Decomposing wage distributions on a large data set

– a quantile regression approach

April 22, 2014

Preliminary version

Karsten Albæk, SFI
Lars Brink Thomsen, SFI

Abstract: This paper presents and implements a procedure that makes it possible to decompose wage distributions on large data sets, which is infeasible with the standard Machado-Mata procedure. We replace bootstrap sampling in the Machado-Mata procedure with "non-replacement subsampling", which is more suitable for large data sets such as the linked employer-employee data applied in this paper. Decompositions shows that about half of the gender wage gap can be attributed to segregation in occupations, industries, workplaces and job-cells, and the other half to differences in reward to observable characteristics (coefficients). Quantile analysis finds that most of the glass ceiling in the Danish labour market is reflected in segregation. Detailed decompositions show that differences in the constant terms (the unexplained part of the gender wage gap) account for more than the whole gender wage gap at the lower end of the wage distribution. The unexplained part of the gender wage gap in the lower end of the wage distribution is larger and quantitatively more important than the glass ceiling in the higher end of the wage distribution.

Thanks to comments from Colin Green, Bo Honoré, Mona Larsen, the participants in the Nordic Econometric Meeting in Bergen, June 2013, the 3th Linked Employer-Employee Data Workshop in Lisbon, June 2013, and the European Meeting of the Econometric Society in Gothenburg, August 2013. A previous version of the content of this paper constituted the last part of a manuscript "Occupational segregation and the gender wage gap – an analysis on linked employer-employee data", which was presented at these conferences. Financial support from the Danish Council for Independent Research, Social Sciences is acknowledged.

Keywords: Gender wage gap, Machado-Mata procedure, segregation, glass ceiling, linked employer-employee data.
JEL classification: J31

SFI – The Danish National Centre for Social Research
Herluf Trolles Gade 11
DK-1053 Copenhagen K.

1. Introduction

Decomposition of wage distributions is an important topic in research areas of considerable policy interest. A leading example is the role of supply, demand and policy interventions in the labour market for the development of wage and income inequality. Another leading example is the role of characteristics and rewards to characteristics for the difference in wage and income between males and females.

This paper investigates the gender wage gap over the wage distribution by quantile regression. We decompose the gender wage gap over the wage distribution by constructing counterfactual wage distributions from female coefficients and male distributions of characteristics. The data is a linked employer-employee data set encompassing most workers in the Danish labour market for 2007. The number of observations is more than one million and the standard Machado and Mata (2005) decomposition procedure is not feasable on such a large data set, see e.g. Fortin et al. (2011).

To solve this problem, this paper presents and implements a procedure that makes it possible to decompose wage distributions on large data sets. The idea of the procedure is to replace the bootstrap sampling (which is sampling with replacement) in the Machado-Mata procedure with a sampling procedure that is suitable for large data sets. This sampling scheme is known as "non-replacement subsampling" and is an alternative to the bootstrap, see Horowitz (2001).

The empirical analysis in the paper includes an assessment of the role of segregation for the gender wage gap. Linked employer-employee data are necessary to construct measures of segregation such as the share of female workers in establishments and job-cells (occupations within establishments). Analysis on linked employer-employee data typically yields the result that segregation plays an important role for wage formation, see Bayard et al. (2003), Gupta and Rothstein (2005) and our companion paper, Albæk and Thomsen (2014).

However, the lack of a procedure to decompose wage distributions on linked employer-employee data has empeded the analysis of the relation between segregation and the gender wage gap over the wage distribution. The methological contribution of this paper makes such an analysis possible.

A recent strand of literature analyses to what extent women face a glass ceiling in the labour market in the sense that the wage gap increases throughout the wage distribution and accelerates in the upper tail. Such glass ceilings are found in the seminal contribution, Albrecht et al. (2003), and in Arulampalam et al. (2007) for 11 European countries. Our quantile analysis confirms the existence of a glass ceiling in the Danish labour market. However, most of the glass ceiling is reflected in segregation in the sense that the magnitude of the glass ceiling is moderate when segregation measures are included in the analysis.

The basis of the decompositions is counterfactual wage distributions that show the distribution of wages if males were remunerated as females. The gender wage gap is subsequently decomposed in the difference between the male wage distribution and the counterfactual distribution, which is the component of the gender wage gap due to differences in coefficients (the 'wage structure' effect), whilst the difference between the counterfactual distribution and the female wage distribution is the component due to differences in characteristics (the 'composition' effect). We first perform aggregate decompositions, where all female coefficients enter the calculations. Then we perform detailed decompositions where coefficients to groups of variables enter the calculations. The detailed decompositions enable us to assess the importance of differences in reward to human capital, the wage penalty associated with segregation and the unexplained part of the gender wage gap. Quantile based decompositions provides a natural way of performing detailed decompositions according to Fortin et al. (2011).

The decomposition procedure of this paper is not confined to the analysis of gender wage differentials but is applicable to other topics. The procedure can also be applied in the analysis of the development of wage inequality and, more generally, on other topics and type of data, where the Machado-Mata procedure is not feasable because of the magnitude of the data sets.

The remainder of the paper is organized as follows. Section 2 presents the data used in the study. Section 3 gives the estimates of gender wage gap in quantile regressions with coefficients to the covariates restricted to be equal for males and females. Section 4 reports results for separate quantile regressions for males and females. Section 5 presents the procedure for decomposing wage distributions on large data sets and applies the procedure in aggregate decompositions, where the whole set of coefficients and characteristics enter the calculations. Section 6 makes disaggregate decompositions where subsets of coefficients enter the decomposition procedure. Section 7 concludes.

2. The data

The data is a linked employer-employee data set encompassing most workers in the Danish labour market for 2007. The matched data are obtained from Statistics Denmark and consist of information from several administrative registers.

The wage information stems from records of wages for individual workers from the private, the state and the local government sectors. In the private sector, the wage register includes companies with 10 or more full-time employees, whilst companies with fewer than 10 employees are exempt from reporting. Enterprises in the agriculture and fishing sector are also exempt from reporting. Some companies not required to report have nonetheless reported, and the wage information from these companies is included in the statistics. In the public sector, all employees are included in the statistics, except for categories such as military conscripts, some temporary teachers and student assistants.

The wage statistics cover employees only when the employment relation lasts more than one month and when the average weekly working hours is least 8 hours. Furthermore, the wage register includes only employees on ‘ordinary’ conditions. Various minor groups are thus omitted from the register (e.g. employees paid at an unusually low rate because of a disability). Included in the statistics, however, are employees for whom the employer receives an employment subsidy from the government. This paper uses a wage measure which includes holiday allowance, payments to pension schemes, fringe benefits and irregular payments but not payment for overtime or absences.

For each employee, in addition to pay, firms report industry and occupation. The classification scheme for occupation is the ‘International Standard Classification of Occupations’ (ISCO, or, more precisely, a Danish variant, DISCO). The classification contains nine major categories (whilst there are actually ten major ISCO groups, group zero, military, is omitted from the analyses). We apply these nine categories of major occupations in the regression analysis. The most detailed level of registration is the 6-digit level. For each of the 6-digit occupations, we calculate the proportion of female workers. We carry out these calculations separately for the private and public sectors for each 6-digit occupation. Furthermore, we calculate the share of female workers in each industry at the 5-digit level, in each establishment and in each 6-digit occupation within establishments, the so called ‘job-cells’.

