Gender Wage Gap in Canada: An Analysis using Counterfactual Distributions Regression

By Hui Lu

(30006165)

Major paper presented to the

Department of Economics of the University of Ottawa

in partial fulfillment of the requirements of the M.A. Degree

Supervisor: Professor Paul Makdissi

ECO6999

Ottawa, Ontario

April 2019

Abstract

In this paper, I estimate and decompose the gender wage gap in Canada using a counterfactual distribution approach based on distribution regression methods developed by Chernozhukov et al (2013). Using the Canadian Labour Force survey of October and April 2018, I show that men earn higher wages than women throughout the distribution. The difference in log hourly wage increases when moving up the distribution and drops at the end of the distribution, which the discrimination is higher between 20^{th} and 70^{th} quantiles. In the decomposition, the structural effect constitutes the majority of the gap while the composition effect reduces the gap. When controlled for industry and occupation, the gap is completely due to discrimination. This contrasts with the positive sign of the explained portion in the canonical Blinder-Oaxaca decomposition. Overall, the result of this paper provides new empirical evidence to support discrimination is the cause of gender wage gap.

Index

1 Introduction

It is generally known that men have a higher average income than women. There are a large number of studies covering this topic, but comparing to decades ago, modern econometric tools can be applied to the labour market to give evidence decomposing the gender wage gap. Decomposing the gender wage gap using these tools gives important information about the Canadian labour market, as well as about the effectiveness of gender equality policies in both the public and private sector.

In terms of methodology, most studies use the Blinder-Oaxaca decomposition. A few studies, such as Boudarbat and Connolly (2013), use unconditional quantile regressions to decompose the entire distribution. Many of them have used the recentered influence function (RIF) to compute for distributional statistics. However, the non-linear characteristic of RIF regression is closely related to the problem of estimating linear regression model for a dichotomous dependent variable according to Firpo, Fortin, and Lemieux (2009). The use of logistic regression in counterfactual distribution approach can thus deal with this special problem of RIF in these studies. To the best of my knowledge, this paper is the first one using this particular counterfactual distributional regression approach to explain the gender wage gap in Canada. Counterfactual distribution regression is very handy and effective as a comprehensive and flexible tool for modeling. The approach adopted from Chernozhukov et al (2013) allows for the overall gender wage gap to be decomposed along the whole range of wage distribution, not only at the mean. This is in contrast to the commonly-used Blinder-Oaxaca decomposition found in Blinder (1973) and Oaxaca (1973).

To better understand the gender wage discrimination in Canada, the purpose of this paper is to estimate and decompose the gender wage gap using a counterfactual distribution approach based on distribution regression methods developed by Chernozhukov et al (2013). This paper uses the Canadian Labour Force survey of October and April 2018 as its data source. In brief, I show that men earn higher wages than women throughout the distribution. The difference in log hourly wage increases when moving up the distribution and drops at the end of the distribution, which the discrimination is higher between 20^{th} and 70^{th} quantiles. In the decomposition, the structural effect constitutes the majority of the gap while the

composition effect reduces the gap. When controlling for industry and occupation, the gap is completely due to discrimination. This contrasts with the positive sign of the explained portion in Blinder-Oaxaca decomposition. Comparing my results to Boudarbat and Connolly (2013), the observable personal characteristics in my decomposition cannot explain the wage gap. But theirs can by a small portion. The total wage gap in my results decreases at the top of the distribution whereas theirs remains high. At the same time, my results are not restricted by age. Comparing to them, I provide a more comprehensive analysis for Canadian labour market as a whole. Despite the great potential of counterfactual distribution regression, it has not previously been used to examine gender income inequality. Thus, this paper provides important new evidence on gender income inequality, and will in turn help to adjust gender income policies.

The rest of the paper is structured as follows: Section 2 reviews recent studies on the relative topics. Section 3 discusses the methodology. Section 4 presents the data source and summary statistics by gender groups. Section 5 explains the covariates. Section 6 shows the regression table and decomposition results. Section 7 summarizes the conclusions.

2 Literature review

An early study by Oaxaca (1973) estimates the average discrimination against female workers in the U.S., using the 1967 Survey of Economic Opportunity as a data source. The percentage of discrimination, as in the unexplained part is 58.4% for white women, 55.6% for black women in the logarithmic wage differential. When personal characteristics like industry and occupation are added, these numbers rise to 77.7% of the wage differentials for white and 93.6% for black. He concludes that part-time employment, marital status, class of worker, industry, and occupation significantly narrow the wage differential between genders. His results show that the phenomenon of unequal pay for equal work does not account for a large part of the gender wage differential; rather, it is the crowding of women in low paying jobs that creates the gap. This suggests that a significant portion of the gender wage gap comes from labour market discrimination.

Shifting the focus to Canada, Baker, Benjamin, Desaulniers, and Grant (1995) is an early Canadian study that examines the male-female earnings differential between 1970 and 1990. They use the 1971, 1981, and 1986 Canadian censuses as their data source adding the individual files of the 1986 and 1991 Surveys of Consumer Finances. Baker et al. (1995) find a decline in the difference between male and female wages from 1970 to 1990, while the earnings ratio rises from 1970 to 1990. A dynamic decomposition in their study shows that most of the change is due to a decrease in the difference of return to personal characteristics across both genders. The study also points out that differential across the population is sensitive to age and education. Changes in wage differential tend to favour younger and less educated women.

Kidd and Shannon (1996) study the gender wage gap between Australia and Canada. They use the 1989 Canada Labour Market Activity Survey as the data set for the Canada section. Their data shows a mean log wage difference for men and women is of 0.287 in Canada and 0.143 in Australia. In percentage terms, this means that women earn 33% less than men in Canada and 15% less in Australia. Kidd and Shannon find that the mean female residual in decomposition is sensitive to both industry and occupation dummies. The extended specification leads to a decline in mean residual from 100% to 80% for Canada, where the unexplained component is decreased by more than 20%. The decomposition suggests that gender-specific factors, as opposed to wage factors, explain 47% to 62% of the inter-country difference in the gender wage gap using their basic model. In the extended model including all industry, occupation and experience controls, the number is 34% to 38%.

In fact, many studies have shown that men and women tend to work in different industries, companies, and positions. For example, Korkeamäki and Kyyrä (2006) find that women are more concentrated in administrative occupations than other job types. They focus on the gender wage gap in the Finnish manufacturing sector using correlated random effects modeling. Their data comes from the records of the Confederation of Finnish Industry and Employers. They find that white-collar women in Finland have an average income of 22% less than white-collar men. Within the same jobs, female workers are paid 6% less than male co-workers assuming both genders are equally productive and qualified for the job. In the

decomposition, a large part of the source of wage differentials comes from the disproportionate concentration of women in lower-paying jobs. However, this dataset only contains information for white-collar workers, thus restricting its application.

