DISCUSSION PAPER SERIES

IZA

IZA DP No. 4783

Gender Wage Gap: A Semi-Parametric Approach with Sample Selection Correction

Matteo Picchio
Chiara Mussida

February 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Gender Wage Gap: A Semi-Parametric Approach with Sample Selection Correction

Matteo Picchio
CentER, ReflecT, Tilburg University and IZA

Chiara Mussida
Università Cattolica, Milan and Prometeia SpA, Bologna

Discussion Paper No. 4783
February 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4783
February 2010

ABSTRACT

Gender Wage Gap: A Semi-Parametric Approach with Sample Selection Correction*

Sizeable gender differences in employment rates are observed in many countries. Sample selection into the workforce might therefore be a relevant issue when estimating gender wage gaps. This paper proposes a new semi-parametric estimator of densities in the presence of covariates which incorporates sample selection. We describe a simulation algorithm to implement counterfactual comparisons of densities. The proposed methodology is used to investigate the gender wage gap in Italy. It is found that when sample selection is taken into account gender wage gap widens, especially at the bottom of the wage distribution. Explanations are offered for this empirical finding.

JEL Classification: C21, C41, J16, J31, J71

Keywords: gender wage gap, hazard function, sample selection, glass ceiling, sticky floor

Corresponding author:

Matteo Picchio
Department of Economics
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands
E-mail: m.picchio@uvt.nl

* We wish to thank Lorenzo Cappellari and Jan van Ours for their comments and suggestions.

1 Introduction

The literature on wage gap has highlighted significant gender differences in most industrialized nations.[1] This topic has been extensively studied with the aim of understanding whether the gender pay gap is explained by gender differences in the distribution of personal characteristics or due to a discriminatory remuneration of apparently identical characteristics. This is a relevant issue from the policy perspective and different implications are drawn if the source of the gender pay gap is found to be either the former or the latter.

Earlier work – see Altonji and Blank (1999) for a survey – focused on decomposing the gender pay gap at the mean of the wage distribution in a part due to differences in characteristics and in a part due to differences in the returns to those characteristics by using the procedure developed by Blinder (1973) and Oaxaca (1973). Later, the attention has shifted to the investigation of the degree to which gender pay gaps might vary across the wage distribution. Bonjour and Gerfin (2001) used 1991–1995 data for Switzerland and the hazard function (HF) approach suggested by Donald et al. (2000) to estimate distributions in the presence of covariates. They found that the gender wage gap is not constant throughout the wage distribution but it is larger at the bottom, even after controlling for gender differences in labour market characteristics. Albrecht et al. (2003), using the quantile regression (QR) decomposition developed by Machado and Mata (2005), showed that in 1998 the Swedish gender wage gap is instead larger at the top of the pay distribution.

Arulampalam et al. (2007) investigated this issue further for eleven European countries. They used harmonized data from the European Community Household Panel (ECHP) for the period 1995–2001. They found that in ten of the eleven countries there is "glass ceiling", i.e. larger gender pay gap at the top of the distribution, and there is evidence of "sticky floor", i.e. larger gender pay gap at the bottom of the distribution, only in Italy and Spain.[2] On the basis of the cross-country comparison of the results, they speculate as to why glass ceilings and sticky floors emerge, suggesting that differences in childcare provision and wage setting institutions are potential determinants.

Different lines of reasoning have been provided to explain why the gender wage gap exists and why it widens at the bottom and/or at the top of the wage distribution, even when the distribution of observed characteristics are the same across gender. It has been argued that women might be perceived to have or might have a smaller work force attachment (Booth and Francesconi, 2003), especially those at the bottom of the wage distribution. Hence, trade unions could be less likely to represent or could differentially represent the interests of female workers (Booth, 2009) and the anticipation of a more intermittent career might affect investments in human capital (Manning and Swaffield, 2008). Alternatively, women at the bottom might have a smaller bargaining power or be more subject to firms' market power

[1] See Weichselbaumer and Winter-Ebmer (2005) for an international comparison of gender wage gaps and Arulampalam et al. (2007) for a recent analysis of the gender pay gap in several European countries.

[2] The sticky floor and glass ceiling phenomena follow the definitions of Booth et al. (2003) and Albrecht et al. (2003), respectively.

than comparable men, as a result of social norms and family commitments whereby the man's career takes the precedence (Arulampalam et al., 2007). Bjerk (2008) theoretically clarifies instead the reason why we should expect gender inequality of opportunity with respect to hiring and promotion to the top jobs and, hence, glass ceilings. He shows that if women have a more intermittent career, they will have fewer opportunities to signal their skills and they will need more time to accumulate the number of positive signals required for promotion to the top jobs.[3]

Most of these explanations for the presence of the gender wage gap and its heterogeneity over the wage distribution hinge on gender differences in the work force attachment and, thereby, in the propensity to participate to the labour market. A sizeable gender gap in employment rates is observed in many countries and, therefore, sample selection into the workforce might be a relevant issue to assess gender wage gaps. Suppose, for instance, that those women that stay out of employment are those who would have got the lowest returns from work and that, in case of participation, these lower returns would have been different across the wage distribution. If so, we would underestimate and misrepresent the gender wage gap compared to the case in which men and women would have the same propensity to be at work. Moreover, as pointed out by Olivetti and Petrongolo (2008) and Albrecht et al. (2009), selection correction is fundamental for meaningful international comparisons of the gender wage gap, especially if gender gaps in employment rates are very different across countries.

Given that sample selection might be an important issue in this framework, it is surprising that, apart from Albrecht et al. (2009) who adjusted the gender wage gap for the selectivity into full-time work, there are no studies that correct the gender gap across the wage distribution for the selection due to work force participation. In the literature there is also lack of studies based on recent data (e.g. Arulampalam et al. (2007) used the ECHP survey which took place at the end of the 90s). In this paper we try to fill these gaps, from both the methodological and the empirical points of view. With regard to the former, we propose a new flexible estimator of distribution functions in the presence of covariates and sample selection. We describe a simulation algorithm to decompose the gender pay gap across the wage distribution into the effect due to different distributions of characteristics and to different returns to those characteristics. Empirically, the proposed approach is illustrated by investigating the recent gender wage gap in Italy, a country where the employment

[3]Other reasons for the glass ceilings can be found in the institutional setting. Albrecht et al. (2003) justify the glass ceiling phenomenon in Sweden on the basis of the parental leave and the daycare system which are likely to provide incentives to be in the work force but to discourage strong career commitment. Sociological and psychological factors might also favour men over women, in particular at the top of the wage distribution. Women might be less willing to bargain over offered wages, since more grateful for the job offer. Because of stereotypes, women might not be seen as potential leaders and, when they are, they might be evaluated less favourably. Women might be less risk adverse, less inclined to competition, and give priority to other job features over the wage. See Booth (2009) for a detailed discussion on the relationship between psychological and sociological factors and the gender pay gap and glass ceiling phenomena.

rate between men and women differs substantially.[4]

How does our approach compare with the methodological literature on gender pay gap? First, we develop a procedure that extends and generalizes the Oaxaca-Blinder decomposition of effects on mean wages to the entire wage distribution in the presence of sample selection. Second, our approach is very much in the spirit of Albrecht et al. (2009), who adapted the Machado-Mata (MM) procedure to take advantage of the Buchinsky (1998) selection correction method for QR.[5] However, in this paper, we operate within the HF framework proposed by Donald et al. (2000). We benefit therefore of a flexible method for density estimation in the presence of covariates that, differently from QR, avoids the potential inconsistency of predicting lower quantiles larger than higher quantiles for a given covariate vector. Lastly, we improve upon the HF approaches of Donald et al. (2000) and Bonjour and Gerfin (2001)[6] in that we correct for sample selection.

This paper is organized as follows. Section 2 discusses the econometric approach used to estimate distributions in the presence of covariates and sample selection and presents the simulation algorithms to recover marginal and counterfactual distributions. Section 3 illustrates our approach by analysing the gender pay gap in Italy on the basis of the European Union Statistics on Income and Living Conditions (EU-SILC) panel data. Section 4 concludes.

2 Modelling Wage Distributions in the Presence of Covariates and Sample Selection

The econometric analysis is designed keeping in mind that we desire an estimation strategy that i) is flexible in the way covariates affect the shape of the wage density function; ii) is flexible in the way sample selection is controlled for; iii) is tractable and allows to test for sample selection when estimating wage densities; iv) returns, for any set of individual characteristics, estimated probabilities in the interval $[0, 1]$ and a wage density function that integrates to one.

Different non/semi-parametric estimators have been proposed to model wage distributions in the presence of covariates. Firstly, DiNardo et al. (1996) adapted kernel density estimator to the case in which sample weights are attached to each observation. Coun-

---

[4]The Italian employment rate for the population aged 25–54 was 86.5% for men and 58.1% for women in 2005. These figures are available in the Internet at http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/-home.

[5]See also Olivetti and Petrongolo (2008) who analyse for EU countries and the US the gender wage gap at the median of the distribution by correcting for selection into the work force. Blundell et al. (2007) deal with nonrandom selection into employment by developing bounds around the wage distribution on the basis of restrictions motivated from economic theory.

[6]See also Behr and Pötter (2009), who analyse the differences in the wage distributions between the USA and Germany by way of the HF approach but less flexible than that suggested by Donald et al. (2000).

terfactual densities are estimated by "re-weighting" functions, which depend on different covariates distribution between populations. Fortin and Lemieux (1998) divided the wage support in small intervals and estimated the probability of being in each wage interval using an ordered response (probit) model. Once the model is estimated, counterfactual distributions can be predicted by playing with covariates and/or estimated coefficients of the ordered response model. Donald et al. (2000) and Bonjour and Gerfin (2001) considered wage distributions as if they were duration distributions and applied HF based estimation techniques. Since there is a one-to-one relationship between hazard and density functions, this approach consists in specifying and estimating a flexible wage hazard function, so as to recover the corresponding conditional wage distribution from the estimated parameters.[7] Finally, the most common method to estimate distributions in the presence of covariates, is QR (e.g. Buchinsky, 1994, 1998) followed by Machado and Mata's (2005) simulation algorithm to carry out counterfactual comparisons of densities.