The paper includes various other variables of relevance for explaining wage differentials. To control for education, we include the length of the education in years calculated from the normal education length for the employees’ highest completed educational level. We use an approximate measure of the individual employees’ actual work experience, namely the number of years the employee has been in the labour market, calculated from contributions to a pension scheme. Further, we include a number of other variables in the analysis such as industry and public sector employment.

For the present purpose we confine the analysis to the employer-employee observation with the longest duration during the year. We exclude observations with unknown occupations and the occupational categories ‘pilots’ and ‘air traffic controllers’ (due to lack of credible information on length of education). Furthermore, we exclude observations with missing values of the variables. Finally, we exclude employees in 6-digit occupations with fewer than 20 workers.

The number of observations in the sample is 1.029.904. We perform the share calculations for 789 6-digit occupations, 541 industrial categories, 22,154 establishments and 152,320 job-cells.

Descriptive statistics for the sample are reported in Table 1. On average, women earn 10.2 per cent less than men.[1] The average share of females in the sample is 46 per cent.

Table 1 around here.

Women are slightly better educated than men, they have 1.5 years less of experience in the labour market, and been employed in their present firm for about the same number of years as men. About half of the women in the sample are employed in the public sector whilst only one out of five men is public employees. Women are more likely to live in the capital (the Copenhagen metropolitan area) than men.

The average share of females in 6-digit occupations is 67 per cent for females (the average of the share of females in the 6-digit occupation that females belongs to) and 28 per cent for males (the average of the share of females in the 6-digit occupation that males belongs to), and the difference on 39 per cent indicates a substantial segregation in the labour market. We also calculate the share of females in industrial categories, with a somewhat smaller difference on 26 per cent between the average share of females for females and males as the result. The difference between the average share of females for female employees and for male employees in establishments is 32 per cent. We also categorize the workforce in each establishment according to 6-digit occupations and calculate the share of females for each of these 'job-cells'. Table 1 shows that females on average work in job-cells with 76 per cent females whilst males work in job-cells with only 20 per cent females, yielding a difference on 56 per cent.

The figures for the nine major occupational groups show that women are underrepresented in the two top groups (1. managers and 2. professionals), they are overrepresented in the three middle groups (3. technicians and associate professionals, 4. clerical support and 5. service and sales) but underrepresented in four lowest groups (6. skilled agricultural, forestry and fishery workers, 7. craft and related trades workers, 8. plant and machine operators, and assemblers, and 9. elementary occupations).

[1] We adopt the convention that a difference of, for example 0.102 log points is stated as a percentage difference.

3. Restricted quantile regressions

This section analyses the gender wage gap over the wage distribution. First, we display how the wage gap varies over the wage distribution; then we perform quantile regression. In this section the gender wage gap is measured by the coefficient to the female dummy in the regressions and the coefficients for the covariates are restricted to be the same for men and women. The next section presents results for separate regressions for men and women.

Figure 1 shows the gender gap at each percentile of the wage distribution. For example to obtain the wage gap in the first percentile, we calculate the average wage at the first percentile in the wage distribution for men, then we calculate the average wage in the first percentile in the wage distribution for women, and then we take the difference. This procedure is repeated for all percentiles up to the 99 per cent percentile. The differences are plotted in Figure 1 as the curve denoted by 'Raw gap' (the explanation of the rest of the curves follows).

Figure 1 around here

Figure 1 shows that the gender gap is small at the bottom of the wage distribution and very large at the upper part. Furthermore, the gender gap increases steadily throughout the wage distribution, tending to accelerate in the upper percentiles. The upper horizontal line in Figure 1 represents the average gender gap over the wage distribution of 10.2 per cent. The tendency of acceleration of the magnitude of the wage gap at the upper quantiles implies that the curve denoted 'Raw gap' crosses the horizontal line around the 60th percentile.

The steady increase and the acceleration of the wage gap in the upper percentiles are properties shared with an analogous distribution on Swedish data for 1992 (see Albrecht et al. (2003), Figure 2, although there are minor differences). The acceleration of the gender gap in upper quantiles appears more pronounced in the Danish labour market than in other European countries (see Arulampalam et al. (2007), Table 2 and Figure 1 (b), according to which only the Netherlands – out of 11 European countries – has a larger difference in the gender gap than Denmark between the 90th and the 50th quantile).

According to Figure 1, men earn less than women below the 5 per cent percentile in the wage distribution. That is, at the very lowest percentiles the gender wage gap is negative in the Danish labour market, a phenomenon that does not appear in the previous literature on the gender wage gap over the wage distribution.

The curve ‘Raw gap’ does not only display the gender wage gap at each percentile of the wage distribution but also the confidence interval of the wage gap at each percentile. The large number of observations implies that the confidence intervals are small and the development of the gender wage gap over the wage distribution is thus statistical significant. The rest of the curves in Figure 1 also display the confidence intervals for the wage gap at each percentile (and the same is the case for the curves in Figure 2).

We proceed with an analysis of how the gender gap varies with observable characteristics over the wage distribution. The method is quantile regressions, which traces the relation between log wage rates, w, and regressors, x, at different quantiles, θ, of the wage distribution. The quantile regression model assumes that the conditional quantile of w, q_θ, is linear in x, $q_\theta = x\beta(\theta)$, see Koenker and Bassett (1978). The vector of coefficients $\beta(\theta)$ is estimated by solving the following programming problem

$$\min_{\beta(\theta)} \left\{ \sum_{i:w_i \geq x_i\beta(\theta)} \theta|w_i - x_i\beta(\theta)| + \sum_{i:w_i < x_i\beta(\theta)} (1-\theta)|w_i - x_i\beta(\theta)| \right\}. \qquad (1)$$

Whilst ordinary least squares (OLS) estimates the impact of various covariates as gender, schooling, etc. on average wage rates, quantile regression estimates the impact of covariates at various points of the wage distribution. The coefficients $\beta(\theta)$ are thus estimates of the marginal impact of the explanatory variables at, e.g. the median ($\theta =$

0.5); at the bottom of the wage distribution, e.g. the 5^{th} quantile ($\theta = 0.05$); and at the top of the wage distribution, e.g. the 95^{th} quantile ($\theta = 0.95$).

Table 2, Panel A, displays the coefficients for the female dummy in various quantile regression models for the wage gap. The first row of Table 2 shows the result of a regression on the female dummy with no other explanatory variables. The coefficient of 0.1 per cent at the 5^{th} quantile corresponds to the height of the 'Raw gap' curve in Figure 1 at the 5^{th} percentile, and the coefficient of 21.5 per cent at the 95^{th} quantile corresponds to the height of the curve at the 95^{th} percentile. The last column is the OLS result of 10.2 per cent, the average gender gap over the wage distribution. The tendency of acceleration of the magnitude of the wage gap at the upper quantiles implies that the wage gap at the median of 8.8 per cent is below the average wage gap.

