TEST
https://doi.org/10.1007/s11749-023-00885-8

ORIGINAL PAPER

Gender wage difference estimation at quantile levels using sample survey data

Mihaela-Cătălina Anastasiade-Guinand[1] · **Alina Matei**[2] · **Yves Tillé**[2]

Received: 9 November 2021 / Accepted: 28 July 2023

Abstract
This paper is motivated by the growing interest in estimating gender wage differences in official statistics. The wage of an employee is hypothetically a reflection of her or his characteristics, such as education level or work experience. It is possible that men and women with the same characteristics earn different wages. Our goal is to estimate the differences between wages at different quantiles, using sample survey data within a superpopulation framework. To do this, we use a parametric approach based on conditional distributions of the wages in function of some auxiliary information, as well as a counterfactual distribution. We show in our simulation studies that the use of auxiliary information well correlated with the wages reduces the variance of the counterfactual quantile estimates compared to those of the competitors. Since, in general, wage distributions are heavy-tailed, the interest is to model wages by using heavy-tailed distributions like the GB2 distribution. We illustrate the approach using this distribution and the wages for men and women using simulated and real data from the Swiss Federal Statistical Office.

Keywords Gender gap statistics · Quantile estimation · Counterfactual distribution · GB2 distribution · Survey weights

Mathematics Subject Classification 62D05

✉ Alina Matei
alina.matei@unine.ch

Mihaela-Cătălina Anastasiade-Guinand
Mihaela-Catalina.Guinand@bfs.admin.ch

Yves Tillé
yves.tille@unine.ch

1 Swiss Federal Statistical Office, Neuchâtel, Switzerland

2 Institute of Statistics, University of Neuchâtel, Neuchâtel, Switzerland

Published online: 19 September 2023

1 Introduction

This paper is motivated by the growing interest in estimating the wage differences between men and women in official statistics. Applications in official statistics deal with random samples drawn from finite populations. Estimation is usually made in what one calls the *design-based framework*, where only the samples are random, while the variables collected from them are fixed. Thus, inference in finite populations may be very different from the one usually used in the classical statistics. In order to accommodate some theory from econometrics, we consider what one calls a *superpopulation framework*, assuming that our finite population is a random sample drawn from an infinite population. Next, the finite population is divided in two groups or subpopulations: men and women.

It is possible that men and women with the same characteristics earn different wages. The wages of men and women are modeled separately using a parametric model. Conditional on some characteristics, we assume that the conditional wage distribution of each individual into a group follows a given theoretical distribution with unknown parameters. Our goal is to capture the shape of the wage distributions and to go beyond the mean differences provided by the regression approach of Blinder (1973) and Oaxaca (1973), which is widely used by the world's national statistical offices. We do this by determining the estimator of the differences between the gender wages 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, using the concept of counterfactual distribution. (For an overview, see Fortin et al. 2011.) The counterfactual distribution is estimated by putting together the parameters of one group and the characteristics of the latter group. This is done in order to estimate what the former group would earn, if they had the characteristics of the other group. We follow this guideline to estimate the wages of women as if they had the same characteristics of men. This leads to the estimation of differences between the gender wages conditionally on fixed covariates at different quantiles. We use a conditional distribution approach similar to the one used by Biewen and Jenkins (2005). First, we estimate the parameters of the distribution of each individual given their characteristics. Next, the marginal wage distribution is fitted based on the individual wage distributions. The parametric approach used in this paper has already been suggested in several papers in the decomposition literature, as, for instance, Biewen and Jenkins (2005), Van Kerm (2013) and Van Kerm et al. (2016).

The novelty of this paper is twofold. First, we use the parametric approach from a survey sampling perspective, underling the specific framework used in the design-based inference. While the main goal is to model wages by using heavy-tailed distributions and survey weights, we also use the parametric approach with a quite different interest. This is specific to survey sampling and was not previously investigated: This approach uses auxiliary information; if this is well correlated with the wages, the variance of the quantile counterfactual estimates may be reduced compared to those of some competitors. We use two parametric methods to estimate quantiles, by assuming a given theoretical distribution of conditional wages of men and women given their characteristics, respectively. While the first parametric method used is similar to one used by Biewen and Jenkins (2005), the second one is new and is

introduced in this paper. The second method has the advantage of allowing an easy construction of confidence intervals of a quantile. Secondly, we want to illustrate the quantile decomposition of wages using data from official statistics and heavy-tailed distributions other than the log-normal distribution usually used in this domain (see, for instance, Leythienne and Ronkowski 2018).

Motivated by a flexible way to model income and wage distributions, we fit in our examples a generalized beta distribution of the second kind (hereafter, GB2) distribution to conditional wages. 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 income distributions. McDonald (1984), Bandourian et al. (2002) and McDonald and Ransom (2008) showed that the GB2 distribution provides a good fit for income. The GB2 distribution can be 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). We illustrate the two parametric methods using the GB2 distribution, with parameters estimated through maximum pseudo-likelihood, when survey weights and characteristics are associated with sampled employees, by expressing the scale parameter of a GB2 distribution as a function of their characteristics. We show in "Appendix" how to estimate the standard errors of the estimated parameters in a GB2 regression model, using a sandwich estimator and a parametric bootstrap approach.

This paper is structured as follows: In Sect. 2, we present the general setup and recall the classical decomposition method of Blinder (1973) and Oaxaca (1973), making the bridge with the estimation in the context of survey data. In Sect. 3, we discuss the concept of counterfactual wage distribution and show a decomposition method at quantiles' level. The two parametric methods to estimate quantiles are presented in Sects. 4 and 5. The two methods are applied for the gender wage distributions and the counterfactual wage distribution, respectively. The Monte Carlo simulation results given in Sect. 6.2 show the methodological interest to use auxiliary information in the quantile counterfactual estimation. We illustrate the decomposition method at quantiles' level in Sect. 6.3, by assuming that the conditional wage distribution for women and men follow, respectively, a GB2 distribution. The data used were obtained from the Swiss Federal Statistical Office and are issued from the Swiss survey on the structure of earnings in 2012. We draw our conclusions in Sect. 7.

2 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 $p(s) = \Pr(S = s), \forall s \subseteq U$. It is assumed that the sampling design is noninformative. 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 with 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 $n_M + n_F = n$.

We work in a superpopulation framework and assume that the finite population is a random sample drawn from an infinite population. Let Y be the variable wage. 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 with each $k \in U$. We assume that the estimation process refers to the infinite population parameter, and is executed in the design-based approach, considering, however, that Y_k associated with unit $k \in U$ is random (see Särndal et al. 1992, p. 516, Case 4).

We also assume that a linear regression model that relates the logarithm of Y to some covariates $X_1, X_2, \ldots, X_c$ holds. The covariates are the same in each $U_g, g \in \{M, F\}$, but for coherence with the subsets' notation we denote by $X_{1,g}, X_{2,g}, \ldots, X_{c,g}$ 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 c covariates are stored in the vector

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

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

2.1 The Blinder–Oaxaca-type decomposition method

We use what is called in econometrics a *decomposition method*. The general idea of decomposition methods is to divide the difference between wages of men and women in two elements: the first one is the part due to the difference in characteristics between them, and thus, it can be explained, while the second one is not. Starting with Blinder (1973) and Oaxaca (1973), many decomposition methods have been proposed, not only to decompose wage means, but also wage densities; for an overview, see Fortin et al. (2011).

Assume that the superpopulation is divided in two subsuperpopulations from where the subsets $U_g, g = \{M, F\}$ are drawn, respectively. In each subsuperpopulation SUP_g, a linear relationship is suitable between the characteristics that are available and the logarithm of the wage. A linear regression model is fitted separately in each subsuperpopulation SUP_g with $g \in \{M, F\}$

$$\log(Y_{k,g}) = \mathbf{X}_{k,g}^\top \boldsymbol{\beta}_g + \varepsilon_{k,g}, k \in \text{SUP}_g, \tag{2}$$

where $\varepsilon_{k,g} \sim N(0, \sigma_g^2)$ are independent and identically distributed (iid), $\boldsymbol{\beta}_g$ represents the vector of regression coefficients and σ_g^2 is the variance of $\log(Y_{k,g}) \mid \mathbf{X}_{k,g}$ in SUP_g. The regression coefficients $\boldsymbol{\beta}_g$ are called the *group wage structure* or the *returns on characteristics*, and they represent the contribution of each characteristic to the logarithm of the wage.

By using Model (2), one obtains the conditional expectation $E(\widetilde{Y}_g \mid \mathbf{X}_g = \mathbf{x}_g) = \mathbf{x}_g^\top \boldsymbol{\beta}_g$ and the unconditional expectation

$$E(\widetilde{Y}_g) = E\left(E\left(\widetilde{Y}_g \mid \mathbf{X}_g\right)\right) = E(\mathbf{X}_g)\boldsymbol{\beta}_g + E(\varepsilon_g) = E(\mathbf{X}_g)\boldsymbol{\beta}_g,$$

where $\widetilde{Y}_g$ represents $\log(Y_g)$, Y_g is the random variable wage in group g, ε_g is the random variable error term in the same group, and $\mathbf{X}_g$ and ε_g are independent.

The difference between the conditional expectations of the logarithm of wages of two groups (it is a Blinder–Oaxaca-type decomposition) can be written as

$$\begin{aligned} \Delta &= E(\widetilde{Y}_M) - E(\widetilde{Y}_F) \\ &= E\left(E\left(\widetilde{Y}_M \mid \mathbf{X}_M\right)\right) - E\left(E\left(\widetilde{Y}_F \mid \mathbf{X}_F\right)\right) \\ &= \left(E\left(\mathbf{X}_M\right) - E(\mathbf{X}_F)\right)\boldsymbol{\beta}_F + E(\mathbf{X}_M)\left(\boldsymbol{\beta}_M - \boldsymbol{\beta}_F\right). \end{aligned} \tag{3}$$

The term $E(\mathbf{X}_M)\boldsymbol{\beta}_F$ that appears in Expression (3) is called the women's *counterfactual mean* of the logarithm of wage. We interpret it as the mean of the logarithm of 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 et al. (2011). Women's counterfactual distribution of logarithm of wage is obtained by using the characteristics of men ($\mathbf{X}_M$) and the wage structure of women ($\boldsymbol{\beta}_F$).

The difference between the average of the logarithm of wages of the groups in Expression (3) contains two elements: an explained part, also called the *composition effect* $(E(\mathbf{X}_M) - E(\mathbf{X}_F))\boldsymbol{\beta}_F$, and an unexplained part, or the *structure effect* $E(\mathbf{X}_M)(\boldsymbol{\beta}_M - \boldsymbol{\beta}_F)$. 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). This is sometimes called "discrimination"; however, the term is no unanimously accepted. Popli (2013) comments that "the unexplained wage gap, which is often termed discrimination, includes the effect of labor market discrimination, unobservable variables and omitted variables." If $\boldsymbol{\beta}_M = \boldsymbol{\beta}_F$, this term is 0.

