deviations, obtained using the linear calibration, the raking-ratio calibration and the weighted DFL method.

Table 6.4
Weights minimum, maximum and standard deviation

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 6.4, 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 1,319 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 $\hat{a}$ is smaller than one, the reweighting factor will shrink. If on the other hand, $\hat{a}$ is larger than one (for instance, for sectors such as the public sector), the reweighting factor might be larger. Table 6.5 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.

Table 6.5
Estimated composition and structure effects in the difference in mean averages

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

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 6.3.

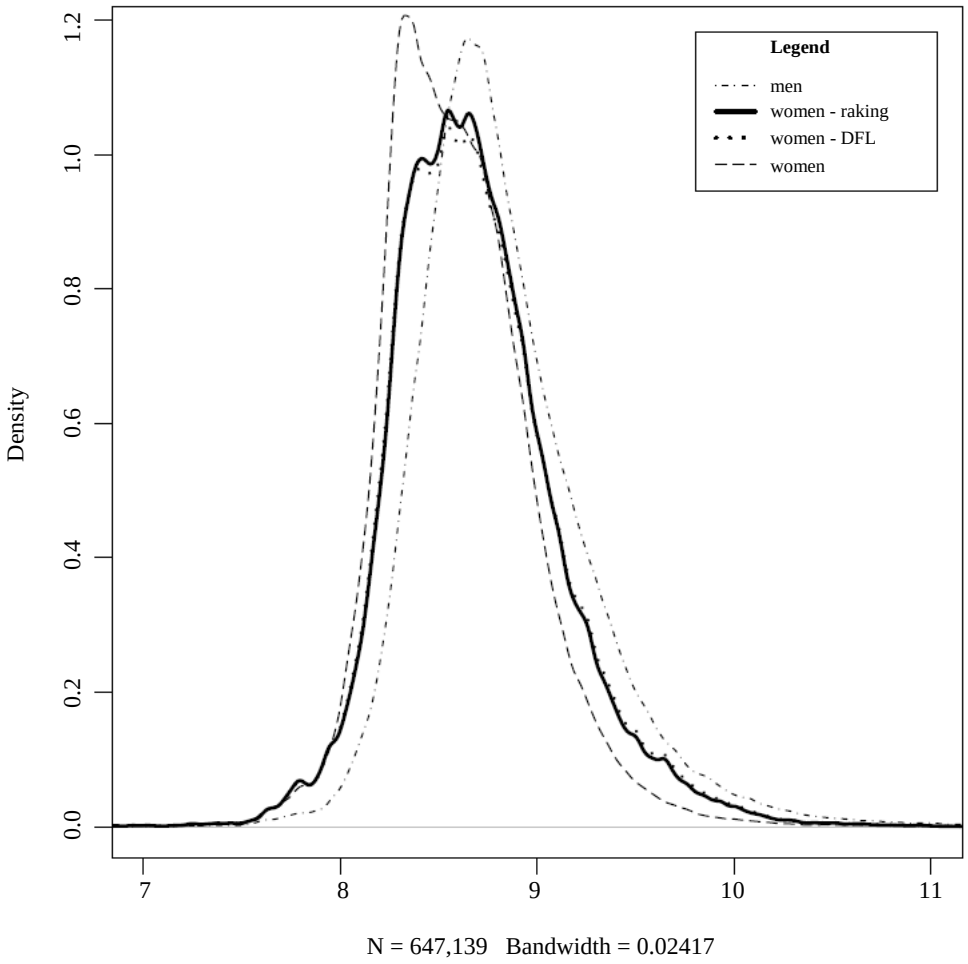

Figure 6.3 Estimated densities of the logarithm of the wage of women and men and the counterfactual distributions of the logarithm of the wage of women constructed using the raking-ratio and the reweighted DFL factor, respectively.

Figure 6.3 shows that the two counterfactual wage distributions are very close to each other around the tails. However, toward the middle, the two methods do 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 6.6 displays the estimated structure and composition effects of the wage differences between men and women computed using the three methods at some selected quantiles.

Table 6.6
Estimated composition and structure effects of the wage difference at selected quantiles

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

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 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 6.4 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.