Table 2 around here

The inclusion of the basic human capital variables (schooling, experience, experience squared, tenure and tenure squared) in row 2 leaves the OLS estimate of the gender gap virtually unaltered. However, as the gender gap increases at the lower quantiles and decreases at the upper quantiles the unchanged average gender gap reflects a twist over the quantiles of the conditional wage distribution.

When extended controls (public sector, residence in the capital and cohabitation) are included, the twist increases as the gender gap at the lower quantiles increases further and the gender gap at the upper decreases. However, the decrease at the upper quantiles is substantial, and the introduction of extended controls implies that the OLS estimate of the gender gap falls to 9.5 per cent.

The last model of Table 2, Panel A, contains the results when measures for occupational segregation are included: dummies of one-digit occupations, the share of females in 6-digit occupations, industries, establishments and job-cells. The result is a reduction of the gender gap throughout the conditional wage distribution. However, there still is a steady increase in the gender gap over the wage distribution from 1.9 per cent at the 5th conditional quantile to 3.2 per cent at the median, up to 6.4 per cent at the 90th quantile.

A result of this section is that a glass ceiling still exists even when controls for occupational segregation are included, although the magnitude is rather moderate. The OLS estimate of the gender dummy of 3.5 per cent in the final model of Table 2 is smaller than the estimate of the wage gap at the 95th quantile of 5.8 per cent.

4. Quantile regressions by gender

This section presents the procedure that makes it possible to perform decompositions of wage distributions on large data sets. We first report quantile and OLS coefficients for the conditioning variables for men and women separately. The estimates from these regressions are used to decompose the gender wage gap in components due to gender differences in characteristics and in gender differences in rewards to characteristics.

Table 3 and 4 contains the quantile and OLS results for men and women, respectively. The coefficients to schooling do not vary much over the quantiles. Moreover, the coefficients are small, as the return to schooling is highly correlated with occupational choice (the return to schooling without the variables for occupational segregation are about twice as high as the returns shown in tables 3 and 4).

Table 3 around here

According to the coefficients experience and tenure, both the experience profile and the tenure profile appear to be most pronounced at the lower quantiles of the wage distribution. However, the coefficients are small and the magnitude of variation is limited. The reward to basic human capital is nearly the same for men and women; the differences between the coefficients in Table 3 and Table 4 are close to zero. According to the OLS results, employment in the public sector implies on average a wage loss for men that is a substantial higher than the loss for women. However, these penalties are the average of moderate wage premiums in the lower quantiles and substantial penalties in the upper quantiles of the conditional wage distributions. Employment in the capital entails a wage premium for both men and women, a premium most pronounced in the upper quantiles. Single men earn less than men living with partners and this wage penalty is most pronounced in the upper quantiles. In contrast, single women in the lower quantiles earn more that women with partners whilst single women in the upper quantiles face a wage penalty.

Table 4 around here

Wages varies considerably with the share of females in occupation, industry, establishment and job-cell. More females within occupations and job-cells imply lower wages for both men and women with a substantial variation over the wage distribution. The relation between wages and the share of females in industry and establishment also varies considerably over the wage distribution.

Average wages for occupational groups, conditional on the covariates, do not vary much between major occupational groups 4 to 9 neither for men nor women. The decompositions in the following sections are relative to the reference group, whose wage level thus corresponds to the level in group 4 to group 9 (these groups constitute more than 50 per cent of the workforce). However, wages increase steeply from the reference group 5 (service and sales workers), to major group 3 (technicians), over group 2 (professionals) up to group 1 (managers). Men enjoy a higher wage premium in these upper occupational groups than women.

In most of the major occupational groups the coefficients for males increase monotonous over the wage distribution. In many cases the coefficients in the upper part of the conditional wage distribution is substantially higher than the coefficients in the lower quantiles of the conditional wage distribution. The coefficients for females do not exhibit the same sharp increase over the wage distribution and in some cases the coefficients exhibit a non-monotonous or declining pattern.

A major difference between the estimates for men and women is the magnitude of the constant terms. All the male constant terms are higher than the female constant terms, but the difference is much larger at the lower quantiles that at the upper quantiles. At the 5th quantile the constant term for men is 11.4 per cent higher than the constant term for women; this difference decreases to 2.4 per cent at the 90th quantile and a level of 5.1 per cent at the 95th quantile. These large differences in the constant terms over the wage distribution have a substantial impact on the decompositions of the wage distributions in the following.

5. Aggregate decompositions

This section decomposes the gender wage gap into components that are due to differences in characteristics between men and women and components that is due to differences in rewards to characteristics. Such a decomposition is not feasible with the available methodology. We present and implement a new procedure that makes it feasible to decompose wage distributions on large data sets.

In this section we consider all male variables and coefficients taken together and all female variables and coefficients taken together, that is, we make aggregate decompositions. In the next section we consider detailed decompositions, where we trace the impact of groups of variables and parameters on the gender wage gap.

Decomposition of the gender gap at different quantiles of the wage distribution is more involved than the Oaxaca-Blinder decomposition of the average wage gap between men and women, since 'all' conditional quantiles are needed to assess one particular marginal quantile (see e.g. Angrist and Pischke (2009), pp. 281-283).

This paper applies an amendment of the decomposition procedure developed by Machado and Mata (2005). The suggested procedure is an innovation, which implies that the decomposition can be carried out for large samples of employees as, e.g. the 1.029.904 employees in our data set. In contrast, the Machado-Mata procedure is practically infeasible for large samples. According to Fortin et al. (2011), p. 62, a main limitation of the Machado-Mata method is that it '…. is computational demanding, and becomes quite cumbersome for data sets numbering more than a few thousand observation'.

We first present the proposed decomposition procedure and then discuss the procedure, including the difference from the Machado-Mata procedure. The estimation is performed for a set of quantiles θ_1, θ_2, ... θ_n that are fixed to $\theta_1 = 0.005$, $\theta_2 = 0.01$, $\theta_2 = 0.015$…,$\theta_{200} = 0.995$ (this set of quantiles serves as an approximation of 'all' conditional quantiles).

The procedure falls in eight steps:

1. Attach a random number to each observation in the data set and sort the data set according to the random number. Carry out a class division of the data set in s disjoint subsets of approximately 5.000 observations (which implies $s = 200$ in the present application).
2. Select a new set of the s disjoint subsets.
3. Divide the data set from (2) in a male data set and a female data set and estimate the male coefficients $\beta_m(\theta)$ and the female coefficients $\beta_f(\theta)$ for each θ.
4. Use the characteristics of the males in the male data set to construct (a) the predicted wage distribution for men using the estimated coefficients $\beta_m(\theta)$ from step 3 and (b) a counterfactual wage distribution for women using $\beta_f(\theta)$ from step 3.
5. Use the two wage distributions from step 4 to estimate the gender wage gap as the difference between the counterfactual wage distribution for women and the predicted wage distribution for men at each quantile.
6. Repeat step 2 to 5 with new selections of the disjoint data sets until all the s subsets have entered the calculations.
7. Perform step 1 to 6 three times.
8. Calculate the average values of the wage gaps in the samples from step 5 as an estimate of the gender wage gaps at the quantiles and compute the associated standard errors.