At 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

$$\boldsymbol{\beta}_g = \left(\sum_{k \in U_g} \mathbf{X}_{k,g}\mathbf{X}_{k,g}^\top\right)^{-1} \sum_{k \in U_g} \mathbf{X}_{k,g}\widetilde{Y}_{k,g}, \quad g \in \{M, F\}, \tag{4}$$

where $\widetilde{Y}_{k,g} = \log(Y_{k,g}), k \in U_g$. The vector $\boldsymbol{\beta}_g$ can be consistently estimated from the subsamples S_g by

$$\widehat{\boldsymbol{\beta}}_g = \left(\sum_{k \in S_g} w_k \mathbf{x}_{k,g} \mathbf{x}_{k,g}^\top \right)^{-1} \sum_{k \in S_g} w_k \mathbf{x}_{k,g} \widetilde{y}_{k,g}, \ g \in \{M, F\}, \tag{5}$$

where $\widetilde{y}_{k,g}$ is the realization of $\widetilde{Y}_{k,g}, k \in S_g$.

The difference Δ can be estimated at the sample level by

$$\widehat{\Delta} = \left(\widehat{\overline{\mathbf{X}}}_M - \widehat{\overline{\mathbf{X}}}_F\right)^\top \widehat{\boldsymbol{\beta}}_F + \widehat{\overline{\mathbf{X}}}_M^\top \left(\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_F\right), \tag{6}$$

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 $\overline{\mathbf{X}}_g$.

Estimating $\widehat{\boldsymbol{\beta}}_M - \widehat{\boldsymbol{\beta}}_F$ allows us to estimate the unexplained part at the mean level, using a log model of the wages. We are interested to estimate it at the quantiles' level, using a more general framework that extends the log model.

3 Quantiles' decomposition

On the superpopulation level, 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.

Recall that a 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. We build the counterfactual wage distribution of women using the characteristics of men. It can be interpreted as the wage distribution of women if they had the characteristics of men. This is done in order to compare the observed and the counterfactual wage distributions to measure the effects of the change on quantiles' levels.

Let $F^C(.)$ be the CDF of the counterfactual distribution of women. Following Chernozhukov et al. (2013), the CDF in the point $y \in \mathcal{Y}_F$, where $\mathcal{Y}_F$ is women's wage support is defined as

$$F^C(y) = \int_{\mathcal{X}_M} F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x}) dF^{\mathbf{X}_M}(\mathbf{x}), \tag{7}$$

where $\mathcal{X}_M$ is the support of $\mathbf{X}_M$. The counterfactual wage distribution is well defined if the support of $\mathbf{X}_F$ ($\mathcal{X}_F$) includes the support of $\mathbf{X}_M$: $\mathcal{X}_M \subseteq \mathcal{X}_F$.

The counterfactual wage is a potential wage of a woman if she matches the characteristics of a man. Expression (7) assumes that to each woman one can match the

characteristics of a man. Under the assumption that $\mathcal{X}_M = \mathcal{X}_F$, DiNardo et al. (1996) re-expressed the counterfactual distribution given in Expression (7) as

$$F^C(y) = \int_{\mathcal{X}_F} F^{(Y_F|\mathbf{X}_F)}(y \mid \mathbf{x})\psi(\mathbf{x})\mathrm{d}F^{\mathbf{X}_F}(\mathbf{x}), \tag{8}$$

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

$$\psi(\mathbf{x}_k) = \psi_k = \frac{P(G_k = 1 \mid \mathbf{x}_k)/P(G_k = 1)}{P(G_k = 0 \mid \mathbf{x}_k)/P(G_k = 0)}, \tag{9}$$

where $G_k = 1$ if individual k is a man and $G_k = 0$ 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 2017); for the calibration method in survey sampling, see Deville and Särndal (1992). The difference between the two methods is discussed by Anastasiade and Tillé (2017).

The classical decomposition on the mean level (Blinder 1973; Oaxaca 1973) is re-expressed at quantile level (Melly 2006; Chernozhukov et al. 2013) as

$$\Delta_{(\alpha)} = Q^M_{(\alpha)} - Q^F_{(\alpha)},$$

with $\alpha \in (0, 1)$, where $Q^M_{(\alpha)}$ and $Q^F_{(\alpha)}$ represent the quantile of order α of the men and women wage distribution, respectively.

The change at quantile level is rewritten as

$$\Delta_{(\alpha)} = Q^M_{(\alpha)} - Q^F_{(\alpha)} = \left(Q^M_{(\alpha)} - Q^C_{(\alpha)}\right) + \left(Q^C_{(\alpha)} - Q^F_{(\alpha)}\right),$$

where $Q^C_{(\alpha)}$ represents the quantile of order α of the counterfactual distribution. The difference $Q^M_{(\alpha)} - Q^C_{(\alpha)}$ is interpreted here as the unexplained part at the α quantile level. Estimation of $\Delta_{(\alpha)}$ results in a quantile estimation problem. To estimate the quantiles $Q^M_{(\alpha)}$ and $Q^F_{(\alpha)}$, we apply the methods shown in Sect. 4, while methods to estimate $Q^C_{(\alpha)}$ are given in Sect. 5.

4 Quantile estimation in finite populations

For simplicity of notation, the index g is suppressed in this section.

Let Y be a random variable defined over a superpopulation. In the classical statistical framework, we assume a joint distribution of $(Y, \mathbf{X})$ and denote by $F^Y(.)$ and $F^{\mathbf{X}}(.)$ the marginal CDF of Y and $\mathbf{X}$, respectively. We also assume that $Y \mid \mathbf{X} = \mathbf{x} \sim D(h(\mathbf{x}^\top\boldsymbol{\beta}), \boldsymbol{\delta})$, where $D(h(\mathbf{x}^\top\boldsymbol{\beta}), \boldsymbol{\delta})$ is a distribution with parameters $h(\mathbf{x}^\top\boldsymbol{\beta})$ and $\boldsymbol{\delta}$. Note that h is a continuous function, and the first parameter is expressed using some characteristics $\mathbf{x}$ and some other parameters $\boldsymbol{\beta}$. The marginal distribution of Y has the CDF

$$F^Y(y) = \int F_{D(h(\mathbf{x}^\top \boldsymbol{\beta}),\boldsymbol{\delta})}(y \mid \mathbf{x}) \mathrm{d}F^{\mathbf{X}}(\mathbf{x}),$$

where $F_{D(h(\mathbf{x}^\top \boldsymbol{\beta}),\boldsymbol{\delta})}(. \mid \mathbf{x})$ is the CDF of the distribution $D(h(\mathbf{x}^\top \boldsymbol{\beta}), \boldsymbol{\delta})$.

Given $\mathbf{X} = \mathbf{x}$, at the U level, the parameters $h(\mathbf{x}^\top \boldsymbol{\beta}), \boldsymbol{\delta}$ are replaced by $h(\mathbf{x}^\top \boldsymbol{\beta}_N), \boldsymbol{\delta}_N$, respectively, where $h(\mathbf{x}^\top \boldsymbol{\beta}_N), \boldsymbol{\delta}_N$ are parameters computed on U. For instance, if a model similar to the one provided by Expression (2) holds, D is the log-normal distribution, $h(.)$ is the exponential function, $\boldsymbol{\beta}_N$ is similarly defined as in Expression (4), and $\boldsymbol{\delta}_N$ is σ_N^2, the error term.

Conditional on U and given $\mathbf{X} = \mathbf{x}$, the CDF of the distribution of Y is expressed at the U level using the following mixture distribution

$$F_N^Y(y) = \frac{1}{N} \sum_{k \in U} F_{D(h(\mathbf{x}_k^\top \boldsymbol{\beta}_N),\boldsymbol{\delta}_N)}(y \mid \mathbf{x}_k), \text{ for any } y \in \mathcal{R}. \tag{10}$$

Note that here we assume that $F_N^Y(y)$ is a model CDF. This is in contrast to the approach given in the topic-related literature (Särndal et al. 1992, p.197), where the estimand is the finite population empirical distribution function $F_{\text{emp}}^Y(y)$ given by

$$F_{\text{emp}}^Y(y) = \frac{\sum_{k \in U} I(y_k \leq y)}{N}, \tag{11}$$

where $I(.)$ is the indicator function, with $I(y_k \leq y) = 1$ if $y_k \leq y$, 0 otherwise. The CDF estimation in finite populations usually concerns the estimation of $F_{\text{emp}}^Y(.)$ and not of $F_N^Y(.)$. In order to compare the usual approach used in survey sampling and the parametric approach, the interest is here to estimate $F^Y(.)$, because both $F_{\text{emp}}^Y(.)$ and $F_N^Y(.)$ are estimators of $F^Y(.)$.

At the sample level, $F_{\text{emp}}^Y(y)$ is estimated by

$$\widehat{F}_{\text{emp}}^Y(y) = \frac{\sum_{k \in S} w_k I(y_k \leq y)}{\sum_{k \in S} w_k},$$

while the quantile $Q_{(\alpha)}$ of the distribution of Y is estimated by

$$\widehat{Q}_{(\alpha),\text{emp}} = \left[\widehat{F}_{\text{emp}}^Y(\alpha)\right]^{-1}, \tag{12}$$

where $\left[\widehat{F}_{\text{emp}}^Y(.)\right]^{-1}$ denotes the inverse of $\widehat{F}_{\text{emp}}^Y(.)$.

For $F_N^Y(.)$, the parameters $\gamma_{k,N} = h(\mathbf{x}_k^\top \boldsymbol{\beta}_N)$ and $\boldsymbol{\delta}_N$ in Expression (10) are estimated, respectively, by $\widehat{\gamma}_{k,N} = h(\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_N)$ and $\widehat{\boldsymbol{\delta}}_N$, where both estimators are computed on the sample, using a weighted approach with weights w_k. The quantile estimation is done using two methods. The first method (denoted as "Method 1") is based on the estimator of $F_N^Y(y)$ given in Expression (13), and it is similar to the one used by Biewen and Jenkins (2005). As an alternative to Method 1, we propose in this paper a second one (denoted as "Method 2") which is a simulation method.