The approach used here for modelling wage distributions in the presence of covariates and sample selection is built on Donald et al. (2000) and Bonjour and Gerfin (2001). Differently from them, we propose a semi-parametric method to correct the wage distribution estimates for endogenous selection process into the work force. Hence, this paper is related to the previous literature by extending Donald et al. (2000) and Bonjour and Gerfin (2001) to control for sample selection, along the same lines as Albrecht et al. (2009) adapted the Machado and Mata (2005) decomposition to account for selection into full-time employment.

In Subsection 2.1 the concept of wage hazard function is introduced. Subsection 2.2 focuses on specification issues and deals with the derivation of the likelihood function. Finally, Subsection 2.3 clarifies how counterfactual densities are derived once we have an estimate of the wage density in the presence of covariates and sample selection.

2.1 Hazard Function Based Approach

Consider a panel data setting where t is the time indicator, with $t = 1, 2, \ldots, T$.[8] We assume that all individual differences in the wage distribution can be characterized by observed characteristics x_t and an individual fixed-effect v. For covariate x_t and individual heterogeneity v there is a positive random variable W_t with associated probability and cumulative density functions $f(w_t|x_t, v)$ and $F(w_t|x_t, v)$. The hazard function is defined as

$$\theta(w_t|x_t, v) = \frac{f(w_t|x_t, v)}{1 - F(w_t|x_t, v)} \equiv \frac{f(w_t|x_t, v)}{S(w_t|x_t, v)}, \tag{1}$$

[7] The duration literature interpretation of the hazard function does not apply in this context. The interest relies rather in the impact of individual characteristics on hazard functions and, thereby, on density functions.

[8] We suppress the individual indicator for the sake of clarity.

where $S(w_t|x_t, v)$ is the survivor function, i.e. the probability of being paid at least a wage w_t. We specify the hazard function as

$$\theta(w_t|x_t, v) = h_t(w_t|x_t)v, \tag{2}$$

where h_t is the structural wage hazard function at time t. We impose now the following regularity conditions on h_t and v:

A1 $H_t(w_t|x_t) = \int_0^{w_t} h_t(s|x_t)ds < \infty$ is non-negative, differentiable, and strictly increasing, $\forall t = 1, 2, \ldots, T$;

A2 v is non-negative with distribution function G.

Under assumptions A1 and A2, Honoré (1993, Theorem 1) shows that, from panel data, the unobserved heterogeneity distribution G and the structural hazard functions h_t, $\forall t = 1, 2, \ldots, T$, are non-parametrically identified. We stress here the importance of non-parametric model identification. If the model is not identified without arbitrary parametric assumptions, parametric specification would play a crucial role in inference and the results would thereby be intrinsically suspect. In the empirical analysis that follows, we do not need to rely on parametric restrictions for identification purposes.

It is noteworthy to remark that while in the duration literature the hazard function is usually assumed to have a mixed proportional hazard (MPH) form, where the structural hazard function factors into separate functions of duration and covariates, here we do need it for identification and we will not impose it in the empirical specification. In this framework a MPH model would imply the strong restriction that covariates have the same impact on the baseline wage hazard function across all the wage support. As in Donald et al. (2000), we will allow the covariate effect to be different over the wage support, at different percentiles of the unconditional wage distribution. Moreover, assumption A2 does not require a particular mixing distribution[9] and the independence assumption between G and x_t is not needed for identification.[10]

2.2 Model Specification and Likelihood Function

An ideal way to avoid any kind of parametric assumption on the functional form of h_t would be to fully non-parametrically estimate it by using kernel-density estimation for each possible combination of covariate values. However, as soon as the number of covariates

[9]Note also that the structural wage hazard function h_t is not required to be the same over time and is in principle allowed to take a completely different shape when t changes. In the empirical application we are however going to put some restrictions on the way in which h_t varies over time for the sake of reducing the dimensionality of the problem: h_t will be allowed to change over time by the introduction of a time dummy among the set of regressors x_t.

[10]As pointed out by Honoré (1993), identification is attained even in the case in which the distribution of v depends on x_t.

increase or a continuous covariate is introduced, the cell-sizes become too small to do useful inference. So we are going to put some parametric restrictions on h_t in order to have a tractable but still flexible conditional density estimator. Following Donald et al. (2000) and Gritz and MaCurdy (1992), we partition the range of W into P intervals and we allow the covariate effects to vary over these intervals. The variation is with respect to a baseline wage hazard function $h^0(w_t)$ that, in order to avoid too strict parametric assumptions, is assumed to be piecewise constant. Formally, denote $\Omega_p = [w_L^p, w_H^p)$ the pth wage interval over which the covariate effect is assumed to be constant (i.e. proportional to the baseline wage hazard function), such that $\Omega_p \cap \Omega_q = 0$ for all $p \neq q$ and $\bigcup_{p=1}^{P} \Omega_p = [0, \infty)$. The structural hazard function is specified as follows

$$h_t(w_t|x_t) = h^0(w_t) \exp\left[x_t'\beta(w_t)\right] = h^0(w_t) \exp\Big[\sum_{p=1}^{P} \mathbb{1}(w_t \in \Omega_p) x_t'\beta^p\Big], \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

As said, $h^0(w_t)$ is the piecewise constant baseline wage hazard function. More in details we divided the wage support into J intervals $I_j = [w_{j-1}, w_j)$, where $j = 1, \ldots, J$, $w_0 < w_1 < \ldots < w_J$, $w_0 = 0$, and $w_J = \infty$. We choose the width of the wage baseline segments by fixing w_{J-1} to the 99^{th} percentile of the unconditional wage distribution and by dividing the wage support between w_0 and w_{J-1} in 70 equally spaced intervals. Our choice of the number of the baseline segments is somewhat arbitrary but it allows us to have a narrow segment width[11] and therefore suitable for flexibly approximating all possible density functions. This reduces the model selection problem to the choice of P (and, as we will see, to the number of points of support of the discrete distribution of the unobserved heterogeneity). In order to reduce the risk of over-fitting bias and to have a reasonable trade-off between model parsimony and fit to the data at hand, we choose P by looking at the Akaike Information Criterion (AIC). In turned out that $P = 8$ for men and $P = 10$ for women is the number of covariates segments with dividing points approximatively at the $[100\frac{1}{P}, 100\frac{2}{P}, \ldots, 100\frac{P-1}{P}]$ percentiles of the unconditional wage distribution.

To construct the contribution to the likelihood function of the wage w_t that lies in the baseline segment $[w_{j-1}, w_j)$, note that the probability of observing a wage in such a segment is

$$\begin{aligned} \Pr(w_{j-1} \leq W_t < w_j|x_t, v) &= S(w_{j-1}|x_t, v) - S(w_j|x_t, v) \\ &= L_{it}^w(w_t|x_t, v; \Theta^w), \end{aligned} \tag{4}$$

[11] The segment width is around 0.27€ for both men and women.

where Θ^w is the set of parameters to be estimated and

$$S(w_j|x_t, v) \equiv \exp\Big[-\int_0^{w_j} \theta(s|x_t, v)ds\Big]$$

is the wage survivor function, i.e. the probability of observing a wage at least as large as the upper limit of the j^{th} segment. Since the wage hazard function is piecewise constant and the covariate effect does not change within the baseline segments,[12] the survivor function reduces to $S(w_t|x_t, v) \equiv \prod_{s=0}^{j} \exp[-\theta(w_s|x_t, v)]$, $s \in \mathbb{N}_0$. Thereby, the difference in survivor functions in (4) is the contribution to the likelihood function of wages lying on the piecewise constant baseline segment $[w_{j-1}, w_j)$. If the wage value is top-coded, i.e. if it is larger than the 99^{th} percentile of the wage distribution, it is treated as a right censored observation. The corresponding contribution to the likelihood function is the probability that the wage is larger than the 99^{th} percentile, i.e.

$$L^w_{it}(w_t|x_t, v; \Theta^w) = S(w_{J-1=99^{\text{th}}}|x_t, v). \tag{5}$$

However, if we are interested in comparing wages of women to those of men, we have so far faced only a part of the story. Since the selection process into the work force is likely to be structurally different for women compared to men, it has to be taken into account when evaluating compositional issues otherwise preventing us from understanding the genuine nature of the gender wage gap. In other words, there are reasons to believe that the probability of being employed is related to the likelihood of being paid a certain wage. Assume for instance that women who decide not to participate are those who would get the lowest earnings from work and that such a relation is ignored. Then the observed gender wage gap across the distribution would be underestimated compared to the case of men and women having the same propensity to be at work.

Correcting for nonrandom sampling has been central to empirical studies in labour economics since the seminal work of Gronau (1974) and Heckman (1974). Parametric and semi/non-parametric methods have been proposed in the selection bias literature.[13] By way of a semi-parametric approach, we correct now the likelihood function for the participation probability.

Let us define y_{it} the scalar indicator variable denoting individual i's employment status at time t and z_{it} the set of explanatory variables explaining participation. The probability model for work participation is empirically specified using a discrete mixture logit model:

$$y_{it} = \mathbb{1}[z'_{it}\delta + \varepsilon_i + u_{it} > 0] \tag{6}$$

where ε_i is unobserved heterogeneity and u_{it} is the idiosyncratic error term which is assumed

[12] Note however that they are allowed to vary across baseline segments.

[13] See, e.g., Vella (1998) for a survey on methods to estimate models with sample selection bias.

to have a logistic distribution, denoted by Λ.