The iterative procedure in Machado and Mata (2005) includes steps 3, 4, 5 and 8 but perform the calculations on new data sets constructed by random draws (with replacement) of the observations. The Machado-Mata procedure is applied in Albrecht et al. (2003), Arulampalam et al. (2007), Fortin et al. (2011) amongst others.

The procedure in this paper is valid for making inference about the counterfactual distributions as the following arguments show. The coefficient estimates of the quantile regressions procedure are consistent and distributed asymptotical normal under conditions stated in Koenker and Bassett (1978). The estimates obtained from a subsample have the same characteristics, that is, the quantile coefficients $\hat{\beta}(\theta)$ in step 3 are consistent and distributed asymptotical normal. These estimates enter the calculations for recovering the counterfactual distributions in both the Machado-Mata procedure and the procedure proposed in this paper.

The difference between the two procedures is that the subsamples in the Machado-Mata procedure are bootstrap samples obtained by a sampling with replacement whilst the subsamples in the present procedure are samples without replacement (step 1 and 2 in the procedure implies random subsampling with replacement). Politis and Romano (1994) analyse this type of sampling as an alternative to bootstrap sampling. In his survey of the bootstrap, Horowitz (2001) includes alternatives to the bootstrap and term the procedure by Politis and Romano "non-replacement subsampling". An advantage of non-replacement subsampling is that asymptotic distributions of statistics are estimated under weaker conditions than are necessary for the bootstrap procedure. A drawback of non-replacement subsampling is that the rate of convergence is slower than under bootstrap sampling. Below we present checks of the convergence of the procedure applied on the present data set.

The results of the procedure are displayed in Table 2, Panel B. The basis for the first row of Panel B is separate quantile estimations for males and females, where the explanatory variables are the basic human capital variables. The first row of Panel B is constructed as the difference between the predicted male wage distribution and the counterfactual wage distribution assuming female reward to basic human capital variables (the wage structure) and male values of basic human capital variables. The total wage gaps between men and women (the first row in Panel A) can thus be decomposed in two components as follows: The difference from zero (male rewards and male characteristics) to first row of Panel B (female rewards and male values of basic human capital variables) is the difference in reward to characteristics between men and women. The remaining difference from the first row of Panel B to the first row Panel A is ascribed to other components, especially differences in characteristics between men and women. The figures in the first row of Panel B are fairly close to those for the raw gender wage gap in the first row Panel A. We can thus conclude that differences in rewards (coefficients) play a major role for the wage gaps between men and women over the wage distribution, whilst differences in basic human capital characteristics play a minor role.

Figure 1 gives a visual impression of the closeness of the estimates between the unconditional gender wage gap over the wage distribution and the counterfactual wage distribution. The curve ‘Basic HC’ is the difference between the predicted male wage distribution and the counterfactual wage distribution assuming female reward to basic human capital variables and male values of basic human capital variables. Instead of wage gaps for the seven quantiles presented in Table 2, we plot the wage gap for all the percentiles from one to 99 from the simulated wage distributions. The difference from horizontal line at 0.00 (that corresponds to the predicted male wage distribution) to the curve ‘Basic HC’ is the difference in reward to characteristics between men and women. The remaining difference from the curve ‘Basic HC’ to the curve ‘Raw gap’ is ascribed to differences in human capital characteristics between men and women. The curve ‘Basic HC’ is very close to the curve ‘Raw gap’ and we thus conclude that the majority of the wage gap between men and women is ascribed differences in coefficients.

We now extend the set of regressors to include not only basic human capital variables but also variables for sector, for living in the province and for being single (the extended human capital variables). When extended controls enter in the construction of the counterfactual wage gap, the second row of Table 2, Panel B, shows a moderate increase in the wage gap in the lower quantiles, a moderate decrease in the upper quantiles and a slight decrease in the OLS estimate to 9.9 per cent. In Figure 1 the curve for the counterfactual wage gap using extended human capital (‘Extended HC’) is very close to the curve for the raw wage gap over most of the wage distribution. We can thus conclude that in the model with basic and extended human capital variables, differences in rewards (coefficients) play a major role for the wage gaps over the wage distribution, whilst differences in characteristics play a minor role.

However, a different picture emerges when we take variables for segregation into account (dummies for the nine major occupational groups and the female share of workers in 6-digit occupations, industries, establishments and job-cells). The quantile regressions that enter this decomposition are the ones where the results are displayed in Table 3 and Table 4 for seven quantiles. In Figure 1 the curve for the counterfactual wage gap including coefficients for segregation (‘All controls’) is substantially below the curve for the raw wage gap. This indicates that differences in characteristics plays an important role for the gender wage gap for the model with all variables included, in contrast to the curves that displays the counterfactual wage gap without taking segregation into account. At the lowest quantiles the curve ‘All controls’ lies below the horizontal line at 0.00 indicating that women earn more than men in the counterfactual wage distribution.

For the model including segregation variables we decompose the gender wage gap numerically into components attributable to characteristics and to wage structure. The basis for the decomposition is the simulated wage gap calculated as the difference between the simulated wage distribution for males (male characteristics and male wage structure) and the simulated distribution for females (female characteristics and female wage structure). The resulting gender wage gap displayed in Figure 2 with the legend ‘Simulated raw gap’. The impression is that this curve has about the same shape and height as the curve for the actual raw wage gap in Figure 1. Figures for seven quantiles of the simulated wage gap appear in Table 2, Panel B, and these figures are fairly close to the actual raw wage gap in the first row of Table 2, Panel A. The OLS estimate is almost the same and the mean absolute prediction error for the seven quantiles is 1.2 per cent. These prediction errors are lower than the prediction errors in the Machado-Mata decomposition presented in Fortin et al. (2011) [2] This close fit between the actual and the simulated gender wage gap indicates the validity of the entire iterative procedure

[2] Fortin et al (2011), Table 4, contains a raw gender wage gap in panel A and a predicted gender wage gap in panel B estimated by the Machado-Mata procedure. The difference yields a mean absolute prediction error on 1.7 per cent.

consisting of steps 1-8, including the novel sampling scheme that makes it possible to decompose wage distributions on large data sets on the basis of quantile regressions.

The counterfactual wage gap for female wage structure and male characteristics is displayed in Table 2, Panel B, in the row 'wage structure' with basic human capital, extended controls and segregation variables included. The gender wage gap is reduced to slightly more than half of the raw gap in the upper quantiles, whilst the gender wage gap is reversed in the lowest quantiles. The numbers in the row 'characteristics' is the part of the gender wage gap attributable to characteristics, which is calculated as the difference between the simulated wage gap and the counterfactual distribution in the row 'wage structure'. In the upper part of the wage distribution, characteristics account of slightly less than half of the gender wage gap, whilst differences in characteristics accounts for more than the whole wage gap in the lower part of the wage distribution.

Overall, the evidence for the regression models without segregation variables is that difference in wage structure (coefficients) between males and females accounts for nearly all the gender wage gap while differences in characteristics plays a close to negligible role. In contrast, in the model including segregation variables, differences in characteristics accounts for almost half of the gender wage gap in the upper quantiles and more than the whole gender wage gap in lowest quantiles of the wage distribution.

