THE CANADIAN GENDER WAGE GAP: SELECTION BIAS, HETEROGENEITY, AND MATCHING

SHENGYI ZHU

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADURATE PROGRAM IN ECONOMICS
YORK UNIVERSITY
TORONTO, ONTARIO

July 2023

Abstract

This dissertation provides a comprehensive analysis of the Canadian gender wage gap over the past two decades, employing modern methodologies and tools.

In the first chapter, selection bias and its impact on the entire earning distribution are examined. A selection-corrected quantile regression is utilized to provide a more accurate depiction of the gender wage gap distribution. The simulation of potential government child care benefits as an instrument helps address the selection bias issue. Findings reveal the persistent but inconsistent effects of selection bias across wage quantiles and time. The presence of negative selection for women entering the workforce is identified, and the absence of this bias would result in an even higher unexplained portion of the gender wage gap.

Moving to the second chapter, a thorough investigation of the heterogeneity of the gender wage gap is conducted. High-dimensional models are employed to explore the diverse factors contributing to the wage gap. Advanced machine learning algorithms are utilized as robustness checks to address possible multicollinearity problems. The analysis reveals significant reductions in the gender wage gap attributed to age and several occupations, while penalties related to family structure persist.

Finally, the third chapter explores the under-researched area of job-education mismatch and its impact on the gender wage gap. The study focuses on the differences in vertical and horizontal matching between women and men. Self-reported measurements of both vertical and horizontal mismatch, as well as an objective index of horizontal mismatch, are utilized. Results indicate that, unlike other countries, vertical mismatch does not contribute significantly to the gender wage gap. Furthermore, the role of horizontal mismatch is economically insignificant in relation to the overall gap.

This dissertation enhances our understanding of the Canadian gender wage gap by addressing important aspects such as selection bias, heterogeneity, and job-education mismatch. The findings contribute to the existing literature on gender inequality, offering valuable insights for policy interventions and future research in this field.

Contents

List of Tables

List of Figures

Chapter 1

Selection Corrected Gender Gap between Earnings Distributions in Canada[1]

1.1 Introduction

The gender wage gap is a topic that has been extensively studied in labor economics but it still remains an area of active research.[2] However, much of the discussion surrounding the gender wage gap has been around the mean pay differences. While there has been remarkable improvements in women's real wage, in both absolute term and relative to men, the gains are not consistent between income quantiles (see Table 1.1). Relying solely on the mean oversimplifies the issue of the gender wage gap; the gap's magnitude and its evolution over time can vary significantly across wage levels. Delving deeper beyond the mean reveals variations that can have implications for policies that targets specific demographics, such as minimum wage adjustments or childcare benefits. Moreover, changes in participation and the composition of the workforce means that addressing selection bias is a critical consideration when estimating gender differences in pay; biases can arise from systematic differences between active and inactive workers, distorting measurements of the gender wage gap. Given the substantial differences in labour participation between income strata, (see Figure 1.2), the effect of selection may

[1] This paper is a collaboration with Prof. Ben Sand

[2] For a history of gender wage gap in industrialized countries, see Olivetti and Petrongolo (2016a).

not be consistent across different wage levels. While some studies have attempted to address this problem by examining select wage percentiles (ex Casey B. Mulligan and Rubinstein (2008b)), few have done so over the entire wage distribution. This paper seeks to offer a more comprehensive understanding of the Canadian gap using selection corrected quantile regression, as proposed by Arellano and Bonhomme (2017). Our analysis will examine the size and changes in the gender gap across the entire earnings distribution, accounting for human capital characteristics and the non-random selection of men and women into the labor market.

The inadequacy of mean wage differentials in understanding the gender wage gap has been highlighted in several recent papers, both internationally (Francine D. Blau and Kahn 2017b; Kunze 2018) and within Canada (Baker and Drolet 2010a; Schirle 2015). For instance, Antonie et al. (2016) documents that, in Canada, the gender wage gap varies by significantly by observable characteristics, such as age, education, and the amount of labour supplied. Similarly, in the US, Bach, Chernozhukov, and Spindler (2018) document significant heterogeneity in the gender pay gap along observable characteristics. In another strand of literature, earnings differentials between men women at different points of the earnings distribution are examined (Maasoumi and Wang 2019; N. M. Fortin 2019; Francine D. Blau and Kahn 2017b). Our paper is closely related to these papers and finds clear differences in the wage gap between wage quantiles.

Figure 1.1 shows the gender wage gap at at various percentiles of the income distribution (the dashed line).[3] The graph shows clear heterogeneity in the wage gap across wage quantiles. In particular, in 2002 the wage-gap profile is U-shaped, with the gap being lower at the extreme ends and larger in the middle of the earnings distribution. The largest gender gap occurred at around the $15th$ percentile of the wage distribution, bottoming out at 26.6%. Fourteen years later, in 2016, the gap narrowed across all wage percentiles. The most significant improvements are seen in at the lower end of the wage distribution which becomes more limited as income increases. In 2016, wage-gap profile becomes more J-shaped, with the lowest earning individuals now having the lowest differences in wage. This finding is consistent with the glass ceiling effect found in other papers (Atkinson, Casarico, and Voitchovsky

[3]Our data comes from the Labour Force Survey from 2002-2016, and our extract contains wage information wage and salary workers between the ages of 25-54 (inclusive). More details can be found in our data section [here].

2018; Francine D. Blau and Kahn 2017b; N. M. Fortin, Bell, and Böhm 2017; Antonie et al. 2016).

Finally, a recent development the gender wage gap literature is correcting for selection bias when examining the gap at a distributional level. If men and women have different labour force participation rates, and self-select into the workforce based on productivity, the classical Heckman (1979) model shows that comparing mean earnings differences between men and women will be biased (such as the case in Boll et al. (2016)). In addition, the direction of the bias cannot be assumed; for example, Casey B. Mulligan and Rubinstein (2008b) has argued that women's selection into the workforce changes overtime. In particular, these authors argue that in the early 1980s women where negatively selected into the labour force, but increases in the price of skill reverses the self-selection of women into the labour force. Furthermore, the gender gap in employment and labour participation have narrowed over time, following the trend of the wage gap.

Figure 1.2 shows the average participation rate for men and women in 2002 and 2016. We predicted (non-log) hourly wage for all individuals using OLS and binned them based on wage quantiles.[4]. The participation rate shown in figure 1.2 is the average form each predicted bin. For both men and women, labour participation rate generally increases with higher wage. However, for women, participation rate plateaus after around 85% while men have near monotonic increase, albeit at a slower rate in the upper income brackets. Women near the bottom of the income distribution have strikingly low participation rate, even when compared to other similarly low income men. Participation rate quickly rises towards the plateau after a small increase in income.

In 2002, participation rate gap is the highest in the lowest wage quantiles (figure 1.1), mainly caused by low participation among low income women. The gap decreased by 2016, but part of it was caused by a drop in participation rate from men. The lowest gap occurred between 15^{th} and 25^{th} percentile, and by 2016, this expanded to the 45^{th}, with women having the same or even higher participation rate than men in a select few quantiles. Participation rate

[4]The actual bins are : (0:0.5, 0.5:0.1...0.95:1), we took the median points of these bins as the wage quantiles used in the graph. Lines shown in graphs have been smoothed. Predicted wages serve as a skill index so that we can compare men and women within the same predicted wage bins. This allows us to examine potential differential selection issues at different points along the earnings distribution.

gap in the upper half of the wage distribution remain relatively even, with the exception of increase at the highest income level. For both years, participation rate gap is concave shaped compared to the convex shape of the wage gap. In 2002, the wage gap is roughly diametrically opposed to the participation gap and it is at its highest when the participation gap is at its lowest. This phenomenon persists to 2016 but to a lesser degree. The differential labour participation rates across skill groups and time indicates potential differential selection bias, which we will seek to examine in this paper. Addressing selection bias quantile regressions is still an active area of research. The first practical method for selection corrected quantile regression is Arellano and Bonhomme (2017) (A&B), which presents the selection problem as biases in the quantiles, and in its absence, becomes a regular quantile regression. Maasoumi and Wang (2019) uses A&B's method in their analysis of the U.S gender wage gap from 1976 to 2013. It found that the magnitude and direction of the effect of selection are different across time and different wage quantiles. Chernozhukov et al. (2018a) is a follow up to Arellano and Bonhomme (2017), which does away with its requirement for a continuous instrument. However, we prefer A&B because conceptually, it is a natural extension to quantile regression, and it is relatively simple to implement and compare. It is combination of Heckman selection and quantile regression by replacing the τ in quantile regressions with one that is corrected for selection. Like Heckman, the method uses probit (or logit) to estimate propensity scores, which is correlated with τ. The copula (joint density) of *tau* and the propensity scores is assumed to have a specific form but the parameters of it are estimated in conjunction with the quantile regression in a "rotated quantile regression". Selection's effect on the outcome is two fold. It changes the distribution of the covariates which will also distort the estimator, β. β is corrected by A&B's quantile selection method while the distribution changes are fixed during decomposition. Our choice of instrument is individual's estimated potential provincial child care benefits, calculated using the tax calculator from Milligan and Stabile (2011). As child rearing is an important factor on women's reservation wage (Keane, Todd, and Wolpin (2011)) and negatively impact Canadian women's labour participation (Latif (2006), Powell (2002)), government policies that targets it can have a significant impact on women's labour participation decision (Lefebvre and Merrigan (2008)), but are plausibly unrelated to potential earnings.

Decomposing the gender wage gap into the part that can be attributed to differences

in characteristics (the explained) and the part is associated differences in the coefficients (often interpreted as discrimination) is important for understanding the proximate causes of gender wage differentials and policies to address gender wage gaps. An important component of decomposition is the counterfactual: what would women earn if they are treated the same as men? The popular Blinder-Oaxaca decomposition, independently developed by Blinder (1973) and Oaxaca (1973), is the decomposition method used in many gender wage gap studies. Unfortunately, it is not suitable for quantile regression decomposition as it is a mean decomposition. There are three main branches of methods to decompose quantile regressions and estimate the counterfactual distribution (see N. Fortin, Lemieux, and Firpo (2011) for details on decomposition methods). The first general method replaces a group (example, women), with a counterfactual of the other group that lies on the same quantile (see Juhn, Murphy, and Pierce (1993),Machado and Mata (2005),David, Katz, and Kearney (2005)). The counterfactual distribution is simply the integral of the conditional distribution of one group's outcome (example, women's wages) over the distribution of the covariates of the other. The second method reweighs observed distribution and uses that as the counterfactual (DiNardo, Fortin, and Lemieux (1995),Firpo, Fortin, and Lemieux (2009)). Finally, the third approach is similar to the first, except the counterfactual distribution is directly estimated (see N. M. Fortin and Lemieux (1998), Chernozhukov, Fernández-Val, and Melly (2013)). We chose the second approach, also used in A&B, as our preferred decomposition method as it can be easily slotted into A&B's quantile selection's output without too much alteration. The decomposition approach examined here will help shed light on the sources of such heterogeneity in both the levels and changes in the gap.

In our results, we have several interesting findings. First, we document that the gender wage gap varies substantially across wage quantiles. The gender gap varies along the earnings distribution in a 'check mark' shaped pattern where the largest gap is skewed toward the lower half of the income distribution. The gap decreased across all income levels but the largest gain is seen in the lowest income brackets. However, by 2016, the general 'check mark' shape persists, albeit flatter and with the nadir moved towards the middle quantiles. Second, we find that when when selection is accounted for, the unexplained gender wage gap would be larger for the first 3 quarters of the wage distribution, and lower for the highest quarter. This remains

true for 2002 and 2016, although the effect is smaller in 2016. Finally, we find that selection decreases the observed wage gap decreases for most quantiles. This effects is lessened in 2016 and not as consistent compared to 2002. This is caused by an uptick in the explained wage gap, which is positive rather than negative. This suggests that women would earn more than men when looking solely at personal characteristics (as women, on average, have higher educational attainment compared to men). Since selection increases this gap, this suggests women are negatively selected into the workforce. That is, women who are less 'skilled' tend to be more likely to participate in the labour market.

Our roadmap for this paper is as follows. We will first show the existence of selection bias through our summary statistics and logit. Next we will show the impact of selection on the gender wage gap through its effect on the coefficients. We do this by comparing the outputs and decomposition of OLS, standard quantile regression, and quantile selection regression. Then, we will repeat this for the covariates. Finally, we will show the selection's entire effect by adding both effects together.

1.2 Literature Review

1.2.1 The gender gap in Canada

Since around late 1980s to early 1990s, the rate of convergence has slowed(eg. Francine D. Blau and Kahn (2007), Baker and Drolet (2010a)). A substantial part (almost two thirds) of the gap still remained unexplained. Schirle (2015) provides a look into the differences in the wage gap between the provinces (from 1997 to 2014). All provinces showed substantial change except for Alberta and Newfoundland. A large portion of the differences can be explained by "individual and job characteristics". Education also played an important role. Boudarbat, Lemieux, and Riddell (2010) focused on the effect of education on the gender gap, especially higher education, along the entire wage quantiles. Returns on education has increased over the years, significantly for men and more moderately for women. While both men and women are more educated than ever, the proportion of women with at least bachelor's degree is now higher than men. Even with relatively lower gains in returns, the larger increase in women's education ultimately

lowered the wage gap. Antonie et al. (2016) examined the heterogeneity of the of gap by various demographics and income. It found that the wage gap increases with age and in part time jobs. It decreases with more education, except for the top 10% earners. It also found evidence of selection bias; women are positively selected into part time jobs. Furthermore, men are more likely to be promoted while women are less likely to work, and work less, when there's children in their care. However, unlike our paper, Antonie et al. (2016) did not correct for selection bias in their analysis and treated their quantile problem as several income cohorts rather than a whole.

1.2.2 The gender gap in an international context

Internationally, the gender wage gap follows similar patterns. Boll et al. (2016) is a study of 23 European nations for the year 2010. All 23 countries have unexplained gender wage gap, which constitutes the largest portion of the observed gap. The magnitude varies between countries; from 5.8% in Belgium to 14.9% in Estonia (Estonia also has the highest gap in all OCED countries). Several have negative explained gap, meaning women have higher wage earning 'abilities' and would earn higher if paid like men. These countries have lower women labour participation rate. It indicates that only those with higher wage potential enters the workforce, and thus a source of selection bias. Francine D. Blau and Kahn (2017b) is a U.S study and have the same general result. It also examined the effect of psychological attributes or non-cognitive skills and found it had only "small to moderate" impact of the wage gap.

The convergence of the wage gap, and its slow down, is not limited to countries with European roots. Hara (2018) looked at the Japanese labour market using a distributional approach from 1980 to 2015. In addition to the usual results, the paper noted while women with middling income fared well, there exists both a sticky floor and glass ceiling effect on women's wage. While the sticky floor effect diminished over time, the glass ceiling effect became more pronounced. While it became easier for women to be promoted, the return on promotion has declined. This phenomenon has also been observed in the UK, U.S and Canada(Francine D. Blau and Kahn (2017b),N. M. Fortin, Bell, and Böhm (2017)). Jong-Wha and Wie (2017) is a study of the gender wage gap in China and India. Like more developed economies, both countries saw increased inequality and increased premium on high skilled labour. Increase in

women's educational attainment help reduce the wage gap in both countries. However, while the gap decreased in India, the opposite occurred in China. This is partly caused by the relative large increase in skill premium in China, which has overwhelmed the effect of the decrease in the education gap. Unexplained gap for both countries remained larger than developed ones, however, while India's decreased over two decades, China's unexplained gap increased, signifying either decrease of women's unobserved characteristics or increased discrimination.

1.3 Data

Data used for this paper comes from the Canadian Labour Force Survey (LFS), years 2002 and 2016. It is a monthly survey containing key employment related statistics relevant to measuring the state of the labour market. We chose it as it is the best, widely available source of hourly wage data for Canada. The survey is conducted in over 54,000 households across Canada (excluding the territories) for those that are 15 years or older. Once selected by the survey, respondents are interviewed monthly over 6 months. Therefore, to avoid repetition of individuals, we used only 2 months (May and November) 6 months apart for each year. A drawback of this data set is that variables' factor levels are binned into buckets, including continuous variables. This rules out using higher dimensions of age and education as variables. Another draw back is the lack of data on number of children.

Selection qualification is not completely random. To ensure a sizable sample for each cohorts of the population, those that are more represented are less likely to be selected. For example, there are significantly less Ontarians in the survey relative to Ontario's proportion of Canada's population. All our results have been weighted to be representative of the original population. 2002 was selected as the beginning year because our instrument requires variation in the provincial benefits. Wage data were added to the LFS in 1997, however, government benefits, especially those pertaining to child care, were lacking. Expansion began in 1998, with all provinces having some provincial child care supplement schemes. By 2001, most provincial child care benefit schemes are similar to 2016, the last year available to us in the tax calculator.

Those that are self-employed do not have wage data and are excluded from our data set. Wages are hourly wage; for salaried workers, Statistics Canada imputed their hourly wage by dividing salary by the usual hours of work during the appropriate time period. We then turned the nominal wage into log real wage using CPI data provided by Statistics Canada with the year 2002 as the baseline. Because self reported wages tend to pool towards whole numbers, to simulate wages as continuous data, we added an error term, $\epsilon \sim \mathcal{N}(0, \frac{mean(wage)}{10,000})$, to each individual's wage observation. Finally, we restricted the data to those that are in the prime working age of 25 to 54. Those that fall outside of those age brackets are significantly less likely to work or receive child care related benefits. [5]

Table 1.1 and Table 1.2 shows the mean real wage and labour participation rate by various demographic factors in 2002 and 2016. In both time periods, men consistently earns more than women in all demographic groups and wage quantiles. However, the gap narrowed over the years, with men earning on average 17.69% more than women in 2002 compared to 14.74% in 2016. Wage of both men and women grew for nearly all demographic and wage levels between 2002 and 2016, with only exceptions being widowed and P.E.I men. Women's wage saw more significant improvements relative to men which is the main driver of the decline in the wage gap. Women's wage growth is higher and more consistent than men's for all quantiles. Although the higher quantiles have nearly the same growth as men, while men in the lower half of the income brackets saw relatively little growth. Since women saw relatively higher growth for lower income groups, the wage gap declined more substantially in those brackets and barely changed for the highest. Table 1.3 shows the unconditional mean real wage of various demographics when the sample is restricted to the lowest 25% of the wage distribution. While gains between the two time periods varies by different demographic groups, no group saw a decline in real wage. This is in contrast to the phenomenon seen in the U.S where low education and income males saw a decline in real wage over time (Binder and Bound (2019)).

Like wage, labour participation rate is higher for men relative to women for all demographics for both years. The gap declined during the two time periods as women's participation rate increased for nearly all demographics, with the exception of low educated

[5] A1 shows the levels of our factor variables. The first factor level is the reference.

women, while men's overall participation rate stagnated. Women's participation rate increases with education level and decreases for married, and having at least one child 3 - 12. On the other hand, men's labour participation rate does not increase much for those that with at least high school degree, and increases with marriage and young child at home. Labour participation gap is the lowest at the lower-middle income levels and the highest at the extreme ends. It mirrors the wage gap, with participation gap low when the wage gap is high, and low when the wage gap is high.

1.4 Methodology

1.4.1 Standard Quantile Regression

In the absence of selection bias, the canonical form of the quantile regression for the estimation of $\tilde{\beta_\tau}$ for quantile τ as described by Koenker and Bassett Jr (1978) is:

$$\tilde{\beta_\tau} = \underset{b}{\operatorname{argmin}} \Big\{ \sum_{Y_i \geq x_i b} (\tau)|Y_i - x_i b_\tau| + \sum_{Y_i < x_i b} (1-\tau)|Y_i - x_i b_\tau| \Big\}, \tag{1.1}$$

where Y is log real wage, x is a set of observable characteristics, and i = 1,2,...,n denotes individual observations. For X, we used age, marital status, education, province, metropolitan area, economic family size, and age of youngest child. The analogous comparison to ordinary least square (OLS) for quantile regression occurs when $\tau = 0.5$. However, as opposed to OLS, which minimizes at the mean, if $\tau = 0.5$, quantile regression minimizes deviations from the median. By construction, the mean of the residual is not zero. Instead, the ratio of positive and negative residuals is equal to $\tau/(1-\tau)$. The effect of X on wage (Y) is allowed to vary across different τ's. The regression coefficient β_τ can be interpreted as the effect of X at the τ^{th} quantile of wage. In practice, β can be solved using linear programming [6].

For this estimate to be unbiased, we must make several strong implicit assumptions. First, those that chooses not to work, if they were to enter the labour force, it would not

[6] See (Koenker et al. 2012) for more details

affect the wage quantile rankings of those that already work (not accounting for the effect of supply/demand change). Additionally, we must also assume that those that didn't work have the same characteristics as those that do work on average. Finally, if those that didn't work were to work, then they would face the same pay structure, (β), as those that didn't work.

1.4.2 Decomposition

For a meaningful analysis of the wage gap, we must also answer the question "What would women's wage be if they have the same characteristics as men". Differences in wages between men and women could be caused by differences in characteristics (explained) and differences in returns from those characteristics (unexplained). The latter can be interpreted as the effect of discrimination and other unobserved variables. We start with the Blinder-Oaxaca decomposition, which can decompose at the mean. The method begin by running a regression of wage on characteristics for female and male separately:

$$Y_{0i} = X_{0i}\beta_0 + u_{0i}$$
$$Y_{1i} = X_{1i}\beta_1 + u_{1i}$$

Where 0 denotes female and 1 denotes male. Using the outputs from these regressions, the mean gender wage gap can be written as:

$$\begin{aligned}
&\overline{Y}_0 - \overline{Y}_1 \\
&= \beta_0\overline{X}_0 - \beta_1\overline{X}_1 \\
&= (\beta_0\overline{X}_0 - \beta_1\overline{X}_0) + (\beta_1\overline{X}_0 - \beta_1\overline{X}_1)
\end{aligned}$$

where "$\overline{\ }$" denotes mean. For discontinuous variables, first one-hot encode the factor levels as separate variables. Remove one from each category to prevent perfect colinearity, then take the mean of each of the new variables as $\bar{X}$. The first part of the above equation, $(\beta_0\overline{X}_0 - \beta_1\overline{X}_0)$, is the unexplained portion of the mean gender wage gap. It is part of the

gap that is caused by differences in the pay structure between male and female. This is the part that is usually considered 'discrimination'. The latter half of the decomposition, $(\beta_1\overline{X}_0 - \beta_1\overline{X}_1)$, is the explained component. It is part of the wage game stemming from the differences in mean characteristics male and female. We choose not to include occupation in our regressions because gender differences in occupation is part of discrimination. We do not seek to untangle the effect of discrimination in hiring practices from occupational choice. Furthermore, individual's choice of occupation is affected by, and is a product of, systematic discrimination and biases throughout one's life, which may be impossible to truly account for.

While Blinder-Oaxaca decomposition is useful in setting a baseline, it is unsuited for quantile regressions as it cannot capture changes in the gender wage gap along the wage quantiles. We use the modified Machado and Mata (2005) used in A&B to decompose our results. For $m = 1, ...19$, we first estimated $\hat{\beta}_m$ with the desired quantile regression for quantiles $0.05, ...0.95$. Wage is then $\hat{Y}_{jjmi} = x_{ji}\hat{\beta}_{mj}$ where the subscript $j = 0, 1$ indicates female and male respectively. The counterfactual female wage, which is what women would earn given men's pay structure, is $Y_{01mi} = x_{0i}\hat{\beta}_{1m}$. Every individual would receive an estimated wage for each desired quantile, which can be interpreted as what would that individual make given the pay structure of quantile m. To find the fitted wage for the whole sample at quantile m, we sorted every estimates for each quantile and individual into a single vector, and then took the m^{th} quantile as Y_{jjm}.

1.4.3 Selection Correction

Under Heckman correction, if there is selection bias in the workforce, then a naive regression will result in biased estimators. Therefore, we will first correct for the bias using the model and method as described in A&B. The sample selection problem is the same as the one used in Heckman correction. Wage is only observed if the individual chooses to work:

$$Y^* = X'\beta + U \tag{1.2}$$

$$D = 1\{V \leq p(Z)\} \tag{1.3}$$

$$Y = Y^* \; if \; D = 1 \tag{1.4}$$

where Y^* is the observed wages, D is participation and is equal to 1 if the person chooses

to work. U and V are the error terms. $Z = (B, X)$ and is a set of covariates that strictly contains observable characteristics X, and exclusion restrictions B. $p(Z)$ is the probability of a person to choose to work given Z and is estimated as the propensity score using logit. A person only chooses to work if $p(Z)$ passes through a threshold and wages are only observed for those that chooses to work. Under these circumstances, if $cov(X, D) \neq 0$, then $cov(U, V) \neq 0$ and $E(Y^*) = E(Y|D = 1) \neq E(Y)$, or in other words, selection bias. The model requires 4 assumptions:

1. (U, V) is jointly statistically independent of Z given X
2. (U, V) is continuous given X and follows some cumulative distribution function (CDF) $C_x(u, v)$
3. Conditional CDF of $Y^*|X$ and its inverse is strictly increasing. $C_x(u, v)$ is also strictly increasing in u
4. Given Z, an individual has positive probability of being treated and non-treated: $0 < P(D = 1|Z) < 1$

The key link between regular quantile regression and the selection corrected quantile regression of A&B is the difference in τ. τ is the proportion of the population with wages lower than Y_τ: $P(Y_\tau < x\beta_\tau)$. Conditioning on participation, that proportion can be written as:

$$P(Y^* \leq (\tau, x)|D = 1, Z = z) = P(U \leq \tau|V \leq (z), Z = z) \tag{1.5}$$

$$= C_x(\tau, p(z); \rho)/p(z) \tag{1.6}$$

$$\equiv G_x(\tau, p(z)) \tag{1.7}$$

Where C is the conditional copula (joint distribution) of U and V and is assumed to follow some particular distribution. We assumed Frank Copula to be our distribution of choice; a part of the family of Archimedean Copulas. This family of copula has the property of being able to be generated with a single parameter, ρ. This property is important as it allows the problem to be solved with linear programming, for which exists easy to use and efficient algorithms. More parameters increases computation time exponentially. If U and V are independent given X,

then $G_x(\tau, p(z)) = \tau$. Therefore, G is conditional copula function that maps the unconditional τ onto selection corrected quantiles. In turn, if G is known, the unconditional latent wage and wage quantiles can be found.

Combining (4) and (7), the selection corrected quantile regression estimation of $\hat{\beta}_\tau$ is:

$$\hat{\beta}_\tau(c) = \underset{b \in B}{\operatorname{argmin}} \Big\{ \sum_{i:Y_i \geq X_i'\beta} D_i \Big\{ |G(\tau, p(Z_i); c)(Y_i - x'\beta_\tau)| + \sum_{i:Y_i < X_i'\beta} |1 - G(\tau, \hat{p(Z_i)}; c)(Y_i - X'\beta_\tau)| \Big\} \Big\} \tag{1.8}$$

The estimation of the selection corrected beta is done in 2 main steps. The first step is to estimate the propensity score $p(Z)$. In our case, we used a standard logit model. The second step is to simultaneously estimate the copula parameter $\hat{\rho}$ (and by extension $\hat{G}$) and $\hat{\beta}_\tau$.

To estimate ρ:

$$\hat{\rho} = \underset{c \in C}{\operatorname{argmin}} \Big\| \sum_{i=1}^{N} \sum_{l=1}^{L} D_i p(Z_i) [1\{Y_i \leq X_i' \hat{\beta}_{\tau l}(c)\} - G(\tau_l, p(Z_i); c)] \Big\| \tag{1.9}$$

where $\|.\|$ is loss function (log likelihood in our case), and $\phi(\tau_l, Z_i)$ are the propensity scores.

1.5 Instrument

For our exclusion restriction, we used provincial child care benefit and its square from the updated tax calculator of Milligan and Stabile (2011). The calculator returns the sum of all provincial-level benefits in all provinces given the input variables. From this, we took only the benefits that are associated with child rearing.[7] Inputs variables are : age, province, gender, married dummy, and yearly earnings.[8] Using all of the above variables, we created a synthetic data set of all possible permutations for each year of our data. For earnings, we estimated mean hourly wage conditioning on age, province, gender, and married. To get the yearly

[7]The variables are: 'Newfoundland and Labrador Child Benefit', 'Newfoundland and Labrador Mother and Baby Nutritional Supplement', 'Newfoundland and Labrador Progressive Family Growth Benefit', ' Nova Scotia Child Benefit', 'Quebec Allowance for Newborn Children', 'Quebec Allowance for Young Children', 'Quebec family allowance', 'Quebec new family allowance', 'Quebec APPORT program', 'Quebec child assistance measure', 'Ontario Child Benefit', 'Ontario Childcare Supplement for Working Families', 'Manitoba CRISP credit', 'Manitoba child benefit', 'Saskatchewan Child Benefit', 'Alberta Family employment Tax Credit', 'Alberta Child Benefit', 'BC Family Bonus', and 'BC Early Childcare Tax Benefit'.

[8]See Appendix A1 for details

earnings, we multiply the hourly earnings by mean usual weekly hours of work (using the same method) and 51. The benefit can be interpreted as the potential dollar amount that someone can expect to receive from their province for child care given their characteristics. Every person would be issued a benefit amount for their potential young child regardless whether or not they have children of valid age. Variations in the benefits received would mainly come from the differences in tax structure between provinces. We believe that this approach is less susceptible to endogeneity issues as long as the input characteristics used to differentiate different individuals are exogenous.

Table 1.4 shows the logit results for our instrument. In the left column, we provide the coefficients, while the right column demonstrates the marginal effects of a $1000 increase in received benefits on workers' labor participation. Both men's and women's engagement in labor participation are influenced by child care benefits. In 2002, we observe diminishing returns in these effects, although not to an extent that would reduce participation at moderate benefit levels. For every $1000 increase in child care benefits, women's likelihood to be in the labour force increases by 5.7% and men's by 2%. This is similar to the findings of Lefebvre and Merrigan (2008). However, the dynamics change by 2016. The impact of the benefit on women's labor participation has reversed; a $1000 increase now leads to a decrease of around 6% in the likelihood of women participating in the labor force. On the other hand, men's participation response remains largely consistent with the 2002 pattern. The results also became more statistically significant. This is a result of an increase in variation between and within provinces' child care benefits.

1.5.1 Latent Wage Decomposition

The selection corrected quantile regression corrects for selection of β , in the decomposition, we correct for the wage distribution. In a latent wage setting, given the the pay schedule, each individual can either work or not at each quantile level. Only those that choose to work are used regressions. To determine whether a person work or not, we first created a copula using ρ_{tj} from the quantile selection regressions; $t = 1997, 2016$. For counterfactual female, we used male's $\rho_{t,1}$. From that copula, we generated 10,000 random pairs of (U, V). For $U = 0.05, ...0.95$, we compare the corresponding V with individual's

propensity score $P(Z)$. [9] If $V < P(Z)$, then $D = 1$ and that person will choose to work at the given wage quantile. Their wage would be calculated using the selection corrected β: $\hat{Y}_{jjmi} = x_{ji}\hat{\beta}_{mj}$. Else, $D = 0$, and that individual is deemed to not work at that quantile of wage and is excluded from the latent wages. The remaining estimated wages represents the latent wage distribution, and from these, we picked out wages from our desired quantiles.

Finally, the gender wage gap for quantile regression, accounted for selection, is very similar to Blinder-Oaxaca decomposition:

$$\begin{aligned} & Y^{ss}_{00m} - Y^{ss}_{11m} \\ & = \beta^s_{0m} X^s_0 - \beta^s_{1m} X^s_1 \\ & = (X^s_0 \beta^s_{0m} - X^s_0 \beta^s_{1m}) + (X^s_0 \beta^s_{1m} - X^s_1 \beta^s_{1m}) \end{aligned}$$

the first half of the decomposition,$X_0\beta_{0m} - X_0\beta_{1m}$, is the unexplained gap while the latter, $X_0\beta_{1m} - X_1\beta_{1m}$, is the explained. The superscript, s, denotes selection correction. Y^{ss} means estimated wage using βs estimated using quantile selection regression and Xs that were corrected for latent effect.

1.6 Results

Figure 1.3 shows the total wage gap calculated from our quantile regression and selection-corrected quantile regression [10]. For both years, men earns more than women for all quantiles and for both models. In 2002, the gap highest at the 20th percentile for both regressions and lowest near the tail ends. Selection correction consistently have lower wage gap compared to no selection, over all income brackets. This suggests that women with higher earning potential is negatively selected for labour participation in 2002. In 2016, the wage gap has decreased for all but the highest wage quantiles. The most noticeable change comes from the lower income levels,

[9] For counterfactual female, we used male logit model to predict female propensity score.

[10] See table A2 in the appendix for a comparison of results at select quantiles and tables A3 to A10 for the coefficients

and the lowest income earners now have the lowest difference in wages. Across most most wage quantiles, the gap has become flat as improvements in the wage gap decreases as income increases. These results is consistent with other papers, such as Boudarbat and Connolly (2013), N. M. Fortin and Lemieux (2015), where improvements in the wage gap is concentrated in the lower half of the income distribution. The effect of selection also become less apparent in 2016 and the total gap is no longer different from the standard quantile regression gap. This indicates that the effect of selection on the total gender wage gap has declined for all quantiles over the two time periods [11].

Figure 1.4 displays the unexplained wage gap and figure 1.5 presents the explained wage gap. The sum of the gaps in the two graphs is equal to the total gap shown in figure 1.3. In 2002, compared to the total wage gap, the regular quantile regression unexplained gap holds the general shape but is higher across all quantiles by roughly 2 percentage points. The explained gap graph shows relatively stable difference, around 2 percentage points, in favour of women for all quantiles. This indicates women having characteristics that are more favourable for higher earnings. Thus, if women have the pay structure of men, then women would have higher wage on average. When selection is accounted for, the unexplained wage gap in 2002 differs substantially compared to the total gap, reaching a zenith of just over 7.5 percentage points (1/3 of the total selection corrected gap) when $\tau = 0.15$. The effect of selection is also not consistent across wage quantiles. For the first 3 quarters of the wage distribution, selection correction makes the unexplained gap higher while the reverse is true for the highest earners. By 2016, the unexplained gap followed the same trend as as the total gap, and has shown improvements across the first three quarters of income. However, unlike the total gap, the effect of selection became more pronounced. While 1.3 suggests women with higher earning potential are no longer negatively selected into the workforce in 2016, according to 1.4, those additional women are not being compensated enough. This becomes clear in 1.5. Unlike the non-selection model, the explained gap decreases with higher income, with the exception of when $\tau < 0.15$. Women that are more 'skilled' are less likely to enter the labour market for nearly all quantiles other than the right tail end. In 2016, the non-selection unexplained gap retained the same general shape as the total, only higher for all quantiles. This is caused by an increase of the explained gap across all wage levels, with higher increase towards the middle; the gap is no longer flat across

[11] All inferences in this section are relative between men and women

income brackets and is now higher around the center. Hence, women's observable characteristics relative to men has improved since 2002, especially for middle income earners. When selection is accounted for, the unexplained gap is consistently lower for the first 3 quarters of income in 2016. The effect of select diminished for the lower quantiles but increases with income. While women still have higher 'skill' compared to men for all income levels, the gap has narrowed for low income workers, and widened for higher wage workers. Like the case with no selection, the middle income brackets now have the highest discrepancy in actual 'skill' as women are the least likely to enter the workforce, relative to men, in those income levels.

Figure 1.6 further decomposes the unexplained wage gap by various factors The only variable that decreases the gap is education. In 2002, the effect of education increases with income. At the highest income levels, education reduces the unexplained gap by nearly 8.5% while having zero, or even negative, effect in the lowest quantiles. From tables A3 to A6, education have the highest impact on wage, with higher education having higher return on income for all quantiles. However, the effect of education is not consistent across wage quantiles. Higher education have relatively low return on the first quarter of income while lower educated workers are penalized at high income. By 2016, the effect of education on the wage gap has changed. It decreases the gap more (relative to 2002) for lower income workers while having a reduced effect for the very high end. The effect of being married increases the wage gap as income decreases in 2002 but is relatively even across the quantiles in 2016. This is mainly caused by relatively high return on being married for men in 2002 in the lower income groups.

1.7 Conclusion

This paper examines the Canadian gender wage gap and the role of selection bias plays on different income groups. Using data from the LFS, we found persistent but inconsistent wage gap across the income distribution. In addition, our analysis reveals selection bias have a noticeable impact on the gap which persists between our sample years. We found that selection bias have conflicting effects for the unexplained and explained wage gap, indicating negative selection of higher skilled female workers into the labour market.

For all wage qantiles, women learn less than men even after controlling for confounding variables. In 2002, gap is particularly noticeable at around the $20th$ quantile in 2002 while the extreme ends saw the lowest difference. By 2016, the gap in the lower half closed dramatically while the high end saw to no change. The unexplained and explained wage gap have opposite effects, which suggests women have characteristics that have higher earning potential and if sharing men's pay structure, would have higher wages than men across all income quantiles. We found that selection bias would increase the observed wage gap for all income levels in 2002 while having no impact in 2016. However, like the case without selection, selection bias have different effects on the unexplained and explained gap. In the presence of selection bias, the unexplained wage gap for the first 3 quarters of income would increase while lowering it for the highest quarter. This remains true for both years. Explained gap bias would reduce the gap for all but the highest wage earners in 2002, and the first 3 quarters for 2016.

In summary, this research has shown that selection bias and quantile regression are important considerations when studying the gender wage gap. Our findings indicate that quantile regression can provide a more nuisances view of the wage gap but must take into consideration the impact of selection. Selection bias can lead to an underestimation of the gender wage gap at the lower 3 quarters of the income distribution and overestimation at the higher end.