Milgrom et al (2001) studies the relative gender wage gap and gap decomposition in Sweden over the period from 1970 to 1990. Similar to Korkeamäki and Kyyrä (2006), the empirical results show that the within-occupation establishment wage gap is relatively small, at less than 4%. This indicats that pay rates are fairly uniform across companies within a job. Even if male and female workers are differently distributed among firms, it does not necessarily cause a large gender wage gap as long as the occupation is held constant. Similar to recent studies, they found a significant reduction in the within-occupation gender wage gap from 1970 to 1990.

According to Waldfogel (1997), the difference in characteristics between men and women such as education and work experience decreased from 1980 to 1991, but marital and parental status decomes increasingly important for both genders. Even if variables such as education and experience are controlled, a woman with a child will have a family penalty of 10% to 15% on income compared to a woman without a child. At the same time, women with children are also less likely to have successful careers. Not only is this punitive effect not found in relation to men, but men who are married with children usually earn more than other men; this increase can range from 10% to 15%, as found in a study done by Loh (1996).

The counterfactual distribution approach actually allows economists to see interesting phenomena. Arulampalam et al (2007) and Albrecht et al (2003) analyze the gender wage gap across wage distributions for Europe. They find strong evidence of a glass ceiling in many countries like Sweden, Britain, Denmark, and France. By glass ceiling, they mean that female workers do quite well in the labour market up to a certain point, after which there will be an effective limit on their prospects. The existence of the ceiling effect shows that women fall behind men more at the upper part of the wage distribution compared to the bottom or middle.

Adopting very similar approaches in the counterfactual distribution regressions developed by Chernozhukov et al (2013), Asplund and Napari (2011) reproduce the method in Melly (2005). They use

unconditional quantile regression techniques to locate important factors underlying the male-female wage gaps in three consecutive steps. Initially, the conditional wage distributions are estimated by quantile regressions. Then, they estimate the corresponding unconditional distributions by integrating the conditional wage distributions over the whole list of characteristics included in quantile regressions. The last step is to decompose the counterfactual wage distribution into three components in groups of gender. One captures the wage gap in differences of coefficients, namely the price effect. Another measures the composition effect of different characteristics between genders. The third captures the residuals left indicating the total effect. The decomposition of the gender wage gap is estimated 100 times of different quantile regressions to be distributed uniformly between 0 and 1. Using data from the full records of the Confederation of Finnish Industries EK, their results show that a great part of the wage differential between both genders can be explained by the price effect, in which men and women are rewarded differently by similar human capital characteristics. Comparing 2002 to 2009, they find that this pattern increases over the time period. Additionally, when going up through wage distribution, the composition effect weakens. In other words, the effect on the gender wage gap of different rewards between genders is relatively strong in the bottom and top of the wage distribution compared to the middle part of the wage distribution.

In the Netherlands, Albrecht, Vuuren, and Vroman (2008) extend the quantile regression decomposition method of Machado and Mata (2005) and establish counterfactual distributions with sample selection adjustments to study the gender wage gap. They believe that a sample selection caused by difference in employment rate is a serious issue to all related studies. In their results, they find the gender wage gap increases when moving up the distribution especially at the end. That is, this is evidence of a glass ceiling effect. In decomposition, most of the gap is accounted for differences in how men and women are rewarded.

As for a recent Canadian study on this topic, Boudarbat and Connolly (2013) decompose the gender wage gap by using the unconditional quantile regression method set forth by Firpo, Fortin, and Lemieux (2009), where the decomposition can be performed by a recentred influence function (RIF). They study

post-secondary graduates using the nice waves of statistics from the Canadaian National graduates survey through the years 1988 to 2007. In their results, they find mean wage gaps of over 6% two years after graduation and of more than 8% after five years. Men earn higher wages than women at every point in the distribution. In lower half of the distribution, the gap shrinks while it increases in the upper half.

3 Methodology

At first, a Blinder-Oaxaca decomposition is executed to break the gap into parts, one due to differences in average characteristics between genders (the explained part) and another part due to differences in returns of the various characteristics (the unexplained part). As usual, estimation is done by ordinary least squares (OLS):

$$Y_{mi} = X'_{mi}\beta_m + u_{mi}$$

for men and

$$Y_{wi} = X'_{wi}\beta_w + u_{wi}$$

for women. Where Y is the log hourly wage, X is a vector of characteristics and u is the error term. *i* represents the individual in the sample. The above can be rewritten as:

$$\bar{Y}_m - \bar{Y}_w = (\bar{X}_m - \bar{X}_w)\hat{\beta}_m + \bar{X}_w(\hat{\beta}_m - \hat{\beta}_w)$$

where $(\bar{X}_m - \bar{X}_w)\hat{\beta}_m$ is the explained part of the gap and $\bar{X}_w(\hat{\beta}_m - \hat{\beta}_w)$ is the unexplained part of the gap.

It is also important to consider decomposition along the whole distribution using counterfactual distributions regressions. The counterfactual distributions are the results of either, a change of covariates *X* that determine the outcome variable of interest *Y*, or a change of the conditional distribution *Y* given *X*. The methodology from Chernozhukov et al (2013) provides important new estimation and inference procedures for the entire marginal counterfactual distribution of *Y*.

In the case of gender wage discrimination, let Y_j denote hourly wage and X_j denote the labour market related characteristics that influence wage for population $j = 0\ or\ 1$, where 0 represent male samples and 1 represent female samples. $F_{Y_j|X_j}(y|x)$ indicates the stochastic assignment of wages to

workers with characteristics X in population $j = 0\ or\ 1$, where 0 represent male samples and 1 represent female samples. Let $F_{Y\langle j,j\rangle}(y)$ represents the observed distribution functions of wages where $j = 0\ or\ 1$, so that $F_{Y\langle 0,0\rangle}(y)$ represents male workers facing male wage schedule and $F_{Y\langle 1,1\rangle}(y)$ represents female workers facing female wage schedule. Therefore $F_{Y\langle 0,1\rangle}(y)$ shows the counterfactual distribution of wages that would have prevailed for female workers facing male wage schedule $F_{Y_0|X_0}(y|x)$. As in:

$$F_{Y\langle 0,1\rangle}(y) = \int_{\mathcal{X}_1} F_{Y_0|X_0}(y|x)dF_{X_1}(x)$$

It is well defined if the support of male characteristics, $\mathcal{X}_0$, includes the support of female characteristics $\mathcal{X}_1$, for example, $\mathcal{X}_1 \subseteq \mathcal{X}_0$. Similar to the Blinder-Oaxaca decomposition, $F_{Y(0,1)}(y)$ can take the form:

$$F_{Y\langle 1,1\rangle}(y) - F_{Y\langle 0,0\rangle}(y) = \left[F_{Y\langle 1,1\rangle}(y) - F_{Y\langle 0,1\rangle}(y)\right] + \left[F_{Y\langle 0,1\rangle}(y) - F_{Y\langle 0,0\rangle}(y)\right]$$

where $\left[F_{Y\langle 1,1\rangle}(y) - F_{Y\langle 0,1\rangle}(y)\right]$ represents the structural effect or, in the Blinder-Oaxaca decomposition the unexplained component. $\left[F_{Y\langle 0,1\rangle}(y) - F_{Y\langle 0,0\rangle}(y)\right]$ represents the composition effect or the explained component. The first term is the structural effect or discrimination and the second term is the composition or the endowment effect.