6. Detailed decompositions

The analysis in this section is an example of a 'detailed decomposition', where we assess the role of groups of variables for the gender wage gap over the wage distribution. In contrast to the 'aggregate decompositions' in the previous section, detailed decompositions assess the contribution of single covariates and parameters (or groups of covariates and parameters).

An alternative to the Machado-Mata procedure is the reweighting method developed in DiNardo et al. (1996). However, as emphasized in Fortin et al. (2011), p. 68, a

'…. limitation of the reweighting method is that it is not straightforwardly extended to the case of the detailed decomposition'. The procedure presented in this paper makes it possible to perform detailed decompositions on large data sets such as the linked employer-employee data set used in this paper.

We perform detailed decompositions on the quantile regression models on the full set of explanatory variables, that is, the regressions in Table 3 and Table 4. The corresponding aggregate decomposition is displayed in Table 2, Panel B, in the row 'wage structure' with all variables included and in Figure 1 as the curve labelled 'All controls'.

The evidence from the aggregate decomposition in Table 2 and Figure 1 is that differences in the wage structure accounts for more than half of the gender wage gap in the upper quantiles and nothing in lowest quantiles of the wage distribution. However, from the evidence presented so far, we are not able asses the role of the reward to different characteristics. We now evaluate the extent to which the wage gap for the model with all controls included is attributable to three sets of components in the wage structure: the coefficients to extended human capital variables (human capital variables and extended control), the coefficients to segregation variables and the constant terms.

We amend the simulations in step 4 that entail multiplying the male data set on the estimates of the female coefficients $\beta_f(\theta) = [\beta_f^c(\theta), \beta_f^{HC}(\theta), \beta_f^{SE}(\theta)]$, where $\beta_f^c(\theta)$ is the constant term, $\beta_f^{HC}(\theta)$ is the coefficients to the extended human capital variables and $\beta_f^{SE}(\theta)$ is the coefficients to the segregation variables. Instead of applying all female coefficients at once, we substitute groups of female coefficients into the set of male parameters $[\beta_m^c(\theta), \beta_m^{HC}(\theta), \beta_m^{SE}(\theta)]$.

We first simulate a counterfactual wage distribution by multiplying the male data set on $[\beta_m^c(\theta), \beta_f^{HC}(\theta), \beta_m^{SE}(\theta)]$, that is, using the male constants and coefficients for the segregation variables (the coefficients in Table 3) but the female coefficients for the extended human capital variables (the coefficients in Table 4). The difference between this counterfactual wage distribution and the simulated wage distribution for males is displayed as the curve 'Extended HC' in Figure 2. This curve is everywhere below the horizontal line at zero (which denotes male characteristics and male coefficients). The coefficients for female extended human capital variables thus draw in the direction of a reduced gender wage gap over the whole wage distribution. The reduction in the gender wage gap is most pronounced in the upper most and the lower most tails of the wage distribution.

Figure 2 around here

Next we assess the impact on the gender wage gap of the difference between male and female reward to the segregation variables. We simulate a counterfactual wage distribution by multiplying the male data set on $[\beta_m^c(\theta), \beta_m^{HC}(\theta), \beta_f^{SE}(\theta)]$, that is, using the male constants and coefficients for extended human capital variables but female coefficients for the segregation variables. The difference between this counterfactual wage distribution and the simulated wage distribution for males is displayed in Figure 2 as the curve 'Segregation'. This curve increases steadily over the percentiles of the wage distribution and is above the zero line from the 40th percentile. That is, differences between male and female coefficients to segregation variable contributes to a decreased gender wage gap in the lower quantiles and draw in the direction of increasing the gender wage gap in the upper quantiles of the wage distribution.

Finally, we evaluated the role of the differences in the constant terms over the wage distribution between males and females. We simulate a counterfactual wage distribution by multiplying the male data set on $[\beta_f^c(\theta), \beta_m^{HC}(\theta), \beta_m^{SE}(\theta)]$, that is, using male coefficients for the explanatory variables but female constants. The curve 'Female constant' in Figure 2 displays the difference between this counterfactual wage distribution and the simulated wage distribution for males. The curve is everywhere above the horizontal zero line, that is, small female constants relative to male constants draws in the direction of a larger gender wage gap. The curve 'Female constant' decreases steadily over the wage distribution from a level of ten per cent at the lower end of the wage distribution to about three per cent at the higher end of the wage distribution.

The 'Female constant' curve lie above the 'Simulated raw gap' curve up to about the 25th percentile and below beyond the 25th percentile. That is, in the lower quantiles of the wage distribution differences in wages between men and women is completely accounted for by differences in the constant terms $\beta_f(\theta)$ and $\beta_m(\theta)$ and more so. In this range of the wage distribution, the combined effects of the other components of the wage distribution, differences in characteristics (other than femaleness) and differences in rewards to these characteristics draws in the direction of a reduced wage differential between males and females.

The difference between the constant terms is the unexplained difference in remuneration. Differences in the male and female constant terms reflect the difference in 'reward' to the characteristic of being a male or a female, which sometimes is taken as an indication of discrimination.

Table 2 contains numerical estimates of the contribution for the three components of the wage structure for seven of the 100 quantiles displayed in Figure 2. There is a close correspondence between the height of the curves Figure 2 and the figures in Table 2 (which are calculated by entering groups of variables sequentially such that the sum adds up to the figures for the wage structure).

It is of interest to compare the magnitude of the difference in reward between males and females in the lower end of the wage distribution with the glass ceiling in the upper end of the wage distribution. Differences in the constant terms account for close to 10 percentage point of the gender wage gap in the lower end of the wage distribution. The magnitude of the glass ceiling that remains after segregation is taken into account is about 6 per cent (according to the last row in Table 2, Panel A). The unexplained part of the wage differential between males and females in the lower end of the wage distribution is thus quantitatively more important than the remaining glass ceiling in the upper end of the wage distribution.

According to the evidence from the detailed decompositions in this section, differences in the reward to human capital (and extended controls) draw in direction of a reduced gender wage gap. Differences in the coefficients for labour market segregation draws in direction of increased wage inequality between males and females in the upper part of the wage distribution and decreased wage inequality in the lower part of the wage distribution. The unexplained part of the gender wage gap (or differences in the constant terms) plays a major role for the magnitude of the gender wage gap in the lower end of the wage distribution and is quantitatively more important than the remaining glass ceiling in the upper end of the wage distribution.

7. Conclusions

The paper presents and implements a procedure to make quantile decompositions of wage distributions on large data sets. The standard Machado-Mata decomposition procedure in not applicable on large data sets. The procedure of this paper replaces the bootstrap sampling in the Machado-Mata procedure with an alternative sampling scheme, “non-replacement subsampling”, that is more suitable for quantile analysis of large data sets. The application of the decomposition procedure of this paper is not limited to decompositions of wage distributions but can be applied in other areas where the Machado-Mata procedure is not feasible because of the magnitude of the data sets.

The paper analyses a linked employer-employee data set with more than one million observations. Such a data set makes it possible to calculate the share of females in occupations, establishments and job-cells, and linked employer-employee data sets is thus especially suited to analyses wage formation in relation to segregation.