Table 1.1: Average wage by demographics

	1998			2018			Change	
	Female	Male	Diff	Female	Male	Diff	Female	Male
Population	17.31	21.03	-17.69	19.96	23.41	-14.74	15.31	11.32
Quantile								
q5	7.38	8.86	-16.7	8.85	9.73	-9.04	19.92	9.82
q25	11.05	14.23	-22.35	12.79	15.04	-14.96	15.75	5.69
q50	15.69	19.4	-19.12	17.86	21.03	-15.07	13.83	8.4
q75	21.68	26	-16.62	25.05	29.73	-15.74	15.54	14.35
q95	32.67	38.88	-15.97	37.79	44.81	-15.67	15.67	15.25
Age								
25-29	15.88	17.67	-11.27	17.74	19.27	-8.62	11.71	9.05
30-34	17.82	20.36	-14.25	20.18	22.92	-13.58	13.24	12.57
35-39	17.7	21.47	-21.3	21.03	24.55	-16.74	18.81	14.35
40-44	17.67	21.99	-24.45	21.25	24.98	-17.55	20.26	13.6
45-49	18.74	23.31	-24.39	20.36	25.96	-27.5	8.64	11.37
50-54	19.15	23.47	-22.56	20.18	24.42	-21.01	5.38	4.05
Education								
< High School	11.68	15.87	-35.87	13.37	17.34	-29.69	14.47	9.26
High School	15.34	18.64	-21.51	15.82	19.5	-23.26	3.13	4.61
College/Certificate	17.42	20.99	-20.49	18.01	22.61	-25.54	3.39	7.72
Bachelor's	21.86	25.33	-15.87	22.81	26.79	-17.45	4.35	5.76
Graduate	25.37	28.92	-13.99	26.14	30.36	-16.14	3.04	4.98
Province								
Ontario	18.41	21.8	-18.41	20.11	23.84	-18.55	9.23	9.36
Newfoundland	14.96	18.02	-20.45	18.41	23.75	-29.01	23.06	31.8
Prince Edward Island	14	16.24	-16	17	16.14	5.06	21.43	-0.62
Nova Scotia	14.78	17.19	-16.31	18.05	20.13	-11.52	22.12	17.1
New Brunswick	14.35	17.1	-19.16	17.91	18.46	-3.07	24.81	7.95
Quebec	17.05	19.95	-17.01	19.39	21.74	-12.12	13.72	8.97
Manitoba	16.48	19.18	-16.38	19.83	21	-5.9	20.33	9.49
Saskatchewan	16.05	18.63	-16.07	20.29	24	-18.28	26.42	28.82
Alberta	16.99	22.91	-34.84	22.25	27.07	-21.66	30.96	18.16
British Columbia	18.5	21.81	-17.89	19.39	23.4	-20.68	4.81	7.29
Marital Status								
Single	17.48	18.88	-8.01	18.78	20.98	-11.71	7.44	11.12
Married	18.01	22.7	-26.04	20.85	25.49	-22.25	15.77	12.29
Living in common-law	17.6	20.68	-17.5	19.82	22.83	-15.19	12.61	10.4
Widowed	15.41	25.04	-62.49	20.72	20.2	2.51	34.46	-19.33
Youngest Child								
<3	18.45	22.17	-20.16	21.88	24.76	-13.16	18.59	11.68
3-5	16.65	21.76	-30.69	21.6	25.38	-17.5	29.73	16.64
6-12	17.21	22.87	-32.89	20.91	26.36	-26.06	21.5	15.26
13-15	17.37	23.27	-33.97	19.84	26.5	-33.57	14.22	13.88
16-17	17.85	25.34	-41.96	20.15	25.98	-28.93	12.89	2.53
18-24	17.96	23.07	-28.45	19.83	25.52	-28.69	10.41	10.62
Other	18.05	19.82	-9.81	19.06	21.66	-13.64	5.6	9.28

Notes: This table shows the unconditional real wage by demographics. Difference and change refers to percentage difference between men and women, and percentage change over time

Table 1.2: Labour participation rate by demographics

	1998			2018			Change	
	Female	Male	Diff	Female	Male	Diff	Female	Male
Population	79.4	90.37	-12.13	80.98	89.77	-9.79	1.99	-0.66
Age								
25-29	80.9	91.13	-12.65	82.75	87.94	-6.27	2.29	-3.5
30-34	78.96	91.12	-15.4	80.87	92.37	-14.22	2.42	1.37
35-39	79.96	92.04	-15.11	79.97	91.22	-14.07	0.01	-0.89
40-44	81.91	91.83	-12.11	81.81	92.01	-12.47	-0.12	0.2
45-49	80.48	89.09	-10.7	81.88	90.03	-9.95	1.74	1.06
50-54	73.31	86.01	-17.32	78.87	85.67	-8.62	7.58	-0.4
Education								
< High School	58.64	79.38	-35.37	51.26	73.02	-42.45	-12.59	-8.01
High School	77.37	90.19	-16.57	72.45	86.46	-19.34	-6.36	-4.14
College/Certificate	85.18	93.87	-10.2	84.88	92.86	-9.4	-0.35	-1.08
Bachelor's	85.24	93.07	-9.19	87.26	94.43	-8.22	2.37	1.46
Graduate	85.32	92.82	-8.79	85.97	92.88	-8.04	0.76	0.06
Marital Status								
Single	82.45	84.13	-2.04	82.05	83.44	-1.69	-0.49	-0.82
Married	77.07	93.56	-21.4	78.26	93.76	-19.81	1.54	0.21
Living in common-law	83.82	92.73	-10.63	88.17	91.98	-4.32	5.19	-0.81
Widowed	81.5	84	-3.07	65.8	80.6	-22.49	-19.26	-4.05
Youngest Child								
<3	81.85	87.08	-6.39	83.78	86.18	-2.86	2.36	-1.03
3-5	66.02	94.71	-43.46	70.64	94.7	-34.06	7	-0.01
6-12	71.33	94.25	-32.13	77.48	95.5	-23.26	8.62	1.33
13-15	81.22	94.08	-15.83	80.55	94.67	-17.53	-0.82	0.63
16-17	83.7	94.04	-12.35	81.83	93.49	-14.25	-2.23	-0.58
18-24	82.55	93.31	-13.03	84.36	91.36	-8.3	2.19	-2.09
Other	81.8	92.72	-13.35	84.05	91.97	-9.42	2.75	-0.81

Notes: This table shows the unconditional labour participation rate by demographics. Difference and change refers to percentage difference between men and women, and percentage change over time

Table 1.3: Average wage by demographics for low income earners

	1998			2018			Change	
	Female	Male	Diff	Female	Male	Diff	Female	Male
Population	9.55	9.99	-4.4	10.68	11.07	-3.52	11.83	10.81
Age								
25-29	9.63	9.85	-2.28	10.84	11.04	-1.85	12.56	12.08
30-34	9.53	10.1	-5.98	10.82	11.17	-3.23	13.54	10.59
35-39	9.66	10.22	-5.8	10.67	11.16	-4.59	10.46	9.2
40-44	9.63	9.87	-2.49	10.46	10.76	-2.87	8.62	9.02
45-49	9.62	10.31	-7.17	10.67	11.17	-4.69	10.91	8.34
50-54	9.78	10.12	-3.48	10.49	11.17	-6.48	7.26	10.38
Education								
< High School	9.05	9.94	-9.83	10.13	11	-8.59	11.93	10.66
High School	9.66	10.14	-4.97	10.34	11.1	-7.35	7.04	9.47
College/Certificate	9.83	10.09	-2.64	10.96	11.12	-1.46	11.5	10.21
Bachelor's	10.04	9.92	1.2	10.79	11.12	-3.06	7.47	12.1
Graduate	9.38	9.78	-4.26	10.61	10.82	-1.98	13.11	10.63
Province								
Ontario	9.79	10.1	-3.17	10.59	10.96	-3.49	8.17	8.51
Newfoundland	8.23	9.88	-20.05	10.3	10.36	-0.58	25.15	4.86
Prince Edward Island	8.92	10.12	-13.45	10.67	10.52	1.41	19.62	3.95
Nova Scotia	9.3	9.6	-3.23	10.66	10.82	-1.5	14.62	12.71
New Brunswick	8.99	9.81	-9.12	10.84	10.95	-1.01	20.58	11.62
Quebec	9.47	10.06	-6.23	10.83	11.2	-3.42	14.36	11.33
Manitoba	10.15	9.84	3.05	10.86	11.18	-2.95	7	13.62
Saskatchewan	9.02	9.71	-7.65	10.67	10.99	-3	18.29	13.18
Alberta	9.55	10.08	-5.55	10.85	11.02	-1.57	13.61	9.33
British Columbia	9.78	9.88	-1.02	10.53	11.32	-7.5	7.67	14.57
Marital Status								
Single	9.6	9.87	-2.81	10.83	11	-1.57	12.81	11.45
Married	9.65	10.13	-4.97	10.5	11.05	-5.24	8.81	9.08
Living in common-law	9.66	10.32	-6.83	10.83	11.3	-4.34	12.11	9.5
Widowed	9.41	11.48	-22	10.08	13.18	-30.75	7.12	14.81
Youngest Child								
<3	9.69	10.05	-3.72	10.67	11.18	-4.78	10.11	11.24
3-5	9.12	10.07	-10.42	10.62	11.15	-4.99	16.45	10.72
6-12	9.56	10.37	-8.47	10.64	11.04	-3.76	11.3	6.46
13-15	9.46	9.89	-4.55	10.51	11.27	-7.23	11.1	13.95
16-17	9.58	10.6	-10.65	10.4	11.4	-9.62	8.56	7.55
18-24	9.86	10.31	-4.56	10.44	10.98	-5.17	5.88	6.5
Other	9.74	9.96	-2.26	10.77	11.06	-2.69	10.57	11.04

Notes: This table shows the unconditional real wage by demographics for the lowest quarter of income earners. Difference and change refers to percentage difference between men and women, and percentage change over time

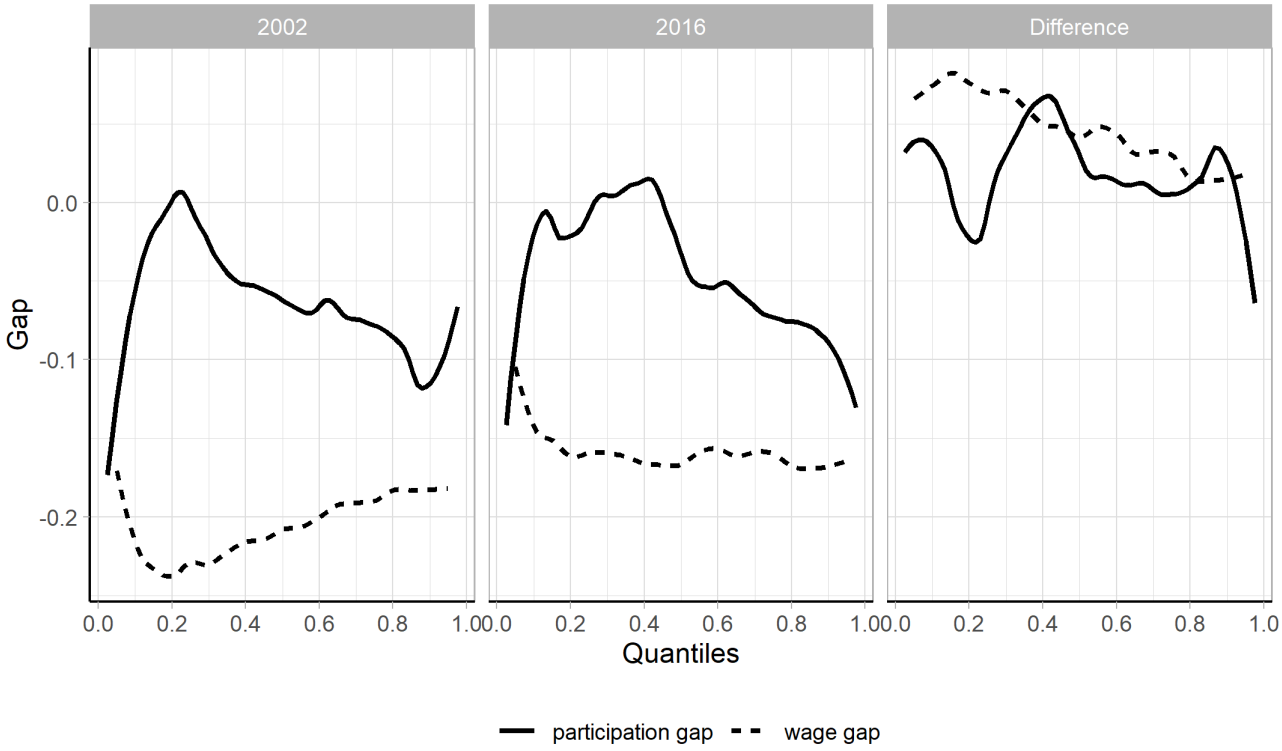
Table 1.4: Provincial child care benefit logit

	Female				Male			
	2002		2016		2002		2016	
	logit	marginal	logit	marginal	logit	marginal	logit	marginal
$Benefits$	0.366*	0.067*	−0.362***	−0.062***	0.217*	0.027*	0.226**	0.028**
	(0.150)	(0.028)	(0.068)	(0.012)	(0.109)	(0.013)	(0.075)	(0.009)
$Benefits^2$	−0.055*	−0.010*			−0.060+	−0.007+		
	(0.024)	(0.004)			(0.036)	(0.004)		
Joint sig	(0.03617) *		($1.2e^{-3}$) ***		(0.144)		(0.0028)**	

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

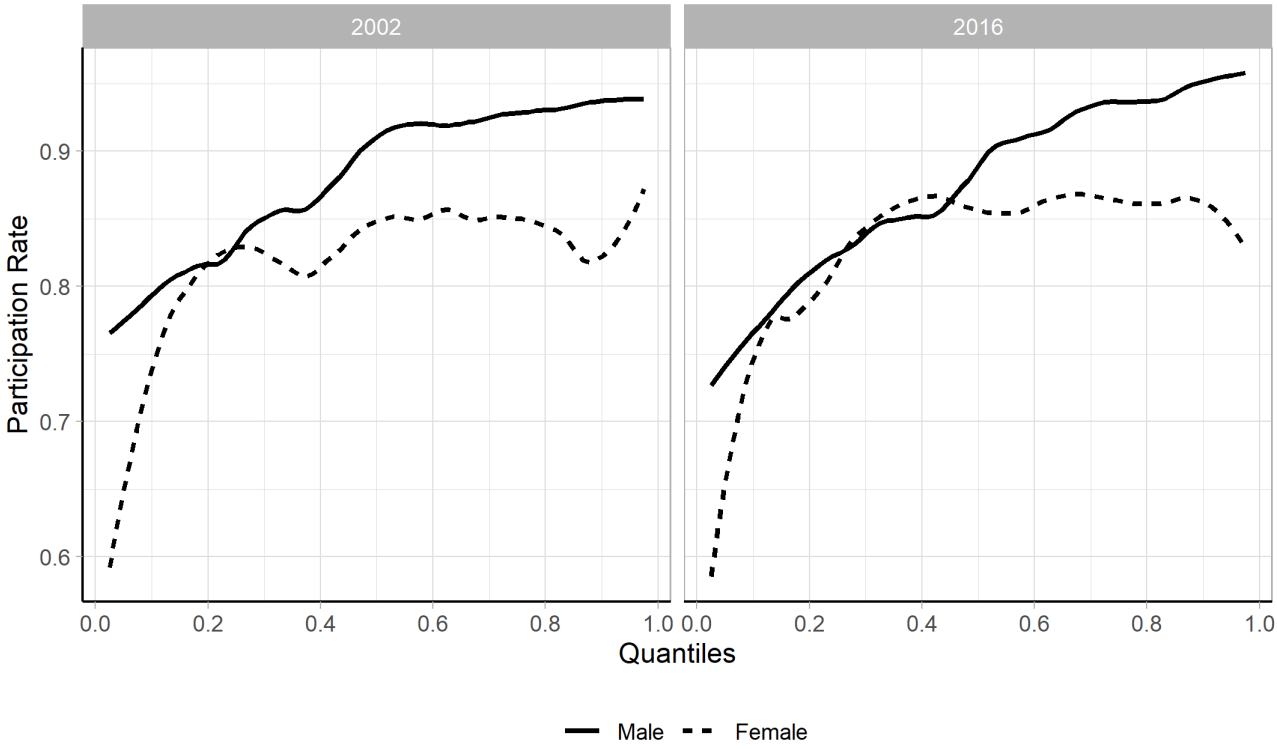
Notes: This table shows the results of the logit regressions for the instrumental variables. The dependent variable is a dummy that is equals to 1 when the person works. Coefficients are the effect of $1000 increase in benefits. The logit columns contain the raw logit coefficients while the marginal columns are the effect on probability of employment when benefits are increased by $1000. Numbers in parentheses denote standard error. $Benefit^2$ is statistically insignificant in 2016 and the results for them are omitted

Figure 1.1: Wage and labour participation gap by income quantiles



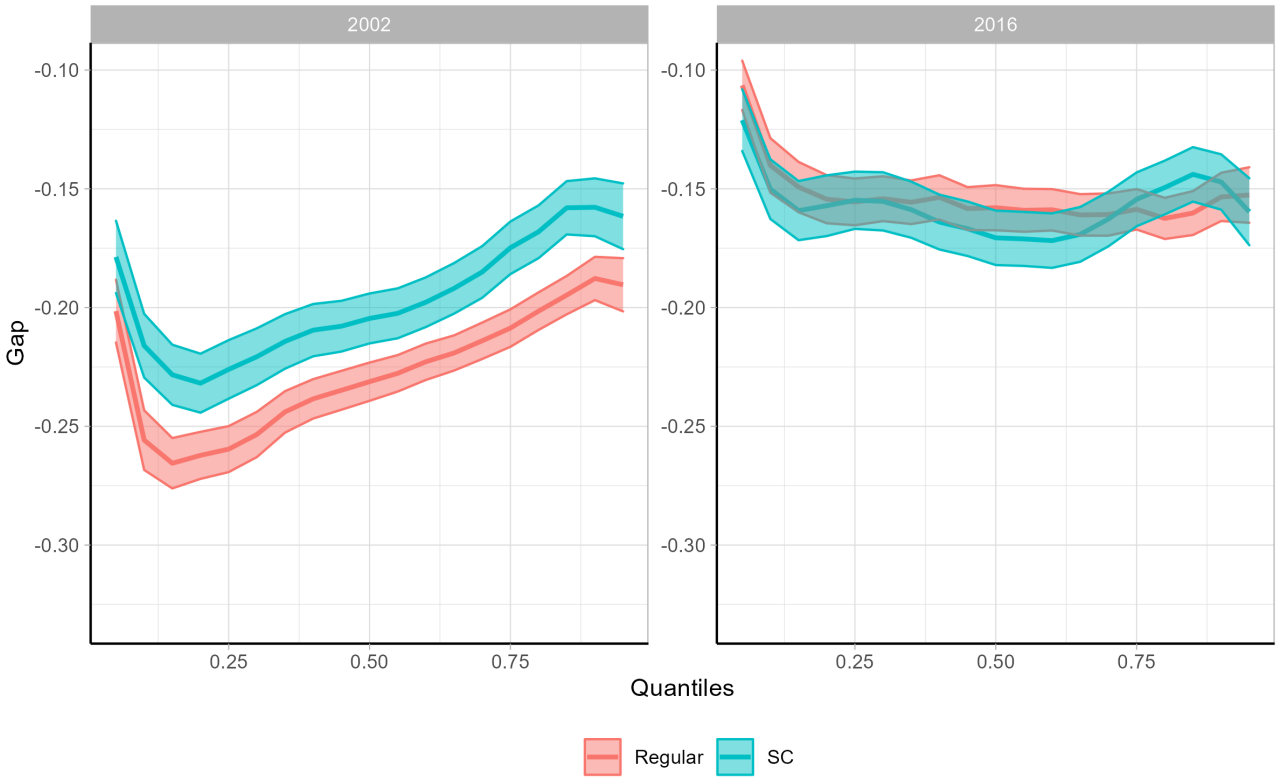
Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). The dash line depicts the total wage gap at different wage quantiles, calculated using the standard quantile regression (see paper for more details). The solid line depicts the labour participation gap between men and women at different wage quantile. Wages for those that don't work are imputed using quantile regression. Participation is the average labour participation rate for each 0.05 estimated quantile interval of wages, from 0 to 1. Difference refers to the change from 2002 to 2016

Figure 1.2: Labour participation rate by income quantiles



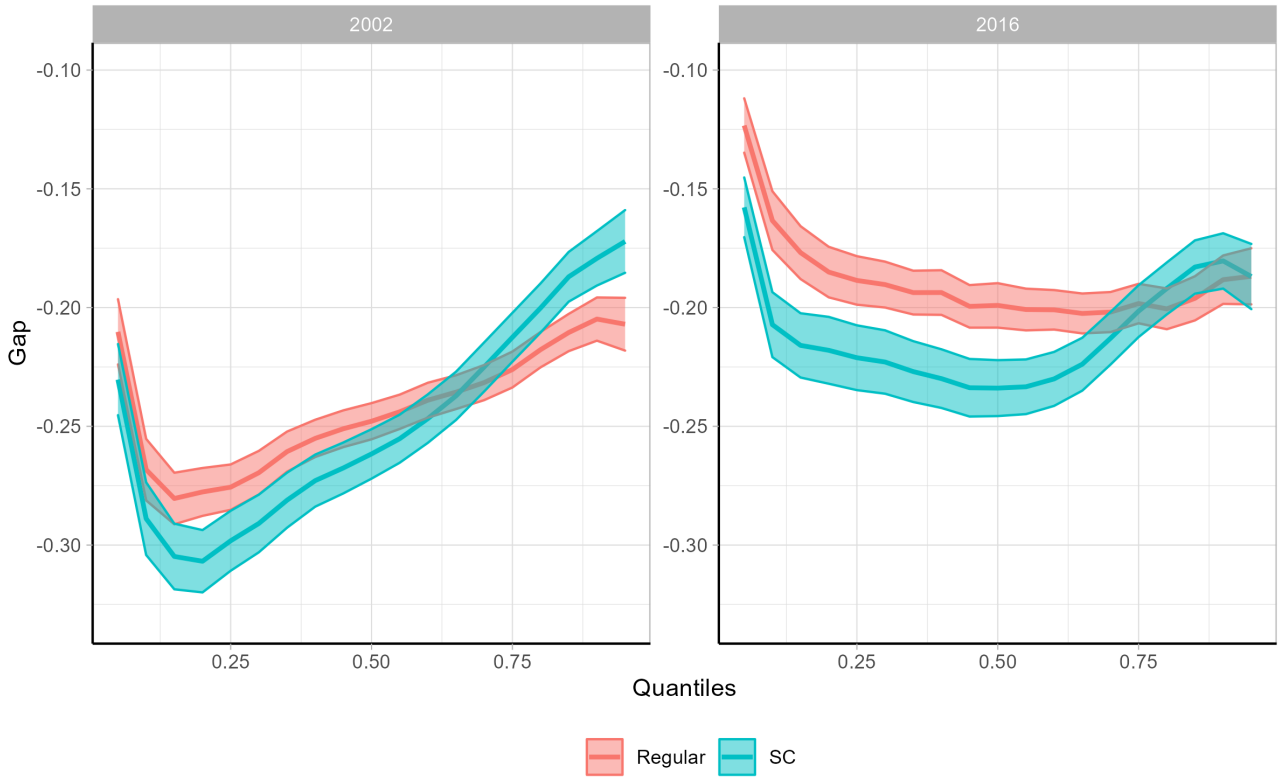
Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). Participation rate is the proportion of workers who are part of the labour force at that wage quantile. Wages for those that don't work are imputed using quantile regression.

Figure 1.3: Decomposition comparison



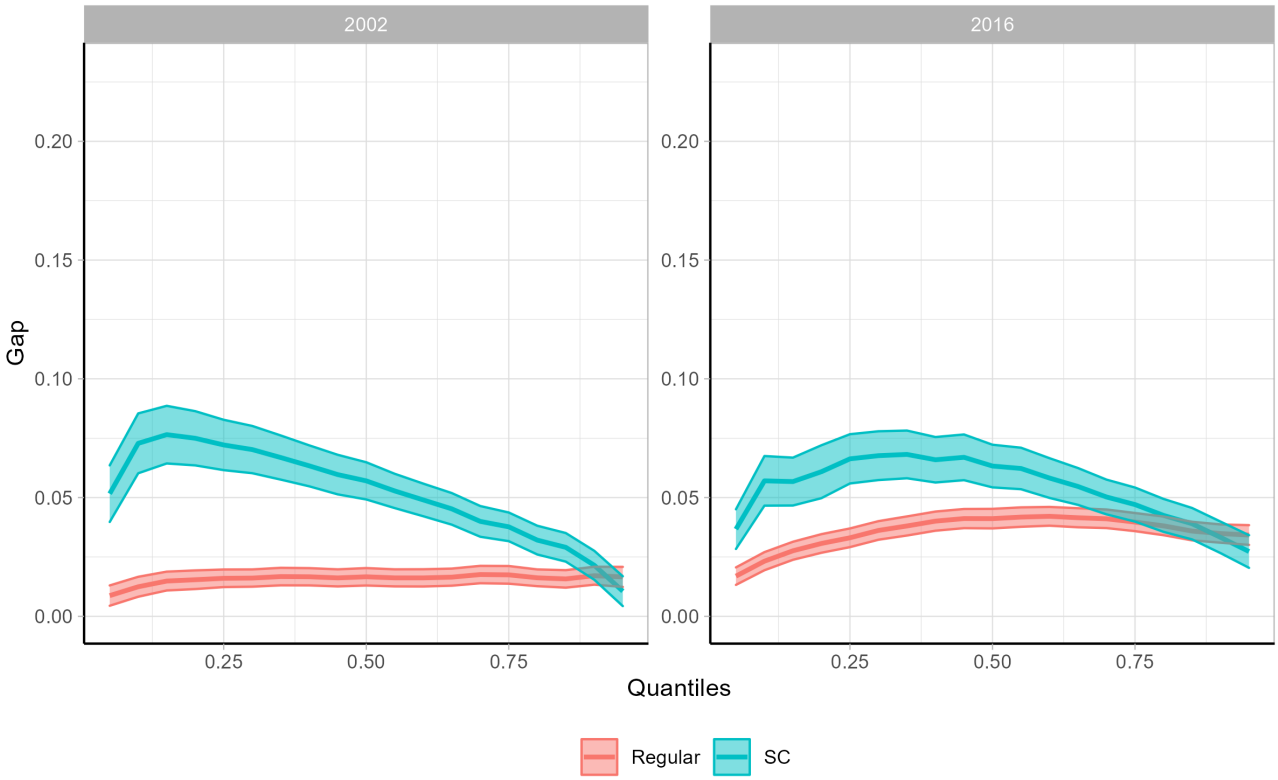
Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). This figure display the total gender wage gap as calculated using regular quantile regression and selection-corrected quantile regression (see paper for more details). The shaded area represents the 95% confidence interval

Figure 1.4: Unexplained gap



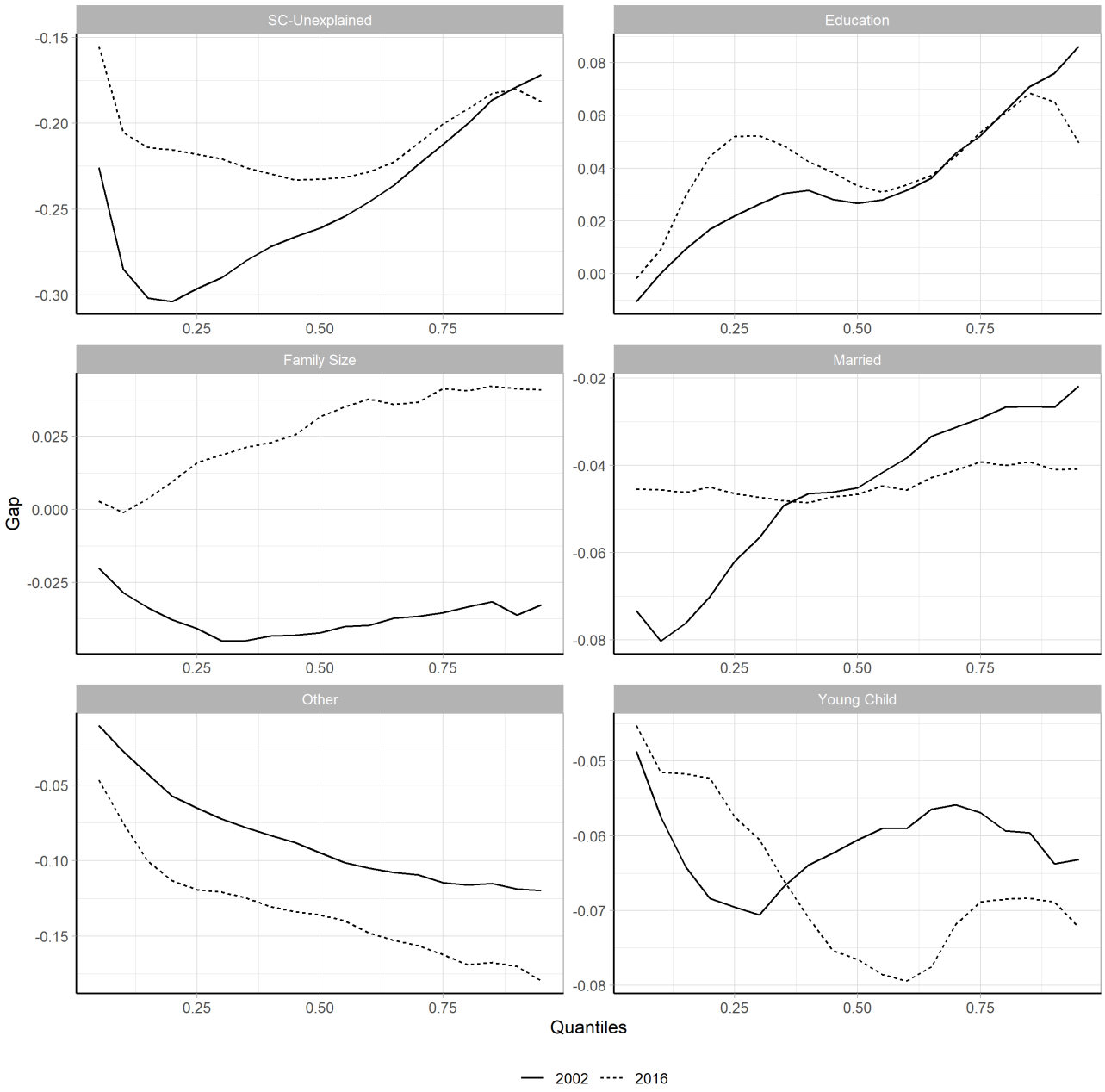
Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). This figure display the unexplained gender wage gap as calculated using regular quantile regression and selection-corrected quantile regression (see paper for more details). The shaded area represents the 95% confidence interval

Figure 1.5: Explained gap



Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). This figure display the explained gender wage gap as calculated using regular quantile regression and selection-corrected quantile regression (see paper for more details). The shaded area represents the 95% confidence interval

Figure 1.6: Detailed decomposition



Notes: The data utilized in this figure originates from the Labour Force Survey (LFS). This figure display contribution of various independent variables to the selection-corrected unexplained gender wage gap by wage quantiles (see paper for more details).

Chapter 2

Heterogeneity in the Canadian Wage Gender Gap[1]

2.1 Introduction

The gender wage gap – average difference in pay between men and women – serves as a critical barometer for gender inequality (Moyser 2019). Despite substantial societal and workplace strides towards gender equity, women persistently receive less pay than their male counterparts on average. However, in this paper, we show that this disparity is not uniform but depends on a variety of factors such as demographic characteristics, job attributes, industries, occupations, and education levels. Given the heterogeneity in the gender gap, understanding the main factors driving wage disparities is vital for devising effective strategies and policies that aim to achieve wage parity.

In this paper, we provide a detailed analysis of the evolution of the gender wage gap in Canada, focusing on the years between 1999 and 2015. We use Double Machine Learning (DML) to evaluate potential contributors to the wage gap while controlling for a high-dimensional set of confounding factors (Chernozhukov et al. (2018b)). Our identification strategy is based on unconfoundedness, and we frame our analysis within the context of the treatment effect literature. Our aim is to estimate 'Average Treatment Effects on the Treated' (ATET), representing the mean difference in hourly wages between genders, while controlling for relevant

[1]This paper is a collaboration with Prof. Ben Sand and Vlad Fenenko

variables. Our model accommodates treatment heterogeneity by allowing the wage gap to depend on a high-dimensional set of observable characteristics. The goal of this investigation is to identify key factors influencing the gender wage gap and trace changes over the study period.

Our approach closely follows Bach, Chernozhukov, and Spindler (2021), who examine the heterogeneity of the gender gap in the US using 2016 ACS data. In particular, their empirical strategy uses a double LASSO procedure to estimate the effect of a large number of characteristics on the gender gap, while flexibly controlling for confounding factors. We build on their approach by using the more recent double/de-biased machine learning tools to estimate causal effects. This procedure uses cross-fitting, which allows for any good machine learning algorithm to be used to estimate nuisance functions. In addition, we examine the heterogeneity of the Canadian gender gap in both 1999 and 2015, allowing us to characterize the major factors associated with the gender gap, as well as how the impact of those factors changes over time.

The double/de-biased machine learning (DML) estimation that we use is a two-step process designed to estimate treatment effects in observational data. In the first step, DML uses machine learning algorithms to predict both the outcome variable (based on control variables) and the treatment assignment (based on covariates). A key is to use cross-fitting, which splits the data into multiple subsets or "folds." Then, the predictive models are fit (both for the outcome and treatment assignment) on one subset of the data and used to compute the residuals on a different subset. The second step involves running a regression model on the cross-fit residuals from the first step predictions to estimate the ATET. This procedure helps to mitigate the risk of bias that could be introduced by overfitting in the prediction models, providing more reliable, unbiased estimates of treatment effects.

Our paper uncovers a range of factors impacting the gender pay gap in Canada from 1999 to 2015. Notably, higher levels of education and employment in unions or the public sector are associated with a diminished gender pay disparity. In terms of industries, we observe higher earnings for men in traditionally male-dominated sectors, such as agriculture and construction, while women tend to earn more in typically female-dominated sectors, like healthcare and accommodation/food services. Importantly, our analysis shows that occupational

choice consistently plays a more pivotal role in shaping the gender wage gap than industry.

By 2015, the data suggests a disappearance of the age gradient in gender pay differentials, and the overall unexplained pay gap has declined from 16.5% to 9.2%. Despite this improvement, family structure variables continue to contribute significantly to wage disparities, showing little progress over the past 15 years. Our findings also reveal the emergence of the influence of working hours on gender pay gaps in 2015. In a hopeful trend, the occupations with the largest pay disparities in 1999 have seen the most significant reductions in these gaps. In contrast, variables such as age and occupation no longer contribute to larger pay gaps, underlining the shifting dynamics of the gender wage gap over this timeframe.

Our estimates allow us to estimate a gender wage gap, conditional on a set of observable features, or the $ATET(Z_i)$. We show that gender wage gap displays substantial variability across different population segments. As an illustration, women situated at the 25th percentile of the $ATET(Z_i)$ distribution earn approximately 25 percent less than their male counterparts with similar characteristics, while the disparity reduces to nearly 10 percent less for women at the 75th percentile. There is a discernible decrease in the gender wage gap throughout the $ATET(Z_i)$ distribution, averaging around a 7 percent reduction between the years 1999 and 2015. However, this reduction isn't evenly distributed across all percentiles, with the most marked improvement occurring at the lower percentiles of the $ATET(Z_i)$ distribution.

To ensure that our results are robust, we utilize an array of machine learning methods within the DML procedure, including the Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest, and Boosted Trees. Our main findings are consistent across different methodologies. We further verify the consistency of our results through alternative machine learning methods and parameterizations, reinforcing the robustness of our findings.

Our paper provides important insights on the evolution of the gender wage gap in Canada over time, offering critical implications for policy makers aiming to further reduce this disparity. The identified factors contributing to the wage gap serve as key areas of focus for devising strategies to address gender pay inequality. In particular, the persistent role of

family-related factors as a driver of the gender wage gap, even as other variables have lessened in impact, emphasizes the importance of policies addressing aspects such as childcare and marital status.

This paper proceeds in a structured manner. In section 2.2, we offer a brief literature review, situating our research within the wider discourse on the gender wage gap. The next section, 2.3, describes our data and provides summary statistics. In section 2.6, we then detail our methodology, describing the DML algorithm that we use. Section 2.8 documents our results and Section 2.12 concludes.

2.2 Literature Review

There is an enormous amount of research examining gender wage disparities. Francine D. Blau and Kahn (2000); Francine D. Blau and Kahn (2017a); Olivetti and Petrongolo (2016b); Goldin (2014), Arntz, Gregory, and Lehmer (2011), and Joseph G. Altonji and Blank (1999) provide excellent reviews. To a great extent, the literature attributes gender differences in pay to various forces, including technology (Welch 2000; Weinberg 2000; Goldin 2014), changing self selection in the labour force (Casey B. Mulligan and Rubinstein 2008a), bargaining ability (Card, Cardoso, and Kline 2016), soft skills (Pető and Reizer 2021) and differences in personal characteristics and their pay. Our paper is most closely related to the set of papers examining decompositions of the changes in the gender wage gap into key characteristics such as education level, age/experience, family composition, industry, occupation, union status and firm size.

Internationally, papers examining wage decompositions find that factors such as education, occupation, industry, and work experience are the primary explanatory factors of the declining gender gap. Our paper focuses on the Canadian experience and is therefore closely related to Pelletier, Patterson, and Moyser (2019b). Their research highlights the overrepresentation of women in part-time work and the distribution of men and women across different industries as the significant contributors to the wage gap. As well as, Baker and Drolet (2010b) who compare wage-based with earnings-based data to reveal that the pay ratios vary

across different geographic regions, education levels, and other demographic characteristics.

Methodologically, our paper is related to a recent set of papers using machine learning to help understand how personal characteristics affect pay differentials. For instance, Angrist and Frandsen (2022) demonstrate the potential benefit of machine learning improving causal inference by accounting for heterogeneity. Similarly, Baiardi and Naghi (2020) replicate previous studies and highlight the improved identification of causal relationships achieved by machine learning methods. In particular, our paper's empirical methodology is very closely related to Bach, Chernozhukov, and Spindler (2021) who acknowledge the heterogeneous nature of the wage gap and use decomposition and regression analysis to analyze the factors contributing to it. While our approach closely follows Bach et. al., we also build on that paper by using more recent methods of cross-fitting and double machine learning to allow for a more flexible set of ML methods to be used.

2.3 Data

The data we use comes from the Canadian Labour Force Survey (LFS) for the years 1998, 1999, 2014 and 2015. The LFS is a representative sample of the Canadian population collected monthly. Since 1997, it has included questions on wage rates for jobs held during the survey week. In the LFS design, respondents are surveyed in six consecutive months. Importantly, a respondent is asked his or her wage on a job either when the person first appears in the survey or when he or she changes jobs. Wage questions are not re-asked for people who report being on the same job in consecutive months. This results in a 'staleness' in the wage observations. To avoid any potential problems with this, we use data only from the May and November surveys in each year, pooling the samples from the two months together to get our annual level data. All calculations are obtained using the LFS weights.

We pool the years 1998 and 1999 into a single year that we refer to as 1999, and similarly for 2014 and 2015. We do so because we wish to study relatively detailed heterogeneity in the gender pay gap based on many observational characteristics, which requires larger sample

sizes. The choice of the years is based on the longest set of years in the LFS in which the occupation and industry codes have not undergone a major revision. This aids in the study of the gender gap across time for various jobs.

Our sample selection includes only those that are working for pay during the reference week and who are not currently enrolled in school. In addition, we restrict our analysis to those working at least 25 weekly hours to focus on men and women with a reasonable attachment to the labour force. Overall our sample includes just over 120,000 observations for each (aggregated) year. Additional details about our sample and variable construction can be found in Appendix B1.

Tables 2.1 and 2.2 show the means of many of the key variables used in our analysis, as well as differences in means between men and women. Starting with hourly earnings, Table 2.1 shows that women in our sample earn about 20 percent less per hour than their male counterparts (difference of .201 log points) in 1999. However, by 2015 the raw gender wage gap fell to 13.5 percent, marking about a 30 percent improvement in relate wages.

Turning to human capital and workplace characteristics, a few broad patterns emerge. In 1999, women are, in general, more highly educated than man. In particular, women are less likely to have less than high school education and more likely to to have a university degree. While men are slightly more likely to have post-graduate education, advanced degrees are relatively rare in both groups. It is also apparent that men and women work in different employment environments. In particular, women are more likely to work in the public sector, less likely to be covered by a union, work fewer hours than men per week, and more likely to work in very large or small firms.

As can be seen in Table 2.2, most of the differences between men and women persist in 2015, with a few exceptions. In particular, women's educational advantage expands – women are more likely to be bachelor and advanced degrees holders and less likely to have high school or less education. Compared to 1999, women are even more likely to work in the public sector and the differential with men has increased. Interestingly, in 2015 women are more likely to be covered by a union than men. It seems that the overall decline in union coverage mainly affected men as men were more exposed when coverage in the private sector fell over the past 20 years. Overall,

both men and women have become more concentrated in very larger firms, but more so for women.

2.4 Treatment effect framework

We start by setting out a model for wages within the potential outcomes framework. This discussion closely follows Słoczyński (2015), Kline (2011) and N. Fortin, Lemieux, and Firpo (2011). For a population of individuals $i \in \{1, \ldots, N\}$, let the treatment group, women, be denoted with $D_i = 1$ and the control group, men, be $D_i = 0$. Potential wage outcomes within Rubin Causal Model (RCM) can be defined by:

1. Y_i^1: The log wage outcome of individual i if they are female.
2. Y_i^0: The log wage outcome of individual i if they are male.

For individual i, the treatment effect, the pay differential between potential outcomes, is given by $Y_i^1 - Y_i^0$. However, for any given individual, we can only observe one of these potential outcomes, depending on whether they receive the treatment or not. Thus, observed outcomes can be denoted by the following equation:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0 \tag{2.1}$$

Our main interest is in estimating the Average Treatment Effect on the Treated:

$$\text{ATET} = E[\tau_i] = E[Y_i^1 - Y_i^0 | D_i = 1]$$

Since we cannot observe both Y_i^1 and Y_i^0 for the same individual, estimating the ATET directly from the data is not possible. To overcome this problem, we make a series of assumptions:

Assumption 1: Unconfoundedness

$$Y_i^0, Y_i^1 \perp D_i | Z_i$$

This assumption states that, conditional on the covariates Z_i, the potential outcomes Y_i^0 and Y_i^1 are independent of the treatment assignment D_i. In other words, after controlling for Z_i, the treatment assignment is as good as random. In our specific application, we assume that men

and women with the same Z_i characteristics are comparable such that differences in outcomes are only due to gender itself.

Assumption 2 Overlap

$$0 < p(D_i = 1|Z_i) < 1$$

This assumes that for each subset of the population defined by the characteristics Z_i, men and women are both contained in that subset such that comparisons between the genders can be made conditional on $Z_i = z_i$.

Assumption 3 SUTVA

A maintained assumption is stable unit treatment value assumption which states that the treatment of individual i can only affect i's outcome. In other words, there are no general equilibrium effects.

Under these assumptions, the ATET can be estimated from the data by conditioning on the covariates Z_i:

$$\text{ATET}(Z_i) = E[\tau_i|Z_i] = E[Y_i^1 - Y_i^0|Z_i, D_i = 1]$$

This is the standard setup for estimating treatment effects based on unconfoundedness in the potential outcomes framework. The critical assumption is unconfoundedness, and if it holds, we can estimate the ATE by comparing the average outcomes of the treated and control groups, conditional on the covariates Z_i.

2.5 Estimation

To make headway towards empirical estimation, let outcomes be determined by:

$$Y_i^d = g^d(Z_i) + \epsilon_i^d \quad \text{for } d \in \{1, 0\},$$

where the functions $g^d(Z_i)$ are unknown functions of observable characteristics and ϵ_i^d contains unobserved determinants of wages. Writing the outcome equation 2.1 in this way implies another assumption.

Assumption 4 We further assume that the observed and unobserved determinants of wages are additively separable. This is a common assumption, particularly in methods that use decomposition (N. Fortin, Lemieux, and Firpo 2011).
The switching regression model can be written as:

$$Y_i = g^1(Z_i) + D_i \cdot (g^1(Z_i) - g^0(Z_i)) + u_i,$$

where $u_i = D_i(\epsilon_i^1 - \epsilon_i^0) + \epsilon_i^1$ and $E(u_i|D, Z) = 0$.

Define the $ATE(Z_i)$ as:

$$ATE(Z_i) = E[Y_i^1 - Y_i^0|Z_i] = g^1(Z_i) - g^0(Z_i).$$

The unknown functions $g^d(Z_i)$ allow the treatment effects (gender wage gap) to depend on Z_i, and thus allow for a great deal of heterogeneity. We restrict this heterogeneity by specifying that it is additively linear in a subset of the coefficient vector Z_i. In particular, we write:

$$ATE(Z_i) = g^1(Z_i) - g^0(Z_i) = \sum_j^{p_1} \beta_j x_{ij}.$$

Assumption 5 Heterogeneity in the gender gap is additively linear in a subset of the coefficient vector Z_i. This describes the gender gap experienced by a woman with characteristics x_{ij} compared to a man with the same characteristics. Negative coefficients for β_j indicate that x_j is associated with lower payments for women, conditional on Z_i. This functional form assumption leads to a partially linear regression model of the form:

$$Y_i = g_0(Z_i) + D_i \cdot \left(\sum_j^{p_1} \beta_j \tilde{x}_{ij} \right) + u_i \tag{2.2}$$

where $g_0(Z_i)$ is an unknown function of characteristics, common between genders. In practice, $X_i = \{1, x_{i1}, \ldots, x_{ip_1}\}$ is a vector of age, education, marital status, age composition of children, job characteristics, hours worked, industry and occupation and includes a constant. Appendix

B1 gives a complete description of these variables. Given the LFS data that we use, each of these variables is categorical, leading $p_1 = 76$ β_j coefficients. We take Z_i to include X_i, plus all two-way interactions, leading to 2423 potential controls. We place no other restrictions on $g_0(Z_i)$ and treat it as a nuisance function and estimate it with machine learning methods, as described below.

2.6 Methods

Our main estimation strategy uses the double/de-biased machine learning (DML) method to estimate causal effects in our high-dimensional setting. It combines supervised machine learning techniques with traditional econometric methods to estimate causal effects in a high-dimensional settings using cross-fitting. In particular, we use the DML2 algorithm (Chernozhukov et al. 2018b; Bach et al. 2023).

1. Split the data into K non-overlapping folds,
2. Let $w_{i\in k} = \{Y_{i\in k}, D_{i\in k}, D_i x_{i\in k,1}, \cdots, D_i x_{i\in k,p_1}\}$ be the $p_1 + 1$ vector of target variables in the ML procedure. For each fold k:
   (a) For each target variable $j \in \{1, \ldots, p_1 + 1\}$ of $w_{i\in -k,j}$, estimate an ML model using data in folds in $-k$ and $z_{i\in -k}$ as predictors.
   (b) Using the fitted model, estimate the predicted values for $\hat{w}_{i\in k,j}$, and form the residuals $\tilde{w}_{i\in k,j} = w_{i\in k,j} - \hat{w}_{i\in k,j}$,
3. Repeat for all k folds.
4. Collecting orthogonalized variables across folds, $\tilde{w}_{i,j}$, estimate an OLS regression of $\tilde{Y}_i$ on $\{\tilde{D}_i, \tilde{D_i x}_{i\in k,1}, \cdots, \tilde{D_i x}_{i\in k,p}\}$, producing a $p_1 \times 1$ vector of estimated treatment effects, $\hat{\beta}$.
5. Repeating this process R times to obtain $\hat{\beta}_r$. We take the median of the $\hat{\beta}_r$ across R estimation repeats.