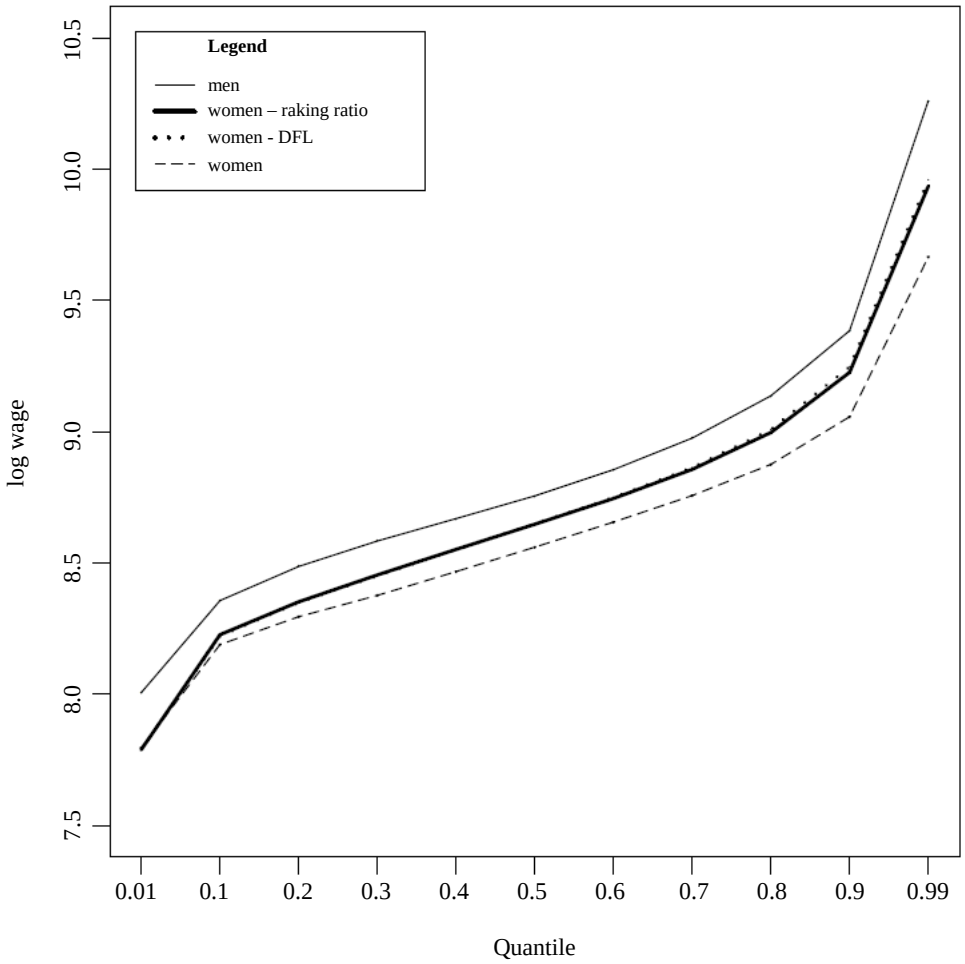

Figure 6.4 Weighted quantiles of the logarithms of the wage of women and men and the weighted quantiles of the counterfactual distribution of the logarithm of the wage of women constructed using the raking-ratio calibration and the weighted DFL factor.

6.4 Further decomposition of the structure effect

A logistic model 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 (4.8), 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 $\hat{\bar{\mathbf{X}}}_M$ and $\hat{\bar{\mathbf{X}}}_{F|M}$. 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.

7 Conclusion

The phenomenon of discrimination has multiple facets and there are many mechanisms that can generate it. However, this paper only examines its estimation from a methodological point of view. The two calibration cases taken into consideration represent a generalization of two existing decomposition methods, the technique of Blinder (1973) and Oaxaca (1973) and the semi-parametric method of DiNardo et al. (1996), both expressed using sampling weights. The original methods can also be obtained, if all the sampling weights are considered to be equal to 1. The linear case yields the same result as the BO method. However, since the resulting weights are unbounded, negative values might be observed. Just as the DFL method, the calibration approach allows for the decomposition of wage differences at other points other than the mean, such as quantiles. However, the raking-ratio calibration is an improvement of the DFL method, in that the estimation of the structure effect will always include a residual effect equal to 0. Therefore, the structure effect will only be composed of the pure effect. Decomposing wage differences along quantiles enables the conclusion that in low-paying jobs, the inequalities are due solely to discrimination. In this article, the emphasis was placed on the generalization of two well-established decomposition methods through the calibration approach.

Acknowledgements

The authors are grateful to the Swiss federal statistical office for the financial support and the LOHN department for providing the data. However, the opinions expressed in this paper do not necessarily reflect those of the Swiss federal statistical office.

Appendix A

Proof of Result 1

$$\begin{aligned}
\hat{\mathbf{X}}_h\hat{\boldsymbol{\beta}}_h &= \hat{\mathbf{X}}_h\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \sum_{j\in S_h} d_j\mathbf{x}_j^\top\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \left(\sum_{j\in S_h}\boldsymbol{\varsigma}^\top d_j\mathbf{x}_j\mathbf{x}_j^\top\right)\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \sum_{l\in S_h}\boldsymbol{\varsigma}^\top d_l\mathbf{x}_l y_l = \sum_{l\in S_h} d_l y_l = \hat{Y}_h.
\end{aligned}$$

By dividing this equation by $\sum_{k\in S_h} d_k$, Result 1 is obtained.