The contribution to the likelihood function of individual i will therefore be given by

$$L_i(y_i, w_i|x_i, z_i, v_i, \varepsilon_i; \Theta) = \prod_{t=1}^{T} [\Lambda(z'_{it}\delta + \varepsilon_i) L^w_{it}(w_{it}|x_{it}, v_i; \Theta^w)]^{y_{it}} [1 - \Lambda(z'_{it}\delta + \varepsilon_i)]^{1-y_{it}}.$$

As a matter of fact, we cannot use directly this density to estimate Θ because v_i and ε_i are not observed. For each individual i, the likelihood contribution $\mathscr{L}_i$ is obtained by integrating out the unobserved heterogeneity under the components:

$$\mathscr{L}_i = \int_{\Re_+ \times \Re} L_i(y_i, w_i|x_i, z_i, v, \varepsilon; \Theta) dG(v, \varepsilon). \tag{7}$$

The log-likelihood function sums the logarithm of this expression over all the individuals (employed and nonemployed) in the sample.[14]

In order to avoid too strict parametric assumptions on the joint distribution of v and ε, it is approximated by means of a bivariate discrete distribution (Heckman and Singer, 1984) with a fixed number of support points, which have unknown locations and probability masses. We assume that (v, ε) has four probability points. The corresponding probability masses are defined as follows:

$$\begin{aligned} p_1 &\equiv \Pr(v = v_1, \varepsilon = \varepsilon_1) \qquad && p_2 \equiv \Pr(v = v_2, \varepsilon = \varepsilon_1) \\ p_3 &\equiv \Pr(v = v_1, \varepsilon = \varepsilon_2) \qquad && p_4 \equiv \Pr(v = v_2, \varepsilon = \varepsilon_2) = 1 - p_1 - p_2 - p_3. \end{aligned}$$

In this case, four points of support and three probability masses are to be estimated. The probabilities associated to the mass points are specified as logistic transforms:

$$p_m = \frac{\exp(\lambda_m)}{\sum_{r=1}^{4} \exp(\lambda_r)} \quad \text{with} \quad \lambda_4 = 0.$$

Note that v and ε are independent if and only if $p_1 p_4 = p_2 p_3$ (see Van den Berg et al., 1994; Van den Berg and Lindeboom, 1998), making it easy to test for sample selection.

2.3 Marginal Densities and Decomposition of Distribution Functions

In the application below, we start by estimating a simpler model that does not take into account sample selection. By doing so, we can compare our decomposition technique to a

[14] In principle, given the error term assumption on ε_{it}, the panel data framework, and the integration over the distribution of unobserved heterogeneity, identification is achieved without exclusion restrictions. However, in the empirical application we use some exclusion restrictions in line with those often used in the labour supply literature (see, e.g., Mroz, 1987): dummy indicators for the number of children by different age groups and the number of members in the household.

standard MM decomposition and we can see what the empirical findings would have been if sample selection had not been taken into account. Subsequently, we estimate the model with sample selection as described above. Therefore, in this subsection, we present two simulation algorithms: the first one to derive the marginal densities implied by the conditional model that omits sample selection; the second one that controls for sample selection.

2.3.1 Marginal and Counterfactual Densities without Sample Selection

Once the male and female hazard functions that fully characterize the conditional wage densities are estimated ignoring sample selection, we generate the corresponding marginal densities by means of the following simulation algorithm, which is run once for men and once for women:

1. Draw a vector of parameter estimates assuming normality around the point estimates $\widehat{\Theta}$ with a variance-covariance matrix equal to the estimated one.
2. Generate N individuals by drawing N times a vector x of covariates from the empirical distribution of covariates.[15]
3. For each baseline wage interval $[w_{j-1}, w_j)$ with $j = 1, \ldots, 70$, we compute the predicted conditional probability of getting a wage within that interval from the following theoretical conditional probability:

$$\Pr(w_{j-1} \leq W_t < w_j | W_t \geq w_{j-1}, x_{it}) = 1 - \exp[-\theta(w_j | x_{it})].$$

   For $i = 1, \ldots, N$ and for $j = 1, \ldots, J$ we draw κ_{ij} from a standard uniform distribution. Starting from $j = 1$, if the predicted probability $1 - \exp[-\widehat{\theta}(w_1|x_{it})] > \kappa_{i1}$, individual i is assigned a random wage in $[w_0, w_1)$. If not, repeat the lottery for the next wage baseline segment, i.e. for $j = 2$. The lottery for individual i is stopped when inequality $1 - \exp[-\widehat{\theta}(w_j|x_{it})] > \kappa_{ij}$ is satisfied.[16]
4. Repeat steps 1 to 3 $R = 999$ times.

With marginal densities in hands for men and women, the basic idea of the decomposition of the gender pay gap is based on recovering the counterfactual density that would have prevailed if female employees had the same distribution of characteristics as male employees, that is

$$\int_{x \in \chi} f_F(w|x; \Theta_F) dF_M(x), \tag{8}$$

[15] As in Machado and Mata (2005), in this step a random sample of the covariates from an appropriate distribution is drawn, allowing to “integrate x out” and get a sample from the marginal.

[16] Note that, by doing so, we can only simulate an interval on which the wage is located. Given that each interval is quite small and equal to 0.27€, a point value is randomly assigned between the lower and upper bound of the interval without losing too much in terms of precision.

where the subscripts F and M stand for female and male, respectively. The counterfactual marginal density in (8) is obtained by repeating the simulation algorithm for women with step 2 revised so that the covariate vectors are drawn from the empirical covariate distribution of male employees.

The basic idea of the procedure is an alternative way of extending Oaxaca's (1973) decomposition of effects on mean wages to the entire wage distribution and is in line with the underlying idea of Machado and Mata (2005). In the next subsection, we explain how, as in Albrecht et al. (2009), we incorporated sample selection due to work participation in the gender pay gap decomposition.

2.3.2 Marginal and Counterfactual Densities with Sample Selection

Let A denote all women (men) and B the women (men) who actually work.[17] Define W_A the counterfactual random variable indicating the wage that a woman would earn if she worked. The conditional wage density is characterized by a parameter vector Θ_A^w, i.e. $f(W_A|x_A, v_A; \Theta_A^w)$. The problem is that W_A is not observed for individuals who do not work. We can only observe the realization of the random variable W_B, which represents the wage earned by those women who actually work. By way of the sample selection correction presented in Subsection 2.2, we recover the true shape of the wage distribution on the basis of a "re-weighted" density for those that in the data are employed, i.e.

$$p(Y_B|z_B, \varepsilon_B; \Theta_A^y) f(W_B|x_B, v_B; \Theta_A^w),$$

where the weights are the probabilities of being at work, Y_B is the participation indicator, and v_B and ε_B are allowed to be correlated.

The extension of the previous algorithm to generate marginal densities in the presence of sample selection proceeds as follows:

1. Draw a vector of parameter estimates assuming normality around the point estimates $\widehat{\Theta}_A$ with a variance-covariance matrix equal to the estimated one.
2. Generate N individuals by drawing N times: i) a vector x_A of covariates from the empirical distribution of covariates among all women (men);[18] ii) unobserved characteristics from the estimated distribution of unobserved heterogeneity $\widehat{G}$.
3. Per each baseline wage interval $[w_{j-1}, w_j)$ with $j = 1, \ldots, 70$, we compute the predicted conditional probability of getting a wage within that interval from the following theoretical conditional probability:

$$\Pr(w_{j-1} \leq W_t < w_j | W_t \geq w_{j-1}, x_{Ait}, v_{Ai}) = 1 - \exp[-\theta(w_j | x_{Ait}, v_{Ai})].$$

[17] The notation in this subsection is inspired by the one in Albrecht et al. (2009).

[18] Work specific covariates are drawn from the empirical distribution conditional on work participation.

For $i = 1, \dots, N$ and for $j = 1, \dots, J$ we draw κ_{ij} from a standard uniform distribution. Starting from $j = 1$, if the predicted probability $1 - \exp[-\widehat{\theta}(w_1|x_{Ait}, \widehat{v}_{Ai})] > \kappa_{i1}$, individual i is assigned a random wage in $[w_0, w_1)$. If not, repeat the lottery for the next wage baseline segment. The lottery for individual i is stopped when inequality $1 - \exp[-\widehat{\theta}(w_j|x_{Ait}, \widehat{v}_{Ai})] > \kappa_{ij}$ is satisfied.

4. Repeat steps 1 to 3 $R = 999$ times.

This procedure simulates (wo)men's wage densities that would have prevailed if all (wo)men had been at work. This is done by taking into account that the distribution of individual characteristics across all the population is not the same as the one across people at work.

As in Subsection 2.3.1, the counterfactual female wage density that we would expect if women had the same distribution of characteristics as men, i.e.

$$\int_{x_A \in \chi_A} \int_{\Re_+} f_F(w_A|x_A, v_A; \Theta_{AF}) dG(v_A) dF_M(x_A), \tag{9}$$

is obtained by repeating the above simulation algorithm for women with step 2 revised so that the covariate vectors are drawn from the empirical covariate distribution of all men.

3 Application: Gender Wage Gap in Italy

3.1 Data

Our data are from the EU-SILC panel. It is a rotating panel survey based on harmonised methodology and definitions across most members of the European Union.[19] The topics covered by the survey are living conditions, income, social exclusion, housing, work, demography, and education. We select data for Italy, where the survey is conducted on a yearly basis by the National Institute of Statistics (ISTAT) under the coordination of Eurostat. The longitudinal EU-SILC data comprise a number of individual records that range from 31,000 to about 43,000 per year over the time window 2004–2006.

Since information on earnings is retrospectively collected (information on income at time t is collected at time $t + 1$), we lose one time period (2006). In order to avoid to get mixed up with early retirement and education enrolment issues, we exclude from our analysis individuals under the age of 25 years and over the age of 55 years. We also drop individuals that are in army, self-employed, inactive, or with missing values for some variables used in the econometric analysis. Finally, individuals lying in the first or last percentiles of the wage or working hours distributions are excluded from the sample. Considering both the non-employed and the employed, we are left with 10,841 female observations and 8,699

[19] See Eurostat (2004) for further and technical details about the EU-SILC data.

male observations over the period 2004–2005. The total number of wage observations is 7,612(5,910) for (wo)men.

Almost all active men, 87.5%, are at work, whilst 54.5% of the active women are employed. The definitions of employment and non-employment do not match the ILO definition. In the EU-SILC questionnaire, the respondents are indeed asked to self-define the main economic status in the current year.[20]

The dependent variable is the net hourly wage in the respondent's main job, deflated to 2004 prices. The deflator is the Consumer Price Index (CPI), gathered by ISTAT.[21] Table 1 reports summary statistics of the net hourly wage disaggregated by gender in 2004 and 2005. Men are paid per hour and on average more than women, €9.4 against €9. Figure 1 plots the kernel estimate of the wage density for men and women. It shows that the hourly wage distribution is positively skewed both for men and women. By contrasting the male and female kernel density estimates, we get a visual inspection of the gender pay gap, which does not seem to be constant over the wage distribution: it is well perceived at the bottom and in the middle of the wage support, but it disappears at the top of the distribution. Table 2 sheds further light on this raw evidence: the raw gender pay gap goes from almost 10% at the bottom of the distribution to 5% at the median and to zero at the 95^{th} percentile of the wage distribution.