Chernozhukov et al. (2018b) prove that DML delivers point estimators approximately unbiased and normally distributed. The two main ingredients to the method are the cross-fitting in step 2(a) and orthogonalization in step 2(b).

Orthogonalization involves residualizing both treatment variables, $D_i \cdot (x_1, \ldots, x_{p_1})$,

and the outcome variable, Y – what Chernozhukov et al. (2018b) call "double machine learning". The result is that the residualized OLS regression produces coefficients whose scaled estimation error contains a bias term that involves the product of estimation errors, which vanishes under general conditions. Thus, we have the residualized (of a cross-fit form) treatment and outcome variables:

$$\tilde{x}_{ij} = D_i x_{ij} - \hat{l}_{0j}(Z_i) \quad \text{for each } j \in 1, \ldots, p_1$$

$$\tilde{Y}_i = Y_i - \hat{g}_0(Z_i)$$

where $\hat{l}_{0j}$ and $\hat{g}_0$ can be estimated by any generic supervised machine learning method. Next, we estimate:

$$\tilde{Y}_i = \sum_{j}^{p_1} \beta_j \tilde{x}_{ij} + \epsilon_i$$

via OLS.

The main purpose of cross-fitting is to ensure that the same data is not used both for generating predictions and for calculating residuals. This separation helps to avoid overfitting that can lead to bias when model errors, such as ϵ_i and estimation errors in the ML procedure are related. Cross-fitting permits the use of a wide range of ML procedures and cross-validation of hyperparameters, while ensuring that biases introduced due to overfitting are negligible.

Finally, we use repeated cross-fitting in step 5. This is to deal with the uncertainty introduced by the random nature of the cross-fitting procedure. Randomly splitting the data into K folds has no impact on the asymptotic behavior of the estimation procedure, but it may be important in finite samples. In particular, the dependence of the estimator on the particular split creates an additional source of variation. We use the median method suggested in Chernozhukov et al. (2018b) for our final estimates and incorporate the uncertainty into our standard errors as they suggest.

In our specific application we chose $K = 5$ (IE, 5-fold cross-fitting) and $R = 50$ (we repeat the procedure 50 times). These choices were mainly to balance robust estimation with computation time.

2.6.1 ML methods

To implement our DML estimation, we rely several ML algorithms and their implementation in `R`.

LASSO

The LASSO regression, an acronym for Least Absolute Shrinkage and Selection Operator, can be implemented in R using the `glmnet` package. The `glmnet` function takes Z_i and one of the target variables in w_i as the input matrix of predictors and response vector respectively, and the alpha argument is set to 1 for LASSO penalty. Determining the optimal value for the regularization parameter lambda, the penalty term, is done using 5-fold cross validation. This is performed using the `cv.glmnet` function, which finds the penalty, `lambda.min`, that minimizes the cross-validation error.

Random Forests

We use the `ranger` package in `R` for the implementation of the Random Forests algorithm. Categorical features are automatically handled in the random forests algorithm. The ranger function then uses these factors to create binary splits, effectively handling the categorical data. The package offers many options to control the behaviour of the forest, including the number of trees to grow (num.trees), the minimum node size (min.node.size), and the number of predictors to use in each tree (`mtry`), among others, which can be customized by the researcher or cross-validated. In our case, we set these values to the `ranger` package default. During training, each tree is built from a bootstrapped sample of the data, and at each split in the tree, a random subset of features is considered. This process helps increase model robustness and reduces overfitting.

Boosted Trees

The `LightGBM` package, short for Light Gradient Boosting Machine, is a gradient boosting framework that uses tree-based algorithms, can estimate boosted trees in `R`, efficiently

handling categorical variables. It builds an ensemble of decision trees in a sequential manner, where each new tree is designed to correct the errors made by the previous trees. Unlike traditional boosting methods that grow trees horizontally, LightGBM grows trees vertically, meaning it chooses to split on the leaf with the highest loss reduction, a strategy called "leaf-wise" growth. This approach makes the algorithm fast and more efficient compared to other boosted tree algorithms. We use the `lgb.train` function and set `num_leaves`, `max_depth`, `learning_rate`, and `n_estimators` to their default values in the `params` list. Boosted trees are estimated in regression mode using L2 regularization.

2.7 Limitations

2.7.1 Categorical Variables

In the publicly available LFS data that we use, every variable in Z_i is categorical. This requires the researcher to choose which group to omit for each categorical variable, so that the other coefficients are interpreted as relative to this reference group created by the combination of omitted categories.

For example, suppose we have a categorical variable for education level with three categories: "high school," "college," and "graduate school." If we use "high school" as the reference group and omit it from the model, the coefficients for "college" and "graduate school" tell us the expected difference in the outcome between those education levels and "high school," holding all other variables constant. In the context of the Blinder-Oaxaca decomposition, this omission of a reference group for categorical variables can cause issues with interpretation.

The Blinder-Oaxaca decomposition breaks down the difference in mean outcomes between two groups into a part due to differences in observable characteristics (endowments) and a part due to differences in the effects of these characteristics (coefficients). Categorical variables with an omitted reference group make the interpretation of the 'endowments' and 'coefficients' effects difficult (see N. Fortin, Lemieux, and Firpo (2011)). The 'coefficients' effect will capture differences in how these characteristics relate to the outcome, again compared to

the reference group, which is essentially arbitrary. In Appendix B1 we explain our variable construction and reference group choice in detail. Briefly, we chose reference groups to be as comparable as possible with Chernozhukov et. al.

However, a limitation of the data and our approach is that if we apply the Blinder-Oaxaca decomposition with different omitted categories of the same categorical variable, we can get different results, and it can be challenging to disentangle the effects. The interpretation of the decomposition then depends critically on which category is used as the reference group. Solutions have been put forward, but none entirely solve this issue because they rely on some sort of normalization which is still arbitrary (see N. Fortin, Lemieux, and Firpo (2011)). Thus, our results should be interpreted in the context of our choice of the reference group.

2.7.2 Failure of identifying assumptions

Our approach involves a series of assumptions. The most important assumption is unconfoundedness. There are several ways that this might fail. For example, differential selection of men and women into the labour market potentially will lead to selection on unobservables that are not accounted for in our framework. Another critical assumption is our choice of variables in Z_i which may not be exogenous to wage formation. For example, men and women might self-select into different education or occupation groups. To the extent that these are problems, our estimates of the ATET can no longer be interpreted as causal. In this case, what we capture is simply conditional associations between gender pay gaps and characteristics, where the conditioning is on a very large set of controls.

2.7.3 Restricting heterogeneity to be linear

Our restriction that the ATET is linear in covariates x_{ij} is mainly for convenience, and still allows for considerable heterogeneity given the size of p_1. However, to the extent that gender gap heterogeneity is not additive, but contains interaction effects, these will be missed by our strategy. One potential solution to this problem is to use methods such as r-learner estimation or causal forests, which we will leave for future work.

2.8 Results

In our first set of estimation results, we focus on all workers in our sample. In following sections, we present results separately by education groups for those with college degrees and those with high school or less. All of our main results focus on using LASSO in our double machine learning estimation. In section 2.11, we present estimates using alternative ML methods and compare their performance.

2.8.1 All Workers

Figures 2.4 and 2.5 illustrate the coefficient estimates of equation 2.2 derived through our Double Machine Learning (DML) approach utilizing LASSO to estimate all nuisance functions. These figures comprise of three columns, showcasing results from 1999, 2015, and the changes between 1999 and 2015.

Each figure is segmented into panels, each dedicated to a specific categorical variable. Within each panel, individual bars denote specific categories. The 'base' panel displays results for the female dummy variable, our chosen baseline reference group formed by excluding a category from every other categorical variable. The bar length in each panel is proportional to the coefficient estimate.

Each bar is overlaid with a black error band signifying the confidence interval. The thicker portion of the band denotes the standard 95% confidence interval calculated using the coefficient estimate's standard error. Bars with shading in red symbolize coefficient estimates that are statistically significant at the 95% level. The slimmer, broader error bands depict the 95% uniform confidence interval, which is calculated using the multiplier bootstrap method, as suggested in Chernozhukov, Chetverikov, and Kato (2014). A red star to the right of the bar indicates coefficients that are statistically significant using the uniform confidence bands.[2]

[2]The uniform confidence interval (or a simultaneous confidence interval) adjusts individual intervals in a way that controls the overall coverage probability for all the intervals combined. The aim is to ensure that, for example, 95% of the time, all the intervals cover their true values simultaneously. This is more stringent than having each individual interval cover its true value 95% of the time, and thus these intervals are wider. These uniform confidence intervals are meant to control for type I errors (false positives) when performing multiple

Shifting attention to the 1999 results in Figure 2.4, certain variables are distinctly linked to larger gender pay gaps, subject to our set of controls. For instance, an evident age gradient is visible in the gender gap, with older workers earning less than their male counterparts with similar characteristics. The magnitude of the gender pay gaps associated with age will be addressed in a subsequent section. The bar length is relative to the omitted group, symbolizing the incremental gender gap compared to the youngest workers (aged 25-29) in our sample.

Further attributes contributing to a wider gender pay gap relate to family structure. For instance, the oldest child's age residing at home is associated with gender pay gaps. The presence of children between the ages of 13-24 is strongly correlated with gender pay differences, where the comparison group is women without children. Marital status also strongly impacts gender pay differences. Compared to single, never-married women, separated and married women earn less compared to men with the same marital status.

On the other hand, some variables are associated with smaller gender pay gaps. For example, the gender pay differences are smaller for college graduates and post-graduate workers, compared to high school workers. Although these differences are not statistically significant with the uniform confidence band, the coefficient size holds economic significance - workers with a graduate degree have a nearly 10 percent smaller gender gap compared to high school workers. Employment within a union or the public sector is also linked to narrower pay gaps.

Figure 2.5 emphasizes the within-occupation and industry pay gaps, taking Sales and Service occupations and the wholesale sector as the reference groups. For the industry categories collectively, there are several sectors with both larger and smaller pay gaps compared to the reference groups. However, only a few of these estimates deviate significantly from zero.

Industries where men earn more than women, such as agriculture and construction, are traditionally male-dominated. Conversely, industries like healthcare and accommodation/food services, where women earn more, are typically female-dominated.

statistical tests simultaneously or estimating multiple parameters.

Compared to industries, occupations have a stronger correlation with gender pay differences. Several occupations, including Trade, Childcare, Machine Operators, Manufacturing, Retail Salespersons, and Administration and Secretarial work, display large, negative pay gaps. Specifically, the pay gap in trades is over 20 percent greater than in sales and service occupations.

The second column of Figure 2.4) illustrates the results for 2015. A noteworthy observation is the disappearance of the age gradient in gender pay differentials by 2015, indicating that age no longer contributes marginally to gender pay gaps. Although our study, based on the publicly available Labour Force Survey (LFS) data, cannot conclusively identify the reasons for this intriguing finding, it may be related to the evolving labor force participation of successive cohorts of women over recent decades (Drolet 2011).

In 2015, similar to 1999, variables linked to family structure continue to play a significant role in the gender pay gap. Women, who are married and have children at home, earn less than their male counterparts with similar attributes.

A distinctive finding in the 2015 results, absent in the 1999 analysis, pertains to the gender pay gaps related to the working hours. In 2015, women working additional hours do not seem to receive equivalent compensation as men working longer hours, leading to a prominent gradient in terms of work hours. This result is conditioned on the occupation and industry of work, suggesting that it isn't merely reflecting the differing work hours typically seen in occupations populated by men and women.

Shifting focus to the second column of Figure 2.5, it's evident that the industry sector doesn't significantly influence pay differences. However, occupation appears to play a more crucial role in shaping gender pay gaps. Interestingly, the large negative gaps previously associated with several occupations, such as Trade, seem to have lessened. Additionally, several occupations now exhibit positive pay differentials, including healthcare professionals, clerical workers, and occupations in culture and arts.

The third column in Figures 2.4 and 2.5 illustrates the differences in the coefficient estimates between 1999 and 2015, thereby helping to pinpoint significant factors influencing the evolution of the gender gap over time. The most notable change in this column is the statistically significant inversion of the age gradient in gender pay differentials, with the gender pay gap decreasing most significantly for older workers. In Figure 2.5, it's also apparent that pay gaps have most substantially declined in the occupations which had the widest pay gaps in 1999. In other words, those occupations where women's relative pay was most deficient in 1999 experienced the greatest reductions in gender pay gaps over time. Notably, variables associated with family structure show no improvement in gender pay gaps over the past 15 years. This suggests that policy interventions might play a crucial role in addressing inequalities tied to child-rearing responsibilities.

Table 2.12 presents the decomposition results linked to the coefficient estimates in Figures 2.4) and 2.5 from the estimation equation 2.2. Similar to the figures, there are three columns in the table related to 1999, 2015, and their difference. Each of these columns offers a detailed decomposition of the unexplained pay gap as well as the aggregate decomposition (to be added later). The 'Effect' column outlines the total impact of a specific variable on the unexplained gap. Alongside this, we provide standard errors and t-statistics that test the null hypothesis of no effect. The standard errors are calculated using equation (22) from (Jann 2008). Statistically significant effects are highlighted in bold typeset.

In 1999, numerous variables contributed to gender wage gaps, with age and family structure being the most influential in leading to more negative gender pay differentials. Age alone accounts for nearly 24 percent of the total unexplained pay gap. Other significant variables include marital status and the age of children living at home, which together account for nearly 50 percent of the unexplained pay gap. Overall, while occupations contribute to a larger unexplained gap, union and public sector jobs help mitigate it.

By 2015, the overall unexplained gap decreased from 16.5 percent to 9.2 percent. Variables such as age and occupation no longer contribute to larger pay gaps, whereas variables

associated with family structure still do. Overall, variables related to family structure continue to account for nearly 50 percent of the unexplained gap.

2.8.2 Summary

The data provided in Figures 2.4 and 2.5, alongside Table 2.12, highlight the changes in gender pay gaps between 1999 and 2015. A key change observed is the diminishing influence of age on the pay gap by 2015, a factor that had considerable impact in 1999. Family structure, on the other hand, consistently contributed significantly to the gender pay gap across both years, with married women and women with children at home earning less than their comparable male counterparts.

Furthermore, occupation emerged as a more substantial determinant of the gender pay gap than industry sector, with pay gaps within certain occupations such as Trade showing improvement by 2015. A new finding in 2015 highlighted that women working longer hours were not compensated equivalently to men, leading to a marked work hours gradient. Over the period from 1999 to 2015, notable reductions in the gender pay gap were observed, particularly amongst older workers and in occupations with the widest initial pay gaps. Yet, there was no improvement in the pay gap linked to family structure over this period, suggesting a potential area for policy focus. According to Table 2.12's decomposition analysis, while age and occupation ceased to contribute significantly to the pay gap by 2015, variables related to family structure persistently accounted for nearly half of the unexplained gender pay gap.

2.9 Results by Education Group

We also analyzed the gender gap separately for two educational groups: those with a high school diploma or less, and those with a college degree or higher. These results are contained in Figures 2.6 to 2.9 and Tables 2.12 and 2.12. The presentation of these subgroup results are identical to our previous presentation for all workers.

Although the gender gap is smaller among college graduates, the trends observed largely mirror those from the broader analysis. For instance, the age coefficient is still contributing to the gender gap among older workers, with older workers earning less than their male counterparts. Similarly, we also find the presence of children to correlate with gender pay differences, however the age range is reduced to 18-24.

When looking at the decomposition results of College graduates in 1999, we observe age, occupation and the age of children living at home, playing a significant role in exacerbating the gender wage gap. With almost 70 percent of the unexplained gender gap alone being accounted by occupation.

In 2015, we see a reduction of the unexplained gender gap from 10.6 percent to 7 percent. While factors like age and occupation no longer account for significant pay gaps, variables related to family structure still contribute to wider disparities in wages. Interestingly the hours worked coefficient emerges as the primary catalyst for intensifying the gender-based wage disparities which we do not observe in the broader analysis.

The second educational group, high school diploma or less, exhibits similar negative effects of age, marital status, and age of the oldest child on the gender wage gap in 1999. By 2015 we observe a disappearance of the significance of these coefficients, indicating that these factors no longer contribute marginally to it.

The decomposition results of the high school group reveal that in 1999, age, marital status and age of children living at home are the biggest contributors to the unexplained gender gap. Combined they explain 64 percent of it. The unexplained gender gap falls from 21.8 percent in 1999, to 13.5 percent in 2015. Age no longer contributes to the pay gap unlike marital status and age of the children. A notable finding is seen in 2015, occupation's sign reverses, from negative to positive, contributing to reducing the gender gap, unlike what we see with the college graduates.

The substantial reduction in the gender pay gap seen among less educated workers compared to college graduate workers implies that education level has a significant effect on

the gender pay gap. With those holding a high school diploma or less experiencing a more pronounced decline in pay disparity over the period from 1999 to 2015. Despite the smaller overall gap among more educated workers, the persistence of factors such as family structure in influencing the pay gap underscores the need for targeted policy interventions across all education levels.

2.10 Average Treatment Effects

In the preceding results sections, we examined the influence of demographic and job characteristics on the gender wage gap. The marginal effects revealed substantial variation, or heterogeneity, in the gender wage gap. This variation is particularly driven by factors such as family structure, occupations, and employment in a union or the public sector. In this section, we further examine the variable $ATET(Z_i)$ to understand how the extent of the gender wage gap varies depending on specific characteristics.

Figure 2.10 presents a visual depiction of the $ATET(Z_i)$, sorted by percentile rank, for each year. This visualization allows us to glean several key empirical insights from our study. Firstly, it's evident from the figure that the gender wage gap varies significantly across the population. For instance, women in the 25th percentile of the $ATET(Z_i)$ earn roughly 25 percent less than their male counterparts, while women in the 75th percentile earn just under 10 percent less than men with similar characteristics.

Secondly, we observed a decline in the gender wage gap throughout the $ATET(Z_i)$ distribution, averaging nearly a 7 percent decrease between 1999 and 2015. However, this reduction was not consistent across all percentiles. The most significant improvement was found at the lower percentiles of the $ATET(Z_i)$ distribution. The black dashed line in the figure depicts the difference in $ATET(Z_i)$ at each percentile between the two years. Starting at over a 10 percent difference, this line converges to zero at the uppermost end of the $ATET(Z_i)$ distribution, where the $ATET(Z_i)$ is predicted to be positive.

Figure 2.11 illustrates the $ATET(Z_i)$ for specific Z_i values. Specifically, the first two panels of the figure represent the anticipated gender wage gap for different marital statuses, segregated by year. For the "single" category, we assume no children at home, while for other marital statuses, the $ATET(Z_i)$ is calculated assuming average values of children's ages within the population. All other characteristics are evaluated at the sample mean. We conduct this analysis separately for different education groups.

For those with a high school education or less, single women in 1999 earned about 12 percent less than men. This wage gap increases to 25 percent for married women, revealing a significant wage penalty for married women compared to men. However, by 2015, all these gender wage gaps had diminished, indicating improvement across all marital statuses. Furthermore, the steepness of the $ATET(Z_i)$ curve became less pronounced, suggesting a decreased marital wage penalty over time. Among college graduates, the $ATET(Z_i)$ curves were lower than for those with a high school education or less, in both years. Specifically, although single college-educated women had negative point estimates, the data does not conclusively prove a gender wage gap. Yet, a gender wage gap persisted for married college graduates, albeit declining over time.

The bottom two panels of the figure present the $ATET(Z_i)$ profiles for age, evaluated at the average sample characteristics. For high school graduates in 1999, an age gradient is apparent, with gender wage gaps increasing with age. This entire profile shifts upwards by 2015, indicating reductions in gender wage gaps across all ages and a lessening age gradient. However, an exception exists for the oldest high school workers where the age gradient persisted. For college graduates in 1999, the age gradient wasn't apparent until workers reached their 40s. By 2015, age was no longer a significant predictor of gender wage gaps for college-educated workers.

Lastly, in Figure 2.12, we present the $ATET(Z_i)$, evaluated at average characteristics within the population, for each occupational group, separating by education status. The figure illustrates gender wage gaps within each occupation for each year. It also includes a grey shaded bar to represent the magnitude and direction of change over time. As depicted in the figure, for high school graduates specifically, the gender wage gaps decreased the most in occupations that had the largest gender pay gaps in 1999. Although gender pay gaps within occupations

generally declined over time, they remain significant.

2.11 Robustness check

We've undertaken several steps to assess the robustness of our findings. Primarily, we've explored using different machine learning methods to estimate nuisance functions. Our main results employ the LASSO method, a parametric, supervised machine learning tool. However, we've also experimented with two tree-based methods, random forest and boosted trees, which allow for a more intricate and flexible estimation of the unknown nuisance functions and enable more complex interactions. These tree-based methods excel at handling categorical variables and complex, non-linear relationships between variables.

Figures 2.13 and 2.14 presents the coefficient estimates derived from these methods, alongside the baseline LASSO method. As evident, the magnitude of the coefficient estimates is remarkably consistent across LASSO, Boosted Trees, and Random Forest, affirming that our results remain robust regardless of the specific machine learning method used in our DML estimation strategy.

In addition to this, in Figures 2.15 and 2.16, we've compared our baseline LASSO model to other parametric machine learning methods as well as to Ordinary Least Squares (OLS). More specifically, we've juxtaposed our baseline method of cross-validating our penalty parameter and selecting the penalty parameter that minimizes out-of-sample cross-validated prediction error, with an approach that employs the 1-standard deviation rule. This latter approach implies more regularization (larger penalty). Additionally, we've also examined results for Ridge regression and OLS (no regularization). Again, the coefficient estimates demonstrate significant similarity across these methods, reinforcing the robustness of our results.

Tables 2.12 and 2.12 contain the decomposition results across all of our estimation methods, for the years 1999 and 2015 respectively, for all workers and both education breakdowns. Not surprisingly, given the similarity of coefficient estimates, the decomposition results are very

similar across ML strategies.

2.12 Conclusion

In conclusion, our study uncovers shifts and enduring patterns in the gender wage gap in Canada between 1999 and 2015. We adopted recent advances in causal estimation that employ machine learning techniques to precisely estimate coefficients and identify key variables influencing wage disparities while flexibly controlling for potential confounding variables. Our findings confirm that age and family structure play prominent roles in contributing to the unexplained gender wage gap, with certain occupations and industries also influencing wage disparities, often echoing traditional gender roles.

Interestingly, our analysis documents a significant contraction of the Canadian gender wage gap from 1999 to 2015, especially among workers with lower educational attainment. We found that age and occupation have diminished in their relevance to wage disparities over time. However, variables related to family structure remain substantial contributors, even among highly educated workers. This highlights the imperative to confront societal norms and policies around family roles and duties.

We assessed the robustness of our results using alternative machine learning methods. The findings were consistent across different techniques, including LASSO, Boosted Trees, and Random Forest. Further, our results demonstrated stability when comparing LASSO with other parametric methods and with Ordinary Least Squares.

Our research underscores the need for more refined and targeted policies to mitigate gender wage disparities. The enduring role of family-related factors in the gender wage gap, even as other variables have lessened in impact, emphasizes the importance of policies addressing aspects such as childcare and marital status. The strong association between family structure and the gender wage gap suggests the necessity for family-friendly workplace policies. Policymakers should contemplate strategies to foster shared domestic and childcare

responsibilities, such as offering parental leave for both parents. Implementing such policies could alleviate career interruptions that disproportionately burden women, thus reducing wage disparities.

Moreover, our findings regarding the narrowing gender wage gap among less educated workers and the decreasing impact of age and occupation over time may indicate the efficacy of existing policies or societal changes. Nonetheless, the persistent significance of family-related factors among all workers, including the highly educated, reveals areas requiring further intervention.

Our study's capacity to pinpoint the most influential factors on the gender wage gap equips policymakers to devise targeted strategies. While addressing societal norms and biases poses a complex and long-term challenge, our findings highlight tangible sectors where policy intervention could bring about tangible change.

Table 2.1: Sample means and comparison between men and women: 1999

	Male (N=70851)		Female (N=56857)		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
A: Earnings						
log hourly wage	2.95	0.002	2.75	0.002	-0.201	-62.82
B: Education						
Less than HS	0.15	0.002	0.11	0.002	-0.044	-18.76
High School	0.20	0.002	0.22	0.002	0.019	6.49
Some post-secondary	0.07	0.001	0.08	0.001	0.006	3.30
Diploma	0.36	0.002	0.36	0.003	0.002	0.71
BA	0.14	0.002	0.17	0.002	0.025	9.05
Graduate Degree	0.07	0.001	0.06	0.001	-0.008	-4.38
C: Age						
age25 to 29	0.16	0.002	0.16	0.002	0.002	0.87
age30 to 34	0.18	0.002	0.17	0.002	-0.009	-3.29
age35 to 39	0.20	0.002	0.19	0.002	-0.006	-2.26
age40 to 44	0.19	0.002	0.19	0.002	0.004	1.43
age45 to 49	0.15	0.002	0.17	0.002	0.011	4.31
age50 to 54	0.12	0.002	0.12	0.002	-0.002	-0.85
C: Marital Status						
Single, never married	0.05	0.001	0.04	0.001	-0.007	-3.98
Married	0.72	0.002	0.70	0.003	-0.020	-5.93
Living in common-law	0.16	0.002	0.14	0.002	-0.024	-8.89
Separated	0.07	0.001	0.12	0.002	0.050	24.21
D: Youngest Child						
No kids	0.47	0.002	0.45	0.003	-0.017	-4.71
less than 3	0.12	0.002	0.09	0.002	-0.030	-13.88
3 to 5	0.08	0.001	0.08	0.001	-0.008	-4.34
6 to 12	0.16	0.002	0.17	0.002	0.010	4.06
13 to 15	0.06	0.001	0.07	0.001	0.010	6.04
16 to 17	0.04	0.001	0.05	0.001	0.010	7.06
18 to 24	0.07	0.001	0.09	0.002	0.024	12.39
E: Employer						
Private employee	0.80	0.002	0.69	0.002	-0.114	-37.08
Public employee	0.20	0.002	0.31	0.002	0.114	37.08
F: Union Status						
Not member or covered	0.62	0.002	0.64	0.003	0.016	4.76
Union	0.38	0.002	0.36	0.003	-0.016	-4.76
G: Hours Worked						
(25,30]	0.02	0.001	0.09	0.002	0.074	44.80
(30,35]	0.07	0.001	0.20	0.002	0.128	51.11
(35,40]	0.69	0.002	0.62	0.003	-0.069	-20.03
(40,50]	0.16	0.002	0.07	0.001	-0.088	-39.84
(50,100]	0.07	0.001	0.02	0.001	-0.046	-33.34
H: Firm Size						
More than 500	0.49	0.002	0.50	0.003	0.013	3.52
Less than 20	0.18	0.002	0.19	0.002	0.009	3.45
20 to 99	0.17	0.002	0.15	0.002	-0.022	-8.06
100 to 500	0.16	0.002	0.16	0.002	0.000	-0.19

Note:
Means and difference in means of key variables for men and women.

Table 2.2: Sample means and comparison between men and women: 2015

	Male (N=64317)		Female (N=58027)		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
A: Earnings						
log hourly wage	3.04	0.002	2.91	0.003	-0.135	-37.75
B: Education						
Less than HS	0.07	0.001	0.04	0.001	-0.034	-20.39
High School	0.19	0.002	0.15	0.002	-0.040	-14.02
Some post-secondary	0.05	0.001	0.04	0.001	-0.008	-5.10
Diploma	0.40	0.003	0.39	0.003	-0.006	-1.50
BA	0.20	0.002	0.27	0.003	0.069	20.08
Graduate Degree	0.09	0.002	0.11	0.002	0.018	7.23
C: Age						
age25 to 29	0.16	0.002	0.16	0.002	-0.002	-0.73
age30 to 34	0.17	0.002	0.17	0.002	-0.005	-1.53
age35 to 39	0.16	0.002	0.16	0.002	-0.008	-2.86
age40 to 44	0.16	0.002	0.16	0.002	-0.004	-1.58
age45 to 49	0.16	0.002	0.17	0.002	0.011	3.94
age50 to 54	0.18	0.002	0.19	0.002	0.008	2.86
C: Marital Status						
Single, never married	0.25	0.002	0.21	0.002	-0.040	-11.86
Married	0.51	0.003	0.51	0.003	-0.002	-0.39
Living in common-law	0.19	0.002	0.19	0.002	0.005	1.52
Separated	0.06	0.001	0.10	0.002	0.037	18.21
D: Youngest Child						
No kids	0.51	0.003	0.46	0.003	-0.050	-13.00
less than 3	0.13	0.002	0.11	0.002	-0.021	-8.45
3 to 5	0.08	0.001	0.08	0.002	-0.006	-2.68
6 to 12	0.14	0.002	0.16	0.002	0.018	6.46
13 to 15	0.05	0.001	0.07	0.001	0.014	8.13
16 to 17	0.03	0.001	0.04	0.001	0.011	7.96
18 to 24	0.06	0.001	0.10	0.002	0.034	16.60
E: Employer						
Private employee	0.82	0.002	0.66	0.003	-0.161	-49.01
Public employee	0.18	0.002	0.34	0.003	0.161	49.01
F: Union Status						
Not member or covered	0.69	0.002	0.64	0.003	-0.051	-14.26
Union	0.31	0.002	0.36	0.003	0.051	14.26
G: Hours Worked						
(25,30]	0.02	0.001	0.08	0.001	0.057	33.49
(30,35]	0.08	0.001	0.20	0.002	0.123	45.68
(35,40]	0.71	0.002	0.63	0.003	-0.077	-21.41
(40,50]	0.14	0.002	0.07	0.001	-0.071	-31.83
(50,100]	0.05	0.001	0.02	0.001	-0.031	-22.29
H: Firm Size						
More than 500	0.51	0.003	0.56	0.003	0.055	14.17
Less than 20	0.17	0.002	0.16	0.002	-0.008	-2.75
20 to 99	0.17	0.002	0.15	0.002	-0.025	-8.84
100 to 500	0.16	0.002	0.14	0.002	-0.022	-8.06

Note:
Means and difference in means of key variables for men and women.

Table 2.3: Decomposition

	1999			2015			2015-1999		
Variable	Effect	s.e	*t*-statistic	Effect	s.e	*t*-statistic	Difference	s.e	*t*-statistic
base	-0.046	0.024	-1.95	-0.031	0.022	-1.44	-0.015	0.032	-0.46
age	**-0.039**	0.007	-5.51	-0.008	0.008	-1.05	**-0.031**	0.011	-2.98
educ	0.006	0.006	1.00	-0.001	0.007	-0.19	0.007	0.009	0.78
hours	-0.014	0.014	-1.05	**-0.049**	0.014	-3.39	0.034	0.020	1.73
union	**0.02**	0.003	6.96	**0.013**	0.003	3.74	0.007	0.004	1.60
public	**0.008**	0.004	2.06	**0.011**	0.005	2.21	-0.003	0.006	-0.43
firmsize	-0.001	0.003	-0.18	0	0.003	0.08	-0.001	0.005	-0.18
occ	**-0.027**	0.010	-2.67	0.016	0.011	1.54	**-0.043**	0.015	-2.95
ind	0.009	0.014	0.70	0.001	0.015	0.10	0.008	0.020	0.39
marstat	**-0.052**	0.013	-4.07	**-0.028**	0.007	-3.91	-0.024	0.015	-1.66
kidage	**-0.029**	0.004	-7.40	**-0.018**	0.004	-4.11	**-0.011**	0.006	-1.96
prov	**-0.009**	0.004	-2.32	-0.003	0.004	-0.77	-0.006	0.006	-0.95
cma	**0.009**	0.003	3.47	0.005	0.003	1.58	0.005	0.004	1.18
Unexplained	**-0.165**	**0.004**	**-42.21**	**-0.092**	**0.004**	**-22.49**	**-0.073**	**0.006**	**-12.92**
Raw Gap	**-0.218**	**0.002**	**-87.96**	**-0.139**	**0.003**	**-54.45**	**-0.079**	**0.004**	**-22.04**
Explained Gap	**-0.053**			**-0.047**			**-0.005**		

Note:
Decomposition results of the gender wage gap.

Table 2.4: Decomposition: College Graduates

	1999			2015			2015-1999		
Variable	Effect	s.e	*t*-statistic	Effect	s.e	*t*-statistic	Difference	s.e	*t*-statistic
base	0.063	0.052	1.22	-0.026	0.044	-0.57	0.089	0.068	1.30
age	**-0.028**	0.013	-2.06	0.004	0.014	0.31	-0.032	0.019	-1.67
educ	0.004	0.004	0.94	**0.011**	0.004	2.71	-0.007	0.005	-1.30
hours	-0.028	0.026	-1.08	**-0.077**	0.028	-2.78	0.049	0.038	1.30
union	**0.03**	0.008	3.56	0.009	0.008	1.17	0.021	0.011	1.86
public	0.011	0.013	0.83	0.019	0.012	1.62	-0.008	0.018	-0.48
firmsize	-0.007	0.006	-1.16	0.002	0.005	0.32	-0.008	0.008	-1.07
occ	**-0.074**	0.035	-2.09	0.032	0.029	1.10	**-0.106**	0.046	-2.31
ind	-0.03	0.034	-0.88	0.009	0.029	0.32	-0.039	0.045	-0.88
marstat	-0.037	0.025	-1.47	**-0.029**	0.013	-2.21	-0.009	0.029	-0.30
kidage	**-0.026**	0.007	-3.61	**-0.02**	0.008	-2.46	-0.005	0.011	-0.50
prov	0.009	0.008	1.15	-0.011	0.008	-1.33	0.02	0.011	1.75
cma	0.006	0.006	0.95	0.005	0.006	0.87	0.001	0.009	0.08
Unexplained	**-0.106**	**0.007**	**-16.25**	**-0.07**	**0.004**	**-18.62**	**-0.036**	**0.008**	**-4.74**
Raw Gap	**-0.155**	**0.005**	**-28.44**	**-0.093**	**0.005**	**-18.46**	**-0.062**	**0.007**	**-8.34**
Explained Gap	**-0.049**			**-0.023**			**-0.026**		

Note:
Decomposition results of the gender wage gap, College Graduates subgroup.

Table 2.5: Decompositoin: High School or less

	1999			2015			2015-1999		
Variable	Effect	s.e	*t*-statistic	Effect	s.e	*t*-statistic	Difference	s.e	*t*-statistic
base	-0.025	0.036	-0.69	-0.064	0.037	-1.74	0.039	0.052	0.75
age	**-0.046**	0.013	-3.58	-0.02	0.015	-1.33	-0.026	0.020	-1.30
educ	0	0.004	-0.02	0.001	0.003	0.46	-0.001	0.005	-0.31
hours	-0.018	0.022	-0.81	**-0.047**	0.022	-2.18	0.03	0.031	0.97
union	**0.012**	0.004	3.31	**0.014**	0.004	3.31	-0.002	0.005	-0.33
public	**0.008**	0.004	2.28	0.004	0.004	1.07	0.004	0.005	0.70
firmsize	-0.003	0.006	-0.55	-0.006	0.007	-0.95	0.003	0.009	0.34
occ	-0.025	0.013	-1.95	**0.032**	0.014	2.24	**-0.058**	0.019	-2.97
ind	-0.015	0.019	-0.78	0.035	0.023	1.52	-0.05	0.030	-1.67
marstat	**-0.061**	0.021	-2.99	**-0.031**	0.013	-2.35	-0.03	0.024	-1.24
kidage	**-0.033**	0.007	-5.07	**-0.023**	0.007	-3.21	-0.01	0.010	-1.01
prov	**-0.018**	0.006	-2.87	**-0.025**	0.008	-3.05	0.007	0.010	0.69
cma	0.007	0.005	1.48	-0.005	0.005	-0.86	0.011	0.007	1.62
Unexplained	**-0.218**	**0.006**	**-33.74**	**-0.135**	**0.008**	**-16.47**	**-0.083**	**0.010**	**-7.96**
Raw Gap	**-0.276**	**0.004**	**-72.93**	**-0.227**	**0.005**	**-49.72**	**-0.049**	**0.006**	**-8.22**
Explained Gap	**-0.058**			**-0.093**			**0.034**		

Note:
Decomposition results of the gender wage gap, High school or less subgroup.

Table 2.6: Comparison of Decomposition across methods: 1999

	All Workers						College Graduates			High school or less		
Variable	lasso	lgb	RF	lasso-1se	ridge	ols	lasso	lgb	ols	lasso	lgb	ols
age	-0.039	-0.038	-0.033	-0.037	-0.034	-0.042	-0.028	-0.032	-0.041	-0.046	-0.036	-0.043
base	-0.046	-0.087	0.015	-0.017	-0.109	-0.042	0.063	-0.037	-0.043	-0.025	-0.087	-0.029
cma	0.009	0.008	0.011	0.009	0.010	0.009	0.006	0.003	0.007	0.007	0.006	0.008
educ	0.006	0.005	-0.005	0.002	0.012	0.008	0.004	0.001	0.005	0.000	0.000	0.001
firmsize	-0.001	-0.005	-0.004	0.000	0.000	-0.002	-0.007	-0.003	-0.003	-0.003	-0.005	-0.005
hours	-0.014	-0.003	-0.027	-0.024	-0.006	-0.020	-0.028	0.009	-0.021	-0.018	-0.015	-0.022
ind	0.009	0.022	0.014	-0.002	0.020	0.014	-0.030	0.001	0.042	-0.015	0.024	-0.008
kidage	-0.029	-0.026	-0.030	-0.031	-0.029	-0.027	-0.026	-0.019	-0.024	-0.033	-0.025	-0.027
marstat	-0.052	-0.050	-0.049	-0.055	-0.046	-0.051	-0.037	-0.044	-0.036	-0.061	-0.062	-0.062
occ	-0.027	-0.004	-0.054	-0.028	-0.005	-0.029	-0.074	-0.026	-0.032	-0.025	-0.011	-0.029
prov	-0.009	-0.011	-0.017	-0.011	-0.007	-0.010	0.009	0.009	0.012	-0.018	-0.018	-0.020
public	0.008	0.008	0.005	0.008	0.007	0.008	0.011	0.010	0.010	0.008	0.007	0.010
union	0.020	0.017	0.016	0.020	0.021	0.019	0.030	0.028	0.027	0.012	0.009	0.011
Total	-0.165	-0.162	-0.159	-0.166	-0.165	-0.165	-0.106	-0.101	-0.098	-0.218	-0.213	-0.215

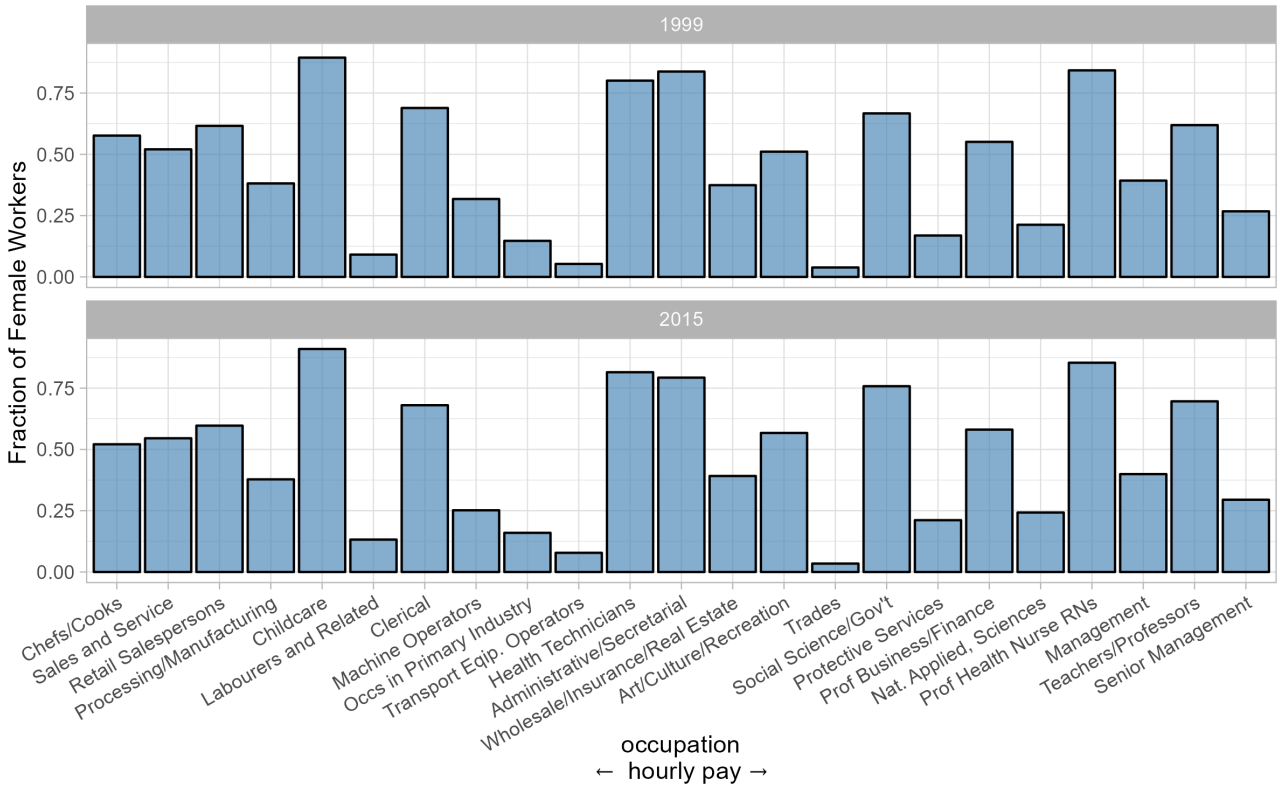
Note:
Decomposition results across estimation methods, for All Workers, College Graduates and High school or less subgroups.