1. **Method 1**
   $F_N^Y(y)$ is estimated on a sample S using a Hájek-type estimator as

$$\widehat{F}_N^Y(y) = \sum_{k \in S} w_k F_{D(\widehat{\gamma}_{k,N}, \widehat{\delta}_N)}(y \mid \mathbf{x}_k) / \sum_{k \in S} w_k. \tag{13}$$

   Next, the quantile $Q_{(\alpha)}$ of the distribution of Y is estimated by

$$\widehat{Q}_{(\alpha)} = \left[\widehat{F}_N^Y(\alpha)\right]^{-1}.$$

   In many cases, $\left[\widehat{F}_N^Y(.)\right]^{-1}$ is computed using a numerical method.
2. **Method 2**
   If the inverse function of $\widehat{F}_N^Y(\alpha)$ cannot be computed (e.g., lack of monotony of $\widehat{F}_N^Y$) or the numerical method is slow, we introduce and use the following Monte Carlo method based on parametric bootstrap:

   (a) Generate a large number m of n independent draws from the distribution $D(h(\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_N), \widehat{\delta}_N)$, $k \in S$, respectively. A matrix M of dimension $m \times n$ of such draws is obtained. Each element (i, k), $i = 1, \ldots, m, k = 1, \ldots n$ in M is the realization $y_{i,k}$ of a random variable $Y_{i,k}$ with $Y_{i,k} \sim D(h(\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}_N), \widehat{\delta}_N)$; given $\mathbf{x}$ all the random variables $Y_{i,k}$ are independent.
   (b) Associate with each element (i, k), $i = 1, \ldots, m, k = 1, \ldots, n$ of M the weight $w_k, k \in S$ and compute the empirical weighted quantile of order $\alpha \in [0, 1]$

$$\widehat{Q}_{\alpha,\text{emp}}^{(i)} = \left[\widehat{F}_{\text{emp},i}^Y(\alpha)\right]^{-1},$$

   where $\widehat{F}_{\text{emp},i}^Y(y) = \sum_{k=1}^n w_k I(y_{i,k} \leq y) / \sum_{k=1}^n w_k$.
   (c) For each $\alpha \in [0, 1]$, compute the mean of $\widehat{Q}_{\alpha,emp}^{(i)}(Y)$, $i = 1, \ldots, m$; this mean represents an estimator of the quantile of order α of the distribution with the CDF given in Expression (10).

Method 2 allows an easy construction of an approximate $100 \times (1-\gamma)\%$ confidence interval ($\gamma \in (0, 1)$) for a quantile using the method of percentile bootstrap confidence intervals. Conditional to the estimated parameters, each column of the previous matrix provides a set of independent estimates $\widehat{Q}_\alpha^{(i)}(Y)$ of a given quantile α. Next, the empirical quantiles of order $100 \times (\gamma/2)\%$ and $100 \times (1 - \gamma/2)\%$ are computed. They form the lower and upper bounds of an approximate $100 \times (1-\gamma)\%$ confidence interval of a quantile of order α. Monte Carlo simulation results (not shown here) indicate coverage rates close to 95% for this method ($\gamma = 0.05$).

Remark 1 *Method 2* can be improved if the CDF is estimated using all $m \times n$ simulated outcomes as follows

$$\sum_{i=1}^{m} \left(\sum_{k=1}^{n} w_k I(y_{i,k} \leq y) / \sum_{k=1}^{n} w_k \right) / m.$$

This CDF estimator can be inverted to obtain the quantile estimation at level α. Step (c) in Method 2 is no longer necessary. The same remark applies to the algorithm given in Sect. 5. Thus, one can improve the quantile estimator precision by using $m \times n$ outcomes instead of m. However, the computation of an approximate $100 \times (1-\gamma)\%$ confidence interval of a quantile is no longer possible. We use in our results the first version of Method 2.

5 Quantile estimation of the counterfactual distribution

We are interested to estimate the quantiles of the counterfactual distribution. This is necessary for a comparison between them and the estimated quantiles of the unconditional distribution of wage of women and those of the men, respectively, as underlined in Sect. 3.

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

$$F^C_{\text{emp}}(y) = \frac{\sum_{k \in U_F} \psi_k I(y_k \leq y)}{\sum_{k \in U_F} \psi_k}. \tag{14}$$

The weighted version of DiNardo et al. (1996) and Anastasiade and Tillé (2017) methods uses the estimated empirical counterfactual CDF defined by

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

where $\widehat{\psi}_k$ is an estimator of ψ_k given in Expression (9). Next, both methods estimate the α-quantile $Q^C_{(\alpha)}$ of the counterfactual distribution using

$$\widehat{Q}^C_{(\alpha),\text{emp}} = \left[\widehat{F}^C_{\text{emp}}(\alpha) \right]^{-1}. \tag{15}$$

An opposite approach is to use the following model counterfactual CDF at the U_F level

$$F^C_N(y) = \frac{1}{N_C} \sum_{k \in U_F} \psi_k F_{D(h(\mathbf{x}^\top_{k,F} \boldsymbol{\beta}_F), \delta_F)}(y \mid \mathbf{x}_{k,F}), \tag{16}$$

where $N_C = \sum_{k \in U_F} \psi_k$ and $\boldsymbol{\beta}_F$, $\boldsymbol{\delta}_F$ are parameters of the distribution $D(.)$ defined on U_F. We estimate $F_N^C(y)$ by

$$\widehat{F}_N^C(y) = \frac{\sum_{k \in S_F} \widehat{\psi}_k w_k F_{D(h(\mathbf{x}_{k,F}^\top \widehat{\boldsymbol{\beta}}_F), \widehat{\boldsymbol{\delta}}_F)}(y \mid \mathbf{x}_{k,F})}{\sum_{k \in S_F} \widehat{\psi}_k w_k},$$

where $\widehat{\boldsymbol{\delta}}_F$ and $\widehat{\boldsymbol{\beta}}_F$ are computed on S_F and $\widehat{\psi}_k$ is computed, for instance, with the help of the calibration approach of Anastasiade and Tillé (2017), using the raking method; this represents a nonparametric estimation of ψ_k in contrast to the method of DiNardo et al. (1996) which uses a logistic regression model. Next, the estimator of $Q_{(\alpha)}^C$ is given by

$$\widehat{Q}_{(\alpha)}^C = \left[\widehat{F}_N^C(\alpha)\right]^{-1}. \tag{17}$$

If the inverse function of $\widehat{F}_N^C(.)$ cannot be computed or its numerical approximation is slow, the following Monte Carlo method based on parametric bootstrap similar to the one given in Sect. 4 is used:

1. Generate a large number m of n_F independent draws from the distribution $D(h(\mathbf{x}_{k,F}^\top \widehat{\boldsymbol{\beta}}_F), \widehat{\boldsymbol{\delta}}_F)$, $k \in S_F$, 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 $Y_{i,k} \sim D(h(\mathbf{x}_{k,F}^\top \widehat{\boldsymbol{\beta}}_F), \widehat{\boldsymbol{\delta}}_F)$; given $\mathbf{x}_F$ all the random variables $Y_{i,k}$ are independent.
2. Associate with 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 compute the empirical weighted quantile of order $\alpha \in [0,1]$ of the counterfactual wage distribution by

$$\widehat{Q}_{(\alpha)}^{(i),C} = \left[\widehat{F}_{\text{emp},i}^C(\alpha)\right]^{-1},$$

   where $\widehat{F}_{\text{emp},i}^C(y) = \sum_{k=1}^{n_F} \widehat{\psi}_k w_k I(y_{i,k} \leq y) / \sum_{k=1}^{n_F} \widehat{\psi}_k w_k$.
3. For each $\alpha \in [0,1]$, compute the mean of the $\widehat{Q}_{(\alpha)}^{(i),C}$, $i = 1, \ldots, m$; this mean represents an estimate of the quantile of order α of the counterfactual wage distribution.

Remark 2 1. The method to compute $\widehat{Q}_{(\alpha)}^{(i),C}$ uses random weights $\widehat{\psi}_k w_k$, $k \in S_F$. Its computation is reliable because $\widehat{\psi}_k w_k$, $k \in S_F$ are fixed in each run of the algorithm.
2. The reweighting factor ψ_k does not allow the computation of the wage variable corresponding to the counterfactual distribution given in Expression (7) (the variable $\psi_k Y_k^F$ has a different CDF), but only the estimation of some of its parameters.
3. As for gender wage quantile estimation, the use of auxiliary information $\mathbf{x}_{k,F}$ in estimating $F_N^C(y)$ is expected to reduce the variance of the estimator given in Expression (17), compared to that of the estimator given in Expression (15). In

Sect. 6.2, we show some Monte Carlo results that sustain the variance reduction of the two parametric methods compared to the other two competitors.

6 Application using the GB2 distribution

6.1 The GB2 regression model

We illustrate the two parametric methods to estimate the structure and composition effects at quantiles' level using the GB2 distribution. The GB2 distribution is characterized by four parameters, namely a, b, p and q. McDonald and Xu (1995) and Kleiber and Kotz (2003) use the following probability density function of a GB2(a, b, p, q) distribution

$$f(y; a, b, p, q) = \frac{\mid a \mid y^{ap-1}}{b^{ap}\mathrm{B}(p, q)\left[1 + \left(\frac{y}{b}\right)^{a}\right]^{p+q}}, y > 0, \tag{18}$$

where $\mathrm{B}(p, q)$ represents the function Beta(p, q) with arguments p and q. Using the notation of Graf et al. (2011) and Graf and Nedyalkova (2015), Equation (18) is rewritten as

$$f(y; a, b, p, q) = \frac{a\left(\frac{y}{b}\right)^{ap-1}}{b\mathrm{B}(p, q)\left[1 + \left(\frac{y}{b}\right)^{a}\right]^{p+q}}, y > 0. \tag{19}$$

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, while the other two shape parameters control for the left and the right tail, respectively. The GB2 distribution can be positively or negatively skewed, depending upon the values of p and q.

We borrow from McDonald and Butler (1990) the idea of changing the scale parameter b, by expressing it as a function of the observed characteristics of the employees. The framework can also be expressed as a regression model

$$\log(Y_k) = \mathbf{X}_k^{\top}\boldsymbol{\beta} + \log(\varepsilon_k), \tag{20}$$

where $\varepsilon_k \sim$ GB2$(a, 1, p, q)$. As $\varepsilon_k \sim$ GB2$(a, 1, p, q)$, we have that $Y_k \mid \mathbf{X}_k = \mathbf{x}_k \sim$ GB2$(a, \exp(\mathbf{x}_k^{\top}\boldsymbol{\beta}), p, q)$ (see McDonald and Butler 1990). Since ε_k follows a GB2 distribution, we refer to the model in Eq. (20) as a GB2 regression model.

In each group $g \in \{M, F\}$, we assume that the conditional wage of $k \in U_g$, $Y_k \mid \mathbf{X}_{k,g} = \mathbf{x}_k \sim$ GB2$(a_g, \exp(\mathbf{x}_k^{\top}\boldsymbol{\beta}_g), p_g, q_g)$. Thus, for each $k \in U_g$, Expression (19) becomes

$$f\left[y_k; a_g, \exp\left(\mathbf{x}_k^\top \boldsymbol{\beta}_g\right), p_g, q_g\right] = \frac{a_g \left[\dfrac{y_k}{\exp\left(\mathbf{x}_k^\top \boldsymbol{\beta}_g\right)}\right]^{a_g p_g - 1}}{\exp\left(\mathbf{x}_k^\top \boldsymbol{\beta}_g\right) \mathrm{B}(p_g, q_g) \left\{1 + \left[\dfrac{y_k}{\exp\left(\mathbf{x}_k^\top \boldsymbol{\beta}_g\right)}\right]^{a_g}\right\}^{p_g + q_g}}. \quad (21)$$

We use the maximum pseudo-likelihood method to fit GB2 regression models using survey weights. The sandwich estimator (Huber 1967; Freedman 2006; Graf et al. 2011) or parametric bootstrap can be used to estimate the standard errors of the estimated parameters. We describe in "Appendix" the entire approach.