A recent strand of literature analyses the extent to which women face a glass ceiling in the labour market in the sense that the wage gap increases throughout the wage distribution and accelerates in the upper tail. Our analysis confirms the existence of a glass ceiling in the Danish labour market. However, most of the glass ceiling is manifested in segregation between males and females in the labour market. Although significant, the magnitude of remaining glass ceiling is limited when segregation variables are taken into account.

Decompositions of the gender wage gap shows that segregation plays a major role for the wage gap. Decompositions without segregation variables indicate that differences in characteristics play a minor role for the gender wage gap. The decomposition with segregation variables included show that characteristics account of slightly less than half of the gender wage gap in the upper end of the wage distribution and for more than the whole wage gap in the lower part of the wage distribution.

A detailed decomposition shows that the reward to human capital draws towards smaller wage differentials between males and females. Different remuneration in female dominated jobs categories constitute an important element in the gender wage gap in the upper part of the wage distribution, whilst the coefficients to the segregation variables draws in direction of a reduced gender wage gap in the lower part of the wage distribution.

A major determinant for the gender wage gap is differences in the constant terms between males and females, or the unexplained part of the gender wage gap. At the lower end of the wage distribution, differences in the constant terms account for more than the whole gender wage gap, so that the combined effects of all other components of the wage distribution draws in the direction of a reduced wage differential between males and females. The magnitude of the component of the gender wage gap due to differences in constant terms between males and females in the lower part of the wage distribution is substantially higher than the gender wage gap that remains in the upper part of the wage distribution after segregation variables have been taken into account. In this sense, the unexplained part of the gender wage gap in the lower end of the wage distribution is larger and quantitatively more important than the glass ceiling in the higher end of the wage distribution.

Reference List

Albæk, K. & Thomsen, L. B. (2014). *Occupational segregation and the gender wage gap - an analysis of linked employer-employee data* SFI.

Albrecht, J., Bjorklund, A., & Vroman, S. (2003). Is there a glass ceiling in Sweden? *Journal of Labor Economics, 21,* 145-177.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Arulampalam, W., Booth, A. L., & Bryan, M. L. (2007). Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution. *Industrial & Labor Relations Review, 60,* 163-186.

Bayard, K., Hellerstein, J., Neumark, D., & Troske, K. (2003). New evidence on sex segregation and sex differences in wages from matched employee-employer data. *Journal of Labor Economics, 21,* 887-922.

DiNardo, J., Fortin, N. M., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica, 64,* 1001-1044.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition Methods in Economics. In D.Card & O. Ashenfelter (Eds.), *Handbook of Labor Economics, Vol. 4A* (pp. 1-102). Amsterdam.: North-Holland.

Gupta, N. D. & Rothstein, S. R. (2005). The impact of worker and establishment-level characteristics on male-female wage differentials: evidence from Danish matched employee-employer data. *LABOUR, 19,* 1-34.

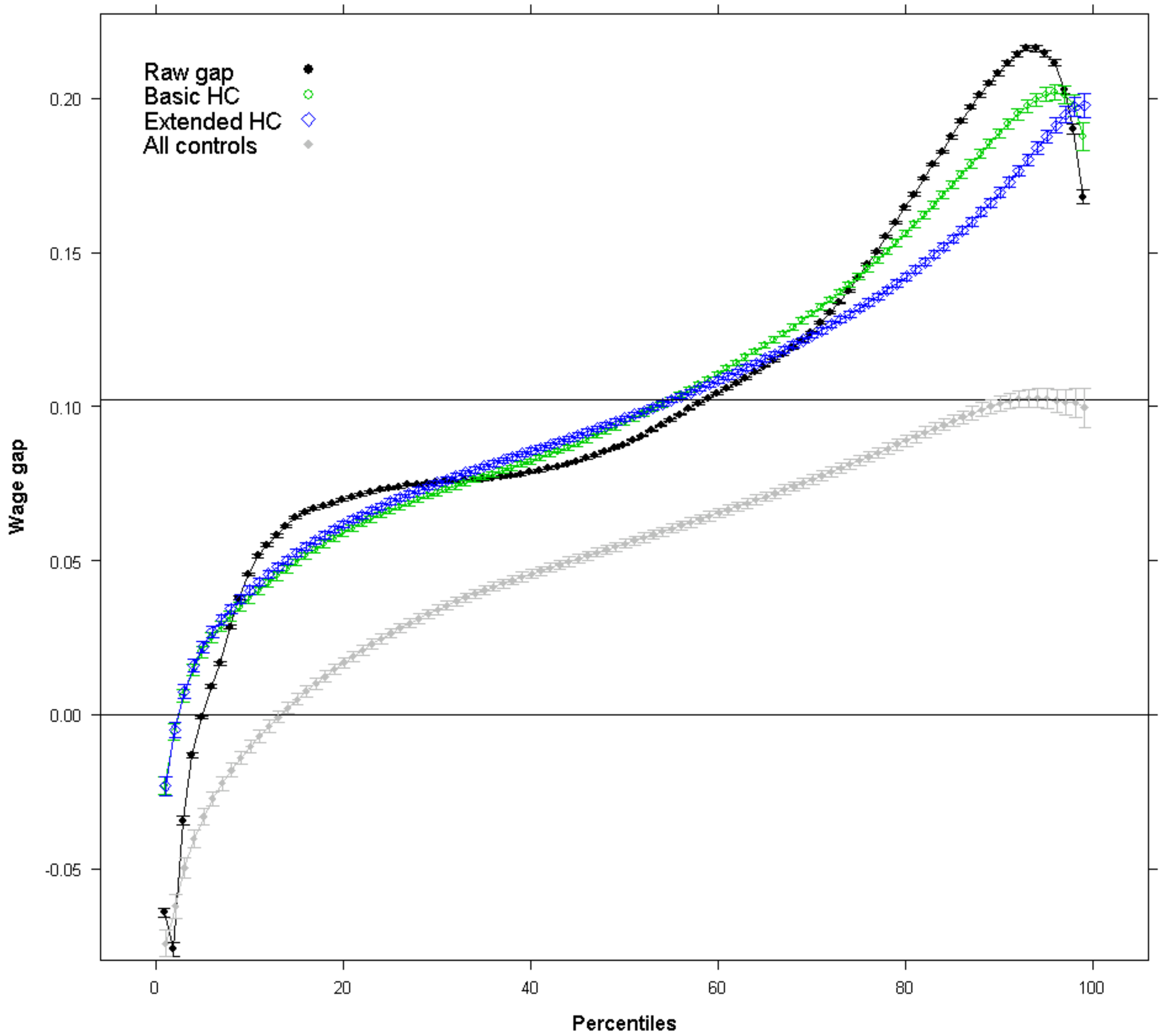
Horowitz, J. L. (2001). The bootstrap. In J.J.Heckman & E. Leamer (Eds.), *Handbook of Econometrics, Vol. 5* (pp. 3159-3228). Amsterdam: Elsevier.

Koenker, R. & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46,* 33-50.

Machado, J. A. F. & Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics, 20,* 445-465.

Politis, D. N. & Romano, J. P. (1994). Large-Sample Confidence-Regions Based on Subsamples Under Minimal Assumptions. *Annals of Statistics, 22,* 2031-2050.