Table 2.7: Comparison of Decomposition across methods: 2015

	All Workers						College Graduates			High school or less		
Variable	lasso	lgb	RF	lasso-1se	ridge	ols	lasso	lgb	ols	lasso	lgb	ols
age	-0.008	-0.009	0.002	-0.010	0.002	-0.007	0.004	-0.003	0.009	-0.020	-0.018	0.003
base	-0.031	-0.078	0.009	-0.026	-0.145	-0.046	-0.026	-0.036	-0.116	-0.064	-0.020	0.019
cma	0.005	0.004	0.002	0.003	0.007	0.006	0.005	0.005	0.013	-0.005	-0.007	-0.006
educ	-0.001	0.006	-0.004	-0.008	0.004	0.002	0.011	0.013	0.012	0.001	0.003	0.002
firmsize	0.000	-0.002	-0.002	0.000	0.003	0.001	0.002	0.003	0.004	-0.006	-0.006	-0.009
hours	-0.049	-0.034	-0.052	-0.044	-0.036	-0.060	-0.077	-0.056	-0.092	-0.047	-0.053	-0.083
ind	0.001	0.014	0.014	0.008	0.021	0.010	0.009	0.005	0.052	0.035	0.019	-0.020
kidage	-0.018	-0.015	-0.019	-0.019	-0.018	-0.016	-0.020	-0.015	-0.021	-0.023	-0.014	-0.024
marstat	-0.028	-0.025	-0.036	-0.030	-0.027	-0.029	-0.029	-0.022	-0.024	-0.031	-0.029	-0.026
occ	0.016	0.031	-0.010	0.019	0.069	0.027	0.032	0.012	0.059	0.032	0.008	0.016
prov	-0.003	-0.005	-0.012	-0.010	0.003	-0.002	-0.011	-0.006	0.006	-0.025	-0.022	-0.020
public	0.011	0.012	0.008	0.010	0.011	0.011	0.019	0.021	0.025	0.004	0.003	0.003
union	0.013	0.012	0.009	0.011	0.015	0.013	0.009	0.005	0.008	0.014	0.012	0.015
Total	-0.092	-0.090	-0.092	-0.096	-0.091	-0.091	-0.070	-0.075	-0.067	-0.135	-0.122	-0.129

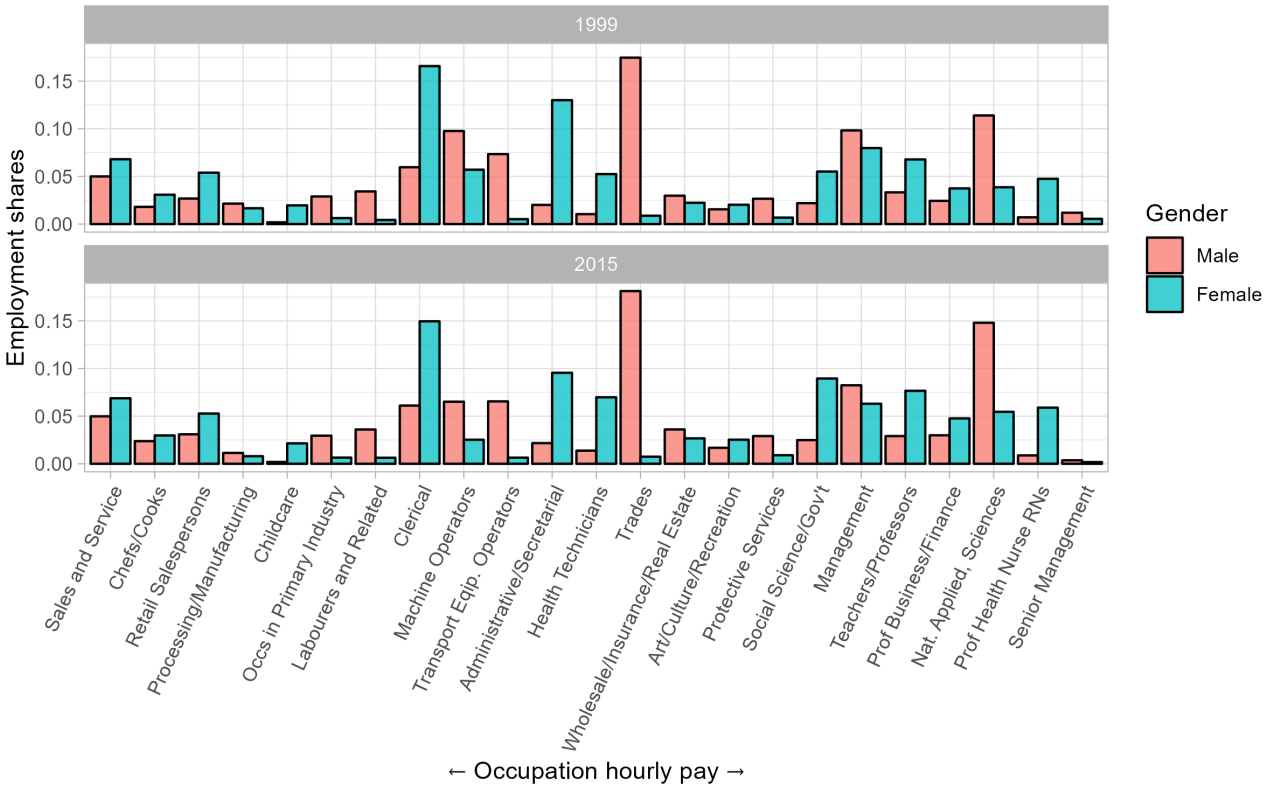
Note:
Decomposition results across estimation methods, for All Workers, College Graduates and High school or less subgroups.

Figure 2.1: Fraction of female workers by occupation



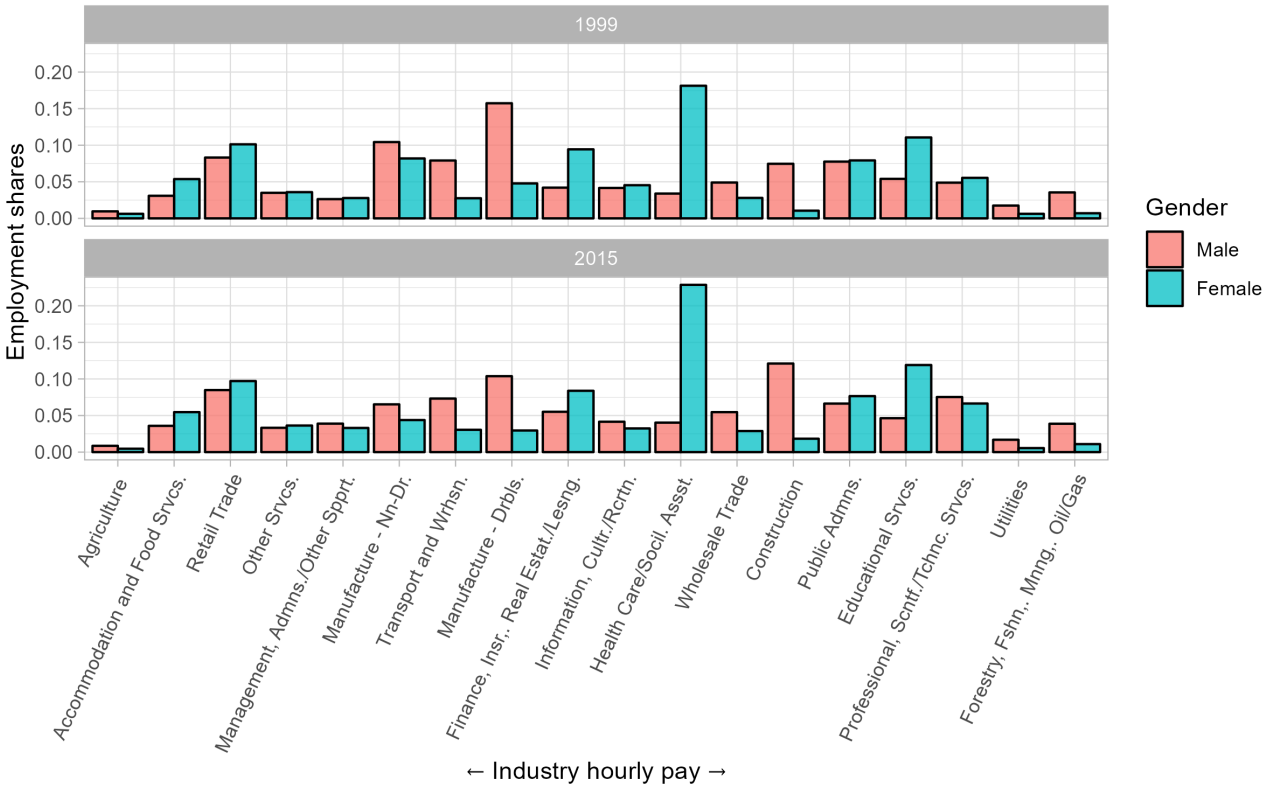
Notes: Occupation is sorted according to hourly wage, with the lowest on the left side and highest on the right.

Figure 2.2: Employment share by occupation



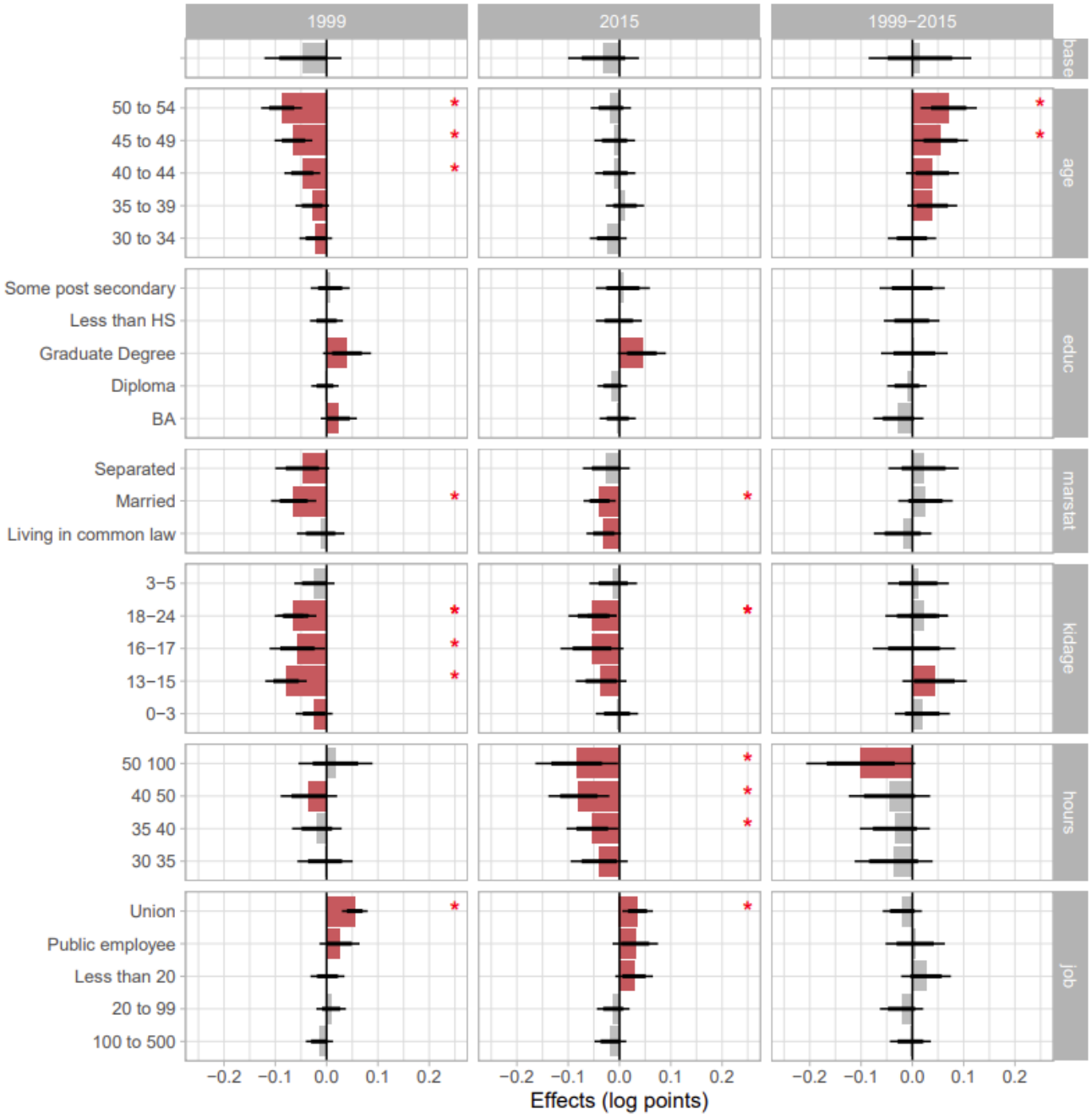
Notes: Occupation is sorted according to hourly wage, with the lowest on the left side and highest on the right.

Figure 2.3: Employment share by industry



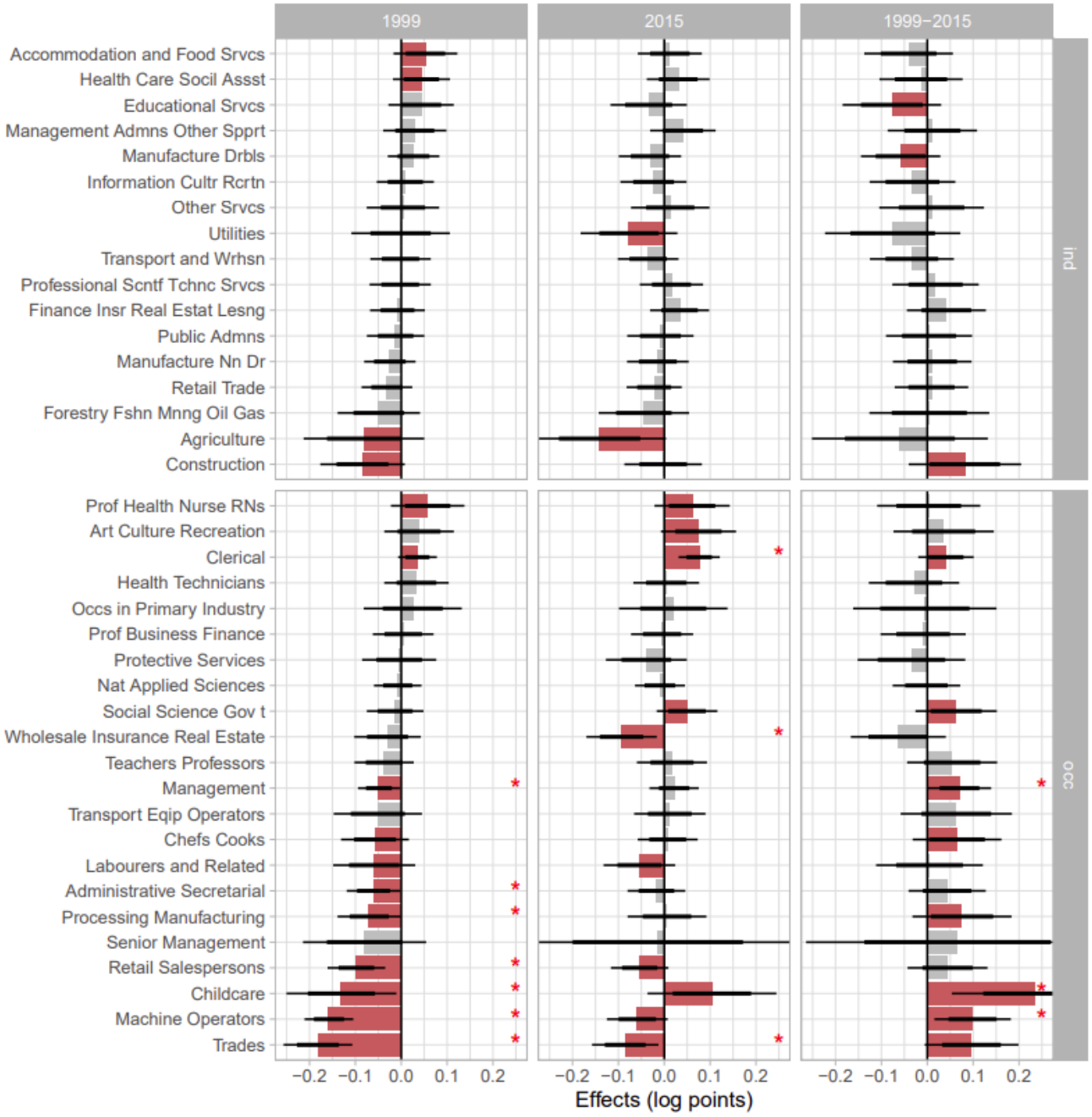
Notes: Industry is sorted according to hourly wage, with the lowest on the left side and highest on the right.

Figure 2.4: Effects of selected variables on the gender wage gap



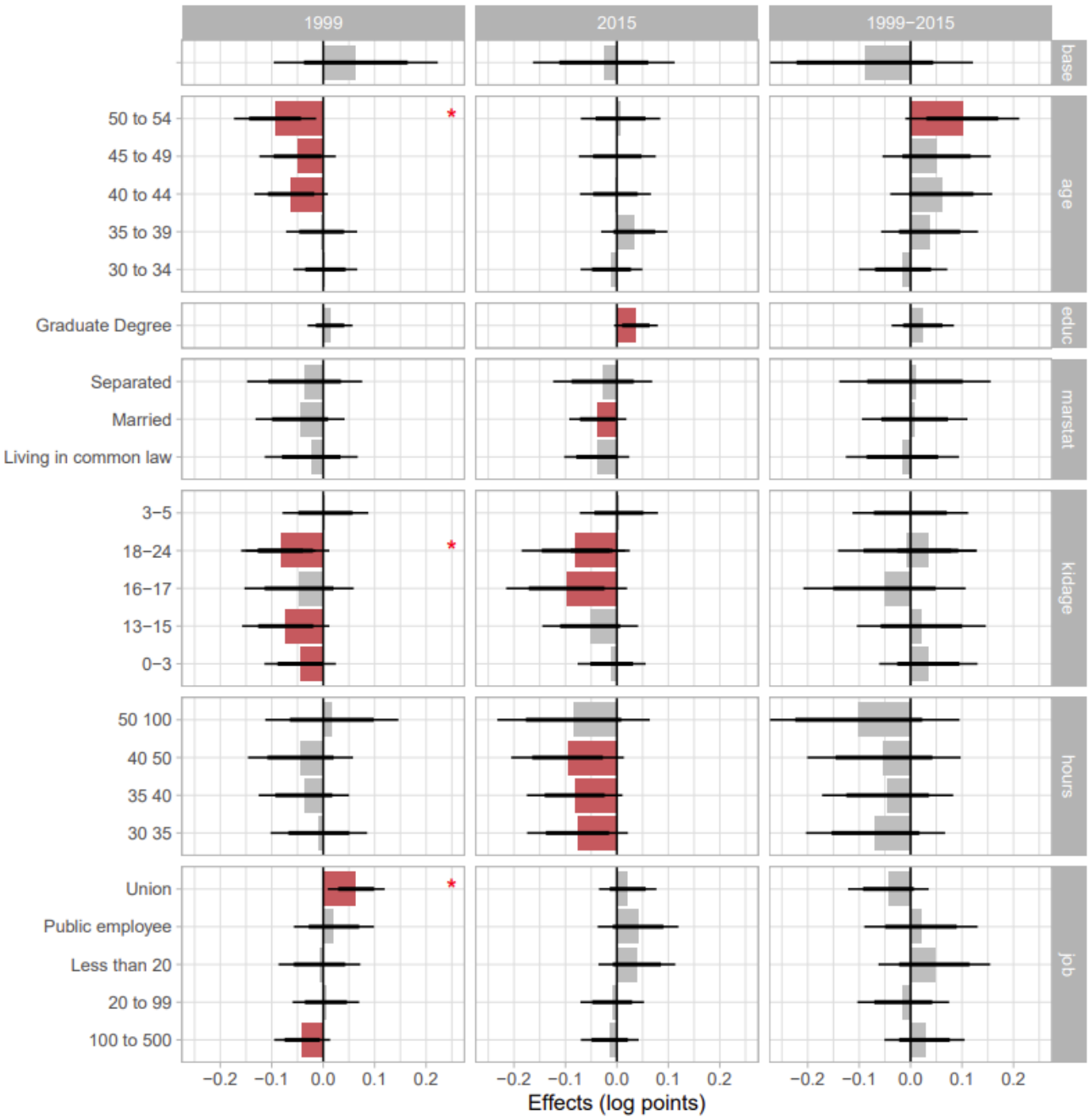
Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.5: Effects of Occupation and Industry on the gender wage gap.



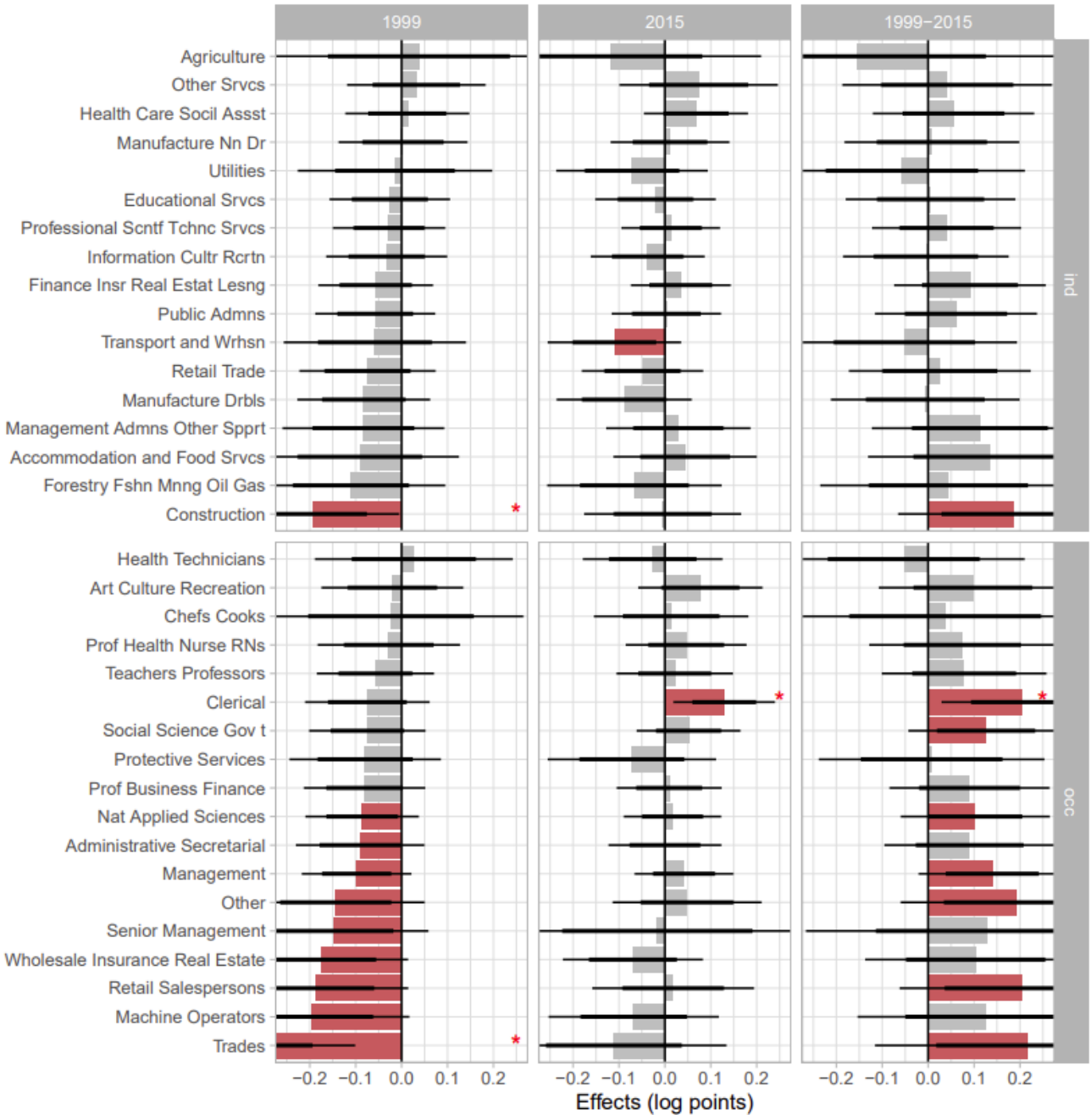
Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.6: Effects of selected variables on the gender wage gap, College Graduates subgroup.



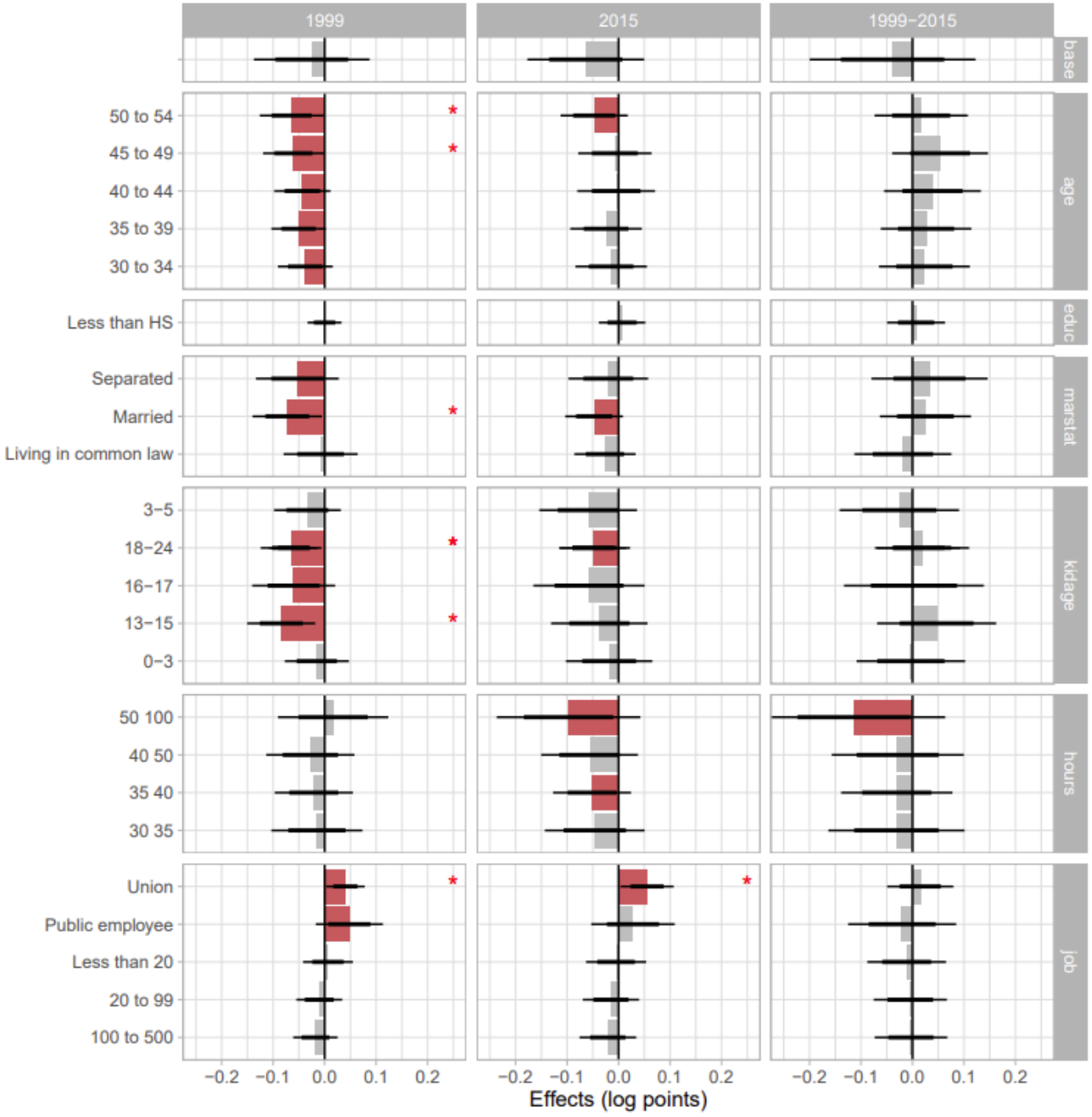
Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.7: Effects of Occupation and Industry on the gender wage gap, College Graduates subgroup



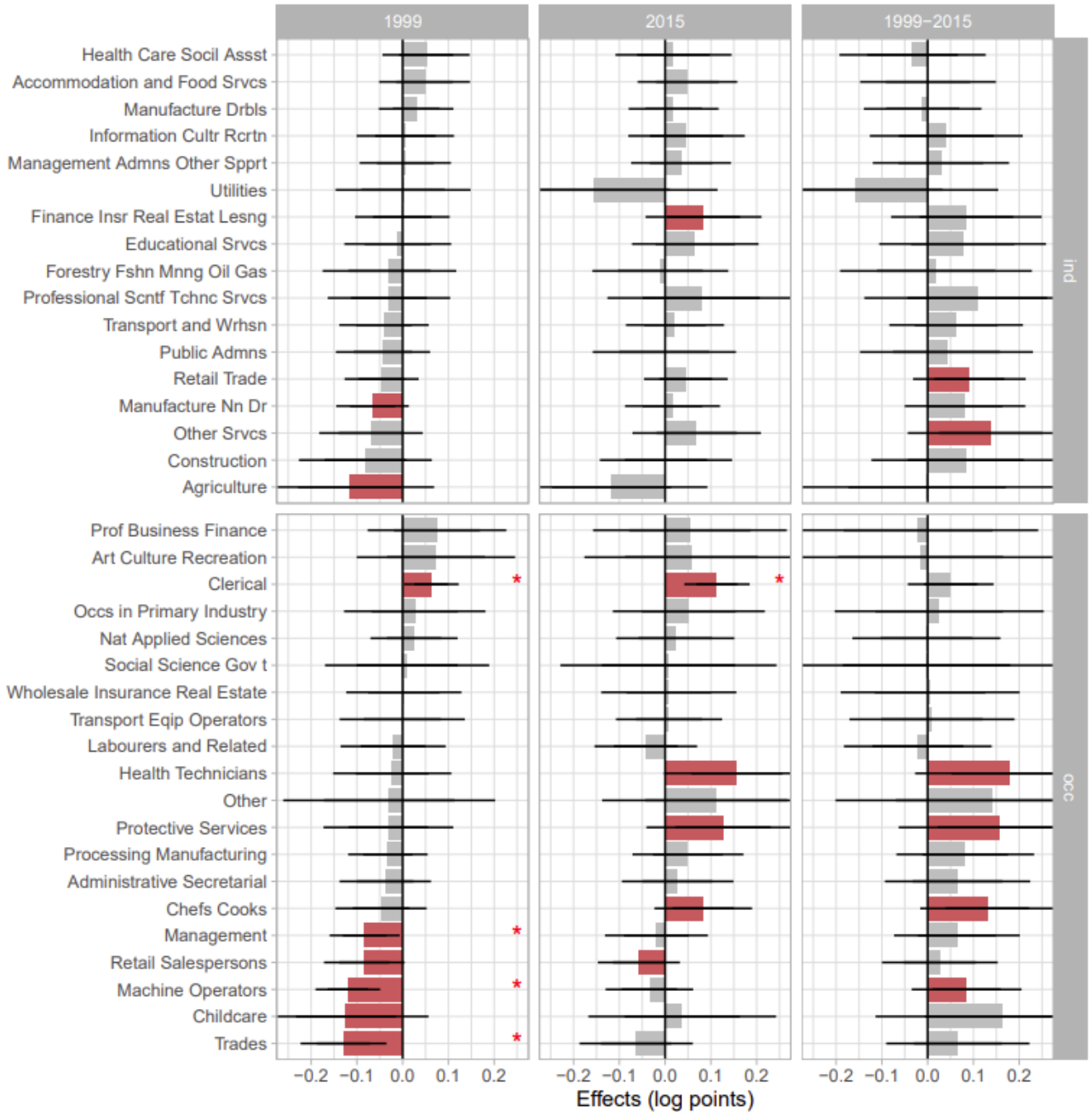
Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.8: Effects of Occupation and Industry on the gender wage gap, High School or Less subgroup.



Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.9: Effects of Occupation and Industry on the gender wage gap, High School or Less subgroup.



Notes: The thicker portion of the black band line represents the 95% confidence interval calculated using the coefficient estimate's standard error. Red bar indicates statistical significance. The thinner portion of the black band is the standard error calculated from Chernozhukov et. al (2014). Star indicates statistical significance from uniform confidence bands.