Biewen and Jenkins (2005) suggested to express all the four parameters of the GB2 distribution as a function of the observed characteristics. However, they note that "there would be too many parameters to be estimated, and variance calculations for the statistics of interest are rather complicated. Also we found that estimation often led to numerical problems." We also note that if the survey weights are skewed, the estimation of the parameters is numerically complicated. In order to avoid all these problems, we express in our examples only the scale parameter as a function of the observed characteristics.

6.2 Monte Carlo studies

Monte Carlo simulation was used to show the performances of the two parametric methods when the quantiles of the counterfactual wage distribution are estimated. Three settings have been employed as follows:

- Setting 1, where we generate a conditional wage distribution for women, $Y_{k,F} = \exp[1.10 + X_{k,F} + \varepsilon_{k,F}]$, where $\varepsilon_{k,F} \sim N(0, 1)$ are iid, $X_{k,F} \sim N(5, 1)$, $k = 1, \ldots N_F$, with $N_F = 50{,}000$. 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 $\varepsilon_{k,F} \sim \mathrm{GB2}(8, 1, 0.50, 0.90)$ are iid, $X_{k,F} \sim \mathrm{Gamma}(9, 2)$, $k = 1, \ldots N_F$, with $N_F = 50{,}000$. The covariate for the men is $X_{k,M} \sim \mathrm{Gamma}(10, 2)$, iid, $k = 1, \ldots N_M$, with $N_M = N_F$. The correlation between $\log(Y_F)$ and X_F is about 0.60.
- Setting 3 is similar to Setting 2, using with the same N_F, $X_{k,F}$, $X_{k,M}$ and $\varepsilon_{k,F}$, but $Y_{k,F} = \exp[1.44 + 0.07X_{k,F} + \log(\varepsilon_{k,F})]$, $k = 1, \ldots N_F$. The correlation between $\log(Y_F)$ and X_F is about 0.30.

At the superpopulation level, the counterfactual distribution uses the factor $\psi(x) = \mathrm{d}F^{X_M}(x)/\mathrm{d}F^{X_F}(x)$. For Setting 1, $F^{(Y_F|X_F)}(y \mid x)$ is the CDF of the log-normal distribution with parameters $\boldsymbol{\mu} = \mathbf{x}_F^\top \boldsymbol{\beta}_1$ and $\sigma^2 = 1$, where $\boldsymbol{\beta}_1 = (1.10, 1)'$. For Setting 2, $F^{(Y_F|X_F)}(y \mid x)$ is the CDF of the distribution $\mathrm{GB2}(8, \exp[\mathbf{x}_F^\top \boldsymbol{\beta}_2], 0.50, 0.90)$, with $\boldsymbol{\beta}_2 = (1.44, 0.15)'$; for setting 3, we have a similar situation. For all settings, the quantile $Q_{(\alpha)}^C$ is computed using the inverse of $F^C(\alpha)$ given in Expression (8).

$F^C(.)$ is used, because two different CDF are employed at the finite population level given, respectively, by Expressions (14) and (16). $F^C(\alpha)$ is computed using Monte Carlo integration, with 10,000,000 runs; its inverse at the point α is computed using a numerical method.

We use r runs and draw in each one a random sample of women and men, respectively. In Setting 1, the number of runs equals 10,000, and in Settings 2 and 3, due to the time-consuming process of fitting a GB2 distribution, we use only 1000 runs. In Setting 1, we select samples of women and men, respectively, by simple random sampling without replacement, with sample sizes $n_F = n_M = 1000$. In Settings 2 and 3, we employ systematic sampling with unequal probabilities for both samples with $n_F = n_M = 10{,}000$, where the inclusion probabilities are proportional to x_F and x_M, respectively; these two settings are close to the framework used by the application given in Sect. 6.3.

In each run of the Monte Carlo simulation, we computed 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 Methods 1 and 2, the method of Anastasiade and Tillé (2017) (with raking calibration; hereafter, the calibration method) and the weighted version of the method of DiNardo et al. (1996) (hereafter, weighted DFL).

For each generic estimator $\widehat{Q}^C_{(\alpha)}$ of $Q^C_{(\alpha)}$, the following Monte Carlo measures were used:

- the Monte Carlo relative bias (in percentages)

$$\mathrm{RB}_{\mathrm{MC}}(\widehat{Q}^C_{(\alpha)}) = 100 \times \left(E_{\mathrm{MC}}(\widehat{Q}^C_{(\alpha)}) - Q^C_{(\alpha)}\right) / Q^C_{(\alpha)},$$

  where $E_{\mathrm{MC}}(\widehat{Q}_{(\alpha)}) = \sum_{i=1}^{r} \widehat{Q}^C_{i,(\alpha)}/r$, and $\widehat{Q}^C_{i,(\alpha)}$ is the quantile estimator of $Q^C_{(\alpha)}$ computed in the ith run;
- the Monte Carlo variance

$$\mathrm{Var}_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right) = \frac{1}{r-1} \sum_{i=1}^{r} \left[\widehat{Q}^C_{i,(\alpha)} - E_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right)\right]^2;$$

- the Monte Carlo root mean square error (RMSE)

$$\mathrm{RMSE}_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right) = \left[\mathrm{Var}_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right) + \left(B_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right)\right)^2\right]^{1/2},$$

  where $B_{\mathrm{MC}}(\widehat{Q}^C_{(\alpha)}) = E_{\mathrm{MC}}(\widehat{Q}^C_{\alpha}) - Q^C_{(\alpha)}$,
- the Monte Carlo coefficient of variation (in percentages)

$$\mathrm{CV}_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right) = 100 \times \left(\mathrm{Var}_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right)\right)^{1/2} / E_{\mathrm{MC}}\left(\widehat{Q}^C_{(\alpha)}\right).$$

For the estimators corresponding to Methods 1 and 2, 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 estimated factor ψ_k given by the method of Anastasiade and Tillé (2017) with raking calibration. The latter 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 weighted DFL method was estimated in each run. We used a weighted logistic regression to compute $P(G_k = 1 \mid x_k)$ and $P(G_k = 0 \mid x_k)$, while $P(G_k = 1)$ and $P(G_k = 0)$ were estimated by weighted means $\sum_{k \in S_g} w_k / \sum_{k \in S} w_k$, $g \in \{M, F\}$; see Expression (9). All the results were computed in R Core Team (2022). The weighted empirical quantiles were computed using the function wtd.quantile from the R package Hmisc (Harrell 2022), while the inverse of a CDF at the point α was computed using the R base function uniroot. We used 1000 bootstrap runs in Method 2.

All the used estimators are biased with respect to the sampling design. The values of the Monte Carlo relative bias in percentages are shown in Tables 1, 5 and 9 for the three settings, while the values of the Monte Carlo variance are given in Tables 2, 6 and 10, respectively; the Monte Carlo root mean square errors are reported in Tables 3, 7 and 11, respectively. The Monte Carlo coefficients of variation are given in Tables 4, 8 and 12, respectively. We note that Method 1 and Method 2 provide very close values of the Monte Carlo measures in all three settings.

The estimator of Anastasiade and Tillé (2017) using calibration is used in Figs. 1, 2 and 3 as a benchmark in order to visualize the behavior of the other estimators at different quantiles. Since Method 1 and Method 2 provide almost identical results, only the results of Method 1 are shown in Figs. 1, 2 and 3. Figure 1 shows the ratio between the Monte Carlo bias B_{MC} obtained by using Method 1, the weighted DFL and that 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%. Figure 2 provides the ratio between the Monte Carlo variance of Method 1, the weighted DFL and that of the calibration method for each quantile. Similarly, Fig. 3 shows the ratio of the Monte Carlo RMSEs.

In Setting 1, the two parametric methods show a smaller value of the Monte Carlo relative bias than the weighted DFL and the calibration estimator at each quantile. The two methods also provide a substantial reduction of the Monte Carlo variance at each quantile, and a good behavior with respect to the RMSE over the calibration and weighted DFL estimators (see Fig. 1). The estimators obtained by the two parametric methods also display a smaller coefficient of variation than the reweighting estimators at all quantiles as shown in Table 4.

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 and 0.91, showing that we provide approximately unbiased estimates under the sampling design, for large sample sizes. In Setting 2, the two parametric methods result in estimators that have a lower Monte Carlo variance than the calibration and the weighted DFL estimators almost at all quantiles (see Fig. 2). Like in Setting 1, the value of the Monte Carlo coefficient of variation of the estimators obtained using the two parametric methods are smaller than of those using the last two methods (see Table 8). The parametric 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 Fig. 2). Compared to Setting

1, note that the correlation between $\log(Y_F)$ and X_F is less important (0.60 compared to 0.70).

Setting 3 shows a smaller correlation between $\log(Y_F)$ and X_F (about 0.30) compared to Setting 2; this is similar to the correlation between the logarithm of women wage and age in the application given in Sect. 6.3. This correlation reduction is visible in the behavior of the Monte Carlo variance and relative bias of the two parametric methods. Thus, the parametric methods still provide a reduction of the Monte Carlo variance for most of quantiles (except for the quantiles of order 30%, 80% and 95%; see also Fig. 3) compared to the two competitors. The value of the Monte Carlo relative bias of the two parametric methods is more important than in Setting 2 for the quantiles of order 20% and 70%. Despite the lower correlation between $\log(Y_F)$ and X_F, the shapes of the Monte Carlo RMSE of the two parametric methods are similar to the ones provided by Setting 2; see the last plot in Figs. 1 and 2, respectively. The Monte Carlo coefficients of variation of Method 1 and Method 2 also show reduced values compared to the other two methods (see Table 12).

6.3 Application to real data

A real dataset with information collected during the Swiss survey on earnings in 2012 by the Swiss Federal Statistical Office is used to illustrate the methods. All the cases where there is missing information are removed. We also removed observations where the monthly wage is less than 1000 CHF 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 2012 in the private sector and are between 18 and 64 years old. The modeled variable is the standardized hourly wage. Standardized refers to the fact that the hourly wages are reported as if all the employees worked full-time. Finally, we use a sample of 144,753 employees: 66,181 women and 78,572 men. The wages of women range between 5 and 149.76 CHF, while those of men between 5 and 299.53 CHF. Figure 4 shows the estimated wage densities of men and women, respectively.

A GB2 regression model was fitted separately for men and women, using the maximum pseudo-likelihood method with survey weights. We used four explanatory variables in the models: the age of the employee (that is a proxy for professional experience, since this information was absent), the education level (8 ordinal categories, with the first category being the most important), the professional position (4 ordinal categories, with the first category being the most important) and the economic sector (38 categories, the first one being the tobacco industry, where the median of the wages is the largest one in each group). The correlation between the logarithm of the wage and age is about 0.30 for both groups, respectively; this value is similar to the one used in Setting 3 given in Sect. 6.2.