Figure 1. Gender wage gap, raw and counterfactuals



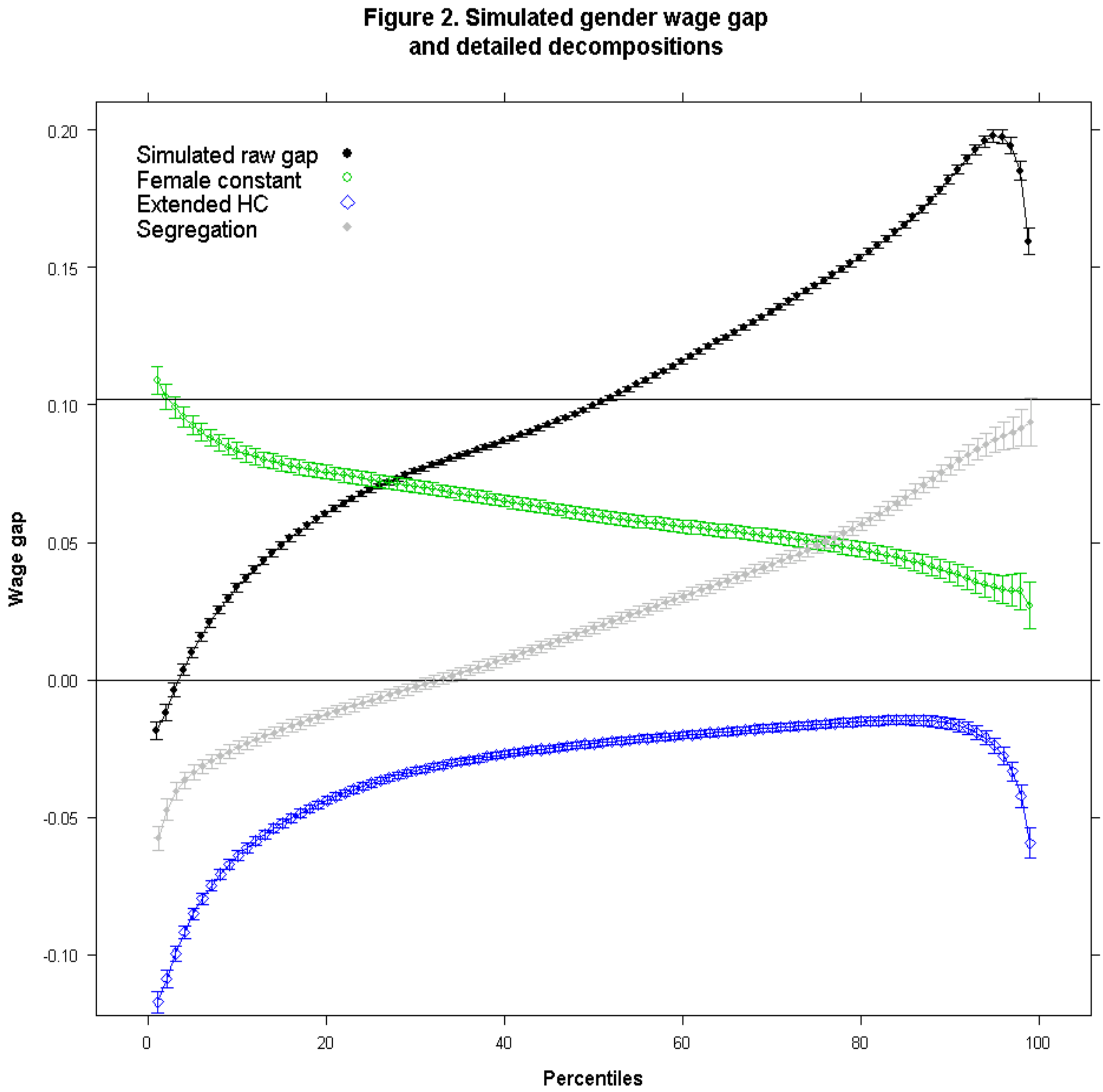


Figure 2. Simulated gender wage gap
and detailed decompositions
Simulated raw gap
Female constant
Extended HC
Segregation
Wage gap
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
0
20
40
60
80
100
Percentiles

Table 1. Descriptive statistics

	All	Men	Women	Difference
Log wage	5.418	5.465	5.363	0.102
Woman	0.460	0.000	1.000	-1.000
Schooling	12.860	12.787	12.945	-0.158
Experience	17.210	17.902	16.397	1.505
Tenure	5.310	5.326	5.292	0.034
Public	0.327	0.197	0.480	-0.284
Capital	0.387	0.354	0.425	-0.071
Single	0.279	0.282	0.275	0.006
Female share in				
Occupation	0.460	0.280	0.672	-0.391
Industry	0.460	0.342	0.600	-0.258
Establishment	0.460	0.313	0.632	-0.319
Job cell	0.460	0.204	0.760	-0.556
Occupation				
1. Managers	0.051	0.074	0.023	0.051
2. Professionals	0.167	0.171	0.163	0.008
3. Technicians	0.230	0.175	0.296	-0.121
4. Clerical support	0.113	0.063	0.171	-0.108
5. Service and sales	0.146	0.091	0.211	-0.121
6. Agriculture	0.002	0.003	0.001	0.002
7. Craft workers	0.103	0.180	0.014	0.166
8. Plant operators	0.089	0.122	0.049	0.073
9. Elementary	0.098	0.121	0.071	0.049
N	1,029,906	555,761	474,145	

Note: Occupation group 3 is Technicians and associate professionals, group 6 is Skilled agricultural, forestry and fishery workers, group 7 is Craft and related trades workers, group 8 is Plant and machine operators, and assemblers, group 9 is elementary occupations

Table 3. Regressions for extended model, quantile and OLS estimates, Men

Explanatory variables:	Quantiles 5th	10th	25th	50th	75th	90th	95th	OLS
Schooling	0.040*	0.041*	0.031*	0.028*	0.028*	0.029*	0.029*	0.035*
Experience	0.018*	0.016*	0.011*	0.010*	0.010*	0.009*	0.009*	0.012*
Exp. squared/100	-0.089*	-0.079*	-0.054*	-0.047*	-0.044*	-0.042*	-0.042*	-0.056*
Tenure	0.015*	0.012*	0.009*	0.007*	0.005*	0.004*	0.003*	0.008*
Tenure squared/100	-0.068*	-0.052*	-0.042*	-0.036*	-0.029*	-0.022*	-0.019*	-0.041*
Public	0.015*	-0.008*	-0.040*	-0.083*	-0.141*	-0.200*	-0.235*	-0.115*
Capital	0.067*	0.077*	0.085*	0.101*	0.110*	0.110*	0.106*	0.099*
Single	-0.009*	-0.011*	-0.021*	-0.028*	-0.036*	-0.043*	-0.045*	-0.026*
Female share in								
Occupation	-0.117*	-0.120*	-0.122*	-0.117*	-0.095*	-0.064*	-0.052*	-0.099*
Industry	-0.059*	-0.074*	-0.093*	-0.096*	-0.046*	0.022*	0.083*	-0.036
Establishment	0.017*	0.038*	0.069*	0.079*	0.079*	0.085*	0.076*	0.087*
Job cell	0.006	-0.026*	-0.047*	-0.064*	-0.081*	-0.117*	-0.142*	-0.066*
Occupation								
1. Managers	0.179*	0.220*	0.290*	0.414*	0.563*	0.722*	0.813*	0.462*
2. Professionals	0.218*	0.254*	0.306*	0.339*	0.354*	0.406*	0.451*	0.336*
3. Technicians	0.121*	0.160*	0.205*	0.254*	0.269*	0.303*	0.335*	0.244*
4. Clerical support	0.002	0.021*	0.020*	0.016*	0.012*	0.037*	0.062*	0.036
6. Agriculture	-0.142*	-0.123*	-0.089*	-0.061*	-0.060*	-0.057*	-0.021	-0.059
7. Craft workers	-0.114*	-0.061*	-0.007*	0.014*	0.010*	0.012*	0.021*	0.002
8. Plant operators	-0.032*	-0.008*	0.008*	0.025*	0.029*	0.037*	0.057*	0.048
9. Elementary	-0.078*	-0.062*	-0.044*	-0.029*	-0.031*	-0.022*	-0.006	-0.011
Constant	5.065*	5.118*	5.220*	5.340*	5.515*	5.696*	5.814*	5.383*