Figure 2.10: Average Treatment Effect on the Treated

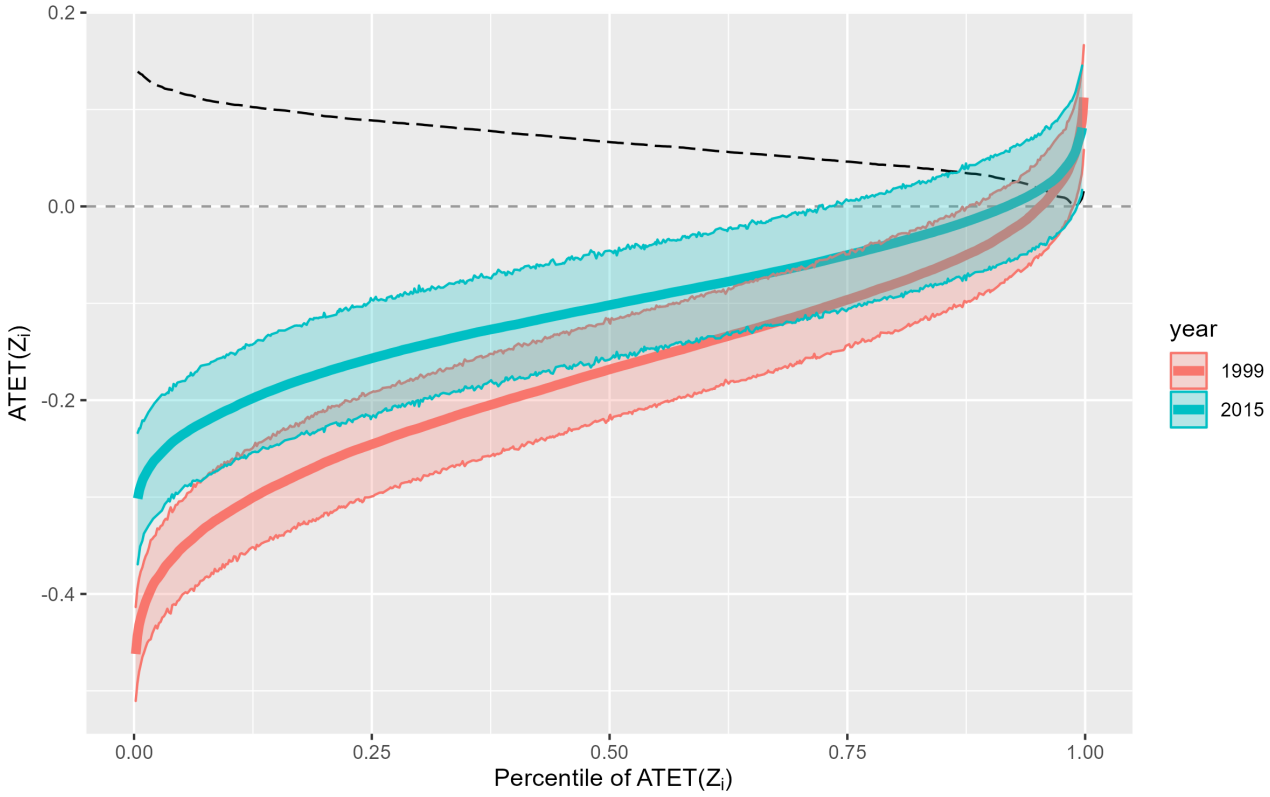


Figure 2.11: Average Treatment Effects for Select Demographic Groups

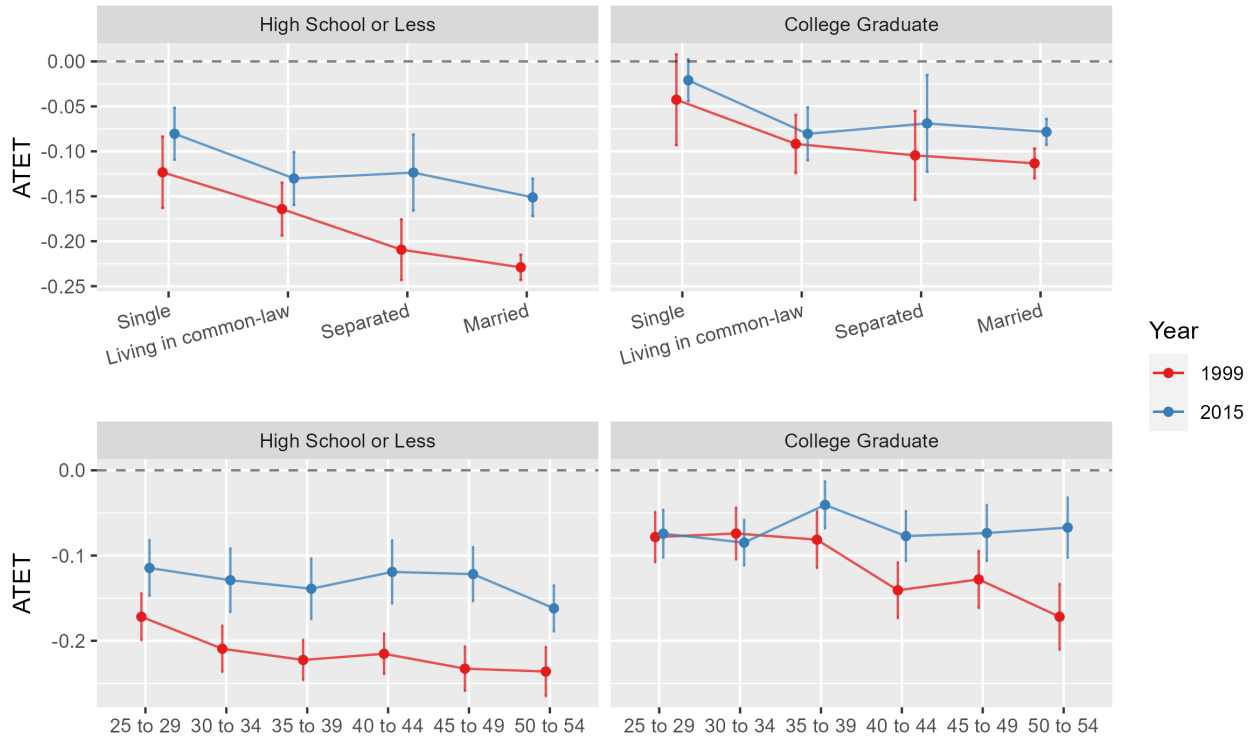


Figure 2.12: Average Treatment Effects on Treated for Occupations

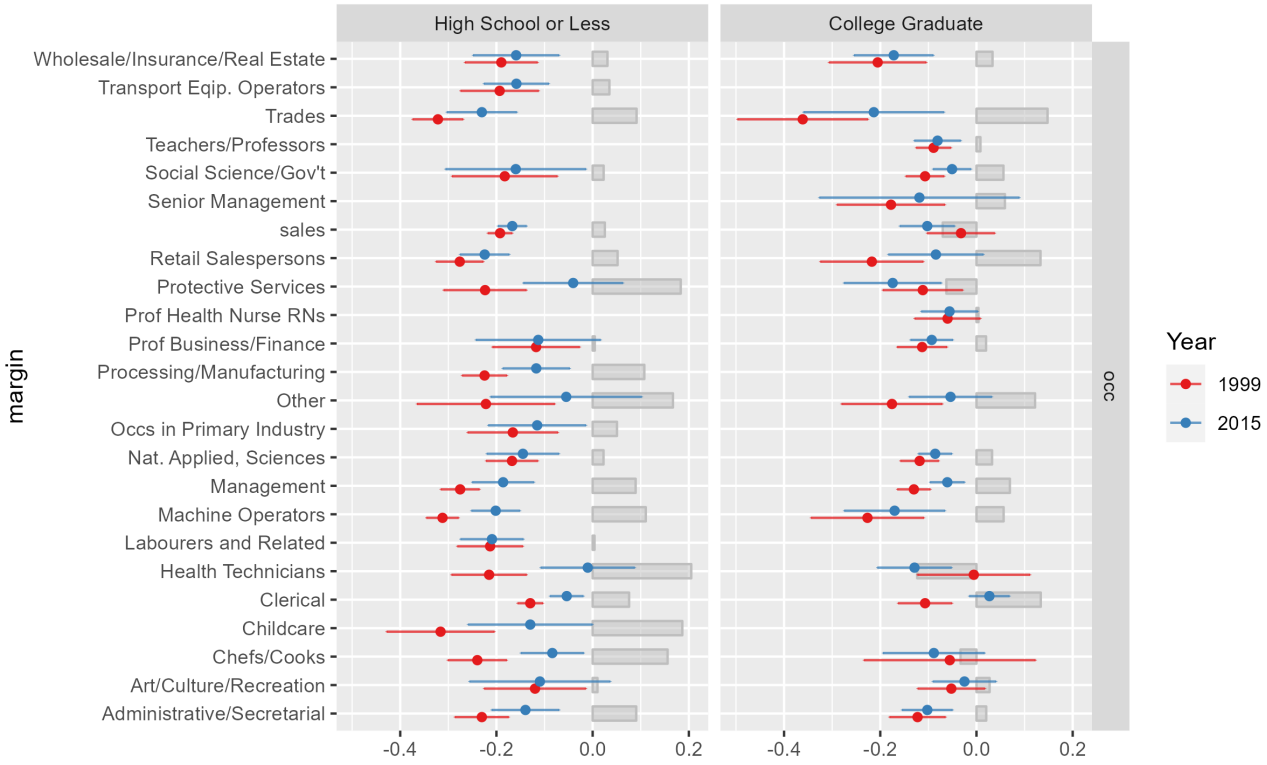


Figure 2.13: Coefficient Estimates for Alternative ML Methods: part 1

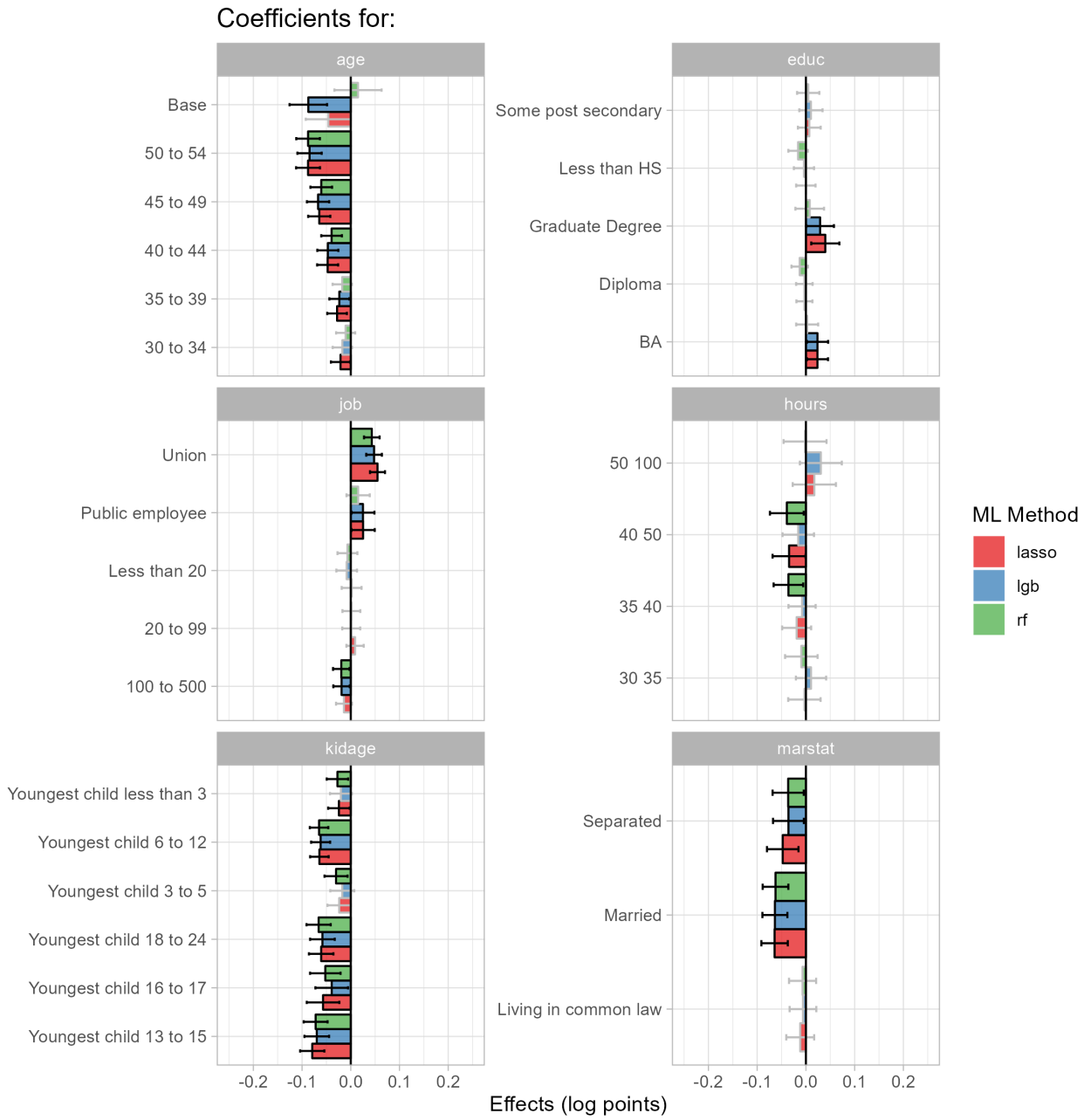


Figure 2.14: Coefficient Estimates for Alternative ML Methods: part 2

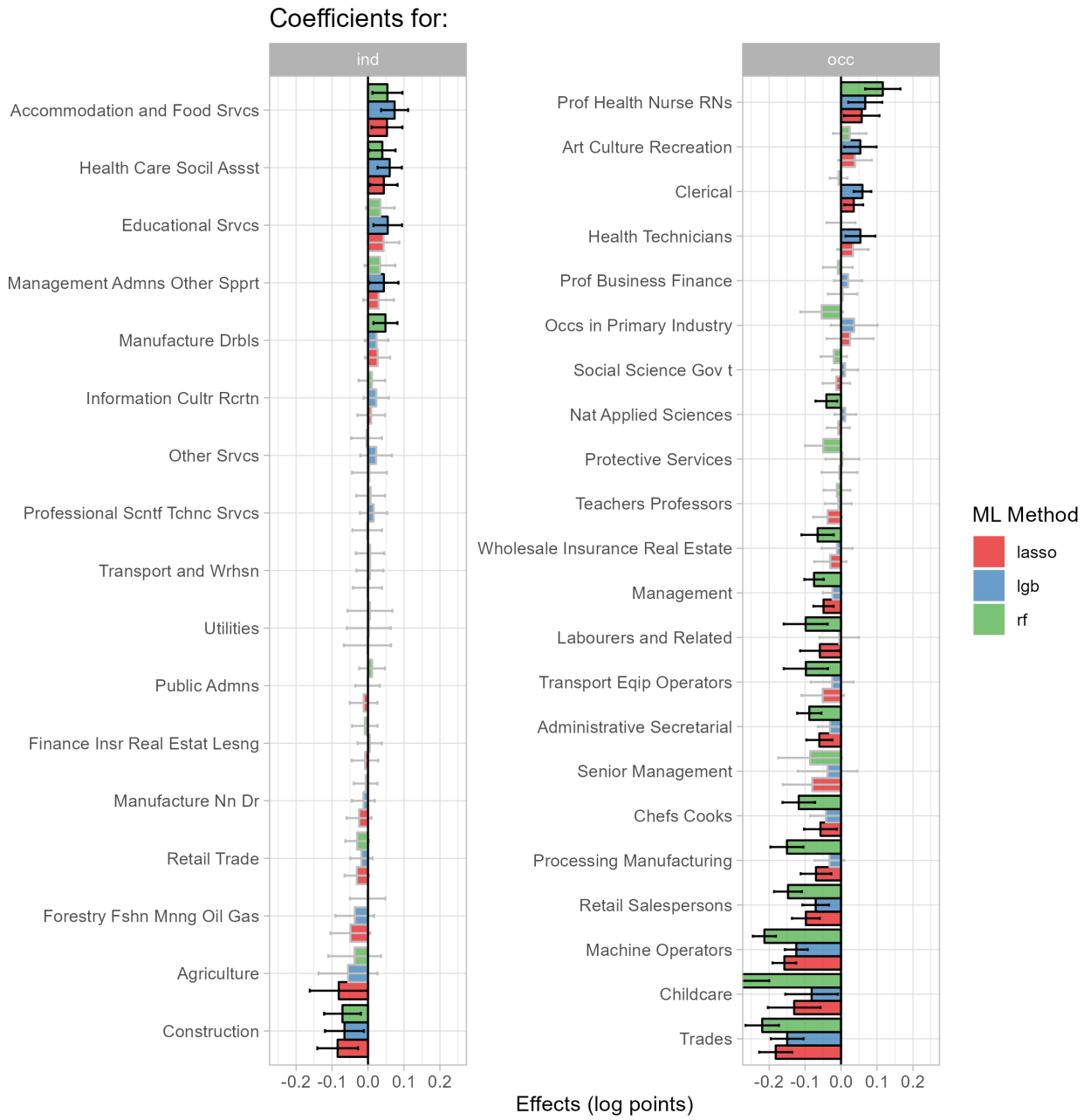


Figure 2.15: Robustness for linear ML methods: part 1

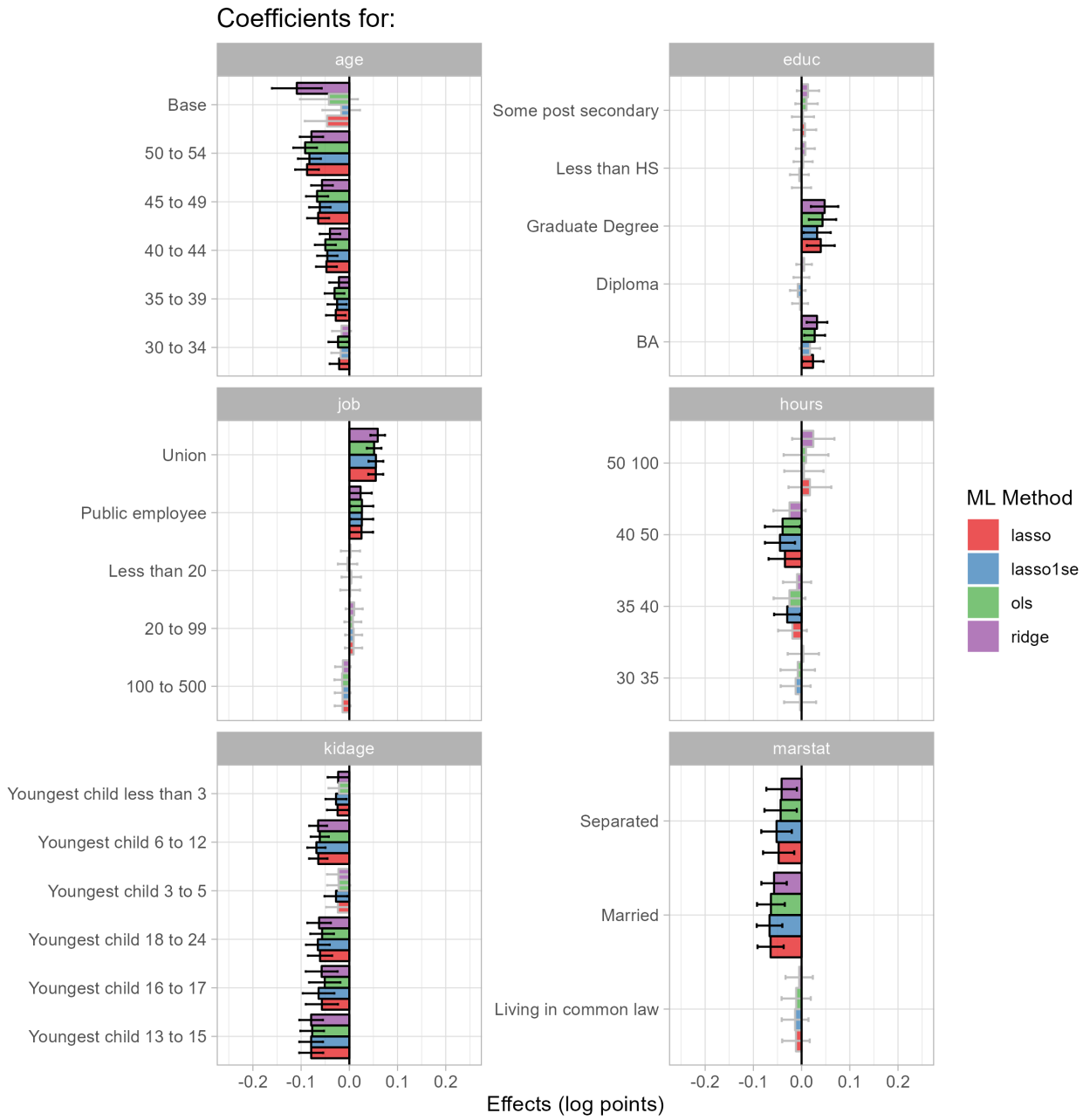
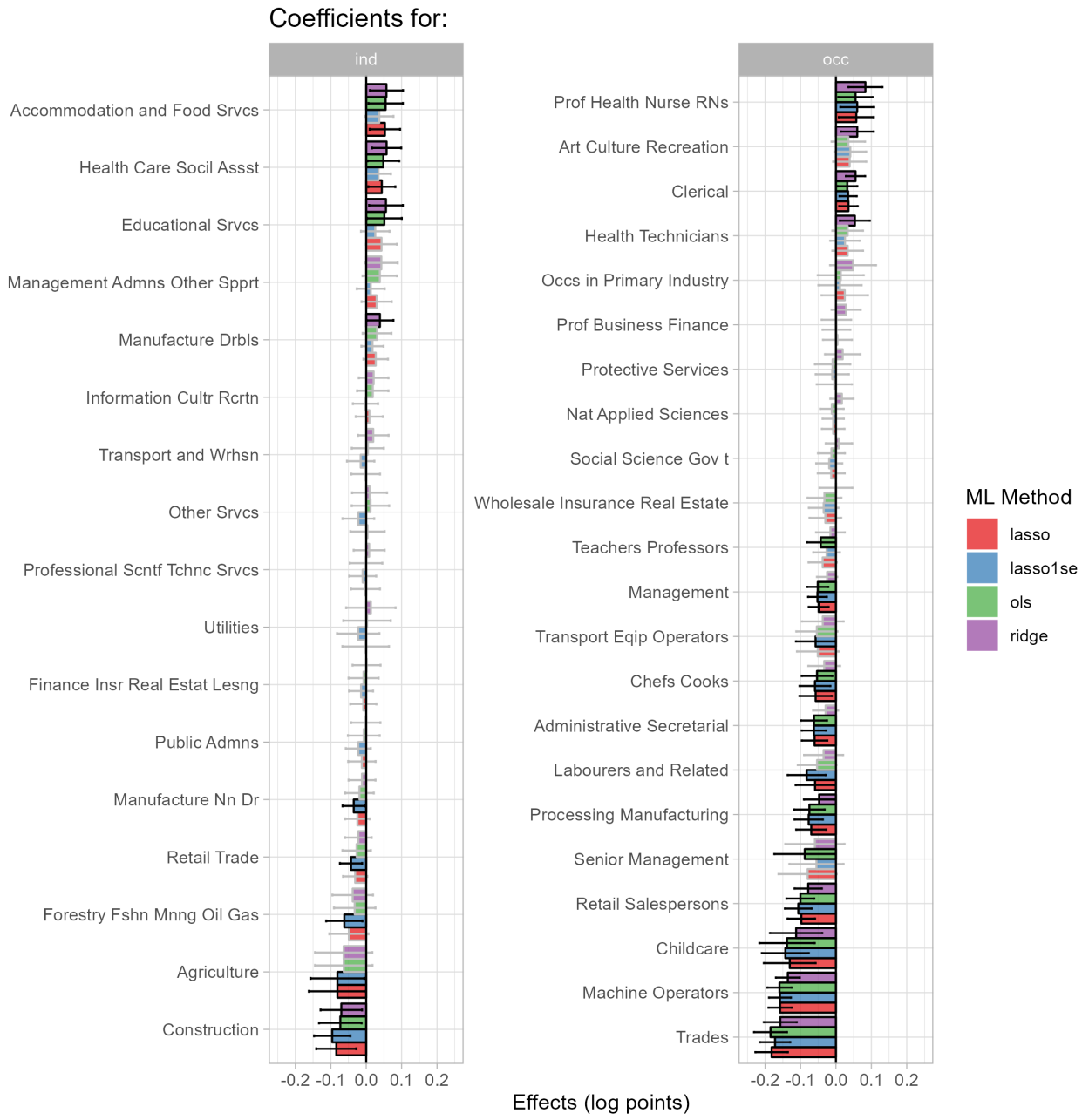


Figure 2.16: Robustness for linear ML methods: part 2



Chapter 3

Job-Education Mismatch and the Canadian Gender Wage Gap

3.1 Introduction

The consensus in economics stipulates higher levels of education lead to higher income. Research over the past few decades has shown that differences in education, rather than innate abilities, are primarily responsible for the earnings gap between those with more or less education. However, we do not fully understand the mechanisms that explain this link. For example, Lemieux (2014) found that the return to education varies significantly depending on occupation, field of study, and the match between the two. This is particularly important in understanding the gender gap, as men and women differ not only in the quantity of education but also in the type of degree earned. Additionally, the penalties for a poor match between quantity and type of education with a job can be different between men and women, further contributing to the wage gap. Such nuances carry potential policy ramifications. While many policies aimed at enhancing women's labor market outcomes advocate for increased education, critics contend that women often enroll in programs with limited returns. If the penalties for mismatch is substantial, then it would be prudent for policy makers to focus on redirecting students in perusing specific degrees rather than higher quantity of education. In this paper, I will explore the effect of matching on the gender wage gap.

The gender wage gap in Canada remains substantial but decreasing. Part of it can be

partly attributed to women's increasing education attainment and their move into higher-paying professions (Pelletier, Patterson, and Moyser (2019a)). However, as documented in Canada (Ferguson (2016))and other industrialized countries (Indicators (2016)), gender segregation in fields of study persists. The choice of degree or major can affect wages in at least two ways: different fields of study can have varying returns in the labour market, and they can differ in matching workers to jobs (Joseph G. Altonji, Bharadwaj, and Lange (2012)). As highlighted by Lemieux (2014), earnings differ significantly by both field of study (e.g. Arcidiacono (2004)) and occupation (e.g. Gibbons et al. (2005)), making the matching process a crucial factor in explaining the variation in returns to education among individuals. If men and women face different rates of education-job mismatch or penalties for mismatch, it could contribute to the gender wage gap. Yet, despite the extensive literature on both gender wage gap and education-job mismatch, there has been little research done that combines both. In this paper, I contribute to this sparse and investigate the impact of education-job mismatch on the gender wage gap.

The primary data source used in this paper is the National Graduate Survey (NGS), which is the only widely available and consistent source in Canada containing information on earnings, field of study, and self-reported measures of mismatch between workers and jobs. However, using this data comes with a major downside: the wage data is discontinuous, with yearly income binned into $10k brackets. To address this issue, this paper employs a newer distributional regression method outlined in Chernozhukov, Fernández-Val, and Melly (2013), which allows for an explanation of the relationship between mismatch and the gender wage gap even when the wage data is discontinuous. As a robustness check, an index of job-relatedness was created using the Canadian Census and the method outlined in Aydede and Dar (2016). The benefit of using the Census is the availability of continuous wage data, which allows for finer income bins when using the distributional regression of Chernozhukov, Fernández-Val, and Melly (2013). However, it lacks data on years of education and qualification required for jobs, which limits the scope to only job relatedness.

The paper finds that despite having very similar characteristics, there still exists an unexplained gender wage gap in Canada, with men's wages stochastically dominating

women's at all income levels. In contrast to prior research, this paper suggests that the impact of mismatch on the gender wage gap is insignificant. While being overly educated or over qualified relative to the job requirement is detrimental to wage, there are no substantial differences between women and men in both the penalty and likelihood to being overly educated. Additionally, the results from the Census data reveal that men working in jobs related to their field of study in 2018 received higher compensation relative to women, whereas women tend to work in jobs that are more related to their education. However, the overall impact on the gender wage gap is economically minor, accounting for only a small proportion of the total gap.

The paper is structured as follows: Section 1 provides a literature review on relevant literature [1]. Section 2 describes the data, Section 3 offers a detailed explanation of the method proposed in Chernozhukov, Fernández-Val, and Melly (2013), and Section 4 provides an analysis of the results.

3.2 Literature Review

There is an enormous amount of literature that attempts to understand the factors that contribute to the gender wage gap. Chapter 1 of my dissertation contains a general overview of the literature. This paper's focus is on the role of education, occupation, and their mismatch. There are two main types of occupation/education mismatch: vertical and horizontal mismatch. Pioneered by Freeman (1976), vertical mismatch occurs when an individual's level of education and skill does not match the requirements of their job . This can be costly for both individuals and the economy, as it can lead to lower return on education and skill, less job satisfaction, reduced productivity, and inefficiencies (Somers et al. (2019),Allen and Van der Velden (2001),Tsang (1987)). Vertical mismatch can be broadly categorized into three groups: overeducation, undereducation, and required education, with the majority of the literature focusing on overeducation (Sloane (2007)) Overeducation/undereducation occurs when the employee's level of education is higher/lower than what is required for their job. Following Duncan and Hoffman (1981), this type of mismatch is often expressed as years over/under the requirement. Required education is when the education qualification matches the job requirement. However, simply

[1] For a study on the gender wage gap or distributional regression in general, see chapter 1

comparing the level of education with the requirement is too simplistic, as noted by Hartog (2000), Leuven and Oosterbeek (2011), and Meroni and Vera-Toscano (2017). Unobserved heterogeneity, such as 'ability', can play an important role in determining whether the employee is truly overqualified. Therefore, in this paper, two measures of vertical mismatch are used for the NGS data: education requirement and self-reported evaluation of qualification. This subjective measurement of mismatch can be preferable to objective measures such as years of education or using job categorization dictionaries, as noted by Agut, Peiró, and Grau (2009) and Linsley (2005).

The second type of mismatch, which directly compares field of study with occupation, is less common and is referred to as education-job horizontal mismatch (see Somers et al. (2019) for a review of the literature). This type of mismatch occurs when an individual's education does not align with the requirements of their current job. Horizontal mismatch is a significant concern as it can lead to lower productivity, lower wages, reduced job security, and increased unemployment for individuals, as well as inefficiencies and resource waste for the economy. As this paper compares differences between recent graduates, the biggest determinant of horizontal mismatch would come from education-related factors. One source of mismatch is the field of education. For example, Wolbers (2003) and Robst (2007) found that graduates of liberal arts degrees experienced the highest level of mismatch, while the lowest was observed in health-related programs. Part of this can be attributed to an excess supply of graduates in a certain field relative to demand, which can create frictions and mismatch (Cosser (2010)). Higher levels of education can lead to less mismatch. Those with higher education can outcompete those with lower levels in the same field (Bender and Roche (2013), Boudarbat and Chernoff (2012), Wolbers (2003)), have more general skills that can be applied to wider jobs, or have specialized training that is harder to find among lower-educated employees (Levels, Van der Velden, and Di Stasio (2014)). Therefore, horizontal mismatch between education and job requirements can have significant implications for both individuals and the economy.

In terms of the effect of mismatch on the gender wage gap, there is currently a lack of research that combines the two fields. The few studies that do combine these topics tend to focus on vertical mismatch and have drawn different conclusions, suggesting that the effect of mismatch is not consistent across countries. For example, Figueiredo et al. (2015) overeducation

of young female graduates relative to men in Southern Europe contributes to the gender wage gap. On the other hand, Boll and Leppin (2013) found that in Germany, there are no differences in overeducation between men and women, and it have no role in the gender wage gap. In Australia, Li and Miller (2012) found that while women are less likely to be overeducated, its contribution to the overall gap was small. Looking at the gender wage gap through vertical mismatch is the most popular approach as it is relatively light on data requirement, needing only years of education. In contrast, for horizontal mismatch and the gender wage gap, Park (2021) is the only one on this topic to the best of my knowledge. It found, in South Korea, substantial mismatch, both vertical and horizontal, in education and jobs for recent graduates, while Women faces higher penalty for mismatch compared to men. This makes up significant portion of the gender wage gap in South Korea.

3.3 Data

The primary data source used in this study is the National Graduate Survey (NGS) from 1998 and 2018. The NGS is a survey that gathers information from those who have completed post-secondary education, including university, college, and skilled trades degrees or diplomas. The survey covers topics such as education, employment, fields of study, and outcomes related to education and work. The NGS is conducted every five years for individuals who have graduated three years prior.

For the purpose of this study, the reference year is the graduation year, and only individuals who are employed full time, have wage data, and are under the age of 40 are included in the data set. In total, remaining data have 19,199 observations in 1998, with 8,668 being women and 10,437 being men. In 2018, there are 12,100 observations remaining, split between 6,269 women and 5,104 men. The NGS provides direct measures of matching by asking interviewees to self-report on the quality, relatedness, and qualification of their education-job matching. One limitation of the NGS is that the wage variable is binned into brackets of $10,000, which prevents direct measurement of the impact of independent variables on hourly wages. Instead, the results are interpreted as changes in the probability of

working under a wage threshold, as described by Chernozhukov, Fernández-Val, and Melly (2013).

The secondary data for my robustness check comes from the Canadian Census, years 1996 and 2016. The Census represents the most comprehensive and unbiased snapshot of the Canadian population. As the Census' reference period is a year prior, they corresponds to the same reference periods as the NGS data. I also restrict the data to only those that work full time, have positive wage data, under the age of 40, and have completed post-secondary education at the college/trade certificate level or more. The 1996 Census have 46,656 individuals left, of which 19,800 are women and 26,856 are men. In the 2018 data set, there are 55,219 observations, with 25,842 women and 29,377 men. Both the NGS and Census data skews towards more men because men have higher labour participation rate compared to women.

Table 3.1 presents the mean and difference of key variables for men and women in 1998, highlighting the educational advantage of women over men on average. Moreover, it shows that women were more inclined to work in positions where they were overqualified. In terms of job relevance to their field of study, both genders had an equal likelihood of working in related occupations, with more than half employed in closely aligned fields. When it came to program of study, men displayed a higher inclination towards math, trades, and engineering On the other hand, women were more likely to enroll in programs in all other categories excluding bio-sciences. Notably, trades-related programs had a significant male enrollment of 34%, compared to a mere 4% for women. Conversely, programs in health and education skewed towards a higher female enrollment relative to others.

Table 3.2 provides summary statistics for 2018, indicating that the education levels of men have nearly caught up with those of women. This shift can be attributed to a higher number of men obtaining bachelor's or graduate degrees and moving away from college and trades. In terms of job qualifications, both genders were less likely to work in positions where they were overqualified compared to 1998, and this change was not statistically significant. In 2018, men and women exhibited similar trends in education, with some minor convergence in specific fields of study. However, fields such as mathematics and engineering, which were predominantly male, became even more so. Conversely, social sciences and agriculture (which

includes bio sciences and health-related fields) became increasingly dominated by women.

Table 3.3 and Table 3.4 shows selected summary statistics from the 1996 and 2016 Census. Compared to the Census, the NGS' skews towards higher levels of education and have different distributions of program studied. This is because the NGS is not a representative of the entire population, but only recent graduates while definition of program studied changes between years and data sets. Despite these disparities, the overall trends between the two data sets remain consistent. In both the Census data and the NGS, women demonstrate increasing educational attainment in absolute terms and relative to men. For men, there is a growing polarization in educational pursuits, with a larger number opting for graduate studies or college/trades rather than pursuing a bachelor's degree. Programs that predominantly attract women are related to education and healthcare in both data sets, while men dominate programs in trades, engineering, and mathematics.

In summary, despite variations in data sources, both the NGS and Census samples reveal consistent patterns. Notably, there are significant gender differences in program choices. Men tend to concentrate in fields associated with trades, engineering, and mathematics. In contrast, women exhibit a more diverse range of program choices, with a noticeable concentration in areas related to health and education. In areas related to matching, both women and men have similar distribution in program requirement, qualification, and relatedness in nearly all except qualification in 1998. This indicates that any effects coming from matching would have to be related to differences in returns rather than differences in probability and quality of matching.

3.4 Methodology

In this section, the methods employed by Chernozhukov, Fernández-Val, and Melly (2013) are outlined and their application to this paper is explained. The objective is to decompose the income differences between men and women into potential contributing factors. The following notations are used: 0 denotes women, 1 denotes men, Y denotes wage, X denotes a set of personal

characteristics[2], and F denotes the CDF of wage. The CDF of wage, conditional on Y and X, is represented as $F_{(Y_0|X_0)}(y|x)$ for women and $F_{(Y_1|X_1)}(y|x)$ for men. The observed distribution functions for wages of women and men can be expressed as $F_{Y(0|0)}$ and $F_{Y(1|1)}$, where $Y(0|0)$ and $Y(1|1)$ represent the observed wages for women/men with relevant personal characteristics of women/men and pay schedules of women/men. Inspired by the Blinder-Oaxaca decomposition, the differences in CDF can be separated into unexplained and explained components:

$$F_{Y(0|0)} - F_{Y(1|1)} = [F_{Y(0|0)} - F_{Y(1|0)}] + [F_{Y(1|0)} - F_{Y(1|1)}] \tag{3.1}$$

Here, $F_{Y(1|0)}$ denotes the distribution function for a counterfactual female worker with the pay schedule of men. The unexplained $(F_{Y(0|0)} - F_{Y(1|0)})$ and explained gap $(F_{Y(1|0)} - F_{Y(1|1)})$ can be further decomposed to determine the influence of a specific variable of interest. This is achieved by substituting the coefficients or characteristics one at a time and calculating the difference. Let $F_{Y(er,q,r,e,p,o|er,q,r,e,p,o)}$ denote the counterfactual distribution function of wages, given the wage schedule and personal characteristics distribution of education requirement, qualification, job relatedness, education, program, and other observable characteristics. When the variable equals 0, it indicates the wage structure/characteristic distribution of women, while 1 represents men. The unexplained gap can be rewritten as [3] :

$$\begin{aligned}
&F_{Y(0,0,0,0,0,0|0,0,0,0,0,0)} - F_{Y(1,1,1,1,1,1|0,0,0,0,0,0)} \\
&= \underbrace{[F_{Y(0,0,0,0,0,0|0,0,0,0,0,0)} - F_{Y(1,0,0,0,0,0|0,0,0,0,0,0)}]}_{\text{Unexplained education requirement effect}} + \underbrace{[F_{Y(1,0,0,0,0,0|0,0,0,0,0,0)} - F_{Y(1,1,0,0,0,0|0,0,0,0,0,0)}]}_{\text{Unexplained qualification effect}} \\
&= \underbrace{[F_{Y(1,1,0,0,0,0|0,0,0,0,0,0)} - F_{Y(1,1,1,0,0,0|0,0,0,0,0,0)}]}_{\text{Unexplained job relatedness effect}} + \underbrace{[F_{Y(1,1,1,0,0,0|0,0,0,0,0,0)} - F_{Y(1,1,1,1,0,0|0,0,0,0,0,0)}]}_{\text{Unexplained education effect}} \\
&= \underbrace{[F_{Y(1,1,1,1,0,0|0,0,0,0,0,0)} - F_{Y(1,1,1,1,1,0|0,0,0,0,0,0)}]}_{\text{Unexplained program effect}} + \underbrace{[F_{Y(1,1,1,1,1,0|0,0,0,0,0,0)} - F_{Y(1,1,1,1,1,1|0,0,0,0,0,0)}]}_{\text{All other unexplained effects}}
\end{aligned} \tag{3.2}$$

[2]The independent variables used in the NGS regressions are: province, age, martial status, dummy for having children, highest level of education, program of study, occupation, education level compared to required, feeling of qualification, and relatedness of program to job.

[3]For exact definitions of the counterfactuals, see equations 7.1 and 7.1 for more details.

Similarly, the explained gap can be decomposed as follows:

$$
\begin{aligned}
&F_{Y(1,1,1,1,1,1|0,0,0,0,0,0)} - F_{Y(1,1,1,1,1,1|1,1,1,1,1,1)} \\
&= \underbrace{\left[F_{Y(1,1,1,1,1,1|0,0,0,0,0,0)} - F_{Y(1,1,1,1,1,1|1,0,0,0,0,0)}\right]}_{\text{Explained education requirement effect}} + \underbrace{\left[F_{Y(1,1,1,1,1,1|1,0,0,0,0,0)} - F_{Y(1,1,1,1,1,1|1,1,0,0,0,0)}\right]}_{\text{Explained qualification effect}} \\
&= \underbrace{\left[F_{Y(1,1,1,1,1,1|1,1,0,0,0,0)} - F_{Y(1,1,1,1,1,1|1,1,1,0,0,0)}\right]}_{\text{Explained job relatedness effect}} + \underbrace{\left[F_{Y(1,1,1,1,1,1|1,1,1,0,0,0)} - F_{Y(1,1,1,1,1,1|1,1,1,1,0,0)}\right]}_{\text{Explained education effect}} \\
&= \underbrace{\left[F_{Y(1,1,1,1,1,1|1,1,1,1,0,0)} - F_{Y(1,1,1,1,1,1|1,1,1,1,1,0)}\right]}_{\text{Explained program effect}} + \underbrace{\left[F_{Y(1,1,1,1,1,1|1,1,1,1,1,0)} - F_{Y(1,1,1,1,1,1|1,1,1,1,1,1)}\right]}_{\text{All other explained effects}}
\end{aligned}
\tag{3.3}
$$

To estimate the counterfactual distributions, a series of OLS regressions was employed. As the income data is discrete, F can be approximated through a series of regressions where the dependent variable is binary and indicates whether an individual's income is equal to or less than an income threshold: $F(Y|Y < \hat{y}, X) = Prob(Y < \hat{y}|X)$, where $\hat{y} \in (y_1, y_2, \ldots y_k)$ are the discrete income bins. Specifically, let $\hat{Y}_i$ be a dummy variable that is equal to 1 if individual i's yearly salary is less than or equal to $\hat{y}$. The output of each regression can be interpreted as the probability of an individual with characteristics X_i having a salary that is less than or equal to $\hat{y} \longrightarrow Prob(Y < \hat{y}|X)$. To compute the counterfactual probability of earning less than or equal to $\hat{y}$ for women with men's wage structure, $F_{Y(1|0)}$, I use women's characteristics X_0 and men's coefficients β_1. Similarly, for the explained gap, I use men's characteristics X_1 and women's coefficients β_0. Because this method is in essence a series of OLS regressions and Blinder-Oaxaca decomposition, stanard assumptions apply[4]. In addition, I assume away any general equilibrium effects, that is, women and men's wage have no effect on the other gender.

Once I have the estimated probabilities for all individuals and for each income category, I can calculate the unexplained and explained effects for each wage class using equations (3.2) and (3.3). These effects help us understand how much of the wage gap between men and women can be attributed to differences in characteristics (explained) or differences in wage structure

[4]For more information on the assumptions of Blinder-oaxaca decomposition, see N. Fortin, Lemieux, and Firpo (2011)

(unexplained). Furthermore, I can assess the impact of individual factors such as education, qualification, job relatedness, and program on the wage gap.

This method provides a comprehensive analysis of the gender wage gap by examining the entire wage distribution, rather than just mean or median wages. This is particularly useful when the wage distribution is not symmetric or when there are substantial differences in the wage distribution between men and women. Through the decomposition of the wage gap into its explained and unexplained components , and further breaking down the effects into individual factors, we can gain a deeper understanding of the underlying causes of the gender wage gap and better inform policy interventions aimed at reducing it.

To create the index variable for the Census data regressions, I used the relatedness index outlined in Aydede and Dar (2016):

$$RI_{of} = \frac{L_{of}/L_f}{L_o/L_T} \tag{3.4}$$

Where L_{of} is the number of workers with occupation o and have field of study f. L_f is the total number of workers that have studied f. L_o is the total number of workers in occupation o, and finally, L_T is the total number of workers. RI can also be thought of as a concentration index. The numerator measures how well fit a field of study is for an occupation. As this number approaches 1, more of individuals with field of study f work in o and the relatedness of that particular o, f pairing increases. The denominator measures the importance of an occupation to the labour force and acts as a weighting variable to the numerator. Thus, each individual is assigned an index number based on their occupation and field of study. To generate more unique fields of studies and achieve greater variation in the index, I created dummy interaction terms between highest obtained level of education and program. For example, a degree in engineering at the university level is considered different to the same degree at graduate level. This resulted in 50 unique field of studies in 1996 and 95 in 2016. Using this index and its square as an additional variable, I applied the same methodology as listed previously to compare with the results from the NGS data. [5]

[5]The independent variables for the Census regressions are: married dummy, age, province, occupation, highest

3.5 Results

Figures 3.1 and 3.2 show the cumulative distribution function (CDF) and the differences in CDF between female and male workers, respectively. The CDF differences represent the differences in probability of earning less than a wage threshold, with the largest gaps expected to be at the lower end of the income distribution and approaching zero at the higher end. For both years, men's income stochastically dominate women for all income levels and years. In 1998, women are more than 12% more likely to be in the lower half of the income distribution compared to men. This gap remained nearly unchanged in 2018,despite higher earnings for both genders.

Figure 3.3 displays the unexplained wage gap, which represents the differences in earning probabilities that are caused by factors other than observable characteristics such as education and experience. For both years, the unexplained gap accounted for nearly all of the total gap. This result is not surprising, given that the data sample was constructed to make women and men highly comparable. In 1998, education played a significant role in reducing the total gap, as women had higher returns for a given level of education. However, this effect disappeared by 2015 as the return on education equalized.

Figure 3.4 illustrates the explained wage gap, which refers to the differences in earning probabilities that are attributable to observable characteristics. The overall magnitude of the explained gap is small for both years. Although women, on average, have slightly higher levels of education, it plays a minor role in reducing the total gap. n contrast, differences in occupation explain why women are 4% more likely to be in the bottom half of the income distribution for both years. [6]

Tables 3.5, 3.6, 3.7, and 3.8 shows the estimated coefficients for select explanatory variables and wage groups. Program in engineering and technology/trades have relatively high returns and are mainly dominated by men while women are more likely to be in lower paid programs in business and fine arts. However, in 1998, the return of other

level of education obtained, program of study, the index and its square

[6]The sequence of decomposition holds significance as it can yield distinct outcomes. To ensure the reliability of the findings, I have incorporated the outcomes of the reverse-order decomposition as well (See figures C11, C12, C13 and, C14 in the appendix). The reverse decomposition produced nearly identical results.

programs are relatively even. Health, and to a lesser extent, education, are women dominated programs with relatively high returns. By 2018, the return on different programs have converged, with very few having abnormally high or low returns.Therefore, while women and men differ in program choices, both have preferred programs that are high paying, and the overall explained gap caused by difference in program choice is small.
Regarding the effects of education requirements, qualifications, and job-relatedness, my results show that they do not have statistically significant effects on the wage gap in either year. Women are not more likely than men to have education or qualifications that differ from job requirements or to work in jobs that are unrelated to their education. Moreover, we do not find any differences in the penalty for being overly educated or working in unrelated jobs. These findings contradict those of Park (2021) and Figueiredo et al. (2015) but are consistent with those of Boll and Leppin (2013) and Li and Miller (2012). These results further suggest that the impact of job mismatch on the gender wage gap differs across countries.

Figure 3.5 and Figure 3.6 present the unexplained and explained wage gap calculated using Census data. The Census data produces similar estimates to those obtained from the NGS, but with tighter confidence intervals. As with the NGS, education reduces the total unexplained gap due to higher returns on education for women. This effect remains statistically significant in 2018 with the Census data due to the more accurate confidence bands. The unexplained gap for program shows signs of statistical significance near the middle income in 2018 with the Census data, while the NGS produces results that are no different from zero. This difference is due to the inclusion of the occupation variable in the Census regressions. The explained gap caused by program and occupation shows that men are more likely to be in programs and occupations that pay more. This result is consistent between the two data sets. Finally, relatedness, as measured by my index, is statistically significant for the unexplained gap in 2018 and for the explained gap in both years. This indicates that women are penalized more for working in fields that are less related to their education and training, and that women are more horizontally mismatched compared to men. However, while the results are statistically significant, they make very little economic difference. At most, lower returns on relatedness cause women to be 3 percent more likely to earn less than $65k compared to men in 2018.

In summary, while not being overly educated and qualified and working in job related to field of study increase wages, there are no significant differences between women and men. While the Census data showed some statistical significance for job relatedness, the economic significance on the gender wage gap is minimal. Furthermore, differences in program of study between women and men have only small impact on the gender wage gap. Men tend to be more prevalent in high-paying programs related to mathematics and engineering, while women are more likely to pursue lower-paying fields such as humanities and social sciences. However, women also show a higher propensity for enrolling in relatively well-paying programs in bio sciences, health, and education. Consequently, both women and men have popular programs that offer higher earning potential, thereby reducing the gender wage gap attributable to program differences.

3.6 Conclusion

In this paper, paper I examined the impact of both vertical and horizontal education-job mismatch on the Canadian gender wage gap, utilizing two different data sets and the decomposition method found in Chernozhukov, Fernández-Val, and Melly (2013). In my results, I find substantial higher probability for women to earn less than men. The majority of the gap cannot be explained by differences in observable characteristics and is evidence for the existence of discriminatory practices. I also found that vertical mismatch have no significant impact on the gender wage gap. While over-education and over-qualification is detrimental to income, there are no substantial differences in both likelihood of matching and return on matching. There are some evidence of horizontal mismatch having an effect on the wage gap in 2018. Women are equality as likely to be horizontally matched compared to men but have lower return on matching. However, this effect is economically small and contribute to only a small portion of the overall gap.

The results of this paper contribute to the sparse literature that combine education-job mismatch and the gender wage gap. While the effect varies across countries, in Canada at least, mismatch have little to no impact on the gender wage gap. Education is the primary driver that lessens the wage gap, with women benefiting more relative to men. In addition, this paper finds

differences in program of study have only small effect on the gap. While men are more likely to be enrolled in higher paying programs, there are also female dominated programs with equally high returns. Thus, for the purpose of reducing the gender wage gap, policies that encourage women to enroll in higher education rather than specific fields of education is sufficient.

Table 3.1: Summary table: 1998 NGS

	Female		Male		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
Education						
College	0.31	0.009	0.29	0.008	-0.017	-1.331
Trade	0.10	0.005	0.20	0.007	0.109	**11.996**
University	0.54	0.011	0.44	0.011	-0.097	**-6.424**
Graduate	0.06	0.004	0.06	0.004	0.005	1.018
Education Requirement						
More	0.33	0.010	0.32	0.009	-0.009	-0.619
Same	0.44	0.011	0.43	0.010	-0.006	-0.392
Less	0.05	0.005	0.05	0.005	0.003	0.514
Not Specified	0.18	0.009	0.20	0.008	0.011	0.972
Qualification						
More	0.34	0.011	0.30	0.010	-0.043	**-3.009**
Same/Less	0.66	0.011	0.70	0.010	0.043	**3.009**
Program						
Education	0.17	0.010	0.08	0.007	-0.090	**-7.34**
Fine Arts	0.03	0.004	0.01	0.003	-0.016	**-3.639**
Humanities	0.07	0.006	0.04	0.005	-0.025	**-3.223**
Sco Sciencess	0.17	0.010	0.12	0.008	-0.059	**-4.809**
Bussniess	0.26	0.010	0.19	0.009	-0.075	**-5.695**
Bio Sciences	0.04	0.003	0.04	0.003	0.000	-0.033
Engineering	0.02	0.002	0.09	0.005	0.075	**14.475**
Trades	0.04	0.003	0.34	0.009	0.301	**32.957**
Health	0.16	0.007	0.03	0.003	-0.134	**-17.45**
Math	0.02	0.002	0.05	0.004	0.032	**8.02**
Other	0.02	0.002	0.01	0.002	-0.008	**-3.07**
Relatedness						
Closely	0.54	0.011	0.55	0.010	0.008	0.516
Somewhat	0.22	0.009	0.23	0.008	0.012	0.967
Not	0.24	0.010	0.22	0.009	-0.020	-1.475

Notes: This table shows the summary statistics of key variables in 1998. Standard deviation is calculated using unconditional OLS regressions. 'Difference' is the male coefficients minus female coefficients

Table 3.2: Summary table: 2018 NGS

	Female		Male		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
Education						
College	0.17	0.008	0.19	0.010	0.023	1.822
University	0.51	0.009	0.50	0.012	-0.018	-1.197
Graduate	0.32	0.008	0.31	0.010	-0.005	-0.396
Education Requirement						
More	0.26	0.009	0.27	0.011	0.004	0.287
Same/Less	0.74	0.009	0.73	0.011	-0.004	-0.287
Qualification						
More	0.27	0.008	0.27	0.010	-0.003	-0.26
Same/Less	0.73	0.008	0.73	0.010	0.003	0.26
Program						
Education	0.09	0.005	0.04	0.004	-0.055	**-8.915**
Fine Arts	0.03	0.004	0.02	0.004	-0.006	-1.204
Humanities	0.04	0.004	0.02	0.004	-0.020	**-3.525**
Sco Sciencess	0.21	0.008	0.10	0.007	-0.109	**-10.16**
Bussniess	0.25	0.008	0.26	0.011	0.010	0.708
Bio Sciences	0.06	0.004	0.05	0.005	-0.003	-0.527
Math/Computer	0.02	0.002	0.09	0.007	0.076	**10.853**
Engineering	0.05	0.004	0.28	0.010	0.228	**20.452**
Agriculture	0.21	0.007	0.08	0.006	-0.135	**-14.287**
Other	0.04	0.004	0.05	0.005	0.014	**2.266**
Relatedness						
Yes	0.90	0.006	0.91	0.007	0.010	1.127
No	0.10	0.006	0.09	0.007	-0.010	-1.127

Notes: This table shows the summary statistics of key variables in 2018 Standard deviation is calculated using unconditional OLS regressions. 'Difference' is the male coefficients minus female coefficients

Table 3.3: Summary table: 1998 Census

	Female		Male		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
Education						
College	0.50	0.004	0.53	0.003	0.036	**7.685**
Bachelor	0.46	0.004	0.41	0.003	-0.054	**-11.7**
Graduate	0.04	0.001	0.06	0.001	0.018	**8.971**
Program						
Education	0.12	0.002	0.04	0.001	-0.078	**-29.928**
Agriculture	0.04	0.001	0.05	0.001	0.005	**2.917**
Engineering	0.01	0.001	0.08	0.002	0.067	**35.871**
Building	0.00	0.000	0.06	0.002	0.063	**40.664**
Comp Science	0.03	0.001	0.04	0.001	0.010	**6.184**
Electronic	0.00	0.000	0.07	0.002	0.063	**39.504**
Applied Tech	0.02	0.001	0.23	0.003	0.213	**78.126**
Nursing	0.05	0.002	0.00	0.000	-0.048	**-29.72**
Other Health	0.08	0.002	0.03	0.001	-0.053	**-24.62**
Math	0.03	0.001	0.05	0.001	0.026	**14.315**
Fine.Arts	0.06	0.002	0.02	0.001	-0.036	**-18.666**
Humanities	0.06	0.002	0.04	0.001	-0.019	**-9.178**
Social Science	0.13	0.002	0.10	0.002	-0.029	**-9.706**
Business	0.07	0.002	0.06	0.001	-0.003	-1.132
Finance	0.08	0.002	0.06	0.001	-0.023	**-9.35**
Administration	0.05	0.002	0.03	0.001	-0.016	**-8.749**
Marketing	0.03	0.001	0.02	0.001	-0.010	**-6.431**
Secretarial	0.14	0.002	0.01	0.000	-0.134	**-53.725**
Index						
Index	0.06	0.039	0.06	0.039	0.000	-0.036

Notes: This table shows the summary statistics of key variables of the Canadian Census: 1996. Standard deviation is calculated using unconditional OLS regressions. 'Difference' is the male coefficients minus female coefficients. Index is education/job relatedness. Higher index indicates education is more similar to job.