In more general cases, the setup of formal counterfactual distribution begins with looking at a population, namely $k \in \mathcal{K}$. For each population k there is a random d_x-vector X_k with support of $\mathcal{X}_k$ of covariates, as well as similarly a random outcome variable Y_k with support of $\mathcal{Y}_k$. The covariate vector is observable for all populations while the outcome is only observable in populations $j \in \mathcal{J} \subseteq \mathcal{K}$. Thus F_{x_k} can be identified from each population $k \in \mathcal{K}$, along with the conditional distribution $F_{Y_j|X_j}$ for all $j \in \mathcal{J}$. The counterfactual distribution and quantile functions of interest can be constructed by combining the conditional distribution in population j with the covariate distribution in population k:

$$F_{Y\langle j|k\rangle}(y) = \int_{\mathcal{X}_k} F_{Y_j|X_j}(y|x)dF_{X_k}(x)$$

$$Q_{Y\langle j|k\rangle}(\tau) = F^{-1}_{Y\langle j|k\rangle}(\tau)$$

where $F^{-1}_{Y\langle j|k\rangle}(\tau) = \inf\{y: F_{Y\langle j|k\rangle} \geq \tau\}$. The counterfactual distribution $F_{Y\langle j|k\rangle}$ is the distribution function of the counterfactual outcome $Y\langle j|k\rangle$ built by first sampling the covariate X_k from the distribution F_{X_k} and followed by sampling $Y\langle j|k\rangle$ from the conditional distribution $F_{Y_j|X_j}(\cdot\,|X_k)$.

The counterfactual distribution provides three types of counterfactual effects. Type 1 shows the effect

of changing the conditional distribution $F_{Y\langle j|k\rangle}(y) - F_{Y\langle i|k\rangle}(y)$. Type 2 shows the effect of changing the covariate distribution $F_{Y\langle j|k\rangle}(y) - F_{Y\langle j|m\rangle}(y)$. Type 3 shows the effect of changing the conditional and covariate distribution $F_{Y\langle j|k\rangle}(y) - F_{Y\langle l|m\rangle}(y)$. In the previous example of the gender wage gap, the wage structural effect is an example of type 1 counterfactual effects while the composition effect is an example of type 2 counterfactual effects.

With an assumption of conditional exogeneity, selection on the observations or unconfoundedness, counterfactual effects can be considered as causal effects. This means that the interpretation only works if group j is random as in:

$$\left(Y_j^*: j \in \mathcal{J}\right) \perp J|X$$

where Y_j^* is the potential outcome vector in j.

The estimation of counterfactual distributions involves three steps as mentioned above. The first stage is estimating the covariate distributions, where:

$$\hat{F}_{X_k}(x) = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbb{I}\left[X_{ki} < x\right]$$

as standard cumulative distribution. In the second stage, conditional distribution is estimated directly by:

$$\hat{F}_{Y_j|X_j}(y|x) = \Lambda\left(\mathrm{P(x)}'\widehat{\beta}_j(y)\right), \forall (y,x) \in \mathcal{Y}_j\mathcal{X}_j$$

so that:

$$\widehat{\beta}_j(y) = \arg\max_{b\in\Re d_p} \sum_{i=1}^{n_j} \{\mathbb{I}[\,Y_{ji} < y]\ln\left[\Lambda\left(\mathrm{P}\left(X_{ji}\right)'\mathrm{b}\right)\right] + \mathbb{I}[Y_{ji} > y]\ln\left[1 - \Lambda(X_{ji}'\mathrm{b})\right]\}$$

where Λ is some link function such as probit or logit estimated using a maximum likelihood estimation. $P(\cdot)$ is a transformation of X_j. With linear expression, $P(\cdot)$ is X_j itself, and $d_p = \dim(X_j)$.

In the third stage, counterfactual distribution is estimated:

$$\hat{F}_{Y\langle j|k\rangle}(y) = \int_{\mathcal{X}_k} \hat{F}_{Y_j|X_j}(y|x)d\hat{F}_{X_k}(x)$$

Finally, the decomposition results are showed in three categories of effect: total, structure and composition. Total effect refers to the total gender wage differential, which can be obtained by adding the structural effect and the composition effect. Structural effect refers to the price effect that the labour

market rewards for characteristics. Composition effect indicates the different characteristics between genders. Besides the above, any functional of the counterfactual distribution can also be estimated:

$$\hat{\Delta}(\omega) = \phi(\hat{F}_{Y\langle j|k\rangle} : (j,k) \in \mathcal{JK})(\omega)$$

and the unconditional counterfactual quantile function is given by:

$$\hat{Q}_{Y\langle j|k\rangle}(\tau) = \hat{F}^{r\leftarrow}_{Y\langle j|k\rangle}(\tau)$$

where $\hat{F}^{r\leftarrow}_{Y\langle j|k\rangle}$ is the rearrangement of $\hat{F}_{Y\langle j|k\rangle}$ if $\hat{F}_{Y\langle j|k\rangle}$ is not monotonic. Chernozhukov et al (2013) show that the counterfactuals and their smooth functionals are unbiased. In addition they also prove the validity of the exchangeable bootstrap. Considering that there are many binary dummy variables in the model, logistic regression is undoubtedly a more precise way of estimating counterfactual distributions between genders. At the same time, it is more direct in processing probability.

At last, counterfactual distribution regression results are hour weight adjusted where $hour\ weight = hours\ worked\ per\ week * statistical\ weight$ since some people work more than others making the distribution biased.