Table 1: Summary Statistics of Net Hourly Wage by Gender in Italy, 2004–2005

	Men		Women		Overall	
Hourly wage (€)[(a)]	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
2004	9.333	3.420	8.898	3.495	9.142	3.459
2005	9.413	3.583	9.013	3.590	9.239	3.591
Overall	9.380	3.517	8.966	3.552	9.199	3.538

[(a)] Wages are in constant prices (2004 prices). They are deflated by using the Consumer Price Index (CPI), gathered by ISTAT.

Table 3 reports summary statistics of the variables used in the econometric analysis computed on the subsample of the employed and disaggregated by year. Table 4 displays instead descriptive statistics over the total population in our sample.

We distinguish between three age groups: young workers (25–35 years old), middle aged workers (35–45 years old), and older workers (45–55 years old). In the empirical analysis below, we do not include among the covariates any measure of experience. In a first stage,

[20]The question PL030 is the variable containing information on the self-defined economic status. People are asked whether they are working, unemployed, students, in retirement, disabled, in military service, or fulfilling domestic task.

[21]The net hourly wage is computed starting from the employees' yearly net cash income, variable PY010N, and using the number of months at work, obtained from variables PL210A-PL210L, and the number of hours usually worked per week, PL060. Considering that on average there are 4.345 weeks in a month, the hourly wage is therefore computed as follows: $w = \text{PY010N}/(\# \text{ of months at work} \times \text{PL060} \times 4.345)$.

Figure 1: Kernel Density Estimates for Net Hourly Wage

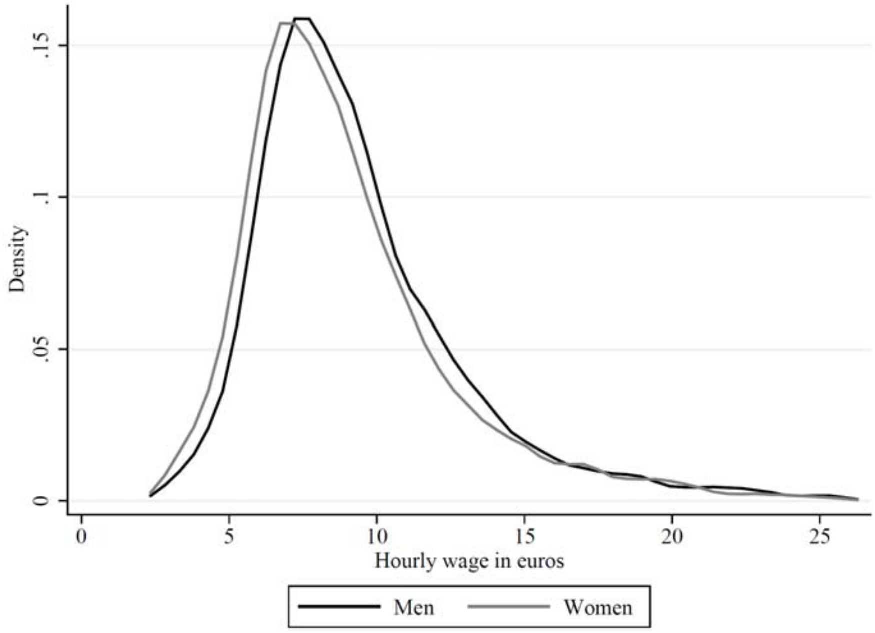


Table 2: Raw Gender Wage Gap in Italy, 2004–2005

	2004	2005	Overall
Proportion of women	.439	.436	.437
Mean	.048	.043	.045
5^{th} percentile	.095	.094	.097
10^{th} percentile	.084	.072	.076
25^{th} percentile	.081	.059	.066
50^{th} percentile	.057	.049	.053
75^{th} percentile	.034	.048	.040
90^{th} percentile	.016	.014	.013
95^{th} percentile	.022	-.009	.001

Note: The raw wage gap is measured as the difference between the log male hourly wage and the log female hourly wage.

we did include in the model specification real experience and/or potential experience, but they turned out to be very correlated with age and they did not generate any improvement in terms of likelihood. They were thereby removed from the set of covariates.

Table 3: Summary Statistics by Gender and Year for Employees

	2004				2005			
	Male		Female		Male		Female	
	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev
Age								
$[25, 35)$	.297	.457	.317	.465	.286	.452	.294	.456
$[35, 45)$	.373	.484	.383	.486	.366	.482	.386	.487
$[45, 55]$	.329	.470	.300	.458	.348	.476	.320	.467
Education[(a)]								
None, elementary, or lower secondary	.447	.497	.294	.456	.431	.495	.290	.454
Upper secondary	.393	.488	.448	.497	.403	.491	.442	.497
Post secondary or tertiary	.160	.367	.257	.437	.166	.372	.268	.443
Area of residence								
North	.521	.500	.582	.493	.493	.500	.548	.498
Centre	.229	.420	.236	.425	.238	.426	.252	.434
South	.250	.433	.182	.386	.270	.444	.200	.400
Bad health[(b)]	.247	.431	.263	.440	.232	.422	.258	.438
Married	.668	.471	.645	.479	.661	.474	.629	.483
City[(c)]	.340	.474	.352	.478	.324	.468	.346	.476
Part time worker	.040	.195	.296	.456	.039	.195	.304	.460
White-collar high-skilled worker[(d)]	.302	.459	.435	.496	.303	.460	.435	.496
White-collar low-skilled worker[(d)]	.219	.414	.330	.470	.215	.411	.346	.476
Blue-collar high-skilled worker[(d)]	.230	.421	.069	.253	.241	.428	.068	.251
Blue-collar low-skilled worker[(d)]	.248	.432	.167	.373	.240	.427	.152	.359
Temporary contract	.083	.276	.136	.343	.075	.263	.128	.334
Observations	3,103		2,432		4,509		3,478	

[(a)] Educational dummy indicators refer to the highest and successfully completed educational attainment of a person. The educational classification used to build these indicators is the ISCED-97.

[(b)] “Bad health” is a dummy indicator based on self-perceived health. It is equal to one if the individual declares that her health is in a fair, rather bad, or bad conditions. It is equal to zero, if the answer is either good or rather good condition.

[(c)] “City” is a dummy indicator equal to one if the individual lives in a densely populated area, i.e. an area with a density superior to 500 inhabitants per square kilometre and a total population larger than 50,000 inhabitants.

[(d)] These occupational indicators are built on the basis of the International Standard Classification of Occupation (ISCO 88). Taking the one-digit ISCO categories, the white-collar high skilled indicator corresponds to categories 1 up to 3; the white-collar low skilled indicator corresponds to categories 4 and 5; the blue-collar high skilled indicator corresponds to categories 6 and 7; the blue-collar low skilled indicator corresponds to categories 8 and 9.

Educational variables are defined according to UNESCO's International Standard Classification of Education (ISCED). The EU-SILC distinguishes between education completed to the lower secondary stage (ISCED 0-2), upper secondary education(ISCED 3), and post-secondary or tertiary education (ISCED 5-7). In our sample working women are more educated than men, and this tendency is even stronger in 2005, when 27% of working women had at least a post secondary education attainment against 17% of working men. The discrepancy almost disappears on the full sample of employed and non-employed individuals: 18% of women has at least a post secondary education attainment against 16% of men. This shows that women with low education attainment have a lower propensity to be at work than their male counterpart.

Table 4: Summary Statistics by Gender for the Whole Population (Employed and Non-Employed)

	Male		Female	
	Mean	St. Dev	Mean	St. Dev
Age				
$[25, 35)$	.309	.462	.281	.450
$[35, 45)$	.363	.481	.379	.485
$[45, 55]$	.329	.470	.340	.474
Education[(a)]				
None, elementary, or lower secondary	.458	.498	.441	.497
Upper secondary	.383	.486	.375	.484
Post secondary or tertiary	.159	.366	.184	.388
Area of residence				
North	.477	.499	.462	.499
Centre	.225	.418	.227	.419
South	.298	.457	.311	.463
Bad health[(b)]	.246	.430	.286	.452
Married	.624	.484	.716	.451
City[(c)]	.334	.472	.341	.474
Part time worker[(d)]	.040	.361	.301	.459
White-collar high-skilled worker[(d)]	.303	.460	.435	.496
White-collar low-skilled worker[(d)]	.217	.412	.339	.473
Blue-collar high-skilled worker[(d)]	.237	.425	.068	.252
Blue-collar low-skilled worker[(d)]	.244	.429	.158	.365
Temporary worker[(d)]	.078	.268	.131	.338
Number of household members	3.312	1.210	3.379	1.158
Number of children by age				
$[0, 3)$ years old	.109	.326	.104	.318
$[3, 6)$ years old	.127	.354	.126	.350
$[6, 12)$ years old	.251	.521	.275	.538
Employed	.875	.331	.545	.498
Time dummy for 2004	.409	.492	.407	.491
Observations	8,699		10,841	

[(a), (b), (c)] See footnotes [(a)], [(b)], [(c)] of Table 3.

[(d)] These are variables observed only for the employed. The corresponding number of observations is not therefore given by the figures at the bottom of the table but equal to 7,612 for men and 5,910 for women.

Three dummy variables for the geographical area of residence (North, Centre, and South) are included in the model specification. Around one half of the sample lives in the North of Italy and the rest is equally distributed in the Centre and South. A dummy indicator accounting for self-perceived health tries to capture the effect of health status (even if subjective or perceived) on wages and propensity to work. Indicators for marital status and living in a densely populated area are included in the model, as likely to affect earnings and job opportunities.