Table 3.4: Summary table: 2016 Census

	Female		Male		Difference	
Characteristic	Mean	SD	Mean	SD	Mean	t-statistic
Education						
College	0.45	0.003	0.56	0.003	0.117	**27.595**
Bachelor	0.41	0.003	0.32	0.003	-0.094	**-22.877**
Graduate	0.14	0.002	0.12	0.002	-0.023	**-8.072**
Program						
Education	0.09	0.002	0.02	0.001	-0.069	**-34.605**
Trades	0.10	0.002	0.32	0.003	0.217	**65.332**
Engineering	0.03	0.001	0.17	0.002	0.139	**57.841**
Business	0.24	0.003	0.18	0.002	-0.057	**-16.288**
Arts	0.09	0.002	0.07	0.001	-0.028	**-12.201**
Soc Science	0.15	0.002	0.07	0.001	-0.083	**-31.388**
Legal	0.03	0.001	0.01	0.001	-0.016	**-13.455**
Health	0.19	0.002	0.03	0.001	-0.158	**-59.947**
Index						
Index	0.09	0.082	0.12	0.039	0.030	**34.644**

Notes: This table shows the summary statistics of key variables of the Canadian Census: 2016. Standard deviation is calculated using unconditional OLS regressions. 'Difference' is the male coefficients minus female coefficients. Index is education/job relatedness. Higher index indicates education is more similar to job.

Table 3.5: Female coefficient table: NGS 1998

	<20k		<30k		<40k	
	Coef	t-val	Coef	t-val	Coef	t-val
Occupation						
Management	0.01	0.52	0.03	1.08	0.03	1.54
Bus/fin/adm	-0.08	-2.6	0.08	2.25	-0.04	-1.57
Nat/appl sci	0.02	0.65	0.03	0.86	-0.06	-2.56
Health	-0.09	-3.97	-0.01	-0.25	0.01	0.59
Socsc/ed/gvt/rel	-0.04	-0.86	0.01	0.25	-0.02	-0.49
Art/culture/rec	-0.02	-0.76	0.01	0.34	0.02	0.87
Transp/equip	0.02	0.25	0.04	0.37	-0.03	-0.38
Primary/ind	0.17	1.26	0.08	0.5	0.03	0.22
Proc/manufac	-0.01	-0.1	-0.03	-0.34	-0.08	-1.39
Education						
University	-0.31	-22.78	-0.37	-24.36	-0.17	-14.29
Graduate	-0.31	-11.9	-0.55	-18.83	-0.47	-20.69
Province						
Atlantic	0.18	8.6	0.13	5.69	0.06	3.48
Quebec	0.09	6.39	0.04	2.74	0.03	2.61
West prov/Terr	0.03	1.87	0	-0.03	0.01	0.56
Educ Requirement						
Same/Less	-0.01	-0.57	-0.02	-1.87	0.01	0.62
Program						
Educ	0.07	3.49	0.02	0.7	0.06	3.33
Fine Arts	0.16	3.79	0.11	2.33	0.05	1.3
Humanities	0.06	2.26	0.12	4.04	0.1	4.21
Soc sciences	0	0.15	0.06	2.75	0.06	3.95
Com/mgt/bus	0.19	6.06	0.13	3.67	0.11	4.06
Agri bio sci	0.03	0.73	-0.31	-6.26	-0.1	-2.68
Engineering	-0.18	-5.07	-0.19	-4.89	-0.01	-0.31
Tech trades	-0.11	-4.13	-0.18	-6	-0.04	-1.85
Health	0	-0.07	-0.16	-3.34	-0.01	-0.38
Math phy sci	0.04	0.89	0.1	2.04	0.06	1.41
Qualification						
Same/less	-0.12	-10.14	-0.17	-12.39	-0.09	-8.72
Job Relatedness						
No	0.21	14.78	0.18	11.19	0.05	4.35

Notes: This table shows the estimated coefficients of key variables the NGS 1998 results.

Table 3.6: Female coefficient table: NGS 2018

	<30k		<50k		<70k	
	Coef	t-val	Coef	t-val	Coef	t-val
Occupation						
Management	0.02	1.37	0.23	7.41	0.26	9.73
Bus/fin/adm	0.02	0.9	0.11	3.14	0.1	3.3
Nat/appl sci	0.01	0.69	0.05	1.38	0.07	2.3
Health	0.04	2.56	0.18	5.78	0.22	8.06
Socsc/ed/gvt/rel	0.03	1.5	0.29	6.39	0.29	7.25
Art/culture/rec	0.12	6.73	0.21	5.35	0.18	5.39
Transp/equip	-0.03	-0.52	-0.13	-1.06	0.12	1.18
Manu/resour	-0.03	-0.74	0.17	1.96	0.19	2.59
Education						
University	-0.04	-3.63	-0.17	-7.65	-0.07	-3.81
Graduate	-0.05	-3.87	-0.26	-9.94	-0.17	-7.65
Province						
Atlantic	0.03	2.01	0.07	2.51	0.03	1.35
Quebec	0.01	1.62	0.06	4.07	0.11	8.79
West prov/Terr	0	0.08	-0.02	-1.19	-0.05	-3.4
Educ Requirement						
Same/Less	-0.05	-4.93	-0.29	-14.22	-0.11	-6.33
Program						
Educ	0.03	2.53	-0.03	-0.99	0.07	3.2
Fine Arts	0.06	2.76	0.07	1.47	0.03	0.75
Humanities	0.05	3.4	0.19	5.77	0.15	5.17
Soc sciences	0.04	4.72	0.1	5.39	0.1	6.23
Com/mgt/bus	0.06	4.18	0.12	4.02	0.16	5.91
Phy/life sci	-0.01	-0.29	-0.07	-1.45	-0.06	-1.46
Math/compu	0.01	0.59	-0.05	-1.37	0.06	1.86
Engineering	0.01	0.54	-0.1	-3.89	0	-0.19
Agri bio sci	0.04	2.25	0.03	0.89	0.13	4.23
Qualification						
Same/less	0	0.57	0.02	0.95	0.01	0.81
Job Relatedness						
No	-0.01	-0.75	0.08	3.52	0.04	2.01

Notes: This table shows the estimated coefficients of key variables the NGS 2018 results.

Table 3.7: Male coefficient table: NGS 1998

	<20k		<30k		<40k	
	Coef	t-val	Coef	t-val	Coef	t-val
Occupation						
Management	0.01	0.75	0.02	0.76	0.04	1.79
Bus/fin/adm	0	0.13	0.05	2.19	0.03	1.16
Nat/appl sci	0.02	0.38	0.03	0.36	0.11	1.69
Health	0	0.21	0.1	3.68	-0.01	-0.22
Socsc/ed/gvt/rel	0.11	3.23	0.04	0.94	0	0.1
Art/culture/rec	-0.01	-0.68	0.05	1.78	0.05	2.13
Transp/equip	0.06	2.94	0.09	3.04	0.05	1.96
Primary/ind	0.05	1.19	0.08	1.46	0.08	1.48
Proc/manufac	-0.02	-0.76	0.04	1.09	0.03	1.1
Education						
University	-0.06	-4.09	-0.2	-10.46	-0.13	-7.46
Graduate	-0.07	-3.25	-0.32	-10.87	-0.38	-13.83
Province						
Atlantic	0.09	5.4	0.14	6.33	0.12	5.68
Quebec	0.07	6.35	0.15	10.29	0.11	7.85
West prov/Terr	0.03	2.47	0.03	2.15	0	-0.01
Educ Requirement						
Same/Less	0.01	1.28	-0.02	-1.83	0.02	2.14
Program						
Educ	0	-0.18	0.09	3.16	0.17	6.02
Fine Arts	0.03	0.57	-0.05	-0.86	0	-0.02
Humanities	0.21	7.89	0.24	7.05	0.23	7.07
Soc sciences	0.05	2.78	0.09	3.9	0.05	2.11
Com/mgt/bus	0.14	5.23	0.26	7.21	0.14	4.16
Agri bio sci	-0.04	-2.18	-0.13	-5.1	-0.17	-6.93
Engineering	-0.01	-0.58	-0.02	-0.9	-0.02	-1.03
Tech trades	-0.04	-0.73	-0.02	-0.32	-0.16	-2.56
Health	0	-0.12	-0.11	-3.62	-0.11	-4.01
Math phy sci	0.08	1.31	0.26	3.18	0.11	1.53
Qualification						
Same/less	-0.1	-9.8	-0.17	-12.54	-0.14	-11.05
Job Relatedness						
No	0.16	13.2	0.16	9.92	0.06	4.03

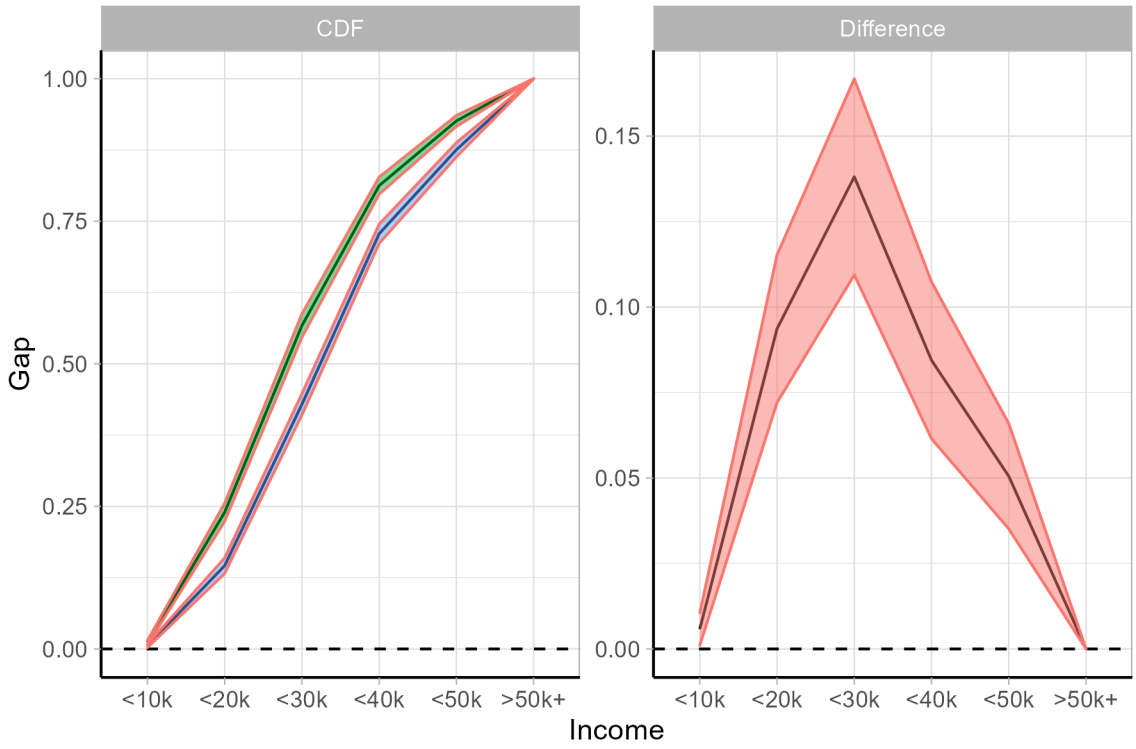
Notes: This table shows the estimated coefficients of key variables the NGS 1998 results.

Table 3.8: Male coefficient table: NGS 2018

	<30k		<50k		<70k	
	Coef	t-val	Coef	t-val	Coef	t-val
Occupation						
Management	0	-0.34	0.1	3.2	0.19	5.68
Bus/fin/adm	0	0.07	0.04	1.23	0.09	2.79
Nat/appl sci	-0.02	-0.8	0.01	0.24	0.04	0.68
Health	0.03	2.54	0.13	3.77	0.16	4.39
Socsc/ed/gvt/rel	0.07	3.1	0.24	4.18	0.22	3.66
Art/culture/rec	0.04	2.44	0.19	4.79	0.18	4.28
Transp/equip	0	0.06	0.04	0.83	0.04	0.7
Manu/resour	0.08	3.75	0.2	3.54	0.06	1.08
Education						
University	0.03	2.52	-0.16	-5.39	-0.04	-1.44
Graduate	0.04	3.12	-0.14	-4.25	-0.14	-3.82
Province						
Atlantic	0.02	1.4	0.15	4.79	0.12	3.66
Quebec	0.01	1.26	0.07	3.89	0.16	8.4
West prov/Terr	0.02	2.01	0.04	2	0.03	1.23
Educ Requirement						
Same/Less	-0.04	-4.58	-0.18	-6.61	-0.15	-5.38
Program						
Educ	-0.01	-0.82	-0.07	-1.65	0.08	1.72
Fine Arts	-0.04	-2.01	-0.08	-1.32	0.03	0.52
Humanities	-0.01	-0.31	0.02	0.33	0.11	2.16
Soc sciences	-0.01	-1.02	0.02	0.77	0.03	1.01
Com/mgt/bus	0.03	1.89	0.11	2.85	0.12	3.04
Phy/life sci	0.01	0.56	-0.04	-1.18	0	-0.05
Math/compu	-0.02	-1.99	-0.1	-4.12	-0.03	-1.25
Engineering	0.01	0.37	-0.01	-0.22	0.01	0.3
Agri bio sci	-0.04	-2.61	-0.05	-1.38	0.01	0.17
Qualification						
Same/less	0.02	2.31	0.02	1.19	0.05	2.3
Job Relatedness						
No	0.04	4.04	0.08	3.17	0.04	1.58

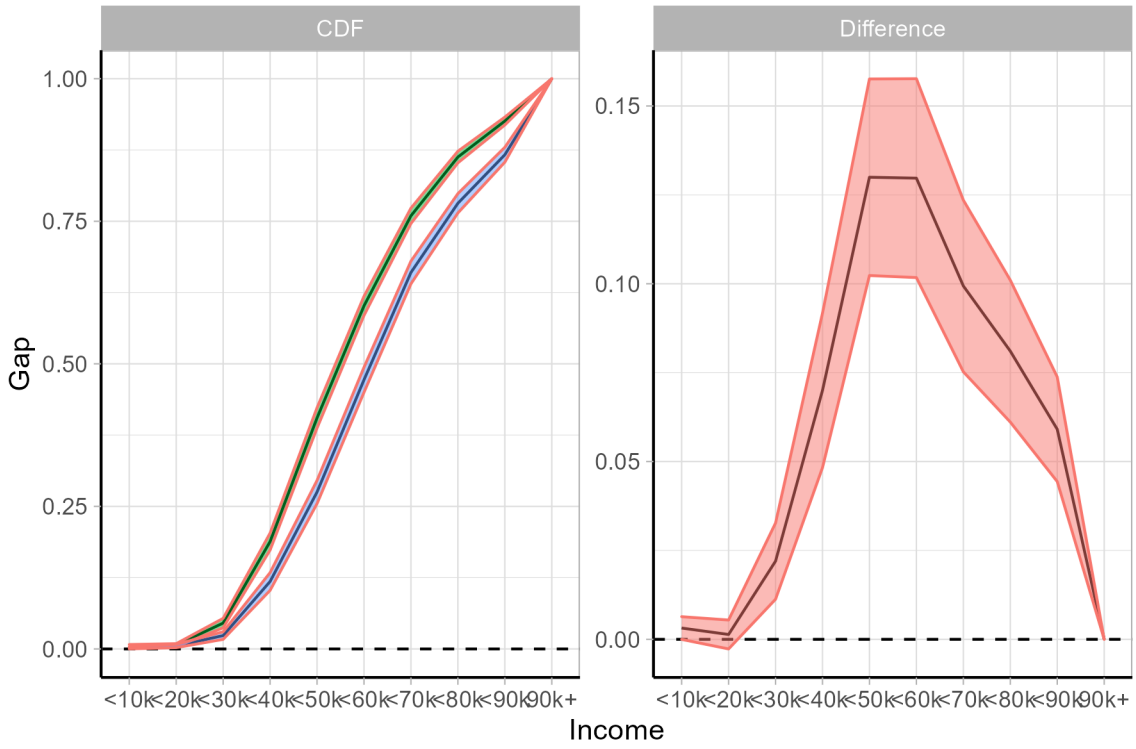
Notes: This table shows the estimated coefficients of key variables the NGS 2018 results.

Figure 3.1: Total wage gap by income quantiles: 1998



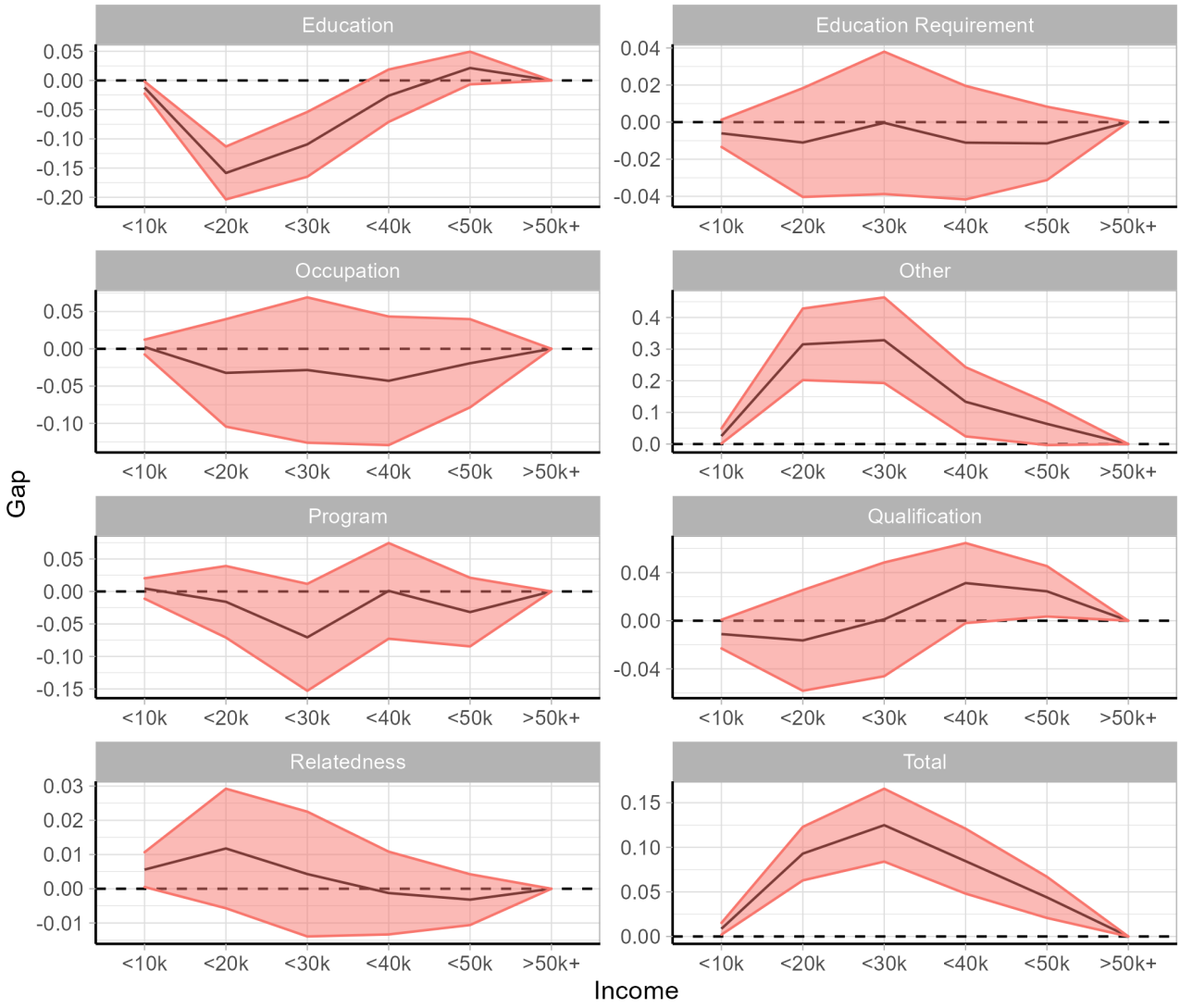
Notes: This figure shows the CDF and total wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes proportions. The shaded area is the 95% confidence interval, calculated using bootstrap. The top line shaded in green is the CDF of men while the bottom line shaded in blue is the CDF of women. Difference refers to the differences in CDF between men and women.

Figure 3.2: Total wage gap by income quantiles: 2018



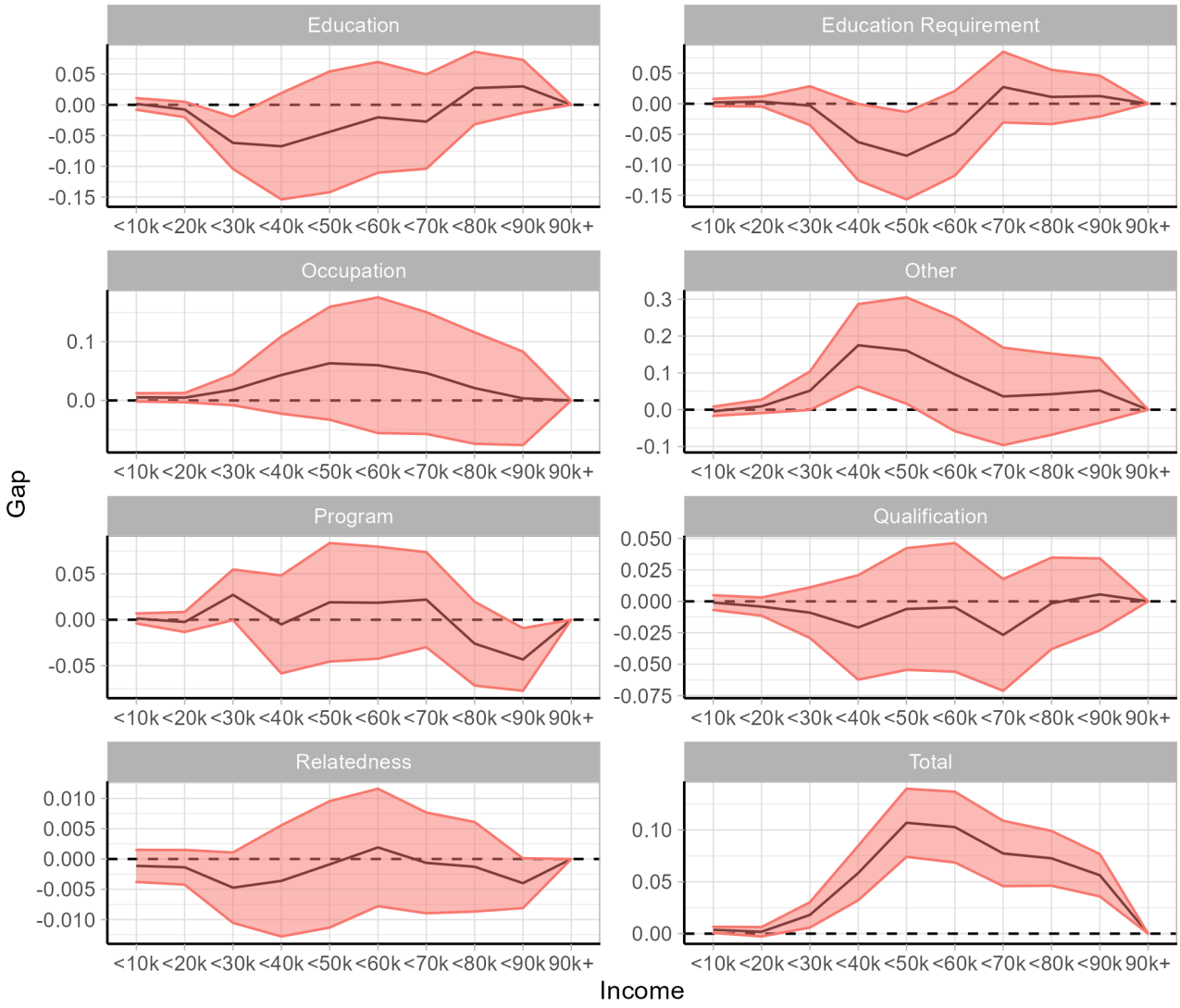
Notes: This figure shows the CDF and total wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes proportions. The shaded area is the 95% confidence interval, calculated using bootstrap. The top line shaded in green is the CDF of men while the bottom line shaded in blue is the CDF of women. Difference refers to the differences in CDF between men and women.

Figure 3.3: Unexplained wage gap by income quantiles: NGS 1998



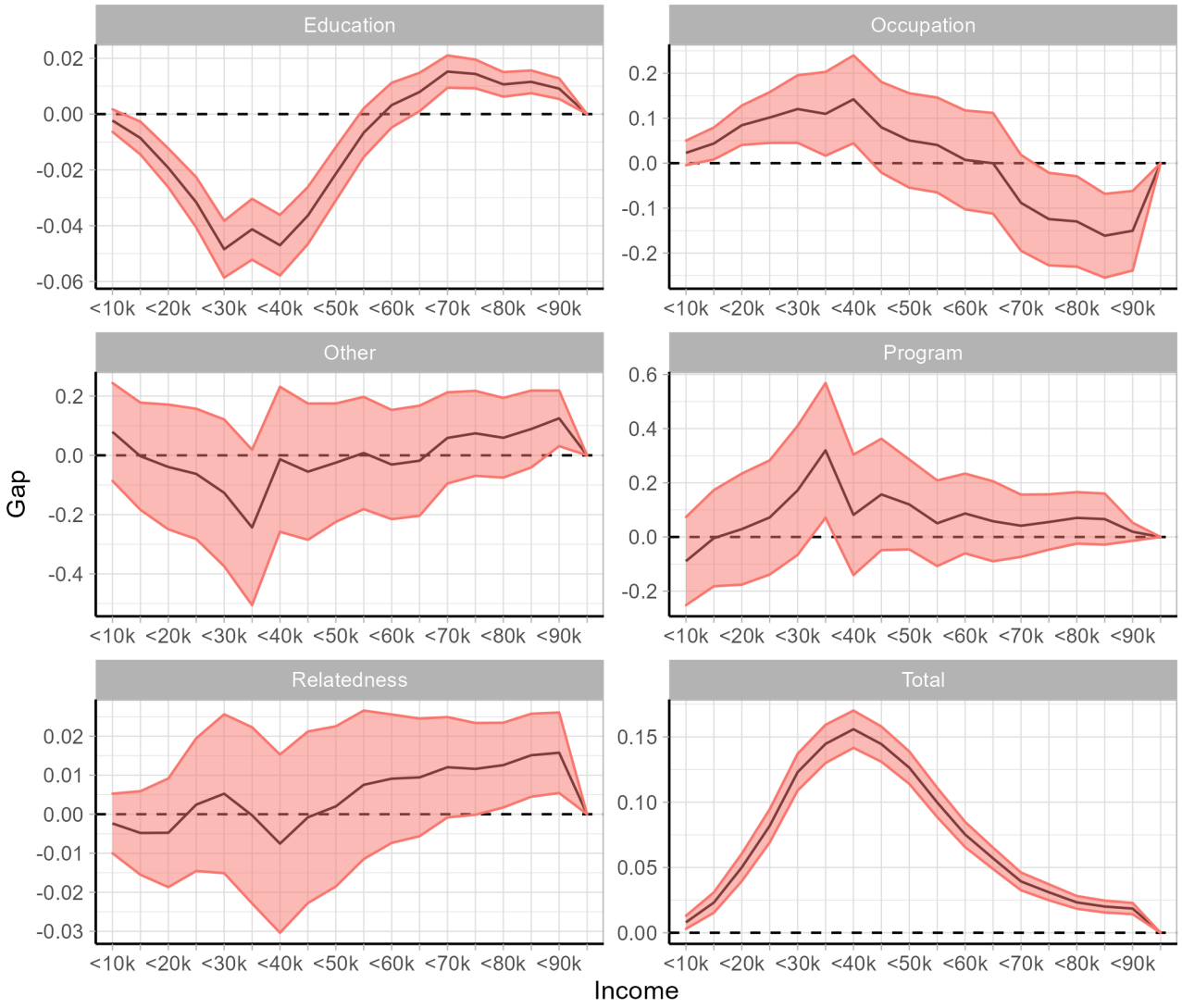
Notes: This figure shows the unexplained wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure 3.4: Unexplained wage gap by income quantiles: NGS 2018



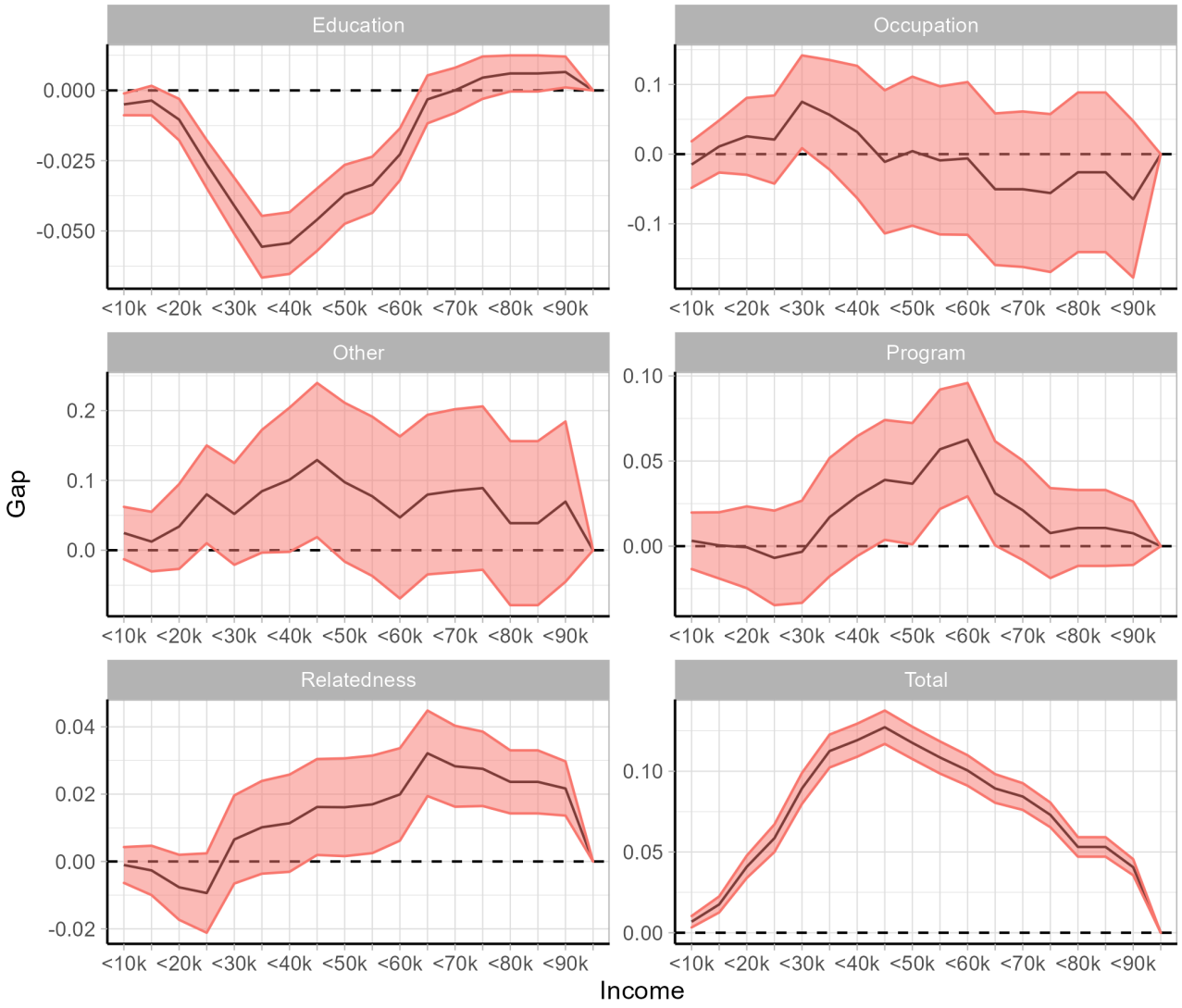
Notes: This figure shows the unexplained wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure 3.5: Unexplained wage gap by income quantiles: Census 1998



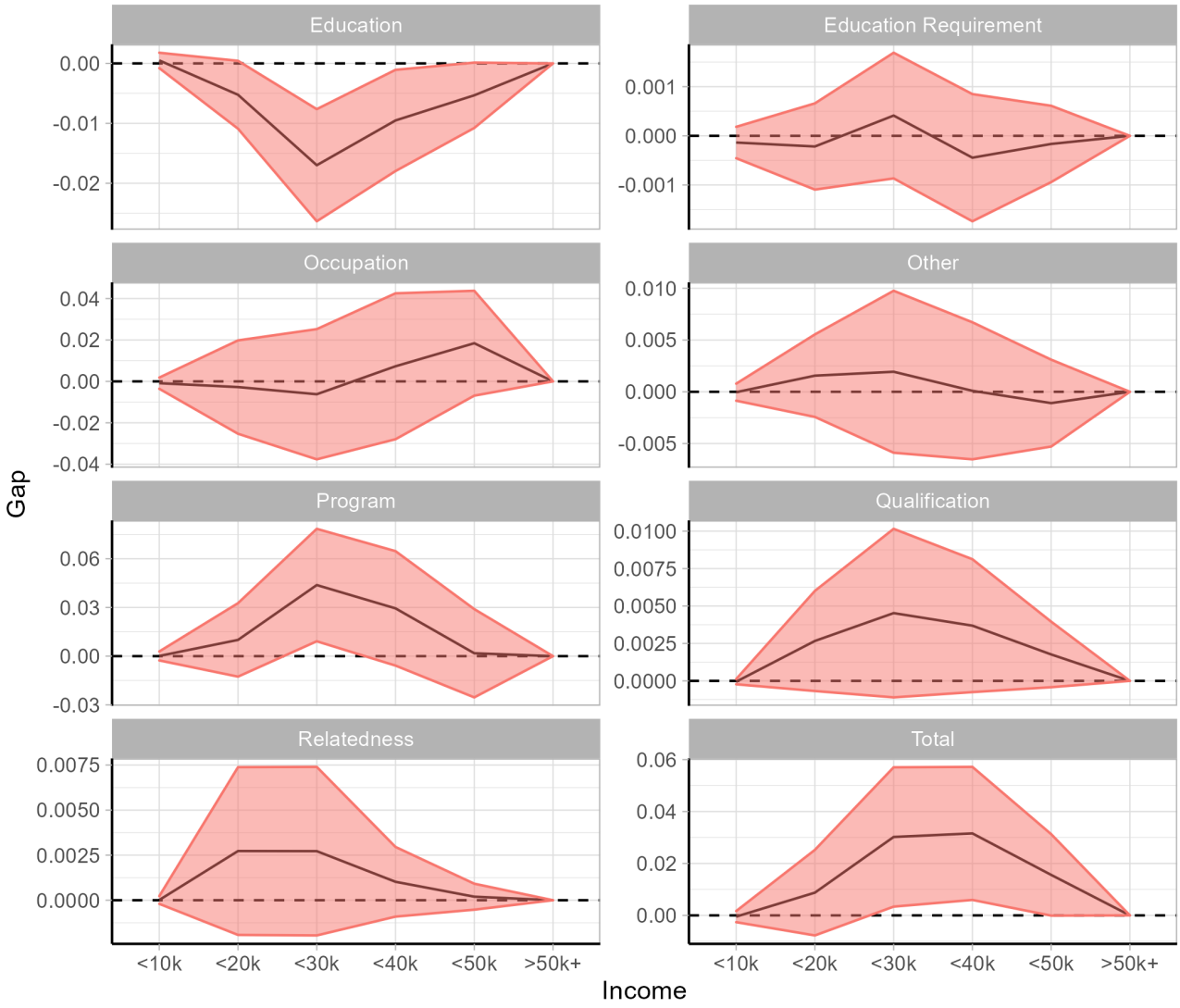
Notes: This figure shows the unexplained wage gap calculated using Census data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure 3.6: Unexplained wage gap by income quantiles: Census 2018



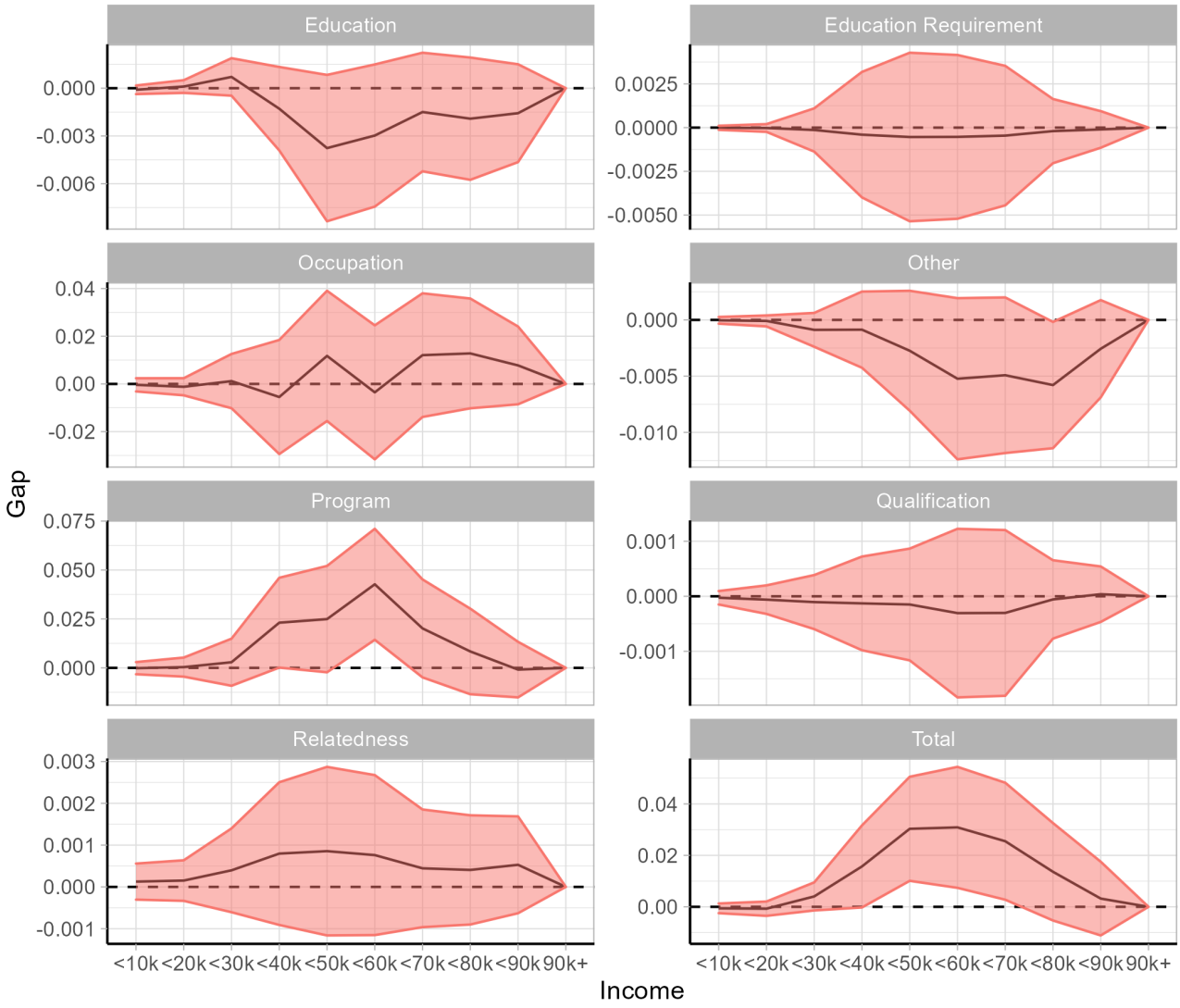
Notes: This figure shows the unexplained wage gap calculated using Census data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure 3.7: Explained wage gap by income quantiles: NGS 1998



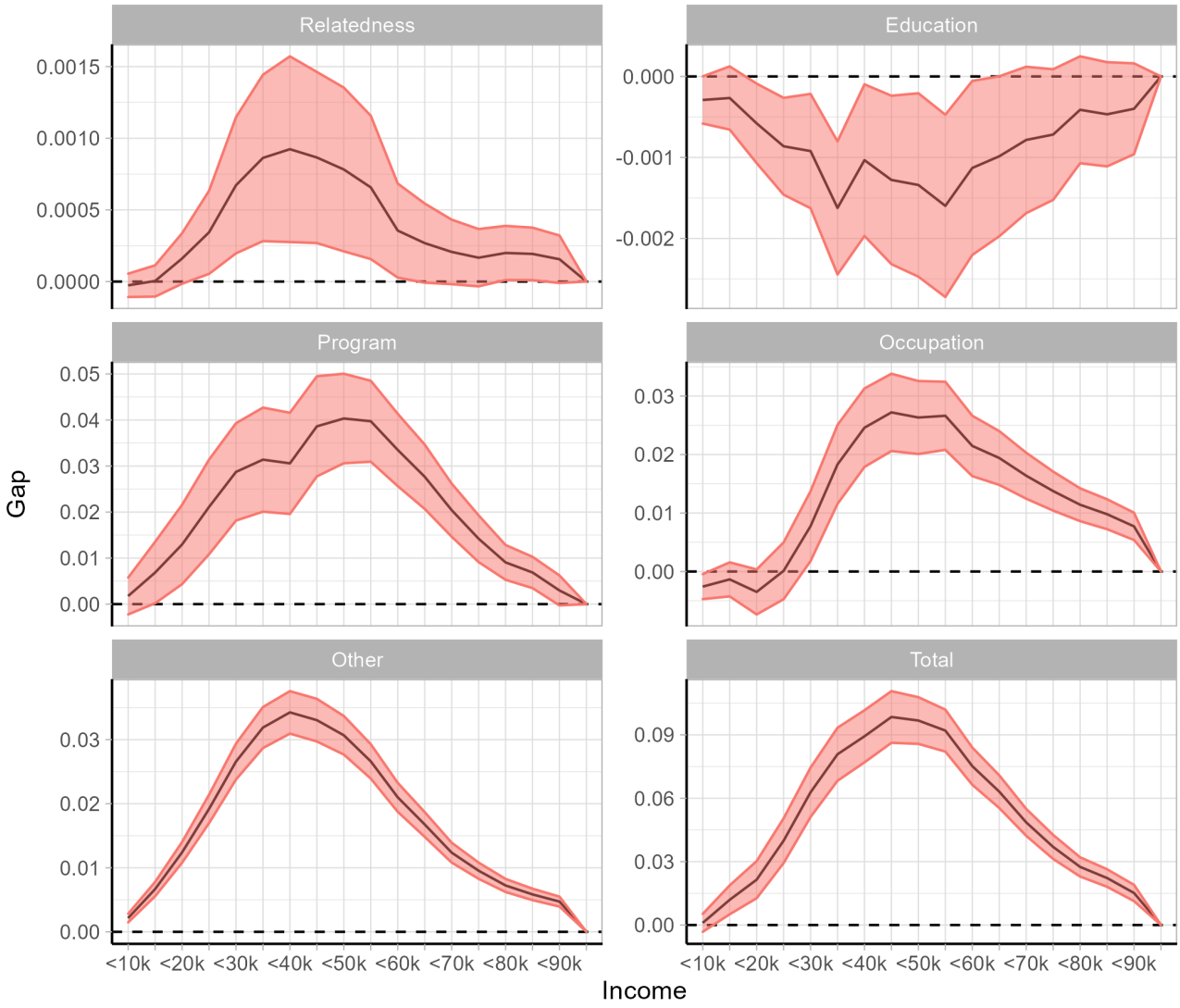
Notes: This figure shows the explained wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total explained gap.