4 Data

This study uses the Canadian Labour Force Survey (LFS) for the data source. Developed after the Second World War, the LFS provides the Canadian government institutions with the most up-to-date and important employment and unemployment data. Each survey report is generated monthly and every observation is tracked for six months to provide long-term measures of the Canadian economy. The LFS is famous for providing unemployment status as well as other standard labour market indicators such as participation rate. Statistics about employees, hourly wage, union status, immigration status, and common personal characteristics are included as well. The LFS target population is 15 or older, lives nationwide (excluding persons living on reserves and other Aboriginal settlements in the provinces), and includes full-time members of the Canadian Armed Forces, the institutionalized population, and households in extremely remote areas with very low population density. When combined these particular reports of April 2018 and October 2018 contain 204,419 observations before dropping samples that are not relevant

to this study. Each observation represents an individual with its cross-sectional information.

One of the advantages the LFS has is actual wage value, instead of intervals. The large sample sizes provide enough room for applying controls and restrictions. The clear industry classification of the generic 10 category variable and the detailed 21 category variable is very convenient for creating controls within the model. Similarly it provides two options of 10 and 40 detailed categories for occupation.

I exclude samples of unemployed, unpaid and self-employed, as their earnings cannot be investigated from the LFS. All observations under age 20 are dropped as they are considered students, as well as because I wish to avoid conflating issues to do with gender. Among older workers there may be differential withdrawal from the labour force, as a result of, for example, planning for retirement. Thus, samples for those age 70 and above are dropped from the dataset. After dropping the data not related to this study, the total number of samples in this data set is 96,798, of which 48,570 are women and 48,228 are men.

Table 1 shows the summary statistics for both genders including all controls, where the average log hourly wage for males is 3.279 and for females is 3.146, thus leaving a log difference of 0.133. In this sample, women are somewhat more educated than men. For a bachelor's degree and above there is a 7.3% point difference between women and men. On the lower side of education level, there is a 3.3% point difference between women and men who are high school dropouts. This indicates that there are more women in the high education level and fewer women in the low education level.

In terms of industry, a great number of women (23.2%) are employed in the Health Care and Social Assistance industry, in contrast to only 4.6% for men. On the other side, 20.6% of men are in the Utility, Construction, Transportation, and Warehousing industry, while only 5.6% of women are in this industry. Looking at the occupations, 25.9% of women are concentrated in sales and services, another 25.9% are in administrative positions, and 17.8% of women work in education, law, and government jobs. In contrast, 20.1% of men are in sales and services, 10.1% are in administrative positions and only 7.5% are in education, law and government jobs. Similarly, 27% of men are engaged in the occupations of Trades, Transport, and Equipment Operators, compare to only 2% of women working in these fields. These

findings make my sample very consistent with the studies of Milgrom et al (2001) and Korkeamäki and Kyyrä (2006), which also show that women and men are concentrated in different industries and occupations.

It is worth noting that, while 91.7% of men are full-time workers, this number is only 79.3% for women, leaving a 12.4% difference. In the weekly working hours, men have an average of 35.6 hours per week, compared with 29.6 hours for women. In the case where the average of other control variables such as age, marital status, and geographical location are similar, women taking more responsibility in the family may be an explanation as mentioned above.

5 Covariates

This section introduces the description of covariates used in the estimations. Education has 4 dummy variables indicating the highest educational attainment for the individual *i*. They are "high school dropouts", "high school graduate and some postsecondary", "postsecondary diploma", and "bachelor's degree and above". The reference group is "high school graduate and some postsecondary". Industry includes 10 dummy variables combined from 21 categories derived from the LFS using the North American Industry Classification System Canada (NAICS). The reference group is manufacturing. Similarly, Occupation includes 10 dummy variables from 10 categories taken from the LFS using the National Occupational Classification (NOC) system. Reference group is manufacturing. There are 6 age groups, starting from age 20 to 29, and going to age 60 to 69. The reference group is age 20 to 29. Marital Status includes 4 categories, "married", "living in common law", "single never married" and "other". The reference group is "single never married". Provincial dummies control for 10 provinces in Canada, where Ontario is the reference group. Full-time is a binary variable that indicates whether if the individual is a full-time worker or part-time at their main job. Union Status indicates if the individual is a union member or covered by a union contract. Finally, immigrant as a binary variable tells whether the individual is an immigrant or not.

6 Empirical results

There are three specifications in this model, where Specification (1) excludes industry and occupation controls, Specification (2) includes 10 industry and occupation controls and Specification (3) pushes it to 21 industries controls and 40 occupation controls.

As a benchmark of the results from counterfactual distributions regressions, Table 2 and Table 3 show the average effects using the Blinder-Oaxaca decomposition for Specifications (1) and (2) respectively. All variables are grouped into few categories to get the impact of each category as a whole on the difference. Education contains all dummy variables of educational attainment except for the reference group; the same is true for the industry, occupation, and province categories. Personal characteristics include variables of age, marital status, full-time status, union status, and immigrant status. The mean log hourly wage for men is 3.279 and for women is 3.146, where a 0.132 log wage difference exists. In Specification (1), the entire gap is contributed to discrimination, with 0% of the contribution from the explained component. While all other variables are significant at 1% level, the explained component is not statistically significant at any level. After controlling for industry and occupation in Specification (2), 80% of the gap is contributed to discrimination and the remaining 20% is contributed to the explained part in a positive sign. Among the explained gap in Specification (2), education contributes to reducing the gap as expected. Industry and personal characteristics increase the gap. These key variables are highly statistically significant and their sign and pattern is consistent. Table 4 shows the Blinder-Oaxaca decomposition for Specification (3). The explained component increases from 20% to 32.6% and the unexplained decreases from 80% to 67.4% in Specification (2). Both Occupation in the explained part and Industry in the unexplained part are not statistically significant. Province under the explained category is significant at 10% level in all three specifications, and it is not economically important. All other figures are statistically significant in this section. The changes from (2) to (3) indicate that the influence of choice of industry on the gap is increased. This result of the Blinder-Oaxaca decomposition supports the theory from Korkeamäki and Kyyrä (2006) that women are concentrated in certain industries, leading to income disparities.

Table 5 and Table 6 show the quantile effects using counterfactual distributions regression on grid of 10 points in Specifications (1) and (2) respectively. These tables show that men earn higher wages than women throughout the distribution. The total effect column shows the total counterfactual gender wage gap, which can be obtained by adding structural effect to composition effect. As mentioned above, a structural effect is considered as the price effect or the discrimination itself, while the composition effect comes from the difference in characteristics. Looking at the estimated point for each quantile along the whole distribution, I can see that the structural effect provides more contribution to the gap than the total effect. At the same time, the sign of the composition effect is negative; in other words, personal characteristics contribute to reducing the gender wage gap even though the amount is relatively small compared to the structural effect. The difference between with and without industry and occupation controls is not significant if only looking at the table.