In each group, there are 104 parameters to estimate (52 coefficients and 52 standard errors). They are reported in Table 15 in Appendix for the women's sample and men's sample, respectively. The standard errors are estimated using the sandwich estimator. Age has a positive effect on the wages for both groups. The first levels for education, professional position and economic sector are taken as reference categories in the

Table 1 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 2 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	20,518.87
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	18,044.12

Table 3 Setting 1: Monte Carlo RMSE 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 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 5 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 6 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 7 Setting 2: Monte Carlo RMSE 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 8 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

Table 9 Setting 3: 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	− 1.35	− 0.78	− 0.54	− 0.31	− 0.19	− 0.09	− 0.01	0.08	0.18	0.31	0.51	0.70	1.08
Method 2	− 1.26	− 0.77	− 0.54	− 0.31	− 0.19	− 0.09	− 0.01	0.08	0.18	0.31	0.51	0.70	1.09
Calibration	− 1.67	− 0.67	− 0.38	− 0.15	− 0.12	− 0.09	0.06	0.08	0.05	0.28	0.42	0.76	1.56
Weighted DFL	− 1.44	− 0.48	− 0.14	0.07	0.09	0.16	0.29	0.32	0.31	0.55	0.68	1.05	1.90

Table 10 Setting 3: Monte Carlo variance of the four estimators of the counterfactual wage quantiles (results multiplied by 100)

Quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Method 1	0.009	0.006	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.009	0.021	0.149
Method 2	0.009	0.006	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.009	0.021	0.149
Calibration	0.022	0.007	0.008	0.005	0.004	0.005	0.007	0.006	0.007	0.005	0.011	0.018	0.345
Weighted DFL	0.023	0.007	0.008	0.005	0.004	0.005	0.007	0.006	0.007	0.006	0.011	0.018	0.342

Table 11 Setting 3: Monte Carlo RMSE 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	0.013	0.011	0.009	0.008	0.007	0.007	0.006	0.006	0.008	0.010	0.016	0.025	0.056
Method 2	0.013	0.011	0.009	0.008	0.007	0.007	0.006	0.006	0.008	0.010	0.017	0.025	0.056
Calibration	0.018	0.011	0.010	0.008	0.006	0.007	0.008	0.008	0.008	0.009	0.015	0.026	0.082
Weighted DFL	0.018	0.010	0.009	0.007	0.006	0.008	0.010	0.010	0.011	0.015	0.021	0.034	0.092

Table 12 Setting 3: 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	1.52	0.83	0.61	0.47	0.42	0.38	0.35	0.32	0.31	0.31	0.36	0.49	1.03
Method 2	1.52	0.83	0.61	0.47	0.42	0.38	0.35	0.32	0.31	0.31	0.37	0.49	1.03
Calibration	2.37	0.89	0.81	0.53	0.39	0.43	0.46	0.41	0.39	0.30	0.40	0.45	1.55
Weighted DFL	2.39	0.89	0.80	0.52	0.40	0.43	0.47	0.40	0.40	0.33	0.40	0.45	1.54

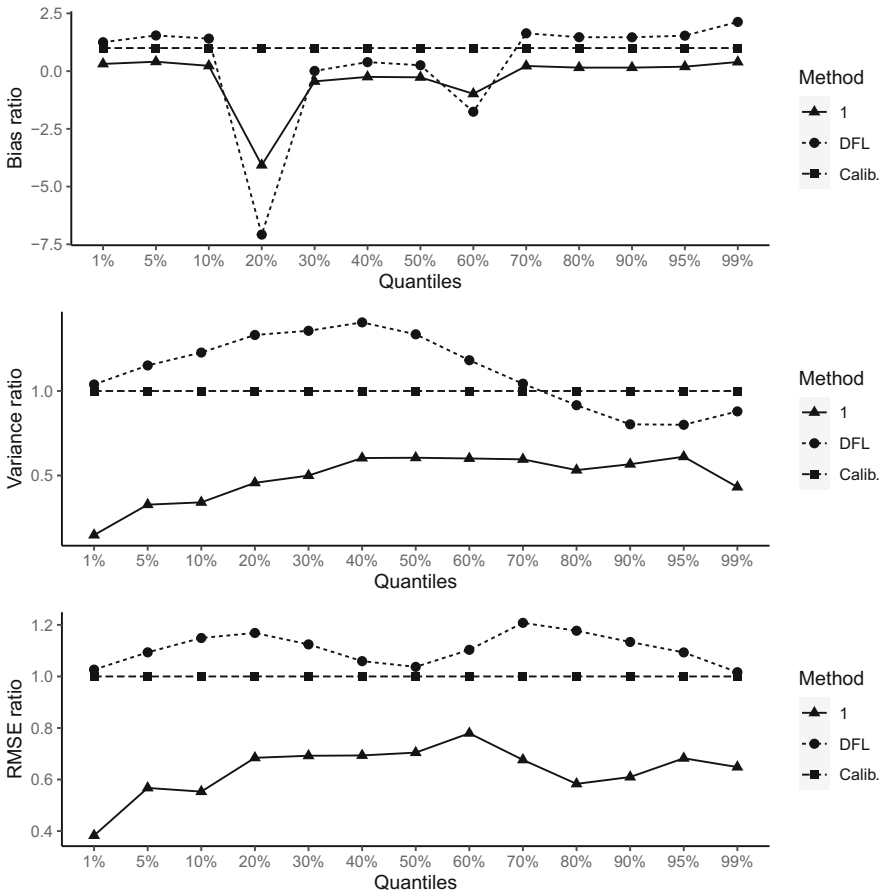

Fig. 1 Setting 1, upper panel: ratio between the Monte Carlo bias obtained by using Method 1, the weighted DFL method and that of the calibration method for each quantile; middle: ratio between the Monte Carlo variance obtained by using Method 1 and that of the calibration method for each quantile; lower panel: ratio between the Monte Carlo RMSE obtained by using Method 1 and that of the calibration method for each quantile. In each panel, the horizontal line drawn at level 1 on the y-axis corresponds to the calibration method. Method 1 and Method 2 provide almost identical ratios

GB2 models. The estimated coefficients associated with each category of the three covariates show negative values, as expected.

Figure 5 shows the histogram of the GB2 residuals and the corresponding P–P plot using the estimated parameters of the GB2 distribution fitted on the women's sample. The P–P plot indicates a good agreement with the GB2$(\widehat{a}_F, 1, \widehat{p}_F, \widehat{q}_F)$ distribution. A similar plot was obtained on the men's sample.

In addition to the previous plot, to test the goodness of fit of the GB2 distribution, we used a nonparametric bootstrap version of the Kolmogorov–Smirnov test (Meintanis and Swanepoel 2007), because the parameters a, p and q are estimated. Recall that in GB2 regression, the residuals should follow the GB2$(a, 1, p, q)$ distribution. We implemented the test by exploiting the relationship between the GB2(a, b, p, q) and

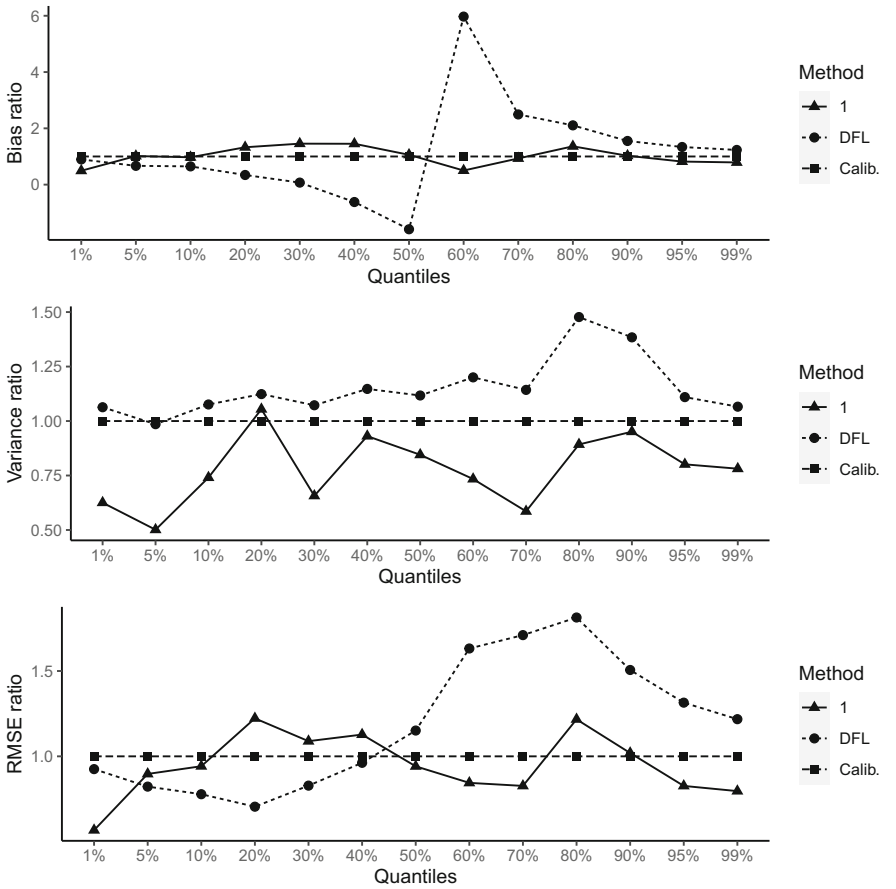

Fig. 2 Setting 2, upper panel: ratio between the Monte Carlo bias obtained by using Method 1, the weighted DFL method and that of the calibration method for each quantile; middle: ratio between the Monte Carlo variance obtained by using Method 1 and that of the calibration method for each quantile; lower panel: ratio between the Monte Carlo RMSE obtained by using Method 1 and that of the calibration method for each quantile. In each panel, the horizontal line drawn at level 1 on the y-axis corresponds to the calibration method. Method 1 and Method 2 provide almost identical ratios

the Beta(p, q) distributions: If $Z \sim$ GB2(a, b, p, q), then $(Z/b)^a/(1 + (Z/b)^a) \sim$ Beta(p, q). Thus, using the transformation $Z^{\widehat{a}_F}/(1 + Z^{\widehat{a}_F})$ (with $b_F = 1$), we tested if the transformed version of the residuals follow the Beta$(\widehat{p}_F, \widehat{q}_F)$ distribution. The test statistic corresponding to the Kolmogorov–Smirnov test for beta distribution is available in majority of statistical software. However, the p-value of the test should be approximated by bootstrap because the parameters a, p and q are estimated. The test was applied on the women's sample using 2500 bootstrap replicates. The approximate p-value of the test was 0.56. A similar value was obtained for the residuals obtained on the men's sample.