Figure 3.8: Explained wage gap by income quantiles: NGS 2018



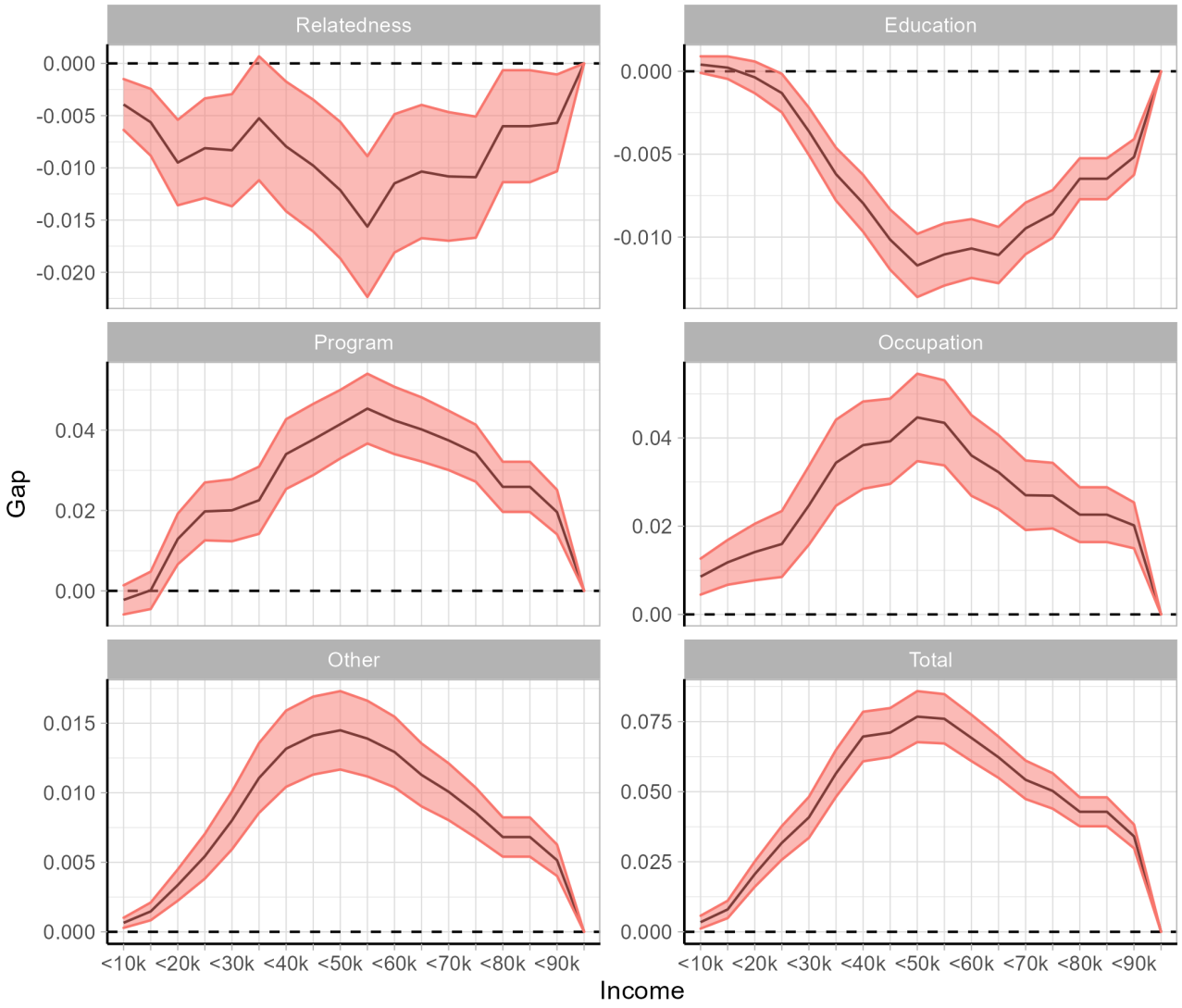
Notes: This figure shows the explained wage gap calculated using National Graduate Survey (NGS) data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total explained gap.

Figure 3.9: Explained wage gap by income quantiles: Census 1998



Notes: This figure shows the explained wage gap calculated using Census data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total explained gap.

Figure 3.10: Explained wage gap by income quantiles: Census 2018



Notes: This figure shows the explained wage gap calculated using Census data. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total explained gap.

References

Agut, Sonia, José M Peiró, and Rosa Grau. 2009. "The Effect of Overeducation on Job Content Innovation and Career-Enhancing Strategies Among Young Spanish Employees." *Journal of Career Development* 36 (2): 159–82.

Allen, Jim, and Rolf Van der Velden. 2001. "Educational Mismatches Versus Skill Mismatches: Effects on Wages, Job Satisfaction, and on-the-Job Search." *Oxford Economic Papers* 53 (3): 434–52.

Altonji, Joseph G, Prashant Bharadwaj, and Fabian Lange. 2012. "Changes in the Characteristics of American Youth: Implications for Adult Outcomes." *Journal of Labor Economics* 30 (4): 783–828.

Altonji, Joseph G., and Rebecca M. Blank. 1999. "Chapter 48 Race and Gender in the Labor Market." In *Handbook of Labor Economics*, 3:3143–3259. Elsevier. https://doi.org/10.1016/S1573-4463(99)30039-0.

Angrist, Joshua D., and Brigham Frandsen. 2022. "Machine Labor." *Journal of Labor Economics* 40 (S1): S97–140. https://doi.org/10.1086/717933.

Antonie, Luiza, Miana Plesca, Jennifer Teng, et al. 2016. "Heterogeneity in the Gender Wage Gap in Canada."

Arcidiacono, Peter. 2004. "Ability Sorting and the Returns to College Major." *Journal of Econometrics* 121 (1-2): 343–75.

Arellano, Manuel, and Stéphane Bonhomme. 2017. "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality." *Econometrica* 85 (1): 1–28.

Arntz, Melanie, Terry Gregory, and Florian Lehmer. 2011. "Unequal Pay or Unequal Employment? What Drives the Skill-Composition of Labor Flows in Germany?" *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1988857.

Atkinson, Anthony B, Alessandra Casarico, and Sarah Voitchovsky. 2018. "Top Incomes and

the Gender Divide." *The Journal of Economic Inequality* 16 (2): 225–56.

Aydede, Yigit, and Atul Dar. 2016. "The Cost of Immigrants' Occupational Mismatch and the Effectiveness of Postarrival Policies in Canada." *IZA Journal of Migration* 5 (1): 1–23.

Bach, Philipp, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. 2023. "DoubleML – An Object-Oriented Implementation of Double Machine Learning in R." arXiv. http://arxiv.org/abs/2103.09603.

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2018. "Closing the US Gender Wage Gap Requires Understanding Its Heterogeneity." *arXiv Preprint arXiv:1812.04345*.

———. 2021. "Closing the U.S. Gender Wage Gap Requires Understanding Its Heterogeneity." arXiv. http://arxiv.org/abs/1812.04345.

Baiardi, Anna, and Andrea Naghi. 2020. "The Value Added of Machine Learning to Causal Inference: Evidence from Revisited Studies." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3759867.

Baker, Michael, and Marie Drolet. 2010a. "A New View of the Male/Female Pay Gap." *Canadian Public Policy* 36 (4): 429–64.

———. 2010b. "A New View of the Male/Female Pay Gap." *Canadian Public Policy / Analyse de Politiques* 36 (4): 429–64. https://www.jstor.org/stable/25782105.

Bender, Keith A, and Kristen Roche. 2013. "Educational Mismatch and Self-Employment." *Economics of Education Review* 34: 85–95.

Binder, Ariel J, and John Bound. 2019. "The Declining Labor Market Prospects of Less-Educated Men." *Journal of Economic Perspectives* 33 (2): 163–90.

Blau, Francine D., and Lawrence M. Kahn. 2017a. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865. https://doi.org/10.1257/jel.20160995.

Blau, Francine D, and Lawrence M Kahn. 2000. "Gender Differences in Pay." *Journal of Economic Perspectives* 14 (4): 75–100. https://doi.org/10.1257/jep.14.4.75.

———. 2007. "Changes in the Labor Supply Behavior of Married Women: 1980–2000." *Journal of Labor Economics* 25 (3): 393–438.

———. 2017b. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal*

of Human Resources, 436–55.

Boll, Christina, and Julian Sebastian Leppin. 2013. "Equal Matches Are Only Half the Story: Why German Female Graduates Earn 27% Less Than Males." HWWI Research Paper.

Boll, Christina, Julian Leppin, Anja Rossen, and André Wolf. 2016. "Magnitude and Impact Factors of the Gender Pay Gap in EU Countries." *Report Prepared for and Financed by the European Commission–Directorate-General for Justice, European Union (Hrsg.), Hamburg.*

Boudarbat, Brahim, and Victor Chernoff. 2012. "Education–Job Match Among Recent Canadian University Graduates." *Applied Economics Letters* 19 (18): 1923–26.

Boudarbat, Brahim, and Marie Connolly. 2013. "The Gender Wage Gap Among Recent Post-Secondary Graduates in Canada: A Distributional Approach." *Canadian Journal of Economics/Revue Canadienne d'économique* 46 (3): 1037–65.

Boudarbat, Brahim, Thomas Lemieux, and W Craig Riddell. 2010. "The Evolution of the Returns to Human Capital in Canada, 1980–2005." *Canadian Public Policy* 36 (1): 63–89.

Card, David, Ana Rute Cardoso, and Patrick Kline. 2016. "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *." *The Quarterly Journal of Economics* 131 (2): 633–86. https://doi.org/10.1093/qje/qjv038.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018b. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68. https://doi.org/10.1111/ectj.12097.

———. 2018a. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2014. "Gaussian Approximation of Suprema of Empirical Processes."

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81 (6): 2205–68.

Cosser, Michael. 2010. "The Skills Cline: Higher Education and the Supply-Demand Complex in South Africa." *Higher Education* 59: 43–53.

David, H, Lawrence F Katz, and Melissa S Kearney. 2005. "Rising Wage Inequality: The Role of Composition and Prices." National Bureau of Economic Research.

DiNardo, John, Nicole M Fortin, and Thomas Lemieux. 1995. "Labor Market Institutions and

the Distribution of Wages, 1973-1992: A Semiparametric Approach." National bureau of economic research.

Drolet, Marie. 2011. "Why Has the Gender Wage Gap Narrowed?" *Statistics Canada*, 3–13.

Duncan, Greg J, and Saul D Hoffman. 1981. "The Incidence and Wage Effects of Overeducation." *Economics of Education Review* 1 (1): 75–86.

Ferguson, Sarah Jane. 2016. "Women and Education: Qualifications, Skills and Technology. Women in Canada: A Gender-Based Statistical Report." *Statistics Canada*.

Figueiredo, Hugo, Vera Rocha, Ricardo Biscaia, and Pedro Teixeira. 2015. "Gender Pay Gaps and the Restructuring of Graduate Labour Markets in Southern Europe." *Cambridge Journal of Economics* 39 (2): 565–98.

Firpo, Sergio, Nicole M Fortin, and Thomas Lemieux. 2009. "Unconditional Quantile Regressions." *Econometrica* 77 (3): 953–73.

Fortin, Nicole M. 2019. "Increasing Earnings Inequality and the Gender Pay Gap in Canada: Prospects for Convergence." *Canadian Journal of Economics/Revue Canadienne d'économique* 52 (2): 407–40.

Fortin, Nicole M, Brian Bell, and Michael Böhm. 2017. "Top Earnings Inequality and the Gender Pay Gap: Canada, Sweden, and the United Kingdom." *Labour Economics* 47: 107–23.

Fortin, Nicole M, and Thomas Lemieux. 1998. "Rank Regressions, Wage Distributions, and the Gender Gap." *Journal of Human Resources*, 610–43.

———. 2015. "Changes in Wage Inequality in Canada: An Interprovincial Perspective." *Canadian Journal of Economics/Revue Canadienne d'économique* 48 (2): 682–713.

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Chapter 1 - Decomposition Methods in Economics." In *Handbook of Labor Economics*, edited by Orley Ashenfelter and David Card, 4:1–102. Elsevier. https://doi.org/10.1016/S0169-7218(11)00407-2.

Freeman, Richard. 1976. "The Overeducated American."

Gibbons, Robert, Lawrence F Katz, Thomas Lemieux, and Daniel Parent. 2005. "Comparative Advantage, Learning, and Sectoral Wage Determination." *Journal of Labor Economics* 23 (4): 681–724.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4): 1091–1119. https://doi.org/10.1257/aer.104.4.1091.

Hara, Hiromi. 2018. "The Gender Wage Gap Across the Wage Distribution in Japan: Within-and

Between-Establishment Effects." *Labour Economics* 53: 213–29.

Hartog, Joop. 2000. "Over-Education and Earnings: Where Are We, Where Should We Go?" *Economics of Education Review* 19 (2): 131–47.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society*, 153–61.

Indicators, OECD. 2016. "Education at a Glance 2016." *Editions OECD* 90.

Jann, Ben. 2008. "A Stata Implementation of the Blinder-Oaxaca Decomposition." *The Stata Journal* 8 (4): 453–79.

Jong-Wha, LEE, and Dainn Wie. 2017. "Wage Structure and Gender Earnings Differentials in China and India." *World Development* 97: 313–29.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy* 101 (3): 410–42.

Keane, Michael P, Petra E Todd, and Kenneth I Wolpin. 2011. "The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications." In *Handbook of Labor Economics*, 4:331–461. Elsevier.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101 (3): 532–37. https://doi.org/10.1257/aer.101.3.532.

Koenker, Roger, and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica: Journal of the Econometric Society*, 33–50.

Koenker, Roger, Stephen Portnoy, Pin Tian Ng, Achim Zeileis, Philip Grosjean, and Brian D Ripley. 2012. "Package 'Quantreg'." Technical Report Last accessed April 21th.

Kunze, Astrid. 2018. "The Gender Wage Gap in Developed Countries." *The Oxford Handbook of Women and the Economy*, 369.

Latif, Ehsan. 2006. "Labour Supply Effects of Informal Caregiving in Canada." *Canadian Public Policy* 32 (4): 413–29.

Lefebvre, Pierre, and Philip Merrigan. 2008. "Child-Care Policy and the Labor Supply of Mothers with Young Children: A Natural Experiment from Canada." *Journal of Labor Economics* 26 (3): 519–48.

Lemieux, Thomas. 2014. "Occupations, Fields of Study and Returns to Education." *Canadian Journal of Economics/Revue Canadienne d'économique* 47 (4): 1047–77.

Leuven, Edwin, and Hessel Oosterbeek. 2011. "Overeducation and Mismatch in the Labor

Market." *Handbook of the Economics of Education* 4: 283–326.

Levels, Mark, Rolf Van der Velden, and Valentina Di Stasio. 2014. "From School to Fitting Work: How Education-to-Job Matching of European School Leavers Is Related to Educational System Characteristics." *Acta Sociologica* 57 (4): 341–61.

Li, Ian W, and Paul W Miller. 2012. "Gender Discrimination in the Australian Graduate Labour Market." *Australian Journal of Labour Economics* 15 (3): 167–99.

Linsley, Ingrid. 2005. "Causes of Overeducation in the Australian Labour Market." *Australian Journal of Labour Economics* 8 (2): 121–43.

Maasoumi, Esfandiar, and Le Wang. 2019. "The Gender Gap Between Earnings Distributions." *Journal of Political Economy* 127 (5): 2438–2504.

Machado, José AF, and José Mata. 2005. "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression." *Journal of Applied Econometrics* 20 (4): 445–65.

Meroni, Elena Claudia, and Esperanza Vera-Toscano. 2017. "The Persistence of Overeducation Among Recent Graduates." *Labour Economics* 48: 120–43.

Milligan, Kevin, and Mark Stabile. 2011. "Do Child Tax Benefits Affect the Well-Being of Children? Evidence from Canadian Child Benefit Expansions." *American Economic Journal: Economic Policy* 3 (3): 175–205.

Moyser, Melissa. 2019. "Measuring and Analyzing the Gender Pay Gap: A Conceptual and Methodological Overview." *Statistics Canada*, no. 45200002 (August): 1–42.

Mulligan, Casey B., and Yona Rubinstein. 2008a. "Selection, Investment, and Women's Relative Wages Over Time *." *Quarterly Journal of Economics* 123 (3): 1061–1110. https://doi.org/10.1162/qjec.2008.123.3.1061.

Mulligan, Casey B, and Yona Rubinstein. 2008b. "Selection, Investment, and Women's Relative Wages over Time." *The Quarterly Journal of Economics* 123 (3): 1061–1110.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review*, 693–709.

Olivetti, Claudia, and Barbara Petrongolo. 2016a. "The Evolution of Gender Gaps in Industrialized Countries." *Annual Review of Economics* 8: 405–34.

———. 2016b. "The Evolution of Gender Gaps in Industrialized Countries." w21887. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w21887.

Park, Kihong. 2021. "Education-Job Mismatch and Gender Wage Gap: Evidence from Recent

College Graduates in Korea." *Asian Women* 37 (1): 1–24.

Pelletier, Rachelle, Martha Patterson, and Melissa Moyser. 2019a. "The Gender Wage Gap in Canada: 1998 to 2018." Statistics Canada= Statistique Canada.

———. 2019b. "The Gender Wage Gap in Canada: 1998 to 2018." *Statistics Canada* 004 (75): 1–15.

Pető, Rita, and Balázs Reizer. 2021. "Gender Differences in the Skill Content of Jobs." *Journal of Population Economics* 34 (3): 825–64. https://doi.org/10.1007/s00148-021-00825-6.

Powell, Lisa M. 2002. "Joint Labor Supply and Childcare Choice Decisions of Married Mothers." *Journal of Human Resources*, 106–28.

Robst, John. 2007. "Education, College Major, and Job Match: Gender Differences in Reasons for Mismatch." *Education Economics* 15 (2): 159–75.

Schirle, Tammy. 2015. "The Gender Wage Gap in the Canadian Provinces, 1997–2014." *Canadian Public Policy* 41 (4): 309–19.

Sloane, Peter J. 2007. "Overeducation in the United Kingdom." *Australian Economic Review* 40 (3): 286–91.

Słoczyński, Tymon. 2015. "The Oaxaca–Blinder Unexplained Component as a Treatment Effects Estimator." *Oxford Bulletin of Economics and Statistics* 77 (4): 588–604. https://doi.org/10.1111/obes.12075.

Somers, Melline A, Sofie J Cabus, Wim Groot, and Henriëtte Maassen van den Brink. 2019. "Horizontal Mismatch Between Employment and Field of Education: Evidence from a Systematic Literature Review." *Journal of Economic Surveys* 33 (2): 567–603.

Tsang, Mun Chiu. 1987. "The Impact of Underutilization of Education on Productivity: A Case Study of the US Bell Companies." *Economics of Education Review* 6 (3): 239–54.

Weinberg, Bruce A. 2000. "Computer Use and the Demand for Female Workers." *ILR Review* 53 (2): 290–308. https://doi.org/10.1177/001979390005300206.

Welch, Finis. 2000. "Growth in Women's Relative Wages and in Inequality Among Men: One Phenomenon or Two?" *American Economic Review* 90 (2): 444–49. https://doi.org/10.1257/aer.90.2.444.

Wolbers, Maarten HJ. 2003. "Job Mismatches and Their Labour-Market Effects Among School-Leavers in Europe." *European Sociological Review* 19 (3): 249–66.

\begin{appendices}
%

Appendix: Chapter 1

A1: Variable table

Table A1

Variable Name	Factor Levels	Description
PROV	Ontario, Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Manitoba, Saskatchewan, Alberta	Province
AGE_12	25-29, 30-34, 35-39, 40-44, 45-49, 50-54	Age
CMA	Other, Montreal, Toronto, Vancouver	Central metropolitan area
MARSTAT	Married or living in common-law, Single	Marital status
EDUC90	High School, < High School, College/Certificate, Bachelor's, Graduate	Highest educational attainment
AGYOWNKN	Other, <3, 3-5, 6-12, 13-15, 16-17, 18-24	Age of yongest child
EFAMSIZE	1,2,3,4,5+	Economic family size

Notes: Name of raw variables from the LFS and their factor levels. The first factor level denotes the reference level

A2: Gender wage gap

Table A2

	2002					2016				
	0.05	0.25	0.5	0.75	0.95	0.05	0.25	0.5	0.75	0.95
Total										
Regular	-0.18	-0.24	-0.22	-0.21	-0.2	-0.1	-0.16	-0.17	-0.16	-0.14
Selection	-0.12	-0.19	-0.18	-0.17	-0.16	-0.07	-0.12	-0.15	-0.14	-0.14
Explained										
Regular	0.01	0.02	0.02	0.02	0.02	0.02	0.04	0.05	0.04	0.03
Selection	0.04	0.04	0.03	0.01	0	0.04	0.05	0.05	0.03	0.02
Unexplained										
Regular	-0.19	-0.26	-0.24	-0.22	-0.21	-0.12	-0.2	-0.21	-0.2	-0.18
Selection	-0.16	-0.22	-0.21	-0.18	-0.16	-0.11	-0.17	-0.19	-0.18	-0.16

Notes: This table shows wage gap at select quantiles for regular and selection corrected quantile regression.

A3: Female coefficients: 2002

Table A3

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.02	0.01	2.21	0.01	2.48	0.01	2.71	0.01	2.94	0.02
Province										
NL	-0.3	0.01	-0.31	0.02	-0.27	0.02	-0.22	0.02	-0.22	0.02
PE	-0.15	0.02	-0.18	0.02	-0.21	0.01	-0.21	0.01	-0.25	0.02
NS	-0.2	0.01	-0.24	0.01	-0.26	0.01	-0.25	0.01	-0.24	0.02
NB	-0.19	0.01	-0.23	0.02	-0.22	0.01	-0.21	0.01	-0.22	0.01
QC	-0.05	0.01	-0.09	0.01	-0.09	0.01	-0.09	0.01	-0.13	0.01
MB	-0.07	0.01	-0.12	0.01	-0.15	0.01	-0.13	0.01	-0.15	0.01
SK	-0.09	0.02	-0.12	0.01	-0.12	0.01	-0.12	0.01	-0.12	0.01
AB	-0.06	0.01	-0.08	0.01	-0.1	0.01	-0.06	0.01	-0.06	0.02
BC	0.06	0.01	0.05	0.01	0.04	0.01	0.03	0.01	0.02	0.01
Age										
30-34	0.02	0.01	0.07	0.01	0.11	0.01	0.12	0.01	0.13	0.01
35-39	0.07	0.01	0.12	0.01	0.17	0.01	0.2	0.01	0.21	0.01
40-44	0.1	0.01	0.17	0.01	0.22	0.01	0.25	0.01	0.26	0.01
45-49	0.08	0.01	0.18	0.01	0.24	0.01	0.27	0.01	0.28	0.01
50-54	0.12	0.01	0.2	0.01	0.26	0.01	0.29	0.01	0.28	0.01
CMA										
Montreal	0.04	0.02	0.07	0.01	0.05	0.01	0.04	0.01	0.03	0.02
Toronto	0	0.02	-0.02	0.02	0.01	0.01	0.04	0.01	0.06	0.01
Vancouver	0.02	0.02	0.02	0.02	0.03	0.02	0.01	0.02	0.01	0.02
Married	0.06	0.01	0.07	0.01	0.07	0.01	0.05	0.01	0.05	0.01
Education										
< High School	-0.1	0.01	-0.19	0.01	-0.28	0.01	-0.27	0.01	-0.26	0.02
College	0.11	0.01	0.19	0.01	0.17	0.01	0.19	0.01	0.21	0.01
Bachelor's	0.34	0.01	0.49	0.01	0.52	0.01	0.49	0.01	0.44	0.01
Graduate	0.45	0.04	0.65	0.02	0.64	0.01	0.58	0.01	0.54	0.02
Yongest Age										
<3	0.06	0.01	0.07	0.02	0.07	0.01	0.06	0.01	0.04	0.02
3-5	0.03	0.01	0.04	0.02	0.05	0.01	0.04	0.01	0.02	0.01
6-12	0.02	0.01	0.02	0.01	0.03	0.01	0.03	0.01	0.02	0.01
13-15	0.02	0.01	0.02	0.01	0.02	0.01	0	0.01	-0.01	0.01
16-17	0.03	0.02	0.01	0.02	0.03	0.02	-0.01	0.01	0	0.02
18-24	0.04	0.01	0.03	0.01	0.04	0.01	0.01	0.01	0	0.01
Family size										
2	-0.03	0.01	-0.06	0.01	-0.07	0.01	-0.07	0.01	-0.06	0.01
3	-0.07	0.01	-0.1	0.02	-0.1	0.01	-0.09	0.01	-0.08	0.02
4	-0.06	0.02	-0.09	0.02	-0.1	0.02	-0.08	0.02	-0.07	0.02
5+	-0.1	0.02	-0.14	0.02	-0.15	0.02	-0.13	0.02	-0.11	0.02

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A4: Selection corrected female coefficients: 2002

Table A4

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	1.99	0.04	2.16	0.06	2.41	0.06	2.67	0.05	2.89	0.06
Province										
NL	-0.31	0.02	-0.42	0.06	-0.34	0.07	-0.25	0.04	-0.24	0.03
PE	-0.15	0.02	-0.2	0.02	-0.2	0.02	-0.23	0.02	-0.25	0.02
NS	-0.2	0.02	-0.25	0.02	-0.26	0.02	-0.23	0.01	-0.25	0.02
NB	-0.19	0.02	-0.25	0.02	-0.23	0.02	-0.23	0.02	-0.22	0.01
QC	-0.04	0.02	-0.1	0.01	-0.1	0.01	-0.1	0.01	-0.13	0.01
MB	-0.04	0.02	-0.09	0.03	-0.11	0.02	-0.11	0.02	-0.12	0.02
SK	-0.08	0.02	-0.1	0.02	-0.1	0.01	-0.1	0.01	-0.11	0.02
AB	-0.05	0.02	-0.07	0.02	-0.08	0.01	-0.05	0.01	-0.05	0.02
BC	0.07	0.02	-0.02	0.03	0.02	0.02	0	0.02	-0.01	0.02
Age										
30-34	0	0.02	0.04	0.02	0.1	0.01	0.13	0.01	0.15	0.02
35-39	0.01	0.02	0.08	0.03	0.14	0.02	0.2	0.01	0.23	0.02
40-44	0.05	0.02	0.12	0.03	0.19	0.02	0.25	0.01	0.28	0.02
45-49	0.03	0.03	0.11	0.04	0.2	0.03	0.27	0.02	0.3	0.02
50-54	0.03	0.05	0.1	0.05	0.19	0.04	0.26	0.03	0.28	0.02
CMA										
Montreal	0.04	0.02	0.08	0.02	0.06	0.02	0.05	0.01	0.04	0.02
Toronto	-0.02	0.02	-0.03	0.02	0	0.02	0.04	0.02	0.06	0.02
Vancouver	0.01	0.02	0.04	0.03	0.01	0.02	0.03	0.02	0.03	0.02
Married	0.03	0.01	0.07	0.01	0.08	0.01	0.07	0.01	0.06	0.01
Education										
< High School	-0.14	0.01	-0.22	0.01	-0.36	0.06	-0.37	0.06	-0.33	0.07
College	0.1	0.01	0.19	0.02	0.22	0.04	0.19	0.02	0.21	0.03
Bachelor's	0.25	0.03	0.43	0.02	0.52	0.03	0.51	0.04	0.45	0.03
Graduate	0.27	0.07	0.57	0.03	0.68	0.04	0.62	0.04	0.56	0.03
Yongest Age										
<3	-0.03	0.05	-0.03	0.06	-0.01	0.05	0.02	0.04	0.02	0.04
3-5	-0.05	0.03	-0.08	0.05	-0.04	0.04	-0.01	0.03	-0.02	0.03
6-12	-0.01	0.02	-0.02	0.02	-0.01	0.02	0.01	0.02	0.02	0.02
13-15	0	0.02	0.03	0.03	0.02	0.02	0.01	0.02	0.01	0.02
16-17	0.02	0.02	0.05	0.03	0.04	0.02	0	0.02	0.01	0.03
18-24	0.07	0.02	0.07	0.03	0.05	0.02	0.04	0.02	0.01	0.02
Family size										
2	0	0.02	-0.06	0.02	-0.08	0.02	-0.08	0.02	-0.07	0.02
3	-0.04	0.02	-0.1	0.02	-0.11	0.02	-0.11	0.02	-0.09	0.02
4	-0.02	0.02	-0.09	0.02	-0.11	0.02	-0.11	0.02	-0.1	0.03
5+	-0.08	0.02	-0.18	0.03	-0.2	0.03	-0.19	0.03	-0.17	0.03

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A5: Male coefficients: 2002

Table A5

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.18	0.01	2.44	0.01	2.68	0.01	2.93	0.01	3.12	0.01
Province										
NL	-0.29	0.03	-0.28	0.02	-0.27	0.02	-0.2	0.02	-0.19	0.02
PE	-0.25	0.02	-0.3	0.02	-0.32	0.01	-0.37	0.02	-0.35	0.02
NS	-0.28	0.02	-0.28	0.01	-0.28	0.01	-0.25	0.01	-0.21	0.01
NB	-0.24	0.02	-0.26	0.01	-0.27	0.01	-0.24	0.01	-0.21	0.01
QC	-0.09	0.01	-0.11	0.01	-0.13	0.01	-0.12	0.01	-0.11	0.01
MB	-0.14	0.02	-0.15	0.01	-0.13	0.01	-0.11	0.01	-0.08	0.01
SK	-0.18	0.02	-0.15	0.02	-0.12	0.01	-0.09	0.01	-0.09	0.01
AB	-0.03	0.02	-0.01	0.01	0	0.01	0.04	0.01	0.08	0.01
BC	0.02	0.02	0.03	0.02	0.05	0.01	0	0.01	-0.01	0.01
Age										
30-34	0.1	0.01	0.09	0.01	0.1	0.01	0.1	0.01	0.1	0.01
35-39	0.12	0.01	0.13	0.01	0.16	0.01	0.17	0.01	0.16	0.01
40-44	0.15	0.01	0.18	0.01	0.21	0.01	0.22	0.01	0.21	0.01
45-49	0.2	0.02	0.23	0.01	0.27	0.01	0.29	0.01	0.28	0.01
50-54	0.2	0.02	0.25	0.01	0.3	0.01	0.32	0.01	0.3	0.01
CMA										
Montreal	-0.02	0.02	0	0.02	0.03	0.02	0.04	0.01	0.06	0.02
Toronto	-0.08	0.02	-0.08	0.02	-0.06	0.01	-0.01	0.01	0.05	0.01
Vancouver	-0.08	0.03	-0.05	0.03	-0.04	0.02	0.01	0.02	0.04	0.02
Married	0.15	0.02	0.14	0.01	0.11	0.01	0.08	0.01	0.07	0.01
Education										
< High School	-0.12	0.01	-0.13	0.01	-0.14	0.01	-0.16	0.01	-0.16	0.01
College	0.11	0.01	0.11	0.01	0.13	0.01	0.12	0.01	0.1	0.01
Bachelor's	0.22	0.02	0.31	0.01	0.37	0.01	0.35	0.01	0.34	0.01
Graduate	0.29	0.02	0.39	0.02	0.47	0.01	0.43	0.01	0.41	0.01
Yongest Age										
<3	0.08	0.02	0.1	0.01	0.08	0.01	0.09	0.01	0.07	0.01
3-5	0.1	0.02	0.1	0.02	0.09	0.01	0.11	0.01	0.08	0.01
6-12	0.12	0.02	0.12	0.01	0.08	0.01	0.08	0.01	0.06	0.01
13-15	0.11	0.03	0.11	0.02	0.09	0.01	0.08	0.01	0.07	0.02
16-17	0.12	0.02	0.11	0.02	0.07	0.01	0.09	0.02	0.07	0.02
18-24	0.11	0.02	0.13	0.02	0.09	0.01	0.08	0.01	0.05	0.02
Family size										
2	-0.03	0.02	-0.03	0.01	-0.05	0.01	-0.04	0.01	-0.01	0.01
3	-0.08	0.02	-0.11	0.02	-0.1	0.01	-0.1	0.01	-0.06	0.01
4	-0.06	0.02	-0.08	0.02	-0.06	0.01	-0.07	0.01	-0.05	0.02
5+	-0.09	0.02	-0.12	0.02	-0.1	0.01	-0.1	0.01	-0.05	0.02

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A6: Selection corrected male coefficients: 2002

Table A6

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.25	0.06	2.5	0.06	2.75	0.06	2.98	0.05	3.18	0.06
Province										
NL	-0.23	0.05	-0.23	0.06	-0.2	0.06	-0.13	0.07	-0.12	0.05
PE	-0.25	0.02	-0.3	0.02	-0.32	0.02	-0.35	0.02	-0.34	0.03
NS	-0.25	0.02	-0.27	0.02	-0.23	0.03	-0.22	0.02	-0.19	0.03
NB	-0.23	0.02	-0.25	0.02	-0.25	0.03	-0.21	0.03	-0.19	0.03
QC	-0.09	0.02	-0.12	0.01	-0.12	0.01	-0.12	0.01	-0.11	0.01
MB	-0.16	0.03	-0.17	0.02	-0.15	0.01	-0.13	0.01	-0.1	0.02
SK	-0.17	0.02	-0.14	0.02	-0.12	0.01	-0.1	0.01	-0.11	0.01
AB	-0.04	0.03	-0.02	0.02	-0.01	0.01	0.04	0.01	0.09	0.02
BC	0.03	0.03	0.05	0.04	0.05	0.02	0.01	0.01	0	0.02
Age										
30-34	0.11	0.02	0.1	0.02	0.11	0.01	0.11	0.01	0.1	0.02
35-39	0.13	0.02	0.14	0.02	0.17	0.02	0.19	0.02	0.17	0.02
40-44	0.15	0.03	0.19	0.02	0.22	0.01	0.23	0.01	0.22	0.02
45-49	0.19	0.04	0.23	0.03	0.29	0.02	0.31	0.02	0.31	0.02
50-54	0.22	0.05	0.28	0.04	0.32	0.03	0.34	0.03	0.33	0.03
CMA										
Montreal	-0.04	0.03	-0.01	0.02	0.02	0.02	0.04	0.01	0.06	0.02
Toronto	-0.08	0.02	-0.08	0.02	-0.05	0.01	0	0.01	0.04	0.02
Vancouver	-0.09	0.03	-0.04	0.03	-0.04	0.02	0.01	0.02	0.02	0.02
Married	0.1	0.04	0.09	0.03	0.05	0.04	0.05	0.03	0.04	0.04
Education										
< High School	-0.11	0.04	-0.1	0.05	-0.1	0.04	-0.11	0.04	-0.14	0.02
College	0.1	0.01	0.11	0.01	0.13	0.01	0.11	0.02	0.09	0.02
Bachelor's	0.18	0.02	0.29	0.02	0.35	0.01	0.35	0.01	0.32	0.02
Graduate	0.29	0.03	0.4	0.02	0.45	0.02	0.43	0.02	0.4	0.02
Yongest Age										
<3	0.08	0.03	0.09	0.03	0.08	0.02	0.08	0.02	0.08	0.02
3-5	0.09	0.03	0.07	0.03	0.07	0.02	0.09	0.02	0.09	0.02
6-12	0.11	0.03	0.11	0.03	0.07	0.03	0.06	0.03	0.07	0.02
13-15	0.08	0.04	0.08	0.04	0.05	0.03	0.05	0.03	0.05	0.03
16-17	0.16	0.04	0.12	0.05	0.05	0.04	0.08	0.04	0.07	0.04
18-24	0.11	0.04	0.08	0.05	0.05	0.04	0.04	0.04	0.06	0.03
Family size										
2	-0.02	0.02	-0.02	0.02	-0.03	0.02	-0.03	0.02	0	0.02
3	-0.07	0.02	-0.08	0.03	-0.08	0.03	-0.07	0.03	-0.05	0.03
4	-0.06	0.03	-0.06	0.02	-0.05	0.02	-0.07	0.02	-0.04	0.03
5+	-0.09	0.03	-0.1	0.02	-0.09	0.03	-0.08	0.03	-0.04	0.03

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A7: Female coefficients: 2016

Table A7

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.14	0.01	2.26	0.01	2.51	0.01	2.76	0.01	3.03	0.02
Province										
NL	-0.04	0.01	-0.1	0.02	-0.08	0.02	-0.07	0.01	-0.06	0.02
PE	-0.03	0.01	-0.08	0.02	-0.11	0.02	-0.16	0.01	-0.17	0.02
NS	-0.06	0.01	-0.12	0.02	-0.16	0.01	-0.16	0.01	-0.14	0.02
NB	-0.06	0.01	-0.09	0.02	-0.12	0.01	-0.14	0.01	-0.15	0.02
QC	0.02	0.01	0.01	0.01	-0.02	0.01	-0.04	0.01	-0.08	0.01
MB	-0.04	0.01	-0.07	0.01	-0.08	0.01	-0.06	0.01	-0.05	0.01
SK	0.03	0.01	0.06	0.01	0.05	0.01	0.05	0.01	0.05	0.01
AB	0.07	0.01	0.07	0.01	0.09	0.01	0.11	0.01	0.11	0.01
BC	0.01	0.01	0.01	0.01	-0.01	0.01	-0.05	0.01	-0.05	0.02
Age										
30-34	0.04	0.01	0.07	0.01	0.08	0.01	0.09	0.01	0.09	0.01
35-39	0.07	0.01	0.12	0.01	0.15	0.01	0.18	0.01	0.17	0.01
40-44	0.08	0.01	0.14	0.01	0.18	0.01	0.21	0.01	0.21	0.01
45-49	0.08	0.01	0.14	0.01	0.19	0.01	0.23	0.01	0.24	0.01
50-54	0.09	0.01	0.15	0.01	0.2	0.01	0.24	0.01	0.25	0.01
CMA										
Montreal	-0.04	0.02	-0.04	0.02	-0.03	0.02	-0.01	0.01	0.02	0.02
Toronto	-0.04	0.01	-0.06	0.02	-0.07	0.02	0	0.01	0.02	0.02
Vancouver	-0.03	0.02	-0.05	0.02	-0.03	0.02	0.01	0.02	0.04	0.02
Married	0.03	0.01	0.06	0.01	0.06	0.01	0.04	0.01	0.04	0.01
Education										
< High School	-0.07	0.01	-0.13	0.01	-0.22	0.01	-0.25	0.01	-0.26	0.02
College	0.1	0.01	0.19	0.01	0.18	0.01	0.17	0.01	0.18	0.01
Bachelor's	0.21	0.01	0.37	0.01	0.48	0.01	0.5	0.01	0.39	0.01
Graduate	0.31	0.02	0.52	0.02	0.65	0.01	0.61	0.01	0.51	0.01
Yongest Age										
<3	0.1	0.01	0.14	0.01	0.12	0.01	0.09	0.01	0.08	0.01
3-5	0.06	0.01	0.08	0.01	0.09	0.01	0.07	0.01	0.08	0.02
6-12	0.05	0.01	0.07	0.01	0.07	0.01	0.05	0.01	0.08	0.01
13-15	0.04	0.01	0.05	0.01	0.06	0.01	0.05	0.01	0.08	0.02
16-17	0.04	0.02	0.06	0.02	0.06	0.02	0.04	0.01	0.06	0.02
18-24	0.04	0.01	0.07	0.01	0.07	0.01	0.04	0.01	0.06	0.02
Family size										
2	-0.01	0.01	-0.02	0.01	-0.04	0.01	-0.04	0.01	-0.05	0.01
3	-0.05	0.01	-0.09	0.01	-0.1	0.01	-0.07	0.01	-0.1	0.02
4	-0.05	0.01	-0.07	0.01	-0.08	0.02	-0.06	0.01	-0.08	0.02
5+	-0.11	0.01	-0.16	0.02	-0.17	0.02	-0.12	0.01	-0.12	0.02

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A8: Selection corrected female coefficients: 2016