Figure 1 and Figure 2 provides a more accurate estimate and a more intuitive perspective for Specifications (1) and (2), since they are re-estimated on grid of 100 points. The y-axis shows the difference in log hourly wage between genders. It is very clear to see that the gender wage gap increases when moving up the distribution. When the gap reaches the upper part of the distribution, it begins to drop. In other words, the structural effect is contributing more to the gap going up the distribution. Similarly, the composition effect negatively contributes more to the gap in the upper part of the distribution, meaning the total gap starts to drop from the highest point in distribution. This indicates that personal characteristics and choice of industries reduce the gender wage gap more in the upper distribution compared to the middle and bottom. Comparing Specifications (1) and (2), it can be found that the industry and occupation controls smooth the composition and structural effects. The gender wage gap increases in the lower and middle part of the distribution when industry and occupation controls. At the same time, there is an increase in the lower part of the distribution in the composition effect, showing that after controlling for industry and occupation, personal characteristics not only reduce the gap in upper distribution, but also in the low and middle distributions. Still, Specification (2) shows that the gender wage gap is at the highest in the middle to upper parts of the distribution and most of the gap comes from

structural effects as in discrimination.

To test the limitation of the econometric specifications in this paper, I increase the industry controls from 10 categories to 21 using the standard of NAICS and occupation controls from 10 categories to 40 using the standard of NOC (Specification (3)). It is believed that detail categorized industry and occupation controls can effectively increase the accuracy of the model. Table 7 and Figure 3 shows the quantile effects for Specification (3). It is very interesting to see that on average the composition effect is contributing nothing to the total effect. Except for the slightly negative effect at the top of the distribution, the distribution from the bottom to the middle has only a small positive effect. The entire composition effect is floating up and down around the 0 lines and it is not statistically significant. This shows that under this specification, the structural effect represents the entire gender wage gap indicating pure discrimination. And in all three specifications, discrimination is higher between 20^{th} and 70^{th} quantiles.

Comparing the results from two estimation and decomposition methods in Specifications (1) and (2), the explained component in the Blinder-Oaxaca decomposition shows a positive sign to the gap, where the sign of composition effect in counterfactual distribution analysis is negative. In other words, the Blinder-Oaxaca decomposition indicates that individual characteristics and choice of industry increases the gap, but the counterfactual distribution shows that characteristics and choice industry reduce the gap. In Specification (3), the decomposition of counterfactual distribution completely denies the results of the Blinder-Oaxaca decomposition. When the latter tries to prove that the gender wage gap is derived from industry choices of women, the former says that the gap is 100% from discrimination. It is very interesting to see that the counterfactual distribution approach tells a completely different story of the gender wage gap. In fact, the existence of discrimination in labour economics has always been controversial. The result of this paper provides solid evidence to support discrimination is the cause of the gender wage gap.

Compared with recent studies, the results of this paper are not the same as those of Asplund and Napari (2011), Albrecht, Vuuren, and Vroman (2008), and Boudarbat and Connolly (2013). Although like other studies I have given similar evidence that most of the gap is caused by discrimination, my results

differ from other studies in that the gender wage gap rises sharply at the lower distribution, and drops sharply at the upper distribution. This indicates that most of the gaps originate from the middle of the distribution rather than the head and tail. Interestingly, this result shows no glass ceiling effect as mentioned in Arulampalam et al (2007) and Albrecht et al (2003) and as featured in all three studies above, especially the Canadian study of Boudarbat and Connolly (2013). By their theory, if there is a glass ceiling effect, wage of women will have a difficult time catching up with wage of men at the upper part of the distribution. This does not happen in my results using Canadian data, where the discrimination does not reduce the rewards of characteristics for those upper distribution female workers. Thus, it appears from the Canadian data that the glass ceiling is not a problem in female careers.

7 Conclusions

This paper uses the Canadian Labour Force survey of October and April 2018 to analyze the wage gap between genders. The counterfactual distribution approach developed by Chernozhukov et al (2013) is used to analyze the effect of various characteristics on the gap. A comparison between the Blinder-Oaxaca decomposition is imposed to identify the difference. In specifications, change of industry and occupation controls are imposed to see their influence on distribution.

The counterfactual distribution analysis shows that men earn higher wages than women throughout the distribution where the difference in log hourly wage increases when moving up the distribution and drops at the end of the distribution. In other words, most of the gaps originate from the middle of the distribution between 20^{th} and 70^{th} quantiles, rather than the head and tail. In the decomposition, the structural effect constitutes most of the gap, and the composition effect reduces the gap by a relatively small amount. When the industry and occupation control in the model increases, the composition effect becomes 0. That is to say, gender wage discrimination is completely derived from discrimination. A follow-up Blinder-Oaxaca decomposition shows controversy in the average effect. Among the explained, industry is the main component of the gap. This result of average decomposition supports the theory that women are concentrated in certain industries, leading to income disparities. When comparing these two

estimation methods, the decomposition of counterfactual distribution completely denies the results of the Blinder-Oaxaca decomposition. Comparing my results to Boudarbat and Connolly (2013), the observable personal characteristics in my decomposition cannot explain the wage gap. But theirs can by a small portion. The total wage gap in my results decreases at the top of the distribution where theirs remain high. Also, this paper finds no glass ceiling effect from the distribution when others do. Overall, the result of this paper provides solid evidence to support discrimination is the cause of the gender wage gap.

The econometric model has a problem of endogeneity since experience is not included in the covariates. Given that the relevant studies provide detailed insights into the impact of work experience on income, lack of these kinds of insights in this paper is problematic. This is mainly because the LFS has age in brackets but not exact numbers, which makes experience as a control variable difficult to compute using the equation $experience = age - school\ years - 6$. Similarly, the number of children is not added to the model because the LFS lacks detailed data. Considering the increasing cost of raising a child today, for example, including education costs and expenditure goods, the income of people with children, regardless of gender, will be greatly affected. Seasonal job adjustments can be another improvement to the model; however, it is not added since there are only 4.1% of men and 1.2% of women working in such industries in my data. Thus, they are considered a minority that will not influence the results.

The estimation results are biased because there exists a difference in the participation of men and women. Generally, it is more difficult for women to enter the labour market than men. It can be inferred that the distribution of workability of women in this dataset is better than the distribution of workability of men. In other words, discrimination from the gender wage gap should be bigger since female samples are a selected group of people who have better work abilities.