Finally, a set of covariates are used to capture job heterogeneity. We included dummy indicators for part-time and temporary jobs. Women are more frequently employed in these jobs than men. Using the International Standard Classification of Occupations (ISCO-88), we built a set of binary indicators aimed at capturing task heterogeneity.[22] Taking the one-digit ISCO categories, the white-collar high skilled indicator corresponds to categories 1 up to 3; the white-collar low skilled indicator corresponds to categories 4 and 5; the blue-collar high skilled indicator corresponds to categories 6 and 7; the blue-collar low skilled indicator corresponds to categories 8 and 9. Women are more likely to be white collar workers. Only 23% of the women at work are blue collar workers, whereas 48% of male employees are blue collar workers.

3.2 Estimation Results

This subsection reports and comments on the impact of the regressors on the male and female wage hazard functions and, thereby, on the wage distribution functions. As mentioned in Subsection 2.2, the impact of covariates on the shape of the wage density functions is flexibly modelled, so that it can be different over the wage support. In other words, analogously to QR, the covariates can have different effects at different percentiles of the wage distribution. Hence, Tables 5 and 6 report the effects of the covariates at selected percentiles for men and women, respectively.

Three points are worthy being remarked before moving through the estimation results. First, Tables 5 and 6 display selection-corrected estimation results.[23] The Log-likelihood Ratio (LR) tests reported at the end of these tables indicate that the null hypothesis of no sample selection can be rejected quite confidently, especially for women. Second, the estimated coefficients inform us about the covariate impact on the wage hazard rate: individual characteristics that have a negative effect on the wage hazard rate reduce the probability of getting a low wage. Individuals holding these characteristics are therefore more likely to get a higher wage with respect to the reference group. Third, a LR test for constant covariates coefficients over the wage support largely rejects the null hypothesis both for men and women: a standard proportional hazard specification like in Behr and Pötter (2009) would have been too restrictive.

[22] See Dumont (2006) for a discussion on the ability of such indicators in capturing skills levels.

[23] Coefficients estimates without sample selection are not reported in the paper but available upon request.

Education plays a significant role in explaining the male and female wage distributions and, as expected, higher education attainments are associated with higher wages. This is especially true at the 25^{th} percentile of the wage distribution, both for men and women. The relevance of education decreases at the top of the wage distributions, especially for men.

Table 5: Coefficient Estimates of the Covariates with Correction for Selectivity at Selected Percentiles of the Wage Distribution – Men

Percentiles	10^{th}		25^{th}		50^{th}		75^{th}		90^{th}	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Education – Reference: None, elementary, or lower secondary										
Upper secondary	-.447***	.070	-.654***	.078	-.299***	.089	-.119	.075	.068	.089
Post secondary or tertiary	-.895***	.127	-1.149***	.126	-.918***	.132	-.608***	.103	-.430***	.102
Age – Reference: $[45, 55]$										
$[25, 35)$	1.115***	.092	1.367***	.101	.936***	.113	.630***	.098	.217*	.114
$[35, 45)$	.300***	.088	.605***	.084	.404***	.092	.440***	.071	.073	.067
Occupation – Reference: High skilled & white collar										
Low skilled & white collar	.203*	.106	.579***	.109	.826***	.112	.463***	.079	.268***	.082
High skilled & blue collar	.837***	.100	1.540***	.108	1.571***	.126	.304***	.112	.258**	.123
Low skilled & blue collar	.695***	.101	1.227***	.106	1.630***	.117	.377***	.099	.187	.124
Area of residence – Reference: North										
Centre	.626***	.083	.373***	.089	.286***	.093	.044	.079	.020	.082
South	1.173***	.077	.540***	.084	.231**	.097	-.209**	.084	.055	.077
Bad health	.100	.080	.135*	.079	.083	.090	.190**	.078	.013	.082
Married	-.698***	.070	-.594***	.079	-.096	.096	-.128*	.077	.139*	.077
Temporary contract	1.013***	.086	.723***	.131	.316*	.174	-.109	.205	-.137	.174
City	-.102	.071	-.279***	.074	-.216**	.084	-.019	.068	-.032	.066
Part-time	-.499***	.161	-1.518***	.246	-1.493***	.249	-1.383***	.205	-.520***	.128
Year 2004	.036	.074	.005	.075	.082	.085	.094	.081	.016	.081
Log-likelihood										-29,020.0
Observations										5,770
LR test for sample selection, H_0: $p_1 p_4 = p_2 p_3$										$\chi^2(1)$=5.8, p-value=.016
LR test against PH assumption, H_0: $\beta^p = \beta$										$\chi^2(105)$=866.4, p-value=.000

Notes: * Significant at the 10% level; ** significant at the 5% level; *** significant at the 1% level.

Getting older increases the probability of getting high wages. This emphasizes a disadvantage for younger men and women to be located at the top of the wage distribution. However, work experience is likely to play a role in partly explaining the age effect. As mentioned in Section 3.1, we do not include work experience among the covariates: since work experience is strongly correlated with age, we were not able to improve in terms of the likelihood by including it among the set of regressors. If follows that the decomposition results presented in the next subsection are not sensitive to the exclusion of work experience from the model specification.

Moving on to task heterogeneity, we note that high skilled and white collar occupations are associated with higher wages, both for men and women. In term of significance, these skill-level indicators seem to be more relevant for men than for women. For women, tasks heterogeneity appears to lose any explanatory power at the top of the wage distribution and the standard errors associated to the task indicators tend to explode, especially if compared to the male counterpart. This is likely due to a different composition of the female work force at the top jobs, which is relatively more homogeneous and richer in high skilled and white collar workers than the male work force.

Table 6: Coefficient Estimates of the Covariates with Correction for Selectivity at Selected Percentiles of the Wage Distribution – Women

Percentiles	10^{th}		25^{th}		50^{th}		75^{th}		90^{th}	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Education – Reference: None, elementary, or lower secondary										
Upper secondary	-.790***	.102	-1.025***	.099	-.680***	.127	-.418***	.119	.203	.213
Post secondary or tertiary	-1.059***	.140	-1.569***	.133	-1.322***	.167	-.887***	.135	-.502**	.217
Age – Reference: $[45, 55]$										
$[25, 35)$	.919***	.126	1.298***	.112	1.131***	.140	.531***	.119	-.166	.139
$[35, 45)$	.351***	.120	.445***	.109	.511***	.129	.182*	.097	.135	.097
Occupation – Reference: High skilled & white collar										
Low skilled & white collar	.805***	.121	.739***	.101	1.058***	.122	.382***	.092	-.219	.132
High skilled & blue collar	1.382***	.159	1.691***	.159	1.594***	.219	-.264	.333	.059	.388
Low skilled & blue collar	1.533***	.132	1.563***	.126	1.526***	.175	.176	.186	.050	.294
Area of residence – Reference: North										
Centre	.538***	.110	.280***	.095	.188	.121	-.045	.098	.104	.102
South	1.414***	.105	.460***	.112	.056	.147	-.270**	.117	-.140	.101
Bad health	-.096	.107	.031	.097	.070	.115	.161*	.096	.026	.100
Married	-.273***	.096	-.162*	.084	-.196*	.107	-.073	.087	-.216**	.100
Temporary contract	.457***	.106	.191	.117	.188	.157	-.355**	.168	-.020	.137
City	-.154	.101	-.189**	.086	-.214**	.107	-.028	.085	-.162*	.088
Part-time	.039	.098	-.826***	.097	-.985***	.128	-1.117***	.103	-.072	.089
Year 2004	.060	.099	-.026	.092	-.052	.107	-.020	.093	.110	.109
Log-likelihood										-25,654.7
Observations										7,156
LR test for sample selection, H_0: $p_1 p_4 = p_2 p_3$										$\chi^2(1)$=12.3, p-value=.000
LR test against PH assumption, H_0: $\beta^p = \beta$										$\chi^2(135)$=859.6, p-value=.000

Notes: * Significant at the 10% level; ** significant at the 5% level; *** significant at the 1% level.

We find geographical differentials in the distribution of wages, characterized by relevant pay disadvantages in the Centre and especially in the South of Italy. The geographical differentials represent a structural feature of the Italian labour market. This is also reflected in other economic indicators, for instance, the sizeable geographical gaps in employment rates.[24] Nevertheless, the geographical pay gaps seem to disappear at the top of the wage distribution, both for men and women. The top jobs are equally paid across the country and this might reflect a different composition of the work force across areas with respect to public/private employment. However, this information is not in the data and this issue cannot be investigated further.

With regard to the remaining covariates, subjective (or perceived) health status, indicators for marital status, and living in a densely populated area do not affect so much earnings.

[24]The Italian unemployment rate for the population aged 25-54 was 78.8% in the North, 75.3% in the Centre, and 56.8% in the South of Italy in 2005. These figures are available in the Internet at http://epp.eurostat.ec.-europa.eu/portal/page/portal/eurostat/home.

Being employed with a temporary contract is associated with low wages at the bottom (for men and women) and the middle (for men) of the wage distribution. Part-time jobs seem to be better paid on hourly basis, especially for men. This is likely to be due to the fact that men are more willing than women to get a part-time job only when it can ensure a satisfactory level of income. Finally, the reference year of our estimates is 2005 and a time indicator for 2004 is instead included in the model specification. This indicator variable has always a small and insignificant impact on the wage hazard rate, both for men and women. It suggests that the shape of the wage distributions did not change significantly between 2004 and 2005.

In general, the estimation results emphasize the relevance of education, age, and geographical partition of residence to explain the shape of male and female wage distributions. These characteristics, together with the gender gap, are classified by the literature as key dimensions of heterogeneity of the Italian labour market.[25].

3.3 Decomposition of the Gender Wage Gap

We present now the decomposition of the gender pay gap, so as to account for gender differences in labour market characteristics and work force attachment. We proceed in two steps. First, we look at the estimated gender wage gap decomposition without correction for selectivity. At this stage, the marginal and counterfactual densities are recovered by following the simulation algorithm described in Subsection 2.3.1; we also compare the empirical findings from our HF based decomposition to those that would have been obtained from a standard MM decomposition based on QR. Second, we look at the decomposition with selection correction, by way of the simulation algorithm described in Subsection 2.3.2. This permits to assess whether and how the gender wage gap, given the same distribution of labour market characteristics, changes after taking into account selection into the work force. Figure 2 reports the decomposition at the first step. Figure 3 displays the decomposition at the second step, i.e. with selection correction.