In order to compare the GB2 regression with the log-normal model usually used in official statistics, we show in Fig. 6 the QQ plots of the standardized residuals of

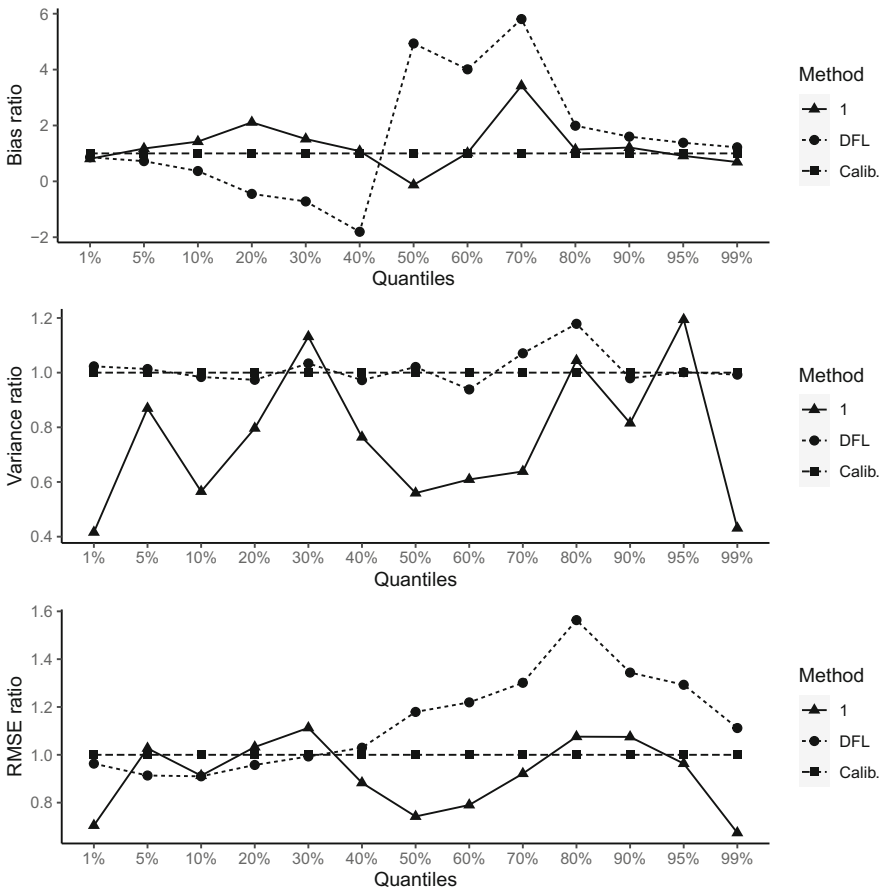

Fig. 3 Setting 3, upper panel: ratio between the Monte Carlo bias obtained by using Method 1, the weighted DFL method and that of the calibration method for each quantile; middle: ratio between the Monte Carlo variance obtained by using Method 1 and that of the calibration method for each quantile; lower panel: ratio between the Monte Carlo RMSE obtained by using Method 1 and that of the calibration method for each quantile. In each panel, the horizontal line drawn at level 1 on the y-axis corresponds to the calibration method. Method 1 and Method 2 provide almost identical ratios

the log-linear model (with the log of the wages as dependent variable and the same characteristics as covariates) and the residuals of the GB2 model, respectively. Both models are fitted on the women's sample. We observe an important improvement of the QQ plot when we fit a GB2 model compared to the log-linear model. A similar result was obtained on the men's sample.

The quantiles of the counterfactual wage distribution were estimated using, respectively, the calibration method and weighted DFL. The factor ψ_k was estimated like in Sect. 6.2: using raking ratio for the calibration method and a weighted logistic regression for the weighted DFL. The empirical distributions of the estimated ψ_k for both methods are very similar and positively skewed (the skewness coefficients are, respectively, 4.81 for weighted DFL and 5.01 for calibration), also due to skewed sur-

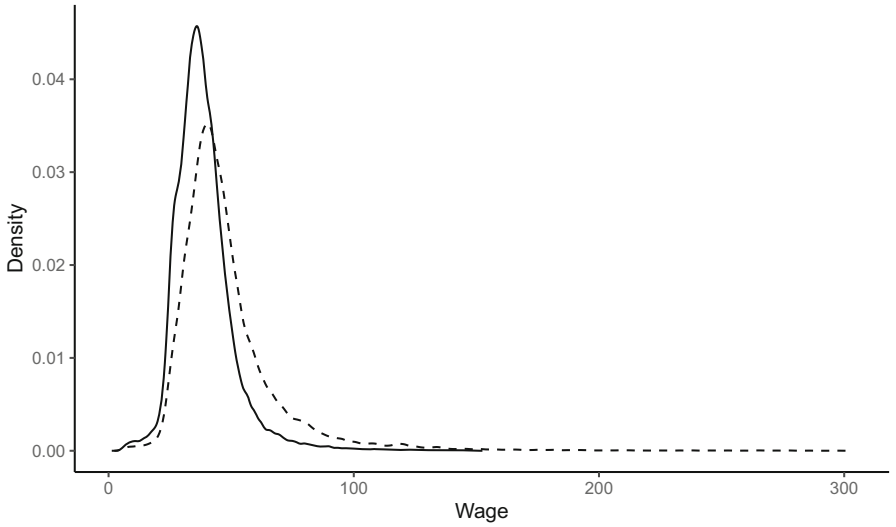

Fig. 4 Application to real data: estimated wage densities of men (dashed) and women (solid)

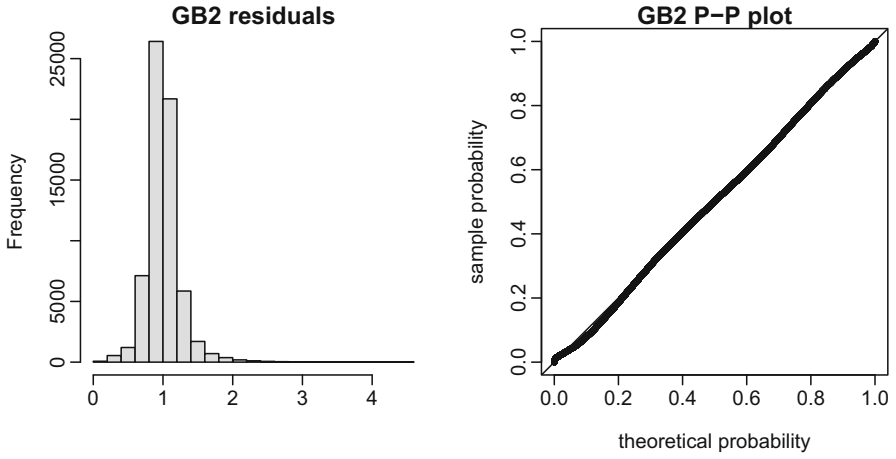

Fig. 5 Application to real data. GB2 regression model fitted on the women's sample: histogram of the residuals (left panel) and P–P plot (right panel)

vey weights (the skewness coefficient equals to 3.56 on the overall sample). Figure 7 shows the boxplots of estimated ψ_k for both methods on the logarithmic scale.

The estimated quantiles of the counterfactual distribution provided by the calibration method and weighted DFL are compared with the estimated quantiles of the wage distribution of men, and with those of the wage distribution of women; they have been computed using Expression (12). Table 13 summarizes the results.

Method 1 and Method 2 have also been applied. In Table 14, we show the corresponding results: the estimated quantiles of men's wages, of the counterfactual wage distribution computed and finally, those of women's wages using Method 1 and Method

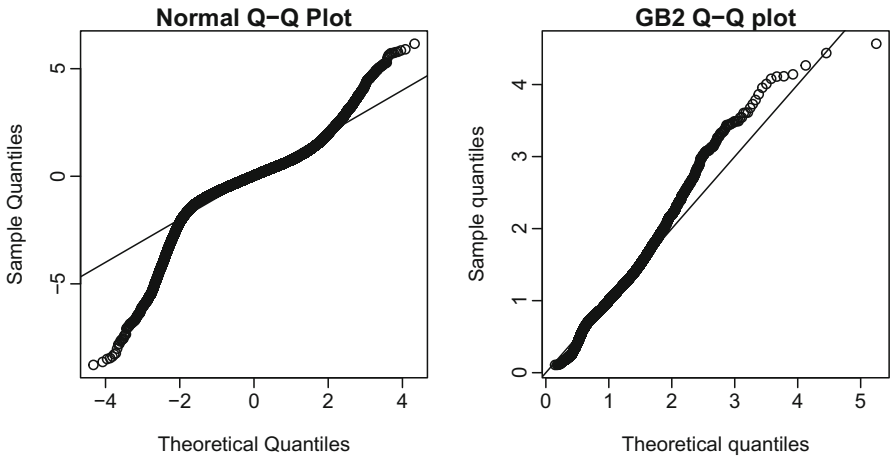

Fig. 6 Application to real data. Women's sample: QQ plot of the standardized residuals for the log-linear model (left panel) and QQ plot of the residuals for the GB2 model (right panel)

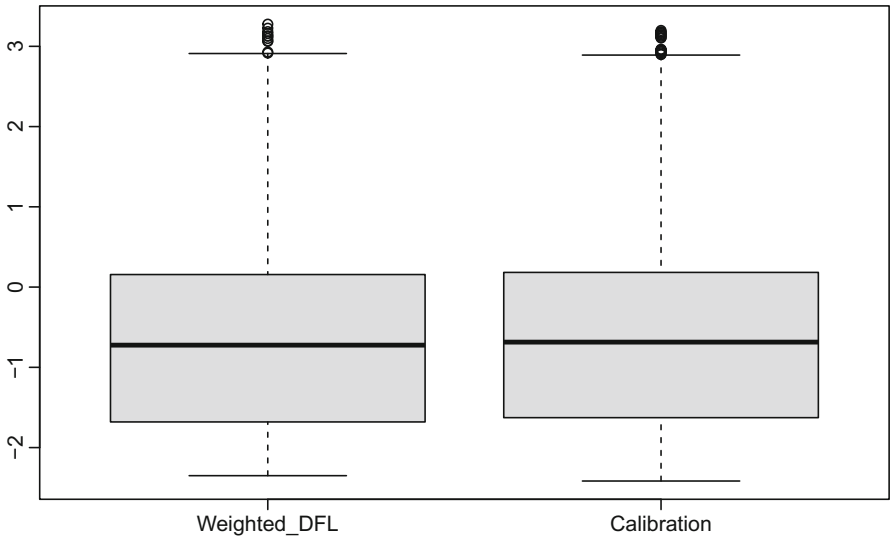

Fig. 7 Application to real data. Boxplots of the estimated $\log(\psi_k)$ using the weighted DFL (left) and calibration method (right). The logarithmic scale is used for a better visualization because both distributions are positively skewed

2. These quantiles were computed using the approach described in Sect. 4 (for the gender wages) and Sect. 5 (for the counterfactual distribution). As in Sect. 6.2, we used 1000 bootstrap runs in Method 2.

Tables 13 and 14 provide the estimated quantiles of the counterfactual distribution for the four estimators (calibration, weighted DFL, Method 1, Method 2). They are within the range of the estimated gender quantiles. The calibration method and the weighted DFL provide similar results; the results for Method 1 and Method 2 are

almost identical. The four methods provide similar values of the quantile estimators of order 2% up to 90% for the counterfactual distribution. More important differences are noted at the lowest and highest levels: 1% and 99%. At these levels, the values of the estimators are, respectively, 14.42.and 93.99 for the calibration method, and 14.25 and 95.27 for the weighted DFL. On the other hand, Method 1 shows, respectively, the values 17.95 and 78.79, and Method 2, 17.94 and 78.85.