9 References

Arulampalam, W., Alison L. Booth., and Mark L. Bryan. (2007) "Is There a Glass Ceiling over Europe? Exploring the Gender Pay Gap across the Wages Distribution" *Industrial and Labour Relations Review* 60, 163-186

Albrecht, J., A. Björklund., and S. Vroman. (2003) "Is there a glass ceiling in Sweden?" *Journal of Labour Economics* 21, 145-177

Albrecht, J., A. Vuuren., and S. Vroman. (2008) "Counterfactual Distributions with Sample Selection Adjustments: Econometric Theory and an Application to the Netherlands" *Labour Economics* 16(4), 383-396

Asplund, R., S. Napari. (2011) "Intangibles and the Gender Wage Gap: An Analysis of Gender Wage Gaps Across Occupations in the Finnish Private Sector" *Journal of Labour Research* 32, 305-325

Baker, M., D. Benjamin., A. Cegep., and M. Grant. (1995) "The Distribution of the Male/Female Earnings Differential, 1970-1990" *Canadian Journal of Economics* 28(3), 479-501

Becker, G.S. (1985) "Human capital, effort, and the sexual division of labour" *Journal of Labour Economics* 3(1), S33-S58

Benjamin, D., M. Gunderson., and W. Riddell. (1998) *Labour Market Economics Theory, Evidence, and Policy in Canada* (Toronto: McGraw-Hill Ryerson)

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

Boudarbat, B., and M. Connolly. (2013) "The gender wage gap among recent post-secondary graduates in Canada: a distributional approach" *Canadian Journal of Economics* 46(3), 1037-1065

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

Chen, M., V. Chernozhukov., I. Fernández-val., and B. Melly. (2013) "Counterfactual: An R Package for Counterfactual Analysis"

Firpo, S., N. Fortin., and T. Lemieux. (2009) "Unconditional Quantile Regressions" *Econometrica* 77(3), 953-973

Kidd, M., and M. Shannon. (1996) “The Gender Wage Gap: A Comparison of Australia and Canada” *Industrial and Labour Relations Review* 49(4), 729-746

Korkeamäki, O., and T. Kyyrä. (2006) “A gender wage gap decomposition for matched employer-employee data” *Labour Economics* 13, 611-638

Loh, E. (1996) “Productivity Differences and the Marriage Wage Premium for White Males” *Journzal of Human Resources* 31(3), 566-89

Melly, B. (2005) “Decomposition of differences in distribution using quantile regression” *Labour Economics* 12, 577-590

Milgrom, E., T. Petersen., and V. Snartland. (2001) “Equal Pay for Equal Work? Evidence from Sweden and a Comparison with Norway and the U.S.” *Scandinavian Journal of Economics* 103(4), 559-583

Oaxaca, R. (1973) “Male-Female Wage Differentials in Urban Labour Markets” *International Economic Review* 14(3), 693-709

Waldfogel, J. (1997) “The Wage Effects of Children” *American Sociological Review* 62, 209-17

Waldfogel, J. (1998) “Understanding the "Family Gap” in Pay for Women with Children” *The Journal of Economic Perspectives*, 12(1), 137-156

Tables

Table 1: Summary Statistics

Variables	Male Mean	Female Mean
Log Hour Wage	3.279 (0.469)	3.146 (0.451)
A. Education		
High School Dropout	0.074 (0.263)	0.041 (0.199)
High School Graduate and Some Postsecondary	0.26 (0.438)	0.22 (0.414)
Postsecondary Certificate or Diploma	0.379 (0.485)	0.379 (0.485)
Bachelor's Degree and above	0.287 (0.452)	0.36 (0.48)
B. Industry		
Resources	0.041 (0.198)	0.012 (0.108)
Utility, Construction, Transportation and Warehousing	0.206 (0.405)	0.056 (0.229)
Manufacturing	0.155 (0.362)	0.061 (0.239)
Trade	0.152 (0.359)	0.147 (0.354)
Finance, Insurance and Real estate	0.052 (0.223)	0.073 (0.26)
Professional and Business Services	0.113 (0.317)	0.093 (0.29)
Education Services	0.051 (0.221)	0.118 (0.323)
Health Care and Social Assistance	0.046 (0.209)	0.232 (0.422)
Food Services	0.049 (0.215)	0.068 (0.251)
Information, Culture, Recreation and Other Services	0.134 (0.341)	0.141 (0.348)
C. Occupation		
Management	0.08 (0.271)	0.055 (0.228)
Business, Finance and Administration	0.101 (0.302)	0.259 (0.438)
Natural and Applied Sciences	0.13 (0.336)	0.041 (0.199)
Health	0.025 (0.156)	0.133 (0.34)
Education, Law and Social, Community and Gove	0.075 (0.263)	0.176 (0.381)
Art, Culture, Recreation and Sport	0.015 (0.123)	0.021 (0.142)
Sales and Service	0.201 (0.401)	0.259 (0.438)
Trades, Transport and Equipment Operators	0.269 (0.444)	0.02 (0.138)
Natural Resources, Agriculture	0.029 (0.168)	0.006 (0.076)
Manufacturing and Utilities	0.075 (0.263)	0.03 (0.17)

(Continued on next page)

Table 1: Summary Statistics (Continued)

Variables	Male Mean	Female Mean
D. Age		
20-29	0.23 (0.421)	0.228 (0.42)
30-39	0.247 (0.431)	0.237 (0.425)
40-49	0.221 (0.415)	0.225 (0.417)
50-59	0.212 (0.409)	0.223 (0.416)
60-69	0.09 (0.286)	0.087 (0.282)
E. Marital status		
Married	0.474 (0.499)	0.475 (0.499)
Common-Law	0.166 (0.372)	0.169 (0.375)
Other	0.061 (0.239)	0.102 (0.302)
Single Never Married	0.298 (0.458)	0.254 (0.436)
F. Province		
Ontario	0.382 (0.486)	0.392 (0.488)
Alberta	0.127 (0.333)	0.119 (0.324)
British Columbia	0.127 (0.333)	0.131 (0.337)
Manitoba	0.035 (0.184)	0.035 (0.183)
New Brunswick	0.02 (0.139)	0.02 (0.141)
Newfoundland and Labrador	0.013 (0.113)	0.013 (0.115)
Nova Scotia	0.024 (0.154)	0.026 (0.159)
Prince Edward Island	0.004 (0.061)	0.004 (0.065)
Quebec	0.238 (0.426)	0.231 (0.421)
Saskatchewan	0.03 (0.171)	0.029 (0.168)
Full-time	0.917 (0.275)	0.793 (0.405)
Hours	35.601 (13.86)	29.564 (13.932)
Union	0.291 (0.454)	0.34 (0.474)
Immigrant	0.245 (0.43)	0.246 (0.431)
Observations	48,228	48,570

Notes. All means are weighted. Standard deviations are in brackets.