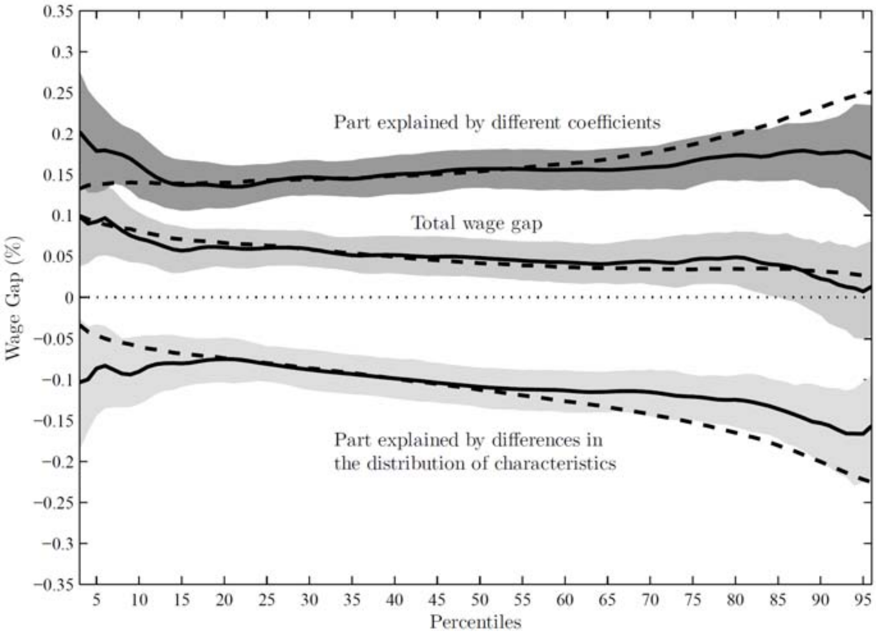
The central lines of Figure 2 plot the total wage gaps after recovering the marginal densities from the corresponding estimated conditional wage distributions. The dotted line is the outcome of a QR based procedure, whilst the solid one is the outcome of an HF based approach. The total wage gap is computed by taking the difference between the log male wage and the log female wage at each percentile of the corresponding estimated marginal distribution. The two lines are very close to each other and they overlap for most of the wage support. The lines at the top of Figure 2 plot instead the wage gap after that gender differences in observed characteristics are taken into account. They show the gender wage gap that would have prevailed if women had had the same distribution of observed characteristics as men, but still paid as women.

Three points are worthy being noted. First, the two techniques for decomposing wage

[25]The same kind of heterogeneity, indeed, has been found for Italian unemployment by, among others, Bertola and Garibaldi (2003)

gaps display very similar results and it would not matter so much whether the MM decomposition after QR or the HF based decomposition were used. Second, the pay gap, once gender differences in the distribution of labour market characteristics are taken into account, becomes larger across all the wage distribution. As it can be seen in Table 7, men are paid 15.7% more than women at the median of the distribution and 17.3%(18%) at the top(bottom) of the distribution.[26] Third, it follows that if women had had the same returns as men but their own distribution of observed characteristics, they would have received higher wages across all the wage distribution, especially at the top (see the lines at the bottom of Figure 2).

Figure 2: Gender Wage Gap Decomposition without Selection Correction: QR approach (Dotted Line) and HF Approach (Solid Line)



Note: The grey areas are the 95% confidence intervals, computed by bootstrapping the results 999 times.

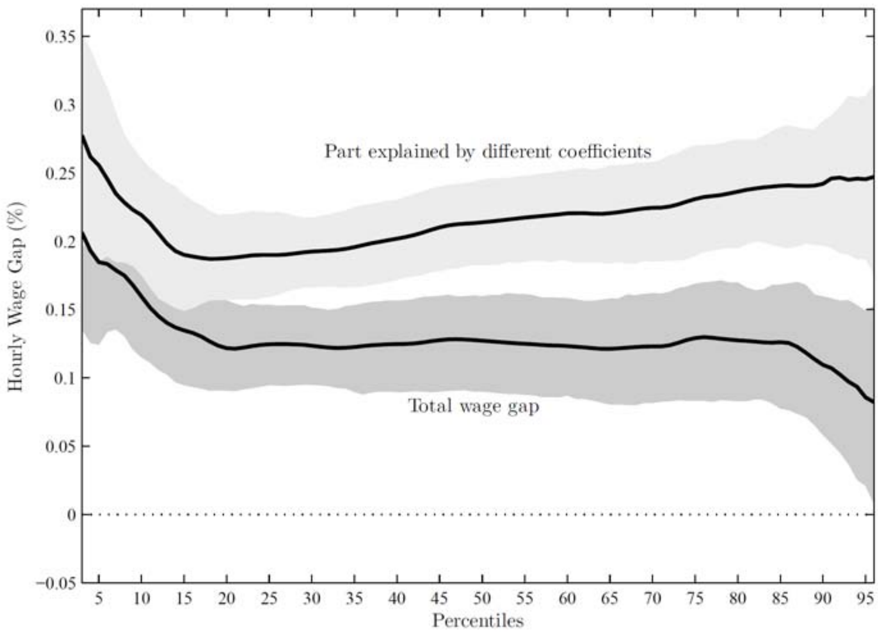
We next investigate the effect of sample selection into the work force on the estimated wage distributions, on the counterfactual distributions, and, thereby, on the gender wage gaps. Figure 3 depicts the total wage gap (lower curve) and the gap if men and women had the same distribution of observed characteristics (upper curve). Once we take into account

[26]Note that the bootstrapped confidence intervals become wider at the tails of the wage distribution. Since the piecewise constant wage baseline hazards are specified by dividing the wage support in many equally spaced intervals, at the tails there are fewer observations per each piece. This implies that the wage density functions and, thereby, the wage gaps are estimated less precisely at the extremities of the wage support.

sample selection, the total wage gap becomes wider across all the distribution. There is evidence therefore that women are overall and strongly more positively selected into the work force than men: women that stay out of employment are those that would have got the lowest returns from the market work with an higher probability than that of men.

The upper curve of Figure 3 shows that, even if men and women had identical distributions of characteristics and after taking into account selection into the work force, the gender pay gap persists and is even wider. The last three columns of Table 7 report the point estimates and confidence intervals of the gender pay gap at selected percentiles of the wage distribution. The pay gap is always significantly different from zero and is characterized by a U-shape. It has maximum at the 5^{th} percentile, where men's wages are 25.6% higher than women's. Then, it decreases and at the 25^{th} percentile it is equal to 19%. Thereafter, it monotonically increases and, at the 95^{th} percentile, it is equal to 24.5%.

Figure 3: Gender Wage Gap Decomposition with Selection Correction



Note: The grey areas are the 95% confidence intervals, computed by bootstrapping the results 999 times.

The U-shape of the gender wage gap suggests that in Italy there are both sticky floors and glass ceilings. The glass ceiling and, especially, sticky floor phenomena become more evident once sample selection is taken into account. Table 8 contrasts measures of sticky floors and glass ceilings, computed from the ratios between the gender gaps at different percentiles, with and without sample selection correction. The gender gap at the 5^{th} percentile is 1.14 larger than the one at the median when sample selection is not taken into account.

Table 7: Wage Gap if Women had Male Characteristics but Women's Returns, with and without Selection Correction

	Without selection correction			With selection correction		
Percentiles	Wage gap (%)	95% confidence interval		Wage gap (%)	95% confidence interval	
5	.179	.128	.240	.256	.185	.326
10	.161	.121	.201	.219	.175	.261
25	.141	.119	.163	.190	.159	.221
50	.157	.134	.180	.214	.183	.245
75	.167	.135	.195	.231	.189	.270
90	.175	.134	.215	.242	.197	.288
95	.173	.113	.235	.245	.187	.306

Table 8: Sticky Floor and Glass Ceiling Measures

	Without selection correction	With selection correction
Sticky floor measured by		
5/50[(a)]	1.142	1.197
5/25[(b)]	1.268	1.346
Glass ceiling measured by		
95/50[(c)]	1.108	1.149
95/75[(d)]	1.042	1.065

[(a)] The sticky floor is measured as the ratio between the 5^{th} percentile wage gap and the 50^{th} percentile wage gap.
[(b)] The sticky floor is measured as the ratio between the 5^{th} percentile wage gap and the 25^{th} percentile wage gap.
[(c)] The glass ceiling is measured as the ratio between the 95^{th} percentile wage gap and the 50^{th} percentile wage gap.
[(d)] The glass ceiling is measured as the ratio between the 95^{th} percentile wage gap and the 75^{th} percentile wage gap.

This figure becomes 1.35 when sample selection is controlled for. Such evidence suggests the existence of obstacles for women in particular at the bottom of the wage distribution.

3.4 Goodness-of-Fit

Given the mixture of parametric and non-parametric assumptions we impose to estimate conditional wage distributions, it might be important to assess the ability of the model in providing quantitative predictions of the statistics of primary interest. We therefore check the goodness-of-fit of the model by contrasting empirical aspects of the data with those predicted by model simulations. The simulation algorithm is described in Appendix A-2. We focus on three aspects: work participation, male and female marginal wage distributions, and the gender wage gap over these distributions.

Table 9 displays, by gender, actual and simulated work participation rates and selected percentiles of the marginal wage distribution. Figure 4 focuses instead on the actual and predicted raw gender wage gap over the wage distribution. The goodness-of-fit can easily be checked by verifying whether the empirical statistics lie within the confidence intervals of the simulated ones.