Appendix B

B.1 Linearization of the means

In order to compute the variance of the average means and of the counterfactual means we have used the linearization method proposed by Graf (2011). The author proposes to compute the partial derivative of the estimator with respect to the sample indicator. This derivative provides the linearized variable that can be plugged in the variance estimator. The average means are defined by:

$$\hat{\bar{Y}}_F = \frac{\sum_{k \in S_F} d_k y_k}{\sum_{k \in S_F} d_k},$$

and

$$\hat{\bar{Y}}_M = \frac{\sum_{l \in S_M} d_l y_l}{\sum_{l \in S_M} d_l}.$$

For the two average wages, we obtain the linearized variables:

$$\frac{\partial \hat{\bar{Y}}_F}{\partial I_j} = \begin{cases} \dfrac{d_j \left(y_j - \hat{\bar{Y}}_F \right)}{\sum_{k \in S_F} d_k} & j \in S_F \\ 0 & j \in S_M \end{cases},$$

and

$$\frac{\partial \hat{\bar{Y}}_M}{\partial I_j} = \begin{cases} \dfrac{d_j \left(y_j - \hat{\bar{Y}}_M \right)}{\sum_{l \in S_M} d_l} & j \in S_M \\ 0 & j \in S_F \end{cases}.$$

B.2 Linearization of the counterfactual

In order to compute the counterfactual mean, we compute the weights v_k defined by the system

$$\mathbf{A} = \sum_{k \in S_F} v_k d_k \mathbf{x}_k = \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \sum_{l \in S_M} d_l \mathbf{x}_l = \hat{\bar{\mathbf{X}}}_M \sum_{k \in S_F} d_k,$$

with

$$v_k = F\left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right).$$

For the linearized variables, we have to consider two cases:

- If $j \in S_F$

$$\frac{\partial \mathbf{A}}{\partial I_j} = v_j d_j \mathbf{x}_j + \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \hat{\bar{\mathbf{X}}}_M .$$

Thus

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = -\mathbf{T}^{-1} d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right),$$

where

$$\mathbf{T} = \sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top .$$

- If $j \in S_M$

$$\frac{\partial \mathbf{A}}{\partial I_j} = \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l} .$$

Thus

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = \mathbf{T}^{-1} d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l} .$$

Since we have supposed that there exists a vector $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}^\top \mathbf{x}_k = 1$ for all $k \in U$, then, we have

$$\boldsymbol{\gamma}^\top \mathbf{A} = \sum_{k\in S_F} v_k d_k = \sum_{k\in S_F} d_k .$$

Now consider

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

Again, two cases must be considered:

- If $j \in S_F$

$$\begin{aligned}
\frac{\hat{\bar{Y}}_{F|M}}{\partial I_j} &= \frac{d_j \left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) + \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j} \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k\right]}{\sum_{k\in S_F} d_k} \\
&= \frac{d_j \left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) - d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{T}^{-1} \sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k} \\
&= \frac{d_j \left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k},
\end{aligned}$$

where

$$\mathbf{B}_F = \mathbf{T}^{-1} \sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k .$$

- If $j \in S_M$

$$\begin{aligned}
\frac{\partial \hat{\bar{\mathbf{Y}}}_{F|M}}{\partial I_j} &= \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j} \frac{\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k \in S_F} d_k} \\
&= d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \mathbf{T}^{-1} \frac{\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k \in S_F} d_k} \\
&= d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{1}{\sum_{l \in S_M} d_l} \mathbf{B}_F .
\end{aligned}$$

Thus the linearized variable is

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

The linearized variable must only be plugged in the variance estimator corresponding to the sampling design. 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 the variance computed for the sample of women for the part that remains unexplained.

References

Bielby, W.T., and Baron, J.N. (1986). Men and women at work: Sex segregation and statistical discrimination. *American Journal of Sociology*, 759-799.

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

Bourguignon, F., Ferreira, F.H. and Leite, P.G. (2002). Beyond Oaxaca-Blinder: Accounting for differences in household income distributions across countries. *Inequality and Economic Development in Brazil*, 105.

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

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. Discussion paper, University of Michigan.

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.

Donzé, L. (2013). Erreurs de spécification dans la décomposition de l'inégalité salariale. In *International Conference Ars Conjectandi*, 1713-2013.

Fortin, N., Lemieux, T. and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, (Eds., O. Ashenfelter and D. Card), 4, 1-102. Elsevier.

Gardeazabal, J., and Ugidos, A. (2005). Gender wage discrimination at quantiles. *Journal of Population Economics*, 18(1), 165-179.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications*, (Eds., V. Pawlowsky-Glahn and A. Buccianti), 114-127. Wiley, Chichester.

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.

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. Women in Canada: A Gender-based Statistical Report, Statistics Canada.