Using all four methods, the differences of type $\widehat{Q}^M_{(\alpha)} - \widehat{Q}^C_{(\alpha)}$ are positive, showing an important unexplained gender wage gap at each quantile (see also Fig. 8; since the results are very similar for the calibration method and the weighted DFL, and, respectively, for Method 1 and Method 2, for a better visualization only the results of the calibration method and Method 1 are plotted). We note that the most important difference (approximately -10) between Method 1 and the calibration method is shown at the quantile of level 99%. The differences $\widehat{Q}^M_{(\alpha)} - \widehat{Q}^C_{(\alpha)}$ also impact the ratio between the estimated value of the unexplained part and the estimated difference between the gender wage quantiles at level α, $\widehat{Q}^M_{(\alpha)} - \widehat{Q}^F_{(\alpha)}$; see also Fig. 9 (since the results are very similar for the calibration method and the weighted DFL, and, respectively, for Method 1 and Method 2, for a better visualization only the results of the calibration method and Method 1 are plotted). The calibration method and the weighted DFL show approximate ratios between 70% and 92%, while Method 1 and Method 2 reduce the ratio range: 79% to 90%. The shapes of the ratio curves provided by Method 1 and by the calibration method are very different as shown in Fig. 9. Method 1 provides a smoother curve than the calibration method, mainly due to the parametric approach used. On the other hand, the ratios of the calibration method are impacted by the skewness of the estimated ψ_k factor and thus impact the shape of their curve.

7 Discussion and conclusions

Our aim is to estimate the unexplained part of the wage gap between men and women at different quantiles of the wage distributions, within the usual context of official statistics, where random samples are selected from finite populations. To do this, we fit parametric models on the conditional wage distributions, using some auxiliary information. The parametric approach can be applied for any conditional distribution of the wages. We illustrate it with a GB2 distribution by modeling wages with covariates and using survey weights. This approach extends the classical framework of a log-normal model of the wages usually used in official statistics. The GB2 distribution includes many distributions as special or limiting cases (for example, the log-normal distribution), and it is expected that its use provides a good fit for wage distributions, which are usually heavy-tailed.

Strictly speaking, the estimators corresponding to Methods 1 and 2 are design-based, 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 the first Monte Carlo simulation (Setting 1), the correlation between the logarithm of the women wage and the covariate is larger compared to Settings 2 and

Table 13 Application to real data. Estimated quantiles of gender wage distributions and of the counterfactual wage distribution computed using the calibration method and weighted DFL

Order of quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Men	18.32	26.07	29.31	33.52	36.79	39.57	42.35	45.62	49.57	55.79	68.51	84.43	137.25
Calibration	14.42	23.50	26.07	29.38	32.29	34.99	37.50	40.41	43.60	48.13	56.82	66.16	93.99
Weighted DFL	14.25	23.43	26.02	29.30	32.19	34.90	37.49	40.38	43.61	48.23	57.06	66.59	95.27
Women	13.74	23.23	25.74	29.02	31.84	34.31	36.58	38.98	41.80	45.40	51.83	59.82	86.17

The estimated quantiles of the gender wage distributions have been computed using the classical approach

Table 14 Application to real data. Estimated quantiles of gender wage distributions and of the counterfactual wage distribution computed using the two parametric methods

Order of quantile	1%	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	99%
Men													
Method 1	20.26	26.15	29.42	33.64	37.03	40.25	43.60	47.36	51.95	58.25	69.09	80.39	111.17
Method 2	20.26	26.16	29.43	33.64	37.03	40.25	43.60	47.36	51.95	58.25	69.09	80.39	111.29
Counterfactual													
Method 1	17.95	23.65	26.70	30.45	33.26	35.83	38.40	41.17	44.39	48.55	55.19	61.76	78.79
Method 2	17.94	23.65	26.70	30.45	33.26	35.82	38.39	41.16	44.39	48.56	55.20	61.79	78.85
Women													
Method 1	17.47	23.02	26.00	29.65	32.38	34.80	37.19	39.73	42.67	46.50	52.76	59.07	75.42
Method 2	17.46	23.01	25.99	29.65	32.38	34.80	37.19	39.73	42.66	46.50	52.76	59.07	75.42

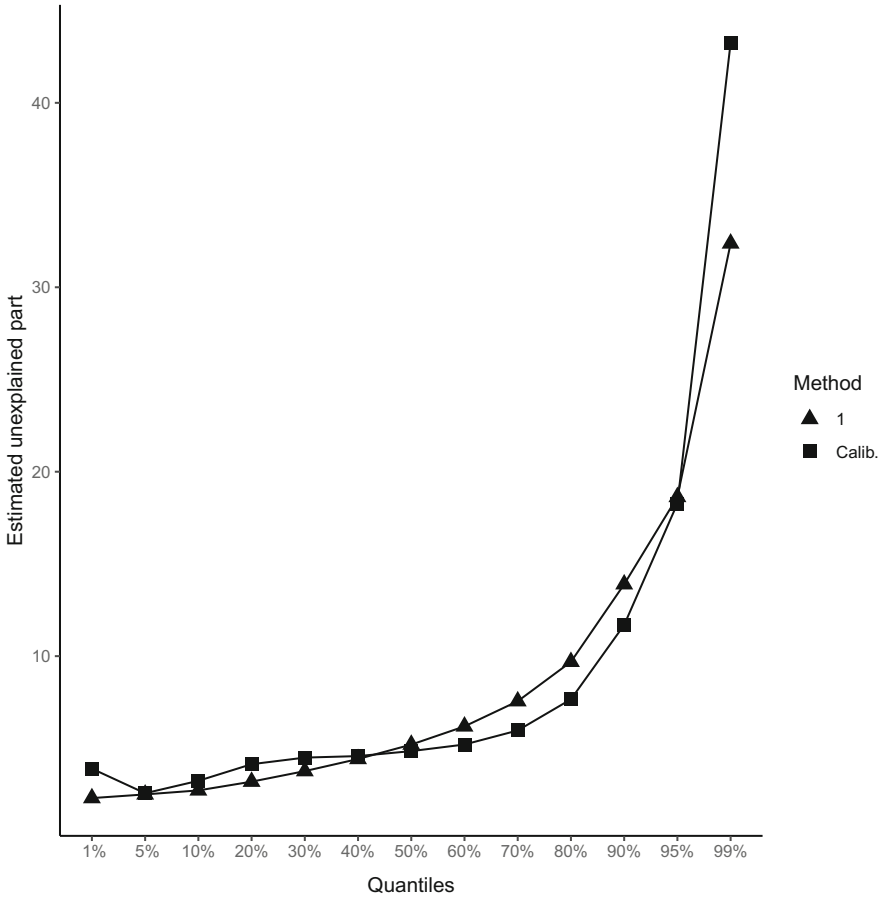

Fig. 8 Application to real data. The unexplained part of the decomposition at each quantile estimated using Method 1 (filled triangle) and the calibration method (filled square)

3 (0.70 versus, respectively, 0.60 and 0.30). This difference could explain that the variance reduction of the parametric methods compared to the calibration method is less important in Settings 2 and 3 than in Setting 1. Nevertheless, the parametric approach provides in all our simulation studies a good behavior in terms of Monte Carlo variance and coefficient of variation compared to the other two competitors.

We also note that there are different ways to estimate a counterfactual distribution as provided in the econometrics literature. The choice of the method used in this paper is determined by the standard use of reweighting estimators in the design-based approach.

We provided an example of GB2 regression model using real data from the Swiss Federal Statistical Office. Compared to the usual log-normal model, the GB2 model showed a better fit for the conditional wage distributions. The unexplained part of the wage differences was estimated using a parametric approach and compared to the results obtained using the methods of DiNardo et al. (1996) and Anastasiade

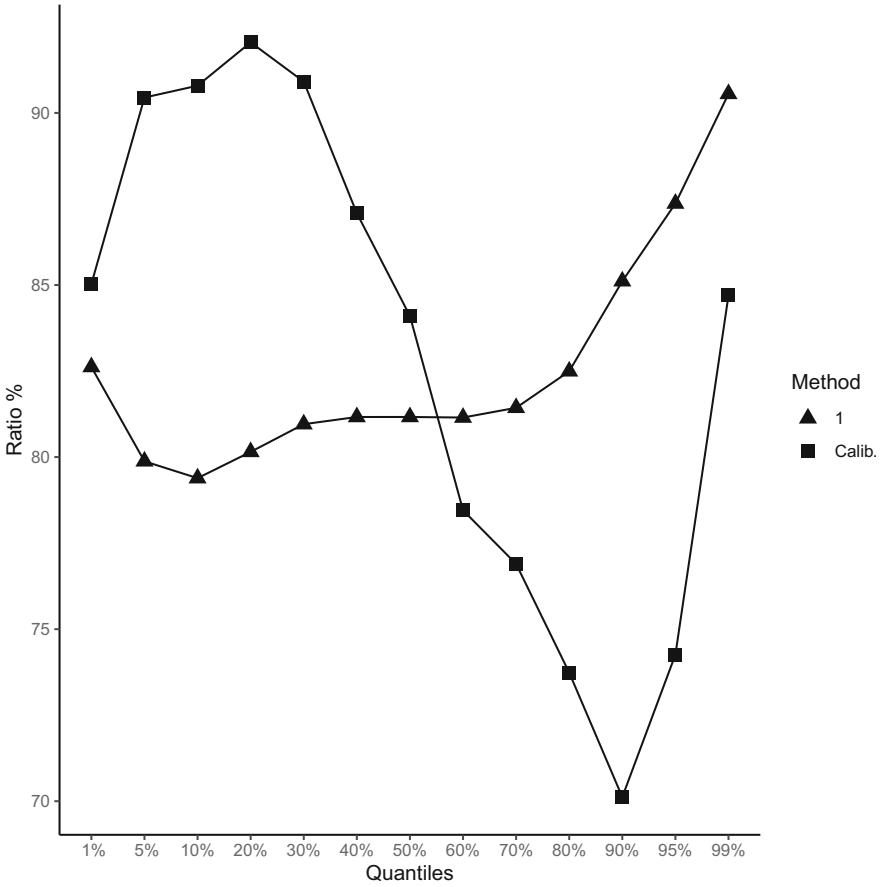

Fig. 9 Application to real data. Ratio of the unexplained part and the difference between the estimated gender quantiles (Method 1—filled triangle and calibration method—filled square). The scale of ratio is expressed in % on the y-axis

and Tillé (2017). The GB2 approach provides smooth results compared to the other two methods, which are impacted by the skeweness of the estimated ψ_k factor. The parametric model used for the conditional distribution "regularizes" the estimates, but impose additional restrictions in the form of a parametric conditional distribution. We used in our computations a reweighting factor involved in the construction of the counterfactual distribution which is estimated using the calibration approach, and not a logistic regression; this represents a nonparametric estimation of this factor. Thus, finally, only the model relating the variable of interest and the covariates should be roughly correct, because the use of the weights in the design-based approach may protect from a misspecification of the model.

Acknowledgements We are grateful to the associate editor and two referees for their constructive comments which have helped us to improve the paper.The research was funded by the Swiss Federal Statistical Office.