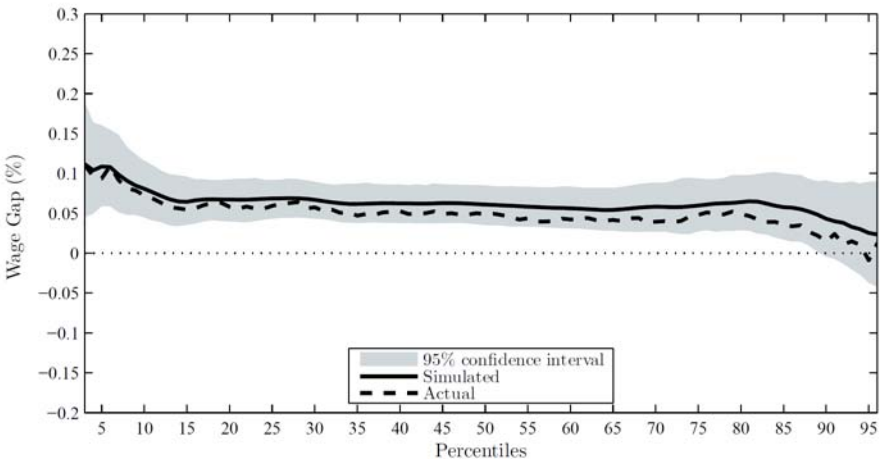
Work participation is perfectly predicted by our models, both for men and women. The model fits also very well the empirical percentiles of the wage distributions. For women,

Table 9: Goodness-of-Fit: Work Participation and Wage Marginal Distribution Conditional on Participating by Gender

	Women				Men			
	Actual	Simulated	95% confidence interval		Actual	Simulated	95% confidence interval	
Work participation	**.541**	.532	.511	.552	**.877**	.865	.847	.879
Percentiles of wage distribution in €								
5	**4.71**	4.75	4.56	4.94	**5.18**	5.30	5.12	5.48
10	**5.48**	5.50	5.33	5.67	**5.89**	5.97	5.85	6.07
15	**5.99**	5.99	5.83	6.12	**6.33**	6.39	6.25	6.51
20	**6.32**	6.31	6.19	6.45	**6.69**	6.75	6.64	6.86
25	**6.63**	6.63	6.51	6.75	**7.03**	7.09	6.98	7.21
30	**6.92**	6.92	6.79	7.05	**7.33**	7.39	7.29	7.49
35	**7.24**	7.21	7.08	7.35	**7.59**	7.67	7.55	7.79
40	**7.53**	7.51	7.37	7.65	**7.94**	7.99	7.86	8.11
45	**7.87**	7.82	7.66	7.97	**8.26**	8.33	8.19	8.47
50	**8.21**	8.15	7.99	8.31	**8.62**	8.66	8.52	8.79
55	**8.59**	8.49	8.32	8.68	**8.97**	9.00	8.86	9.14
60	**8.99**	8.87	8.69	9.06	**9.38**	9.38	9.22	9.54
65	**9.40**	9.27	9.08	9.46	**9.80**	9.79	9.63	9.97
70	**9.90**	9.71	9.49	9.95	**10.30**	10.30	10.09	10.49
75	**10.50**	10.29	10.02	10.54	**11.01**	10.92	10.67	11.17
80	**11.21**	10.92	10.64	11.21	**11.77**	11.64	11.37	11.90
85	12.21	11.82	11.44	12.20	**12.65**	12.52	12.23	12.86
90	**13.67**	13.16	12.66	13.67	**13.87**	13.74	13.39	14.13
95	16.62	15.66	14.92	16.47	**16.48**	16.06	15.45	16.72

Note: Actual frequencies lying in the 95% confidence interval of the simulated frequencies are in bold.

Figure 4: Goodness-of-Fit: Gender Wage Gap



Note: The grey area is the 95% confidence intervals, computed by bootstrapping the results 999 times.

the model tends to slightly underpredict high percentiles of the wage distribution. However, the size of the misalignment is small and the ability of the model to predict female wage distribution is nonetheless very good. The ability of the model in replicating the data is confirmed by Figure 4, which shows that the empirical wage gap over the wage distribution always lies within the 95% confidence interval of the simulated one and is furthermore very close to the point predictions.

4 Conclusions

A sizeable gender gap in employment rates is observed in many countries. Non-random sample selection into the workforce might be a potentially important issue when evaluating gender wage gaps. This is especially true if those women that stay out of employment are the ones who would get the lowest returns from work. If so, we would underestimate the gender wage gap compared to the case in which women and men had the same propensity to be at work.

On the basis of these considerations, this paper is aimed at offering and implementing a flexible estimator of distribution functions in the presence of covariates and sample selection. We have operated within the HF framework proposed by Donald et al. (2000) and improved upon them, in that we incorporate in the model sample selection. We have described a simulation algorithm to decompose the gender pay gap across the wage distribution into the effect due to different distributions of individual characteristics and to different returns on those characteristics, once gender differences in work force selection are taken into account.

Our approach has been illustrated by investigating the gender wage gap in Italy, a country persistently characterized by a large gender gap in terms of work participation. The empirical analysis, based on 2004–2005 EU-SILC panel data, shows that women are paid less than men. Not controlling for sample selection yields gender wage gaps across the wage distributions that are in line with those recently found for Italy, e.g. Arulampalam et al. (2007) and Gannon et al. (2007).

We find that gender wage gap widens when we control for non-random sample selection into the work force, even if the distribution of observed characteristics is the same across genders. There is therefore evidence that women are more strongly and positively selected into the work force than men: women that stay out of employment are those that would have got the lowest returns from the labour market with an higher probability than that of comparable men. It is also found that, when we control for sample selection, the gender wage gap becomes wider at the top, where it jumps from 17.3% to 24.5%, and especially at the bottom of the wage distribution, where it moves from 17.9% to 25.6%.

We now speculate on the possible explanations that can be drawn from the fact that in Italy the sticky floor phenomenon becomes more evident when sample selection is taken into account. This is consistent with explanations based on the interaction between work and family. Italy has a relatively not well developed system of work/family reconciliation policies,[27] characterized by short maternity leave, poorly paid optional parental leave, and limited part-time jobs opportunities (Del Boca, 2002; Del Boca et al., 2005; Del Boca and Pasqua, 2005). This institutional setting makes Italian women less attached to the labour market and more likely to commit to family than to career than comparable men, especially at the bottom of the wage distribution where financial constraints are more likely to be binding. This also suggests that family-friendly policies might increase the female attachment to the work force, especially at the bottom of the wage distribution.

Note that similar reasoning might also explain why the gender pay gap widens, even if less evidently, also at the top of the distribution. It might be that the absence of a structured child care makes Italian women at the top jobs more oriented towards less demanding jobs than the male counterpart.

Other factors might however explain our findings. There could be, for instance, sociological barrier for women to be at work, like social norms according to which man's career has the priority. Then, only those women with relatively high wage characteristics might be able to overtake the threshold imposed by the social norm and enter the work force, whilst those that have relatively low wage characteristics remain out of the work force.

Appendix

A-1 Further Estimation Results: Marginal Wage Distributions and Work Participation

The attributes of the estimated marginal wage distributions for men and women are reported in Table A-1. The values of these distributions at selected percentiles confirm the existence of a gender gap: the figures for male component are higher than the ones for females through the overall pay distribution.

Scale, skewness, and kurtosis give insights on the scale and shape of the wage distributions. More precisely, the scale is a measure of the "statistical dispersion" of the probability density function (pdf). If it is large, the distribution will be more spread out; if it is small, it will be more concentrated. Both the scales of the estimated distributions are small, suggesting more concentrated pdf between the 25^{th} and the 75^{th} percentiles of wage distributions.

The skewness is instead a measure of the asymmetry of the pdf. A negative skewness implies longer left tail and the mass of the distribution is concentrated on the right of the figure. A positive skewness implies longer right tail and the mass of the distribution is concentrated on the left of the figure. The distributions in Table A-1 are therefore left-skewed, but the values are however low.

Finally, the kurtosis is a measure of the peakedness of the pdf. The Normal distribution has kurtosis equal to 0. The Laplace distribution has excess kurtosis equal to 3. In our case, both for

[27] See the OECD's (2001, § 4) work/family reconciliation composite index for a comparison between Italy and other OECD countries.

men and women, the excess kurtosis is very close to 2, about the same peakedness as the hyperbolic secant distribution.

Table A-1: Attributes of the Estimated Marginal Wage Distributions for Men and Women

	Men	Women
5^{th} percentile	5.269	4.381
10^{th} percentile	5.950	5.073
25^{th} percentile	7.100	6.269
50^{th} percentile	8.682	7.646
75^{th} percentile	10.977	9.648
90^{th} percentile	13.803	12.372
95^{th} percentile	16.127	14.804
Scale[(a)]	.446	.442
Skewness[(b)]	-.746	-.748
Kurtosis[(c)]	2.026	2.161

[(a)] Denote Q_p the wage corresponding to quantile p, then scale is defined as $(Q_{0.75} - Q_{0.25})/Q_{0.50}$.
[(b)] Denote Q_p the wage corresponding to quantile p, then skewness is defined as $(Q_{0.75} - Q_{0.25} - 2Q_{0.50})/(Q_{0.75} + Q_{0.25})$.
[(c)] Denote Q_p the wage corresponding to quantile p, then kurtosis is defined as $(Q_{0.90} - Q_{0.10})/(Q_{0.75} - Q_{0.25})$.

Figure A-1 plots the marginal wage distributions described so far. Here, we can see a clear wage gap across the distribution, apart from the top. The gap between the wage distribution for men and women, indeed, becomes narrow from the 90^{th} percentile onwards.

Finally, Table A-2 reports the estimation results of the discrete mixture logit model for the probability of being employed. The results are in line with the expectations. The probability of being employed is increasing with the education attainment and is lower in the Centre and especially in the South of Italy. Family structure and married status have opposite effects on work participation between men and women: married (wo)men have a (lower)higher probability of being employed and work participation is increasing in the number of children for men and decreasing in the number of children for women. The number of household members reduces the employment probability both for men and women. Men living in a highly urbanized area are less likely to work, whereas whether or not a woman lives in an urbanized area has little impact on the propensity to be at work. It is interesting to note that young men are less likely to be employed than old men, whilst young women are more likely to be at work than old women.