Table 2: Blinder-Oaxaca Decomposition Specification (1)

Variables	Coefficients
Overall	
Mean Log Hourly Wage Male	3.279*** (0.003)
Mean Log Hourly Wage Female	3.146*** (0.003)
Difference	0.132*** (0.004)
Explained	0 (0.003)
Unexplained	0.132*** (0.004)
Explained	
Education	-0.027*** (0.001)
Personal Characteristics	0.026*** (0.002)
Province	0.001* (0.001)
Unexplained	
Education	-0.069*** (0.012)
Personal Characteristics	0.092*** (0.011)
Province	0.022*** (0.005)
Constant	0.088*** (0.018)

Notes. Standard errors are in brackets. All coefficients are weighted. * Significant at 10% level, ** Significant at 5% level, *** Significant at 1% level.

Table 3: Blinder-Oaxaca Decomposition Specification (2)

Variables	Coefficients
Overall	
Mean Log Hourly Wage Male	3.279*** (0.003)
Mean Log Hourly Wage Female	3.146*** (0.003)
Difference	0.132*** (0.004)
Explained	0.027*** (0.004)
Unexplained	0.105*** (0.005)
Explained	
Education	-0.015*** (0.001)
Industry	0.042*** (0.003)
Occupation	-0.014*** (0.004)
Personal Characteristics	0.012*** (0.002)
Province	0.001* (0.001)
Unexplained	
Education	-0.018* (0.011)
Industry	0.029** (0.015)
Occupation	-0.15*** (0.019)
Personal Characteristics	0.067*** (0.011)
Province	0.02*** (0.004)
Constant	0.157*** (0.02)

Notes. Standard errors are in brackets. All coefficients are weighted. * Significant at 10% level, ** significant at 5% level, *** significant at 1% level.

Table 4: Blinder-Oaxaca Decomposition Specification (3)

Variables	Coefficients
Overall	
Mean Log Hourly Wage Male	3.279*** (0.003)
Mean Log Hourly Wage Female	3.146*** (0.003)
Difference	0.132*** (0.004)
Explained	0.043*** (0.004)
Unexplained	0.089*** (0.004)
Explained	
Education	-0.007*** (0.001)
Industry	0.042*** (0.004)
Occupation	0.001 (0.004)
Personal Characteristics	0.005*** (0.002)
Province	0.001* (0.001)
Unexplained	
Education	-0.025** (0.011)
Industry	0.021 (0.017)
Occupation	-0.058** (0.028)
Personal Characteristics	0.061*** (0.01)
Province	0.019*** (0.004)
Constant	0.073** (0.03)

Notes. Standard errors are in brackets. All coefficients are weighted. * Significant at 10% level, ** Significant at 5% level, *** Significant at 1% level. Industry reference group is manufacturing in durable goods. Occupation reference group is manufacturing assemblers.

Table 5: Quantile Effects Specification (1)

Quantile	Structure	Composition	Total
0.1	0.063(0.012)	0.000(0.016)	0.063(0.017)
0.2	0.173(0.014)	-0.026(0.000)	0.147(0.014)
0.3	0.159(0.000)	-0.046(0.005)	0.113(0.005)
0.4	0.152(0.011)	0.000(0.010)	0.152(0.002)
0.5	0.188(0.003)	-0.036(0.003)	0.150(0.003)
0.6	0.182(0.008)	-0.032(0.008)	0.163(0.006)
0.7	0.193(0.008)	-0.030(0.006)	0.163(0.006)
0.8	0.180(0.004)	-0.042(0.003)	0.138(0.006)
0.9	0.140(0.003)	-0.031(0.006)	0.109(0.005)

Notes. Standard errors are in brackets. Conditional model is logistic. A total number of 100 regressions are estimated. The variance has been estimated by bootstrapping the results 100 times. Reference group is female. Counterfactual group is male.

Table 6: Quantile Effects Specification (2)

Quantile	Structure	Composition	Total
0.1	0.063(0.017)	0.000(0.022)	0.063(0.017)
0.2	0.201(0.014)	-0.054(0.000)	0.147(0.014)
0.3	0.183(0.010)	-0.070(0.011)	0.113(0.005)
0.4	0.193(0.006)	-0.041(0.007)	0.152(0.006)
0.5	0.188(0.011)	-0.036(0.011)	0.152(0.002)
0.6	0.182(0.011)	-0.032(0.011)	0.150(0.003)
0.7	0.177(0.010)	-0.014(0.007)	0.163(0.006)
0.8	0.180(0.010)	-0.042(0.010)	0.138(0.006)
0.9	0.140(0.010)	-0.031(0.009)	0.109(0.098)

Notes. Standard errors are in brackets. Conditional model is logistic. A total number of 100 regressions are estimated. The variance has been estimated by bootstrapping the results 100 times. Reference group is female. Counterfactual group is male.

Table 7: Quantile Effects Specification (3)

Quantile	Structure	Composition	Total
0.1	0.063(0.005)	0.000(0.018)	0.064(0.018)
0.2	0.121(0.017)	0.026(0.012)	0.147(0.014)
0.3	0.113(0.011)	0.000(0.012)	0.113(0.005)
0.4	0.152(0.014)	0.000(0.014)	0.152(0.007)
0.5	0.152(0.006)	0.000(0.006)	0.152(0.003)
0.6	0.132(0.009)	0.018(0.009)	0.150(0.003)
0.7	0.145(0.011)	0.018(0.011)	0.163(0.007)
0.8	0.138(0.009)	0.000(0.009)	0.138(0.005)
0.9	0.109(0.009)	0.000(0.009)	0.109(0.005)

Notes. Standard errors are in brackets. Conditional model is logistic. A total number of 100 regressions are estimated. The variance has been estimated by bootstrapping the results 100 times. Reference group is female. Counterfactual group is male.