Funding Open access funding provided by University of Neuchâtel.

Appendix

Estimation of the parameters in GB2 regression with survey weights

We estimate the parameters in GB2 regression using the maximum pseudo-likelihood method (Chambers 2003, p. 22) on a sample S_g of size n_g. For simplicity of notation, the index g is suppressed in this section. The pseudo-log-likelihood function is

$$l = \frac{\sum_{k=1}^{n} w_k \log f\left[y_k; a, \exp\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right), p, q\right]}{\sum_{k=1}^{n} w_k}, \tag{22}$$

where w_k is the survey weight allotted to individual k and $f(\cdot)$ is the density defined in Expression (21). The function l in Expression (22) is maximized with respect to the parameters. The number of parameters in the log-likelihood function depends on the number of covariates in the model. If there are c covariates, then there are $c + 3$ parameters to be estimated.

When covariates and weights are introduced, the maximization of the pseudo-log-likelihood function can show multiple local maximum points. In such cases, the choice of the starting points of the algorithm used to maximize l have a crucial importance. As for i.i.d. GB2 fits (already underlined by Graf et al. 2011), the sample sizes should be large.

We use the sandwich estimator (or the Huber estimator) to estimate the standard errors of the estimated parameters (Huber 1967; Freedman 2006; Graf et al. 2011) in the GB2 distribution. To compute the sandwich estimator, we first estimate the Fisher matrix as

$$\widehat{\mathbf{V}} = \sum_{i=1}^{n} w_i^2 l'[f(y_i)] \left(l'[f(y_i)]\right)^\top,$$

where $l'[f(y_i)]$ is a $(c + 3) \times 1$ vector containing the first-order derivatives of the pseudo-log-likelihood function l with respect to the parameters. The vector of estimated standard errors is provided by

$$\left(\mathbf{l}''[f(y_i)]\right)^{-1} \widehat{\mathbf{V}} \left(\mathbf{l}''[f(y_i)]\right)^{-1},$$

Table 15 Application to real data. Values of the estimated parameters of the GB2 regression in the women's sample and men's sample and their estimated standard errors given in parentheses

Parameter	Estimates for women		Estimates for men	
a	26.285	(0.224)	28.362	(0.341)
β_0	4.264	(0.062)	4.530	(0.042)
β_{age}	0.092	(0.001)	0.112	(0.001)
β_{educ1}	− 0.110	(0.009)	− 0.093	(0.007)
β_{educ2}	− 0.124	(0.008)	− 0.174	(0.006)
β_{educ3}	− 0.215	(0.017)	− 0.268	(0.018)
β_{educ4}	− 0.210	(0.010)	− 0.285	(0.008)
β_{educ5}	− 0.258	(0.008)	− 0.288	(0.006)
β_{educ6}	− 0.400	(0.012)	− 0.381	(0.009)
β_{educ7}	− 0.354	(0.010)	− 0.375	(0.008)
β_{prof1}	− 0.049	(0.009)	− 0.132	(0.005)
β_{prof2}	− 0.117	(0.009)	− 0.221	(0.005)
β_{prof3}	− 0.204	(0.008)	− 0.281	(0.005)
β_{sect1}	− 0.272	(0.093)	− 0.338	(0.048)
β_{sect2}	− 0.362	(0.062)	− 0.436	(0.042)
β_{sect3}	− 0.447	(0.063)	− 0.481	(0.045)
β_{sect4}	− 0.287	(0.062)	− 0.394	(0.042)
β_{sect5}	− 0.200	(0.062)	− 0.126	(0.042)
β_{sect6}	− 0.104	(0.062)	− 0.178	(0.041)
β_{sect7}	− 0.306	(0.062)	− 0.380	(0.041)
β_{sect8}	− 0.327	(0.062)	− 0.419	(0.041)
β_{sect9}	− 0.233	(0.061)	− 0.307	(0.041)
β_{sect10}	− 0.265	(0.062)	− 0.342	(0.041)
β_{sect11}	− 0.245	(0.061)	− 0.350	(0.041)
β_{sect12}	− 0.309	(0.066)	− 0.392	(0.043)
β_{sect13}	− 0.271	(0.062)	− 0.361	(0.041)
β_{sect14}	− 0.182	(0.062)	− 0.308	(0.041)
β_{sect15}	− 0.295	(0.069)	− 0.362	(0.044)
β_{sect16}	− 0.265	(0.062)	− 0.311	(0.041)
β_{sect17}	− 0.217	(0.061)	− 0.347	(0.041)
β_{sect18}	− 0.387	(0.061)	− 0.444	(0.041)
β_{sect19}	− 0.191	(0.061)	− 0.261	(0.041)
β_{sect20}	− 0.316	(0.063)	− 0.413	(0.045)
β_{sect21}	− 0.453	(0.062)	− 0.598	(0.041)
β_{sect22}	− 0.173	(0.062)	− 0.251	(0.042)
β_{sect23}	− 0.182	(0.062)	− 0.310	(0.042)
β_{sect24}	− 0.205	(0.062)	− 0.293	(0.042)
β_{sect25}	− 0.158	(0.061)	− 0.184	(0.041)
β_{sect26}	− 0.150	(0.061)	− 0.194	(0.041)
β_{sect27}	− 0.223	(0.062)	− 0.364	(0.042)

The standard errors are estimated using the sandwich estimator

Table 15 continued

Parameter	Estimates for women		Estimates for men	
β_{sect28}	− 0.208	(0.061)	− 0.332	(0.041)
β_{sect29}	− 0.126	(0.063)	− 0.166	(0.044)
β_{sect30}	− 0.374	(0.063)	− 0.451	(0.045)
β_{sect31}	− 0.343	(0.061)	− 0.447	(0.042)
β_{sect32}	− 0.249	(0.065)	− 0.352	(0.049)
β_{sect33}	− 0.266	(0.062)	− 0.445	(0.043)
β_{sect34}	− 0.215	(0.061)	− 0.402	(0.041)
β_{sect35}	− 0.348	(0.064)	− 0.500	(0.047)
β_{sect36}	− 0.269	(0.062)	− 0.416	(0.043)
β_{sect37}	− 0.581	(0.072)	− 0.571	(0.078)
p	0.223	(0.001)	0.226	(0.001)
q	0.255	(0.004)	0.177	(0.003)

The standard errors are estimated using the sandwich estimator

where $\mathbf{l}''[f(y_i)]$ is a $(c+3) \times (c+3)$ matrix containing the second-order derivatives of the pseudo-log-likelihood function l with respect to the parameters.

Alternatively, the standard errors can be estimated using a parametric bootstrap method as follows:

1. compute the estimated parameters $\widehat{a}$, $\widehat{p}$, $\widehat{q}$ and $\widehat{\boldsymbol{\beta}}$ from the sample S.
2. generate a large number m of draws from the distribution GB2($\widehat{a}$, $\widehat{p}$, $\widehat{q}$, $\widehat{\boldsymbol{\beta}}$) using the estimated parameters at Step 1; for each draw $j = 1, \ldots, m$, re-estimate the parameters $\widehat{a}_j^*$, $\widehat{p}_j^*$, $\widehat{q}_j^*$ and $\widehat{\boldsymbol{\beta}}_j^*$,
3. compute, respectively, the standard deviation of the estimated parameters $\widehat{a}_j^*$, $\widehat{p}_j^*$, $\widehat{q}_j^*$ and $\widehat{\boldsymbol{\beta}}_j^*$, $j = 1, \ldots, m$, obtained in the m runs.

References

Anastasiade MC, Tillé Y (2017) Decomposition of gender wage inequalities through calibration: application to the Swiss structure of earnings survey. Surv Methodol 43(2):211–234

Bandourian R, McDonald J, Turley RS (2002) A comparison of parametric models of income distribution across countries and over time. Technical Report 305, Luxembourg Income Study

Biewen M, Jenkins SP (2005) A framework for the decomposition of poverty differences with an application to poverty differences between countries. Empir Econ 30(2):331–358

Blinder AS (1973) Wage discrimination: reduced form and structural estimates. J Hum Resour 8(4):436–455

Chambers RL (2003) Introduction to part A. In: Chambers RL, Skinner CJ (eds) Analysis of survey data. Wiley, Hoboken, pp 13–28

Chernozhukov V, Fernández-Val I, Melly B (2013) Inference on counterfactual distributions. Econometrica 81(6):2205–2268

Deville J-C, Särndal C-E (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87(418):376–382

DiNardo J, Fortin NM, Lemieux T (1996) Labor market institutions and the distribution of wages, 1973–1992: a semiparametric approach. Econometrica 64(5):1001–1044

Fortin N, Lemieux T, Firpo S (2011) Decomposition methods in economics. In: Ashenfelter O, Card D (eds) Handbook of labor economics, volume 4 of handbook of labor economics, chapter 1. Elsevier, Amsterdam, pp 1–102

Freedman DA (2006) On the so-called "Huber sandwich estimator" and "robust standard errors". Am Stat 60(4):299–302

Graf M, Nedyalkova D (2015) GB2: generalized beta distribution of the second kind: properties, likelihood, estimation. R Package Ver 2:1

Graf M, Nedyalkova D, Münnich R, Seger J, Zins S (2011) Parametric estimation of income distributions and indicators of poverty and social exclusion. Research Project Report, FP7-SSH-2007-217322 AMELI, European Commission

Harrell Jr FE (2022) Hmisc: Harrell miscellaneous. R package version 4.7-1

Huber PJ (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Berkeley, CA, vol 1, pp 221–233

Kleiber C, Kotz S (2003) Statistical size distributions in economics and actuarial sciences. Wiley, Hoboken

Leythienne D, Ronkowski P (2018) A decomposition of the unadjusted gender pay gap using structure of earnings survey data. Technical report statistical working papers, Eurostat

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

McDonald JB, Butler RJ (1990) Regression models for positive random variables. J Econom 43(1–2):227–251

McDonald J, Ransom M (2008) The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. In: Chotikapanich D (ed) Modeling income distributions and Lorenz curves, vol 5. Economic studies in equality, social exclusion and well-being. Springer, New York, pp 147–166

McDonald JB, Xu YJ (1995) A generalisation of the beta distribution with applications. J Econom 66:133–152

Meintanis S, Swanepoel J (2007) Bootstrap goodness-of-fit tests with estimated parameters based on empirical transforms. Stat Prob Lett 77(10):1004–1013

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

Oaxaca R (1973) Male–female wage differentials in urban labor markets. Int Econ Rev 14(3):693–709

Popli GK (2013) Gender wage differentials in Mexico: a distributional approach. J R Stat Soc Ser A Stat Soc 176(2):295–319

R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Särndal C-E, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer, New York

Thurow LC (1970) Analyzing the American income distribution. Am Econ Rev 60(2):261–269

Van Kerm P (2013) Generalized measures of wage differentials. Empir Econ 45(1):465–482

Van Kerm P, Yu S, Choe C (2016) Decomposing quantile wage gaps: a conditional likelihood approach. J R Stat Soc Ser C Appl Stat 65(4):507–527