A-2 The Simulation with regard to the Goodness-of-Fit

Predictions are computed by implementing the following simulation algorithm for men and women:

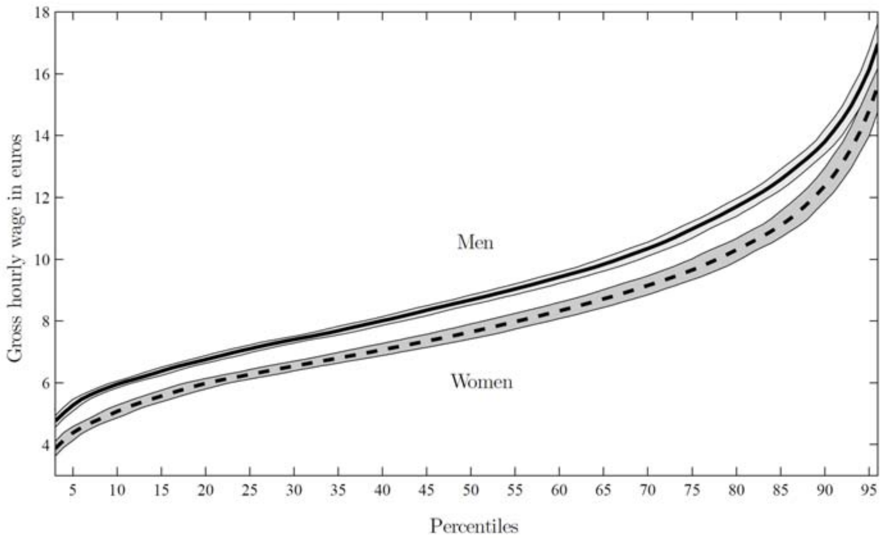
1. Draw a vector of parameter estimates assuming normality around the point estimates $\widehat{\Theta}$ with a variance-covariance matrix equal to the estimated one.
2. Generate N individuals by drawing N times: i) a vector z of covariates from the empirical distribution of covariates; ii) unobserved characteristics from the estimated distribution of unobserved heterogeneity $\widehat{G}$.

Table A-2: Estimation Results of the Logit Model in (6) for Labour Market Participation

	Men		Women	
	Coeff.	S.E.	Coeff.	S.E.
Bad health	-1.562***	.203	-.205	.133
Married	3.576***	.347	-1.499***	.154
City	-.938***	.180	-.192	.117
Education – Reference: None, elementary, or lower secondary				
Upper secondary	1.739***	.209	4.310***	.353
Post secondary or tertiary	2.440***	.306	5.006***	.371
Age – Reference: $[45, 55]$				
$[25, 35)$	-.837***	.268	.409***	.162
$[35, 45)$	-.395	.262	.198	.143
Area of residence – Reference: North				
Centre	-.801***	.258	-1.475***	.203
South	-3.372***	.300	-4.723***	.274
Number of children by age				
$[0-3)$ years old	.937***	.286	-.907***	.181
$[3-6)$ years old	.301	.297	-.435***	.179
$[6-12)$ years old	.434**	.217	-.190	.125
Number of household members	-.232***	.076	-.294***	.055
Year 2004	.310	.241	.427***	.170
Individual heterogeneity distribution				
Location of the support points				
ε_1	5.857***	.458	-3.434***	.436
ε_2	-2.458***	.500	6.441***	.351
Logistic transform coefficients of the probability masses				
λ_1	-.031	.050	-.737**	.330
λ_2	-2.617***	.520	-.357***	.125
λ_3	-1.301***	.147	.130	.106
λ_4	.000	–	.000	–
Resulting probability masses				
	$p_1 = .419$		$p_1 = .145$	
	$p_2 = .032$		$p_2 = .211$	
	$p_3 = .118$		$p_3 = .343$	
	$p_4 = .431$		$p_4 = .301$	

Notes: * Significant at the 10% level; ** significant at the 5% level; *** significant at the 1% level.

Figure A-1: Estimated Marginal Distributions for Men and Women



Note: The grey areas are the 95% confidence intervals, computed by bootstrapping the results 999 times.

3. Simulate work participation by a lottery based on the predicted probability of work participation. More in details, for $i = 1, \ldots, N$ we draw η_i from a standard uniform distribution. If $\Lambda(z'_{it}\widehat{\delta}_F + \widehat{\varepsilon}_i) > \eta_i$, individual i participates. Otherwise, she does not participate and her simulation is halted.
4. Simulate the wage distribution over the subsample of participating individuals. For each baseline wage interval $[w_{j-1}, w_j)$ with $j = 1, \ldots, 70$, we compute the predicted conditional probability of getting a wage within that interval from the following theoretical conditional probability:
$$\Pr(w_{j-1} \leq W_t < w_j | W_t \geq w_{j-1}, x_{it}, v_i) = 1 - \exp[-\theta(w_j | x_{it}, v_i)].$$
For $i = 1, \ldots, N$ and for $j = 1, \ldots, J$ we draw κ_{ij} from a standard uniform distribution. Starting from $j = 1$, if the predicted probability $1 - \exp[-\widehat{\theta}(w_1 | x_{it}, \widehat{v}_i)] > \kappa_{i1}$, individual i is assigned a random wage in $[w_0, w_1)$. If not, repeat the lottery for the next wage baseline segment. The lottery for individual i is stopped when the inequality is satisfied.
5. Repeat steps 1 to 4 $R = 999$ times.

References

Albrecht, J., A. Bjorklund, and S. Vroman, "Is There a Glass Ceiling in Sweden?," *Labour Economics*, 2003, *21* (1), 145–177.

— , **A. van Vuuren, and S. Vroman**, "Counterfactual Distributions with Sample Selection Adjustments: Econometric Theory and an Application to the Netherlands," *Labour Economics*, 2009, *16* (4), 383–396.

Altonji, J.G. and R.M. Blank, "Race and Gender in the Labor Market," in O.C. Ashenfelter and D. Card, eds., *Handbook of Labor Economics, Volume 3C*, Amsterdam: Elsevier Science, 1999, chapter 48, pp. 3143–3259.

Arulampalam, W., A.L. Booth, and M.L. Bryan, "Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap acroos the Wages Distribution," *Industrial and Labor Relations Review*, 2007, *60* (2), 163–186.

Behr, A. and U. Pötter, "Analysing Wage Differences between the USA and Germany Using Proportional Hazards Models," *Labour*, 2 2009, *23*, 319–347.

Bertola, G. and P. Garibaldi, "The Structure and History of Italian Unemployment," 2003. CESifo Working Paper Series No. 907.

Bjerk, D., "Glass Ceilings or Sticky Floors? Statistical Discrimination in a Dynamic Model of Hiring and Promotion," *Economic Journal*, 2008, *118* (530), 961–982.

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

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition using Bounds," *Econometrica*, 2007, *75* (2), 323–363.

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

Booth, A.L., "Gender and Competition," *Labour Economics*, 2009, *16* (6), 599–606.

— **and M. Francesconi**, "Union Coverage and Non-Standard Work in Britain," *Oxford Economic Papers*, 2003, *55* (3), 383–416.

— , **M. Francesconi, and J. Frank**, "A Sticky Floors Model for Promotion, Pay, and Gender," *European Economic Review*, 2003, *47* (2), 295–322.

Buchinsky, M., "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 1994, *62* (2), 405–458.

— , "The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach," *Journal of Applied Econometrics*, 1998, *12* (1), 1–30.

Del Boca, D., "The Effect of Child Care and Part Time Opportunities on Participation and Fertility Decisions in Italy," *Journal of Population Economics*, 2002, *15* (3), 549–573.

— **and S. Pasqua**, "Labor Supply and Fertility in Europe and the US," in T. Boeri, D. Del Boca, and C. Pissarides, eds., *Women at Work: An Economic Perspective*, Oxford: Oxford University Press, 2005, chapter 8, pp. 125–153.

— , — , **and C. Pronzato**, "Fertility and Employment in Italy, France, and the UK," *Labour*, 2005, *19* (Special Issue), 51–77.

DiNardo, J., N.M. Fortin, and T. Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 1996, *64* (5), 1001–1044.

Donald, S.G., D.A. Green, and H.J. Paarsch, "Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates," *Review of Economic Studies*, 2000, *67* (4), 609–633.

Dumont, M., "The Reliability–or Lack Thereof–of Data on Skills," *Economics Letters*, 2006, *93* (3), 348–353.

Eurostat, *Description of Target Variables: Cross-sectional and Longitudinal, Doc. EU-SILC 065/2004* 2004.

Fortin, N.M. and T. Lemieux, "Rank Regressions, Wage Distributions, and the Gender Gap," *Journal of Human Resources*, 1998, *33* (3), 610–643.

Gannon, B., R. Plasman, F. Rycx, and I. Tojerow, "Inter-Industry Wage Differentials and the Gender Wage Gap: Evidence from European Countries," *Economic and Social Review*, 2007, *38* (1), 135–155.

Gritz, R.M. and T. MaCurdy, "Unemployment Compensation and Episodes of Nonemployment," *Empirical Economics*, 1992, *17* (1), 183–204.

Gronau, R., "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 1974, *82* (6), 1119–1143.

Heckman, J.J., "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 1974, *42* (4), 679–694.

— **and B. Singer**, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 1984, *52* (2), 271–320.

Honoré, B.E., "Identification Results for Duration Models with Multiple Spells," *Review of Economic Studies*, 1993, *60* (1), 241–246.

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

Manning, A. and J. Swaffield, "The gender gap in the early-career wage growth," *The Economic Journal*, 2008, *118* (530), 983–1024.

Mroz, T.A., "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 1987, *55* (4), 765–799.

Oaxaca, R., "Male-Female Differentials in Urban Labor Markets," *International Economic Review*, 1973, *14* (3), 673–709.

OECD, *Employment Outlook*, Paris: OECD, 2001.

Olivetti, C. and B. Petrongolo, "Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps," *Journal of Labor Economics*, 2008, *26* (4), 621–654.

Van den Berg, G.J. and M. Lindeboom, "Attrition in Panel Survey Data and the Estimation of Multi-State Labor Market Models," *Journal of Human Resources*, 1998, *33* (2), 458–478.

— , — , and G. Ridder, "Attrition in Longitudinal Panel Data and the Empirical Analysis of Dynamic Labour Market Behaviour," *Journal of Applied Econometrics*, 1994, *9* (4), 421–435.

Vella, F., "Estimatimg Models with Sample Selection Bias: A Survey," *Journal of Human Resources*, 1998, *33* (1), 127–169.

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