Figures

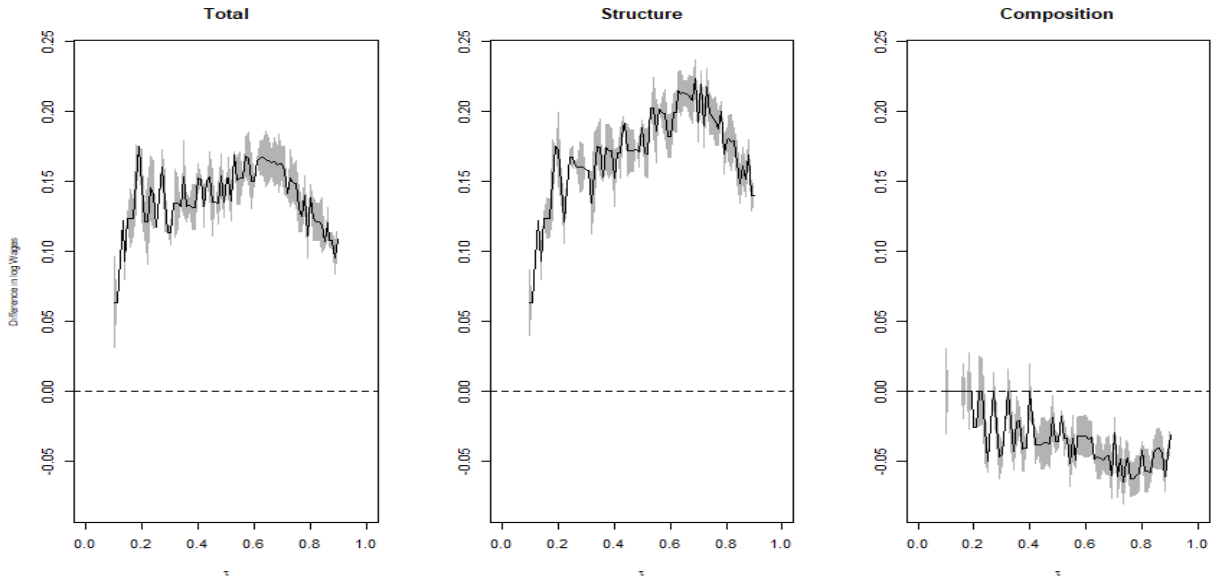

Fig.1 Decomposition of gender wage gaps for Specification (1). Figure is hour weight adjusted. Quantile effect functions estimated on grid of 100 points.

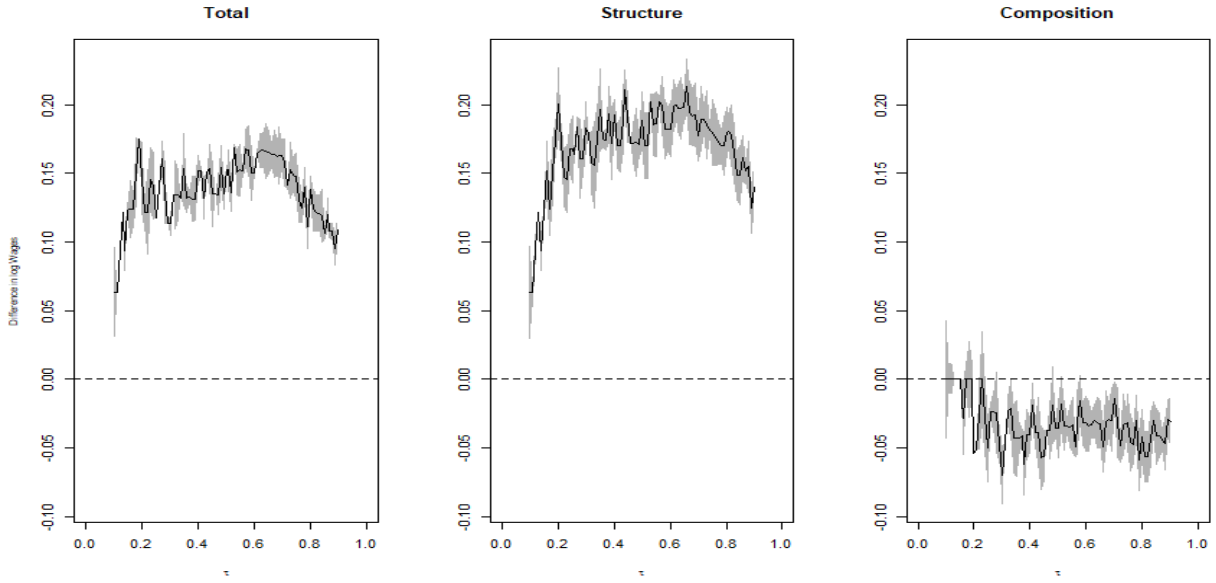

Fig.2 Decomposition of gender wage gap for Specification (2). Figure is hour weight adjusted. Quantile effect functions estimated on grid of 100 points.

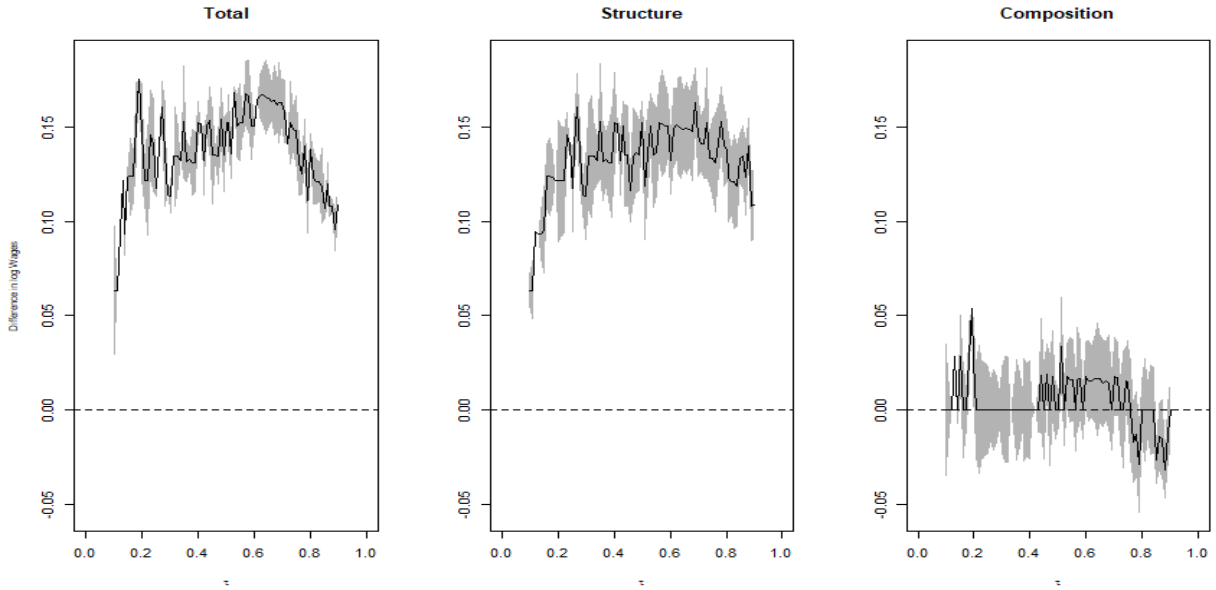

Fig.3 Decomposition of gender wage gap for Specification (3). Figure is hour weight adjusted. Quantile effect functions estimated on grid of 100 points.