Note: * denotes significance at 5 per cent level. The reference group is a man with 13 years of schooling, 17 years of experience, 5 years of tenure, employed in the private sector, living in the province, married, works together with 46.0 per cent females and employed as a service and sales worker, major occupation group 5.

Table 4. Regressions for extended model, quantile and OLS estimates, Women

Explanatory variables:	Quantiles							OLS
	5th	10th	25th	50^{th}	75th	90th	95th	
Schooling	0.032*	0.031*	0.023*	0.023*	0.026*	0.030*	0.033*	0.030*
Experience	0.015*	0.012*	0.009*	0.009*	0.009*	0.008*	0.007*	0.010*
Exp. squared/100	-0.067*	-0.055*	-0.037*	-0.037*	-0.037*	-0.038*	-0.035*	-0.043*
Tenure	0.016*	0.012*	0.008*	0.005*	0.004*	0.004*	0.005*	0.008*
Tenure squared/100	-0.068*	-0.052*	-0.040*	-0.035*	-0.030*	-0.032*	-0.039*	-0.047*
Public	0.055*	0.051*	0.020*	-0.022*	-0.080*	-0.115*	-0.098*	-0.018
Capital	0.070*	0.071*	0.070*	0.086*	0.101*	0.115*	0.123*	0.094*
Single	0.006*	0.007*	0.002*	-0.002*	-0.007*	-0.017*	-0.039*	-0.008*
Female share in								
Occupation	-0.028*	-0.031*	-0.021*	-0.033*	-0.064*	-0.083*	-0.066*	-0.043
Industry	-0.029*	0.005	0.030*	0.051*	0.100*	0.150*	0.179*	0.053
Establishment	0.032*	0.015*	-0.010*	-0.027*	-0.037*	-0.048*	-0.050*	-0.011
Job cell	-0.103*	-0.085*	-0.065*	-0.077*	-0.100*	-0.116*	-0.102*	-0.092*
Occupation								
1. Managers	0.273*	0.293*	0.324*	0.382*	0.456*	0.544*	0.638*	0.400*
2. Professionals	0.270*	0.283*	0.303*	0.296*	0.274*	0.298*	0.351*	0.289*
3. Technicians	0.151*	0.170*	0.180*	0.177*	0.150*	0.148*	0.203*	0.171*
4. Clerical support	0.105*	0.112*	0.095*	0.084*	0.046*	0.027*	0.055*	0.079*
6. Agriculture	-0.090*	-0.078*	-0.028	0.00	-0.041*	-0.093*	-0.034	-0.038
7. Craft workers	-0.029*	0.008	0.025*	0.019*	-0.037*	-0.054*	-0.001	-0.001
8. Plant operators	0.093*	0.103*	0.068*	0.057*	0.020*	0.021*	0.095*	0.082*
9. Elementary	0.015*	0.017*	-0.006*	-0.020*	-0.064*	-0.073*	-0.035*	-0.007
Constant	4.951*	5.015*	5.133*	5.279*	5.480*	5.672*	5.763*	5.318*

Note: * denotes significance at 5 per cent level. The reference group is a woman with 13 years of schooling, 17 years of experience, 5 years of tenure, employed in the private sector, living in the province, married, works together with 46.0 per cent females and employed as a service and sales worker, major occupation group 5.

Table 2. Gender wage gap in quantile regressions and counterfactual decompositions.

	Explanatory variables			Quantiles							
	Basic HC	Extended controls	Segregation variables	5th	10th	25th	50th	75th	90th	95th	OLS
Panel A. Regressions											
	No	No	No	0.001	-0.046*	-0.073*	-0.088*	-0.142*	-0.208*	-0.215*	-0.102*
	Yes	No	No	-0.035*	-0.043*	-0.064*	-0.101*	-0.146*	-0.175*	-0.180*	-0.103*
	Yes	Yes	No	-0.047*	-0.057*	-0.071*	-0.094*	-0.122*	-0.146*	-0.149*	-0.095*
	Yes	Yes	Yes	-0.019*	-0.016*	-0.021*	-0.032*	-0.050*	-0.064*	-0.058*	-0.035*
Panel B. Decompositions											
Wage Structure	Yes	No	No	-0.020*	-0.038*	-0.066*	-0.095*	-0.142*	-0.189*	-0.201*	-0.103*
Wage Structure	Yes	Yes	No	-0.024*	-0.042*	-0.069*	-0.095*	-0.131*	-0.168*	-0.186*	-0.099*
Simulated wage gap	Yes	Yes	Yes	-0.010*	-0.034*	-0.069*	-0.100*	-0.143*	-0.182*	-0.198*	-0.103*
Charactistics	Yes	Yes	Yes	-0.041*	-0.042*	-0.042*	-0.045*	-0.062*	-0.083*	-0.098*	-0.054*
Wage Structure	Yes	Yes	Yes	0.031*	0.008*	-0.027*	-0.055*	-0.081*	-0.099*	-0.100*	-0.049*
Extended HC	Yes	Yes	Yes	0.086*	0.064*	0.038*	0.024*	0.016*	0.017*	0.024*	0.034*
Segregation	Yes	Yes	Yes	0.038*	0.027*	0.008*	-0.019*	-0.047*	-0.077*	-0.090*	-0.021*
Constant	Yes	Yes	Yes	-0.093*	-0.083*	-0.073*	-0.060*	-0.050*	-0.039*	-0.034*	-0.062*

Note: * denotes significance at 5 per cent level. Basic Human Capital variables are number of years of schooling, experience, experience squared, tenure in firm and tenure squared. Extended controls are dummies for the public sector, for residence in the capital and for cohabitation. Segregation variables are dummies for 9 occupational categories at the one digit ISCO level and the share of female workers in 789 occupational categories, 662 industrial categories, 22,154 establishments and 152,320 job cells. The counterfactual decompositions are constructed from female coefficients and the data set for males.