Table A8

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.15	0.02	2.26	0.05	2.51	0.07	2.76	0.08	3.02	0.08
Province										
NL	-0.04	0.02	-0.09	0.02	-0.08	0.03	-0.07	0.02	-0.04	0.03
PE	-0.03	0.02	-0.09	0.02	-0.11	0.02	-0.16	0.02	-0.15	0.03
NS	-0.06	0.01	-0.09	0.02	-0.14	0.02	-0.14	0.01	-0.12	0.02
NB	-0.05	0.01	-0.08	0.02	-0.12	0.02	-0.13	0.01	-0.14	0.02
QC	0.03	0.01	0.02	0.02	-0.02	0.02	-0.04	0.02	-0.08	0.02
MB	-0.03	0.01	-0.03	0.02	-0.07	0.02	-0.06	0.01	-0.04	0.02
SK	0.04	0.02	0.07	0.02	0.04	0.02	0.06	0.01	0.06	0.01
AB	0.07	0.02	0.09	0.02	0.09	0.02	0.12	0.01	0.1	0.02
BC	0.01	0.02	0.03	0.02	-0.02	0.01	-0.05	0.01	-0.06	0.02
Age										
30-34	0.05	0.01	0.08	0.02	0.09	0.02	0.11	0.01	0.1	0.02
35-39	0.07	0.02	0.11	0.02	0.16	0.02	0.2	0.01	0.18	0.02
40-44	0.06	0.02	0.12	0.02	0.18	0.02	0.22	0.02	0.22	0.02
45-49	0.06	0.02	0.14	0.02	0.2	0.02	0.26	0.02	0.26	0.02
50-54	0.08	0.02	0.14	0.03	0.2	0.03	0.26	0.03	0.28	0.02
CMA										
Montreal	-0.04	0.02	-0.04	0.03	-0.02	0.02	0	0.02	0.02	0.03
Toronto	-0.03	0.02	-0.04	0.02	-0.06	0.04	0.01	0.02	0.03	0.03
Vancouver	-0.03	0.02	-0.05	0.03	-0.03	0.03	0.01	0.02	0.04	0.03
Married	0.02	0.01	0.04	0.01	0.05	0.01	0.04	0.01	0.04	0.02
Education										
< High School	-0.07	0.02	-0.12	0.02	-0.21	0.04	-0.24	0.07	-0.23	0.08
College	0.1	0.01	0.19	0.02	0.18	0.04	0.16	0.04	0.15	0.04
Bachelor's	0.2	0.02	0.37	0.02	0.46	0.04	0.47	0.06	0.36	0.06
Graduate	0.28	0.04	0.48	0.02	0.61	0.04	0.58	0.06	0.49	0.05
Yongest Age										
<3	0.1	0.04	0.15	0.04	0.14	0.03	0.09	0.03	0.1	0.03
3-5	0.05	0.03	0.07	0.03	0.09	0.03	0.08	0.02	0.08	0.02
6-12	0.05	0.02	0.07	0.02	0.08	0.02	0.06	0.02	0.1	0.02
13-15	0.02	0.02	0.05	0.02	0.02	0.02	0.01	0.02	0.06	0.02
16-17	0.02	0.03	0.05	0.03	0.02	0.03	0.04	0.02	0.06	0.03
18-24	0.02	0.02	0.06	0.02	0.06	0.03	0.03	0.02	0.05	0.02
Family size										
2	0	0.02	-0.01	0.02	-0.03	0.02	-0.03	0.02	-0.03	0.02
3	-0.06	0.02	-0.09	0.02	-0.09	0.03	-0.07	0.02	-0.1	0.02
4	-0.03	0.02	-0.05	0.02	-0.06	0.03	-0.06	0.02	-0.07	0.03
5+	-0.11	0.02	-0.18	0.03	-0.18	0.05	-0.15	0.04	-0.13	0.04

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A9: Male coefficients: 2016

Table A9

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.23	0.01	2.44	0.01	2.69	0.01	2.98	0.01	3.25	0.01
Province										
NL	-0.06	0.02	-0.05	0.02	-0.03	0.02	-0.02	0.02	0	0.02
PE	-0.15	0.02	-0.22	0.02	-0.27	0.01	-0.29	0.02	-0.23	0.03
NS	-0.14	0.02	-0.18	0.01	-0.18	0.02	-0.16	0.01	-0.15	0.02
NB	-0.09	0.02	-0.16	0.01	-0.21	0.01	-0.21	0.01	-0.17	0.02
QC	0.02	0.01	-0.03	0.01	-0.06	0.01	-0.07	0.01	-0.07	0.01
MB	-0.08	0.01	-0.11	0.01	-0.11	0.01	-0.07	0.01	-0.06	0.01
SK	0.07	0.02	0.08	0.01	0.1	0.01	0.08	0.01	0.06	0.01
AB	0.12	0.02	0.16	0.01	0.2	0.01	0.21	0.01	0.22	0.01
BC	0.05	0.02	0.08	0.01	0.09	0.01	0.05	0.01	0.03	0.01
Age										
30-34	0.06	0.01	0.07	0.01	0.08	0.01	0.08	0.01	0.05	0.01
35-39	0.07	0.01	0.1	0.01	0.12	0.01	0.12	0.01	0.1	0.01
40-44	0.09	0.01	0.13	0.01	0.15	0.01	0.16	0.01	0.15	0.01
45-49	0.11	0.01	0.16	0.01	0.17	0.01	0.19	0.01	0.18	0.01
50-54	0.13	0.01	0.17	0.01	0.18	0.01	0.2	0.01	0.2	0.01
CMA										
Montreal	-0.08	0.02	-0.06	0.02	-0.01	0.02	0	0.01	0	0.02
Toronto	-0.04	0.02	-0.07	0.02	-0.02	0.01	0.02	0.02	0.05	0.02
Vancouver	-0.04	0.03	-0.09	0.02	-0.1	0.02	-0.07	0.02	-0.05	0.02
Married	0.06	0.01	0.1	0.01	0.11	0.01	0.1	0.01	0.08	0.01
Education										
< High School	-0.07	0.01	-0.08	0.01	-0.09	0.01	-0.13	0.01	-0.13	0.01
College	0.1	0.01	0.14	0.01	0.18	0.01	0.17	0.01	0.14	0.01
Bachelor's	0.14	0.01	0.24	0.01	0.35	0.01	0.34	0.01	0.3	0.01
Graduate	0.19	0.02	0.37	0.02	0.47	0.01	0.44	0.01	0.4	0.01
Yongest Age										
<3	0.15	0.02	0.16	0.01	0.16	0.01	0.12	0.01	0.1	0.01
3-5	0.16	0.02	0.2	0.02	0.18	0.01	0.15	0.01	0.13	0.02
6-12	0.14	0.02	0.19	0.01	0.18	0.01	0.15	0.01	0.13	0.01
13-15	0.16	0.02	0.18	0.02	0.17	0.02	0.15	0.01	0.11	0.02
16-17	0.15	0.03	0.15	0.02	0.16	0.02	0.14	0.02	0.13	0.02
18-24	0.14	0.02	0.15	0.02	0.16	0.02	0.12	0.02	0.1	0.02
Family size										
2	0.04	0.01	-0.01	0.01	-0.03	0.01	-0.04	0.01	-0.05	0.01
3	-0.06	0.01	-0.11	0.01	-0.14	0.01	-0.12	0.01	-0.11	0.02
4	-0.04	0.01	-0.08	0.01	-0.11	0.01	-0.11	0.01	-0.1	0.02
5+	-0.09	0.02	-0.16	0.01	-0.18	0.01	-0.15	0.02	-0.12	0.02

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

A10: Selection corrected male coefficients: 2016

Table A10

	0.10		0.25		0.5		0.75		0.90	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept	2.28	0.04	2.51	0.05	2.79	0.07	3.1	0.08	3.34	0.05
Province										
NL	-0.05	0.03	-0.03	0.03	0	0.03	0.04	0.03	0.01	0.02
PE	-0.15	0.02	-0.24	0.02	-0.27	0.02	-0.28	0.03	-0.23	0.03
NS	-0.14	0.02	-0.18	0.02	-0.16	0.03	-0.16	0.02	-0.14	0.02
NB	-0.11	0.02	-0.18	0.02	-0.2	0.02	-0.21	0.02	-0.15	0.03
QC	0.01	0.02	-0.06	0.02	-0.08	0.01	-0.1	0.01	-0.11	0.02
MB	-0.08	0.02	-0.12	0.02	-0.11	0.01	-0.09	0.02	-0.1	0.02
SK	0.07	0.02	0.07	0.02	0.08	0.02	0.06	0.02	0.03	0.02
AB	0.11	0.02	0.14	0.02	0.18	0.01	0.19	0.01	0.18	0.02
BC	0.05	0.02	0.08	0.02	0.08	0.01	0.03	0.02	0.01	0.02
Age										
30-34	0.07	0.02	0.1	0.02	0.1	0.01	0.08	0.02	0.07	0.02
35-39	0.09	0.02	0.12	0.02	0.15	0.02	0.15	0.02	0.14	0.02
40-44	0.09	0.02	0.15	0.02	0.18	0.02	0.18	0.02	0.18	0.02
45-49	0.12	0.03	0.18	0.02	0.19	0.02	0.23	0.02	0.23	0.02
50-54	0.13	0.03	0.2	0.03	0.22	0.02	0.24	0.02	0.26	0.03
CMA										
Montreal	-0.07	0.03	-0.04	0.02	0.01	0.02	0.03	0.02	0.02	0.02
Toronto	-0.06	0.02	-0.08	0.02	-0.03	0.02	0.03	0.02	0.03	0.02
Vancouver	0	0.03	-0.08	0.03	-0.08	0.02	-0.05	0.02	-0.04	0.02
Married	0.03	0.02	0.03	0.04	0.03	0.04	0.02	0.04	0.02	0.04
Education										
< High School	-0.04	0.03	-0.04	0.04	-0.07	0.04	-0.03	0.06	-0.02	0.04
College	0.1	0.01	0.13	0.01	0.15	0.02	0.13	0.03	0.11	0.02
Bachelor's	0.16	0.02	0.26	0.02	0.33	0.02	0.29	0.03	0.3	0.02
Graduate	0.19	0.03	0.39	0.03	0.45	0.02	0.41	0.03	0.38	0.03
Yongest Age										
<3	0.14	0.02	0.15	0.02	0.16	0.02	0.14	0.02	0.1	0.03
3-5	0.13	0.03	0.16	0.03	0.16	0.03	0.14	0.03	0.09	0.03
6-12	0.11	0.03	0.15	0.03	0.17	0.03	0.13	0.03	0.09	0.03
13-15	0.12	0.04	0.15	0.03	0.15	0.04	0.13	0.03	0.1	0.04
16-17	0.15	0.04	0.13	0.04	0.18	0.03	0.13	0.04	0.09	0.04
18-24	0.13	0.04	0.13	0.03	0.15	0.03	0.08	0.04	0.05	0.04
Family size										
2	0.03	0.02	0.02	0.02	0	0.02	-0.02	0.02	-0.03	0.03
3	-0.06	0.02	-0.09	0.03	-0.12	0.03	-0.1	0.03	-0.06	0.04
4	-0.05	0.02	-0.06	0.02	-0.08	0.03	-0.1	0.03	-0.08	0.03
5+	-0.1	0.02	-0.14	0.03	-0.14	0.03	-0.11	0.03	-0.07	0.04

Notes: Results from the selection-corrected quantile regression at various quantiles. The left column denotes coefficients while the right column denotes standard error.

Appendix: Chapter 2

List of Variables

Table B1: List of Variables

Variable	Type	Omitted Group
Log hourly wage	continuous	
Year	binary	
Female	binary	
Province (smaller provinces lumped together)	5 categories	Ontario
Marital Status	4 categories	Single, never married
Weekly work hours (>25)	5 categories	[25 to 30)
CMA (Census metropolitan areas)	4 categories	CMA
Firm size (Number of employees, all locations)	4 categories	More than 500
Public sector	binary	Private sector
Union	binary	Non-union
Age [25,55)	6 categories	25 to 29
Age of Youngest Child (<24)	7 categories	No kids
Education	6 categories	High School
Industry	18 categories	Wholesale trade
Occupation	25 categories	Sales and Service

ML mean squared errors

All ML methods performed similarly. However, the best average fit, measured in terms of the mean squared errors of the target variables, was produced with the boosted tree method in both years and in all education breakdowns. The boosted trees preformed slightly better than our baseline LASSO estimator. This performance could potentially be improved had we cross-validated the the boosted tree parameters for fit. Given that the DML were so consistent across specifications, we did not pursue this.

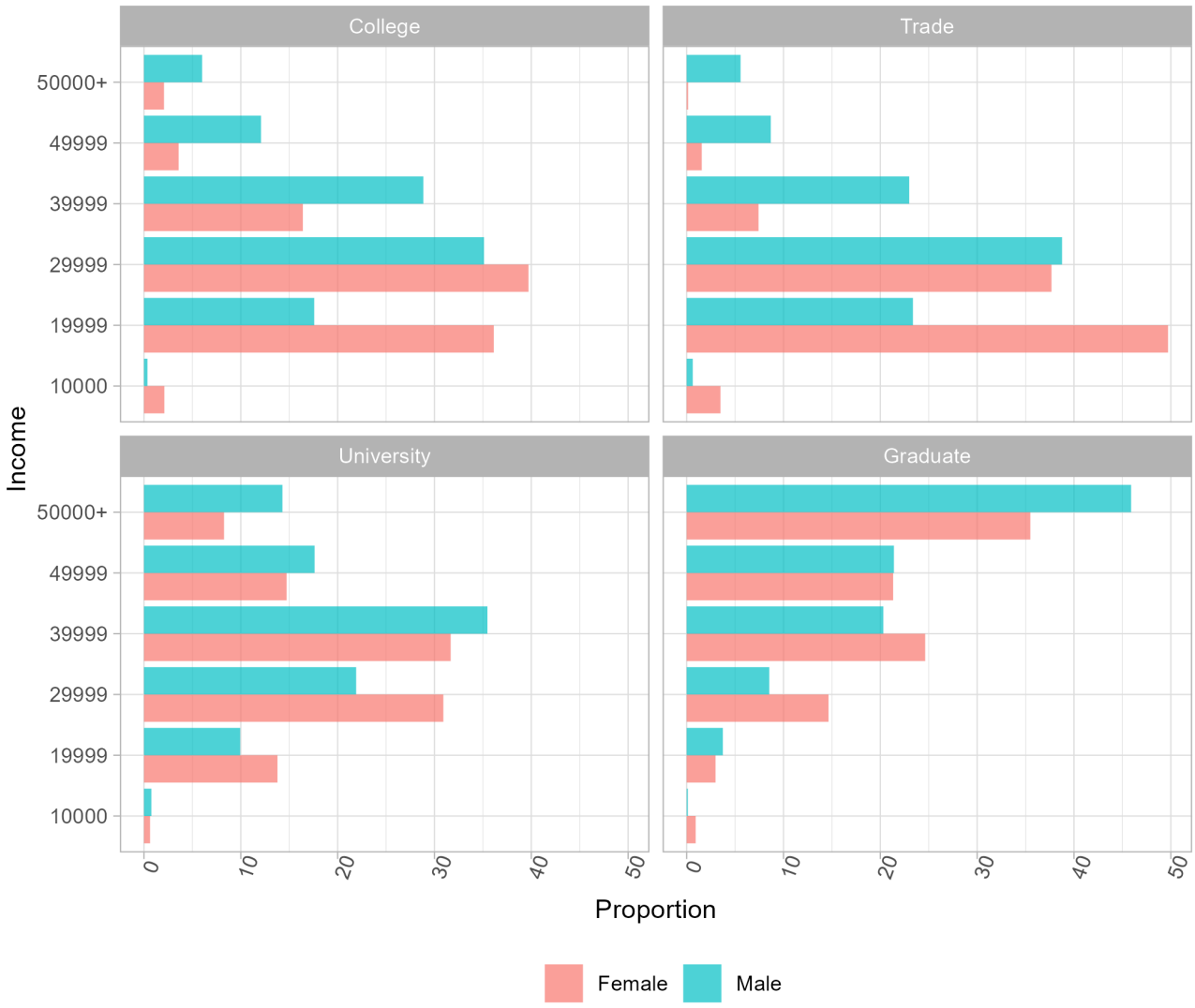
Table B2: ML mean squared errors, 1999

	All Workers					College Graduates		High school or less	
Variable	lasso	lgb	RF	lasso-1se	ridge	lasso	lgb	lasso	lgb
1999									
lnwage	0.099	0.098	0.105	0.100	0.100	0.116	0.112	0.097	0.096
Average	0.019	0.018	0.019	0.019	0.019	0.028	0.028	0.019	0.018
2015									
lnwage	0.107	0.106	0.113	0.108	0.108	0.127	0.125	0.103	0.100
Average	0.019	0.019	0.020	0.020	0.020	0.026	0.027	0.018	0.018

Note:
Average and log wage mean squared errors ML estimation methods, for All Workers, College Graduates and High school or less subgroups.

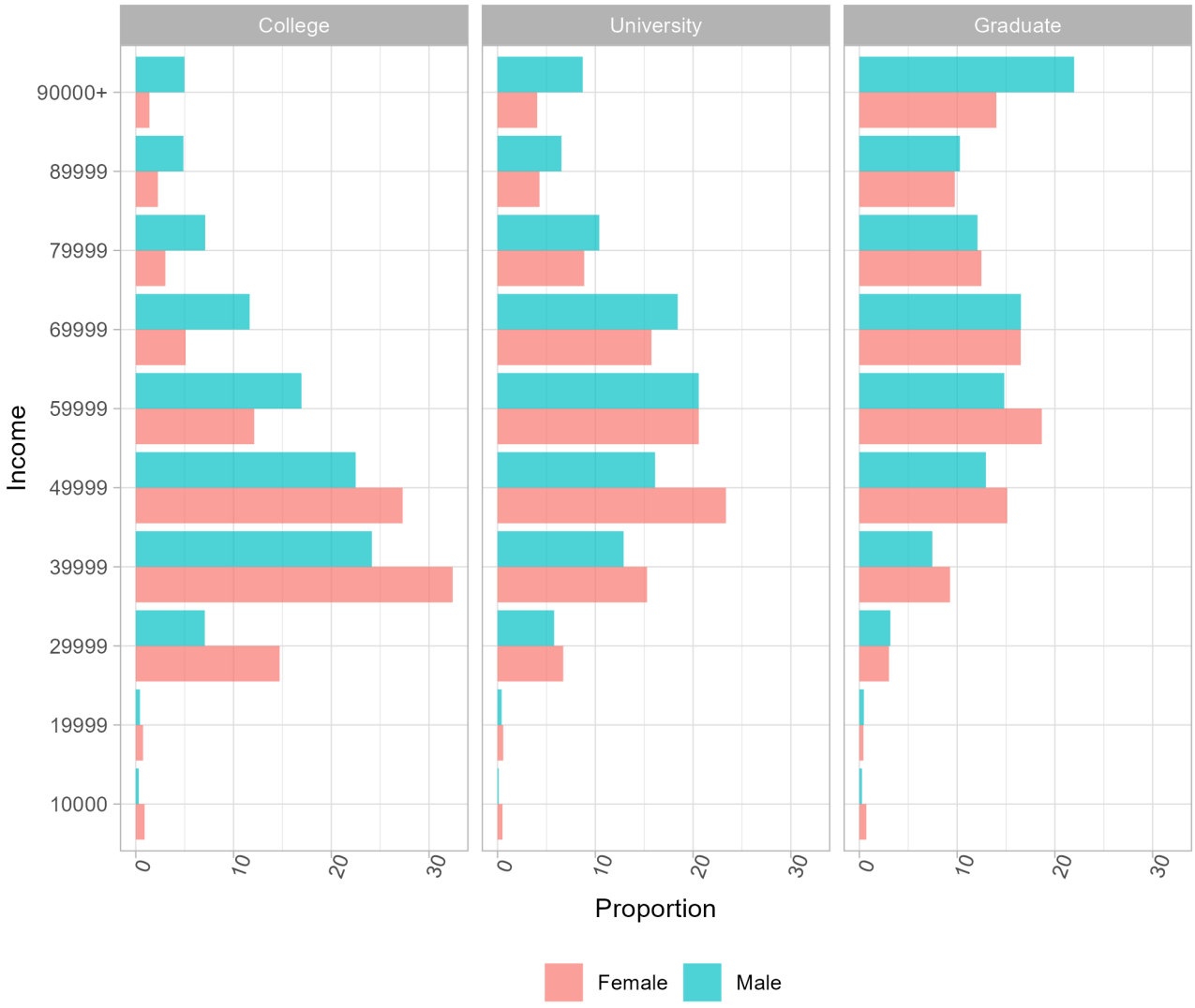
Appendix: Chapter 3

Figure C1: Wage distribution by education: 1998



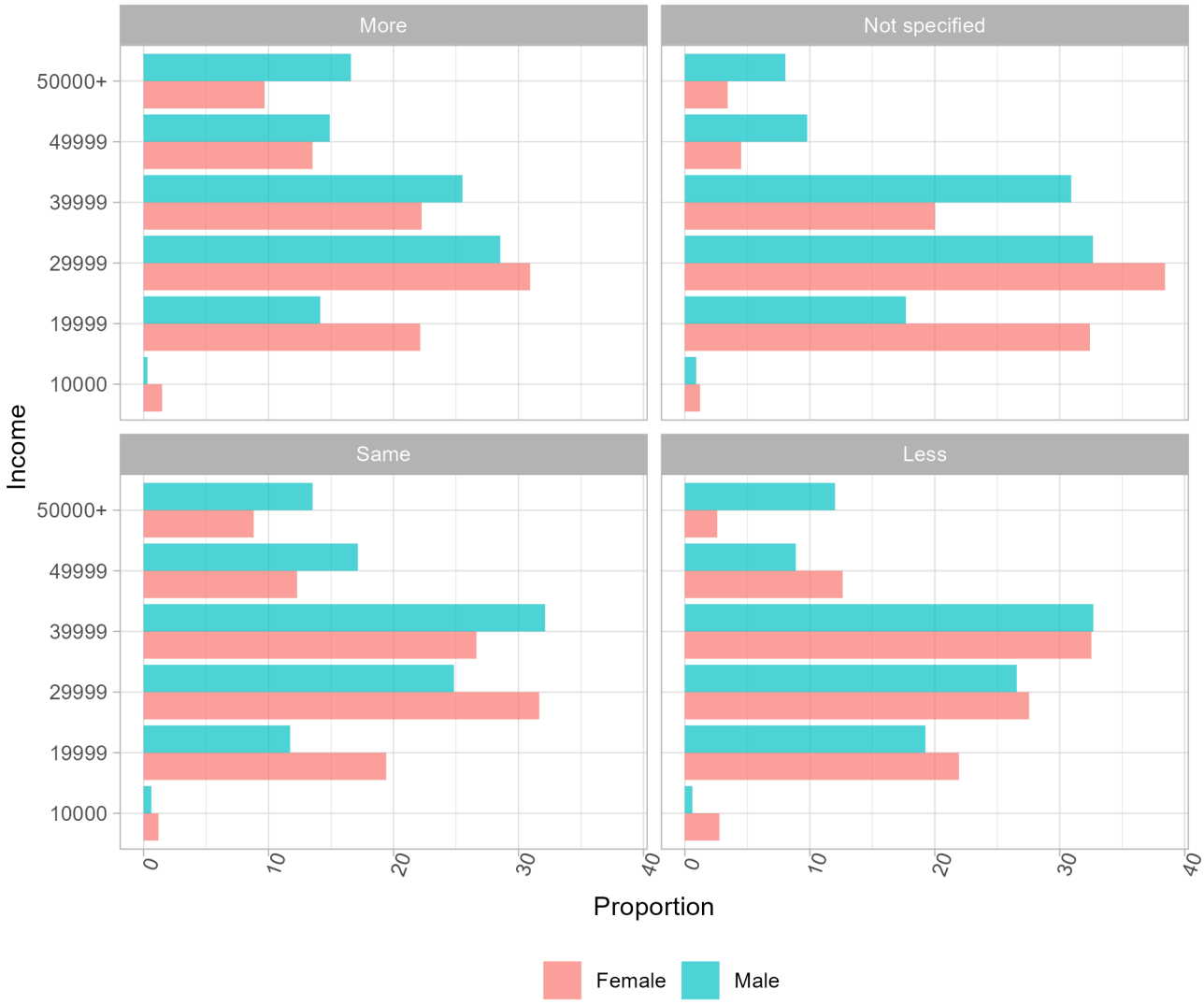
Notes: This figure shows the distribution of wages by different education levels. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C2: Wage distribution by education: 2018



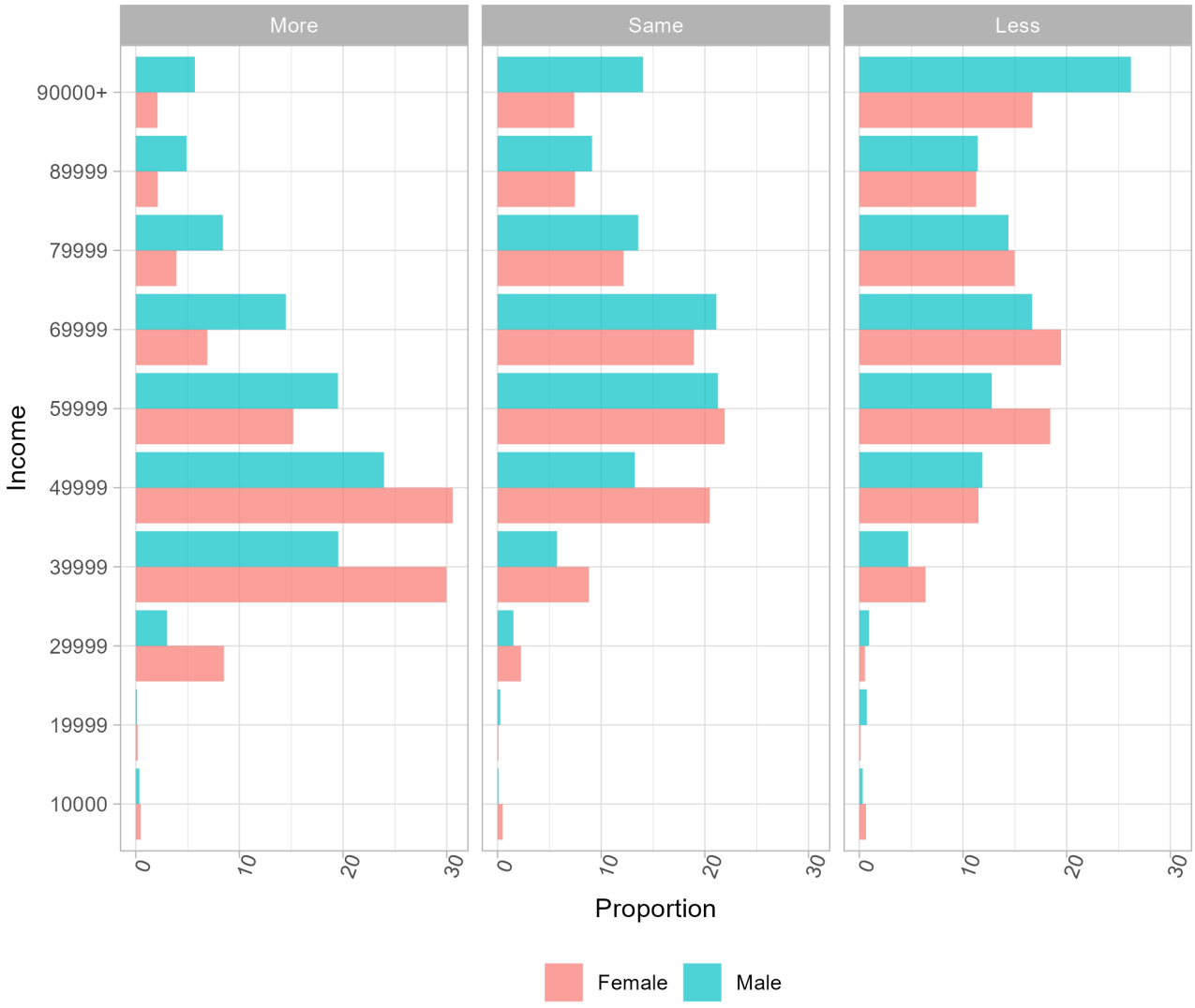
Notes: This figure shows the distribution of wages by different education levels. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C3: Wage distribution by education requirement: 1998



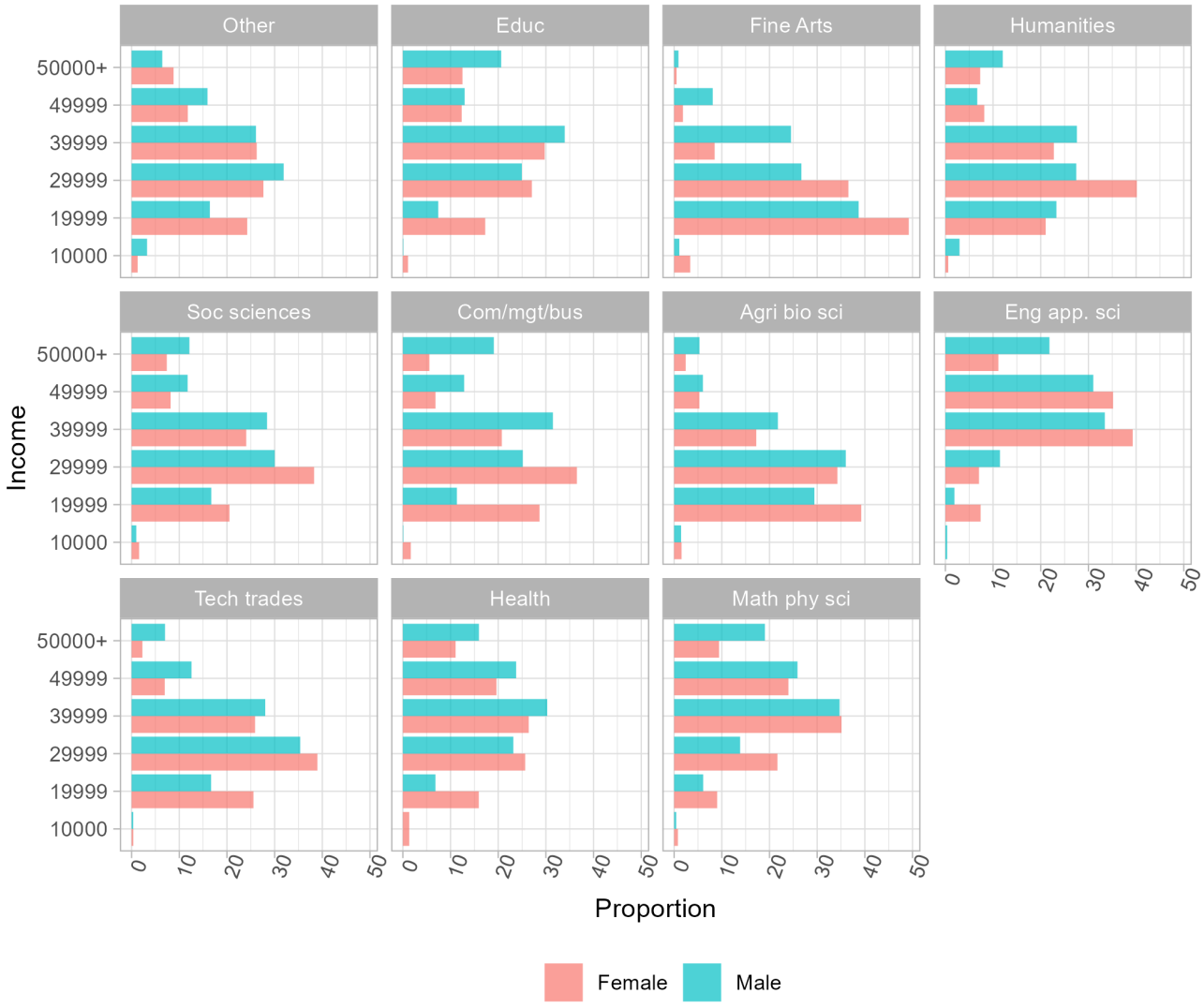
Notes: This figure shows the distribution of wages by varying educational requirements of a worker's job relative to their own education. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C4: Wage distribution by education requirement: 2018



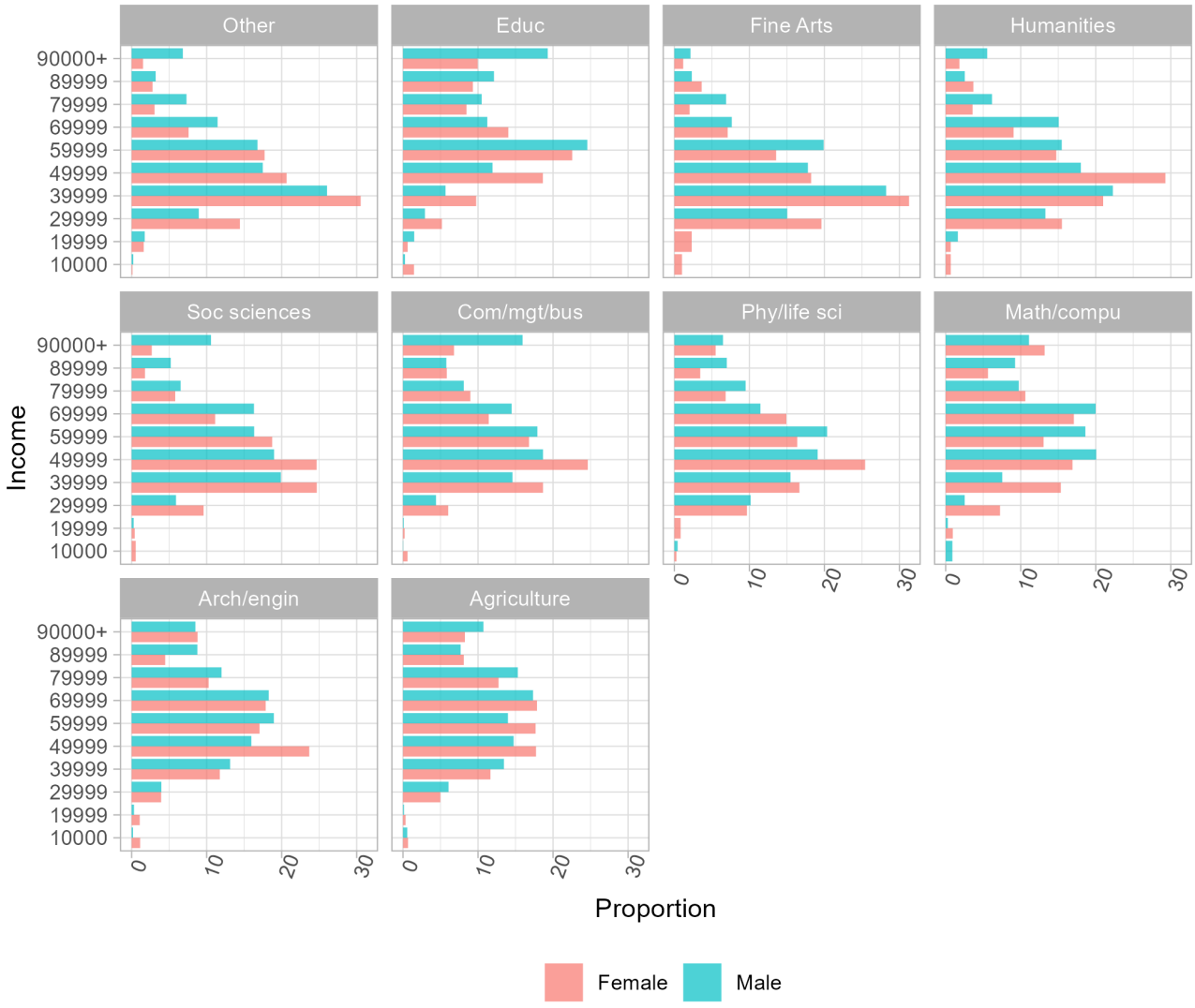
Notes: This figure shows the distribution of wages by varying educational requirements of a worker's job relative to their own education. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C5: Wage distribution by program: 1998



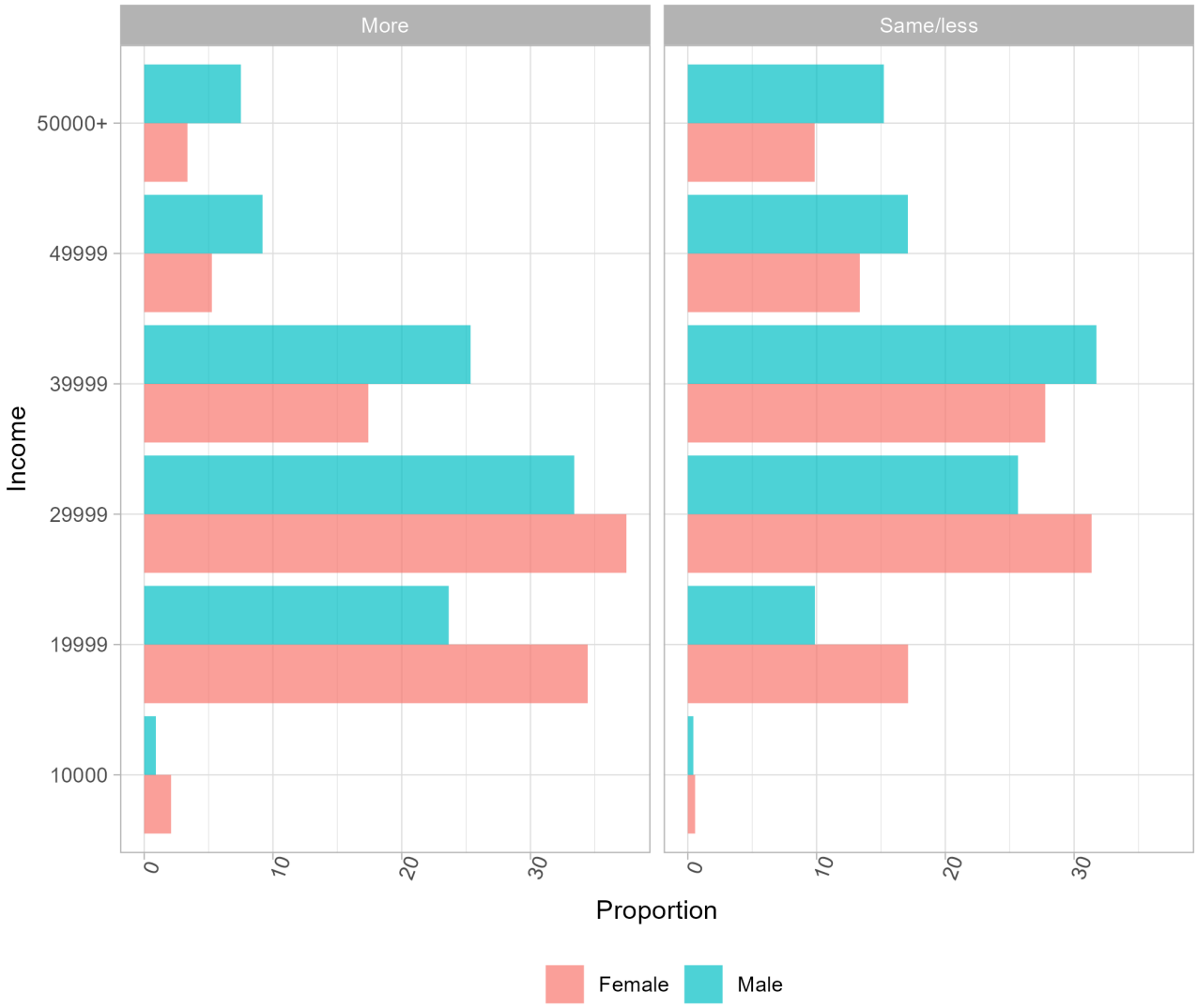
Notes: This figure shows the distribution of wages by program of study. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C6: Wage distribution by program: 2018



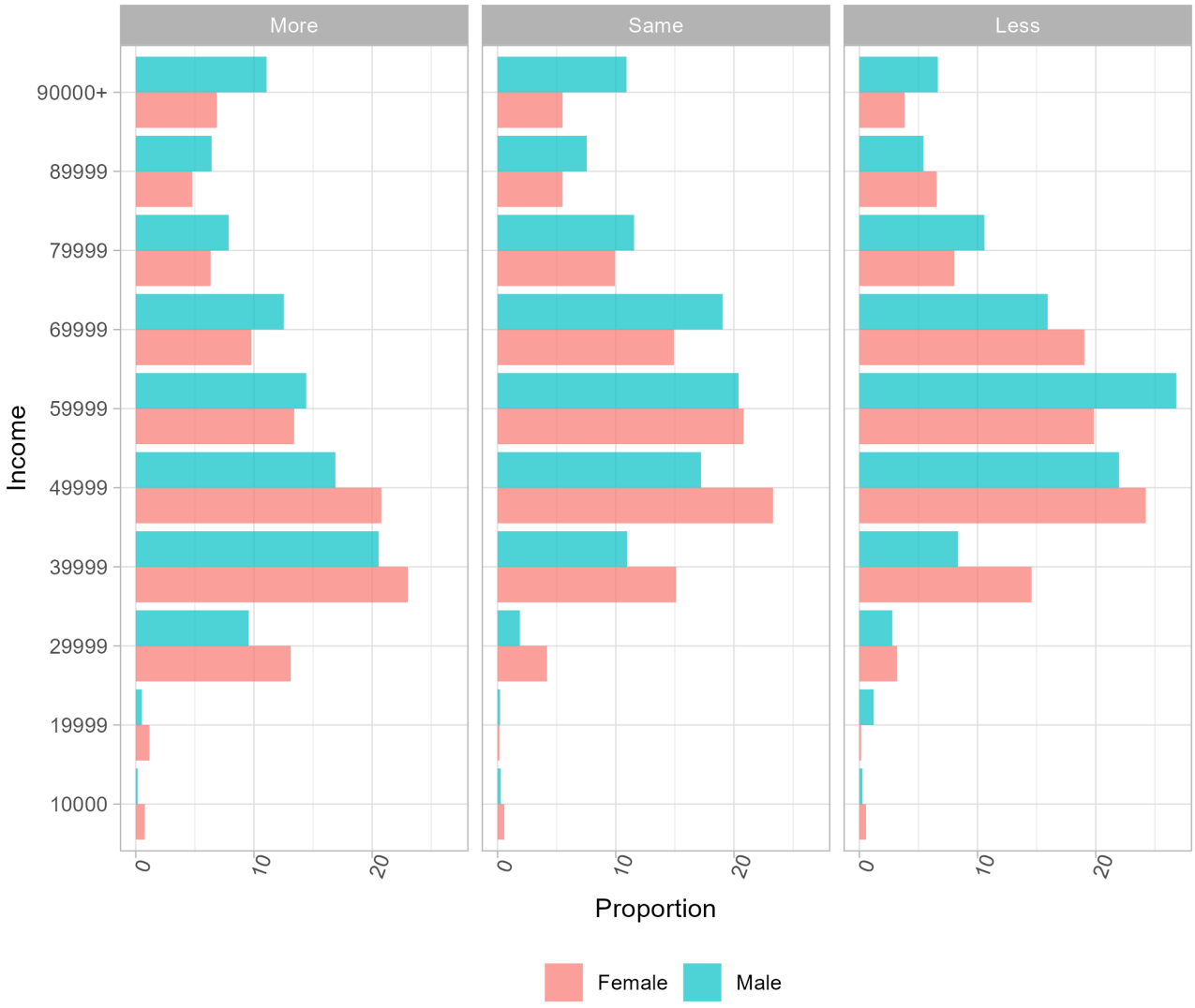
Notes: This figure shows the distribution of wages by program of study. Data used in this figure comes from the National Graduate Survey (NGS). The y-axis denote income brackets: '<10k', '10-20k', etc

Figure C7: Wage distribution by qualification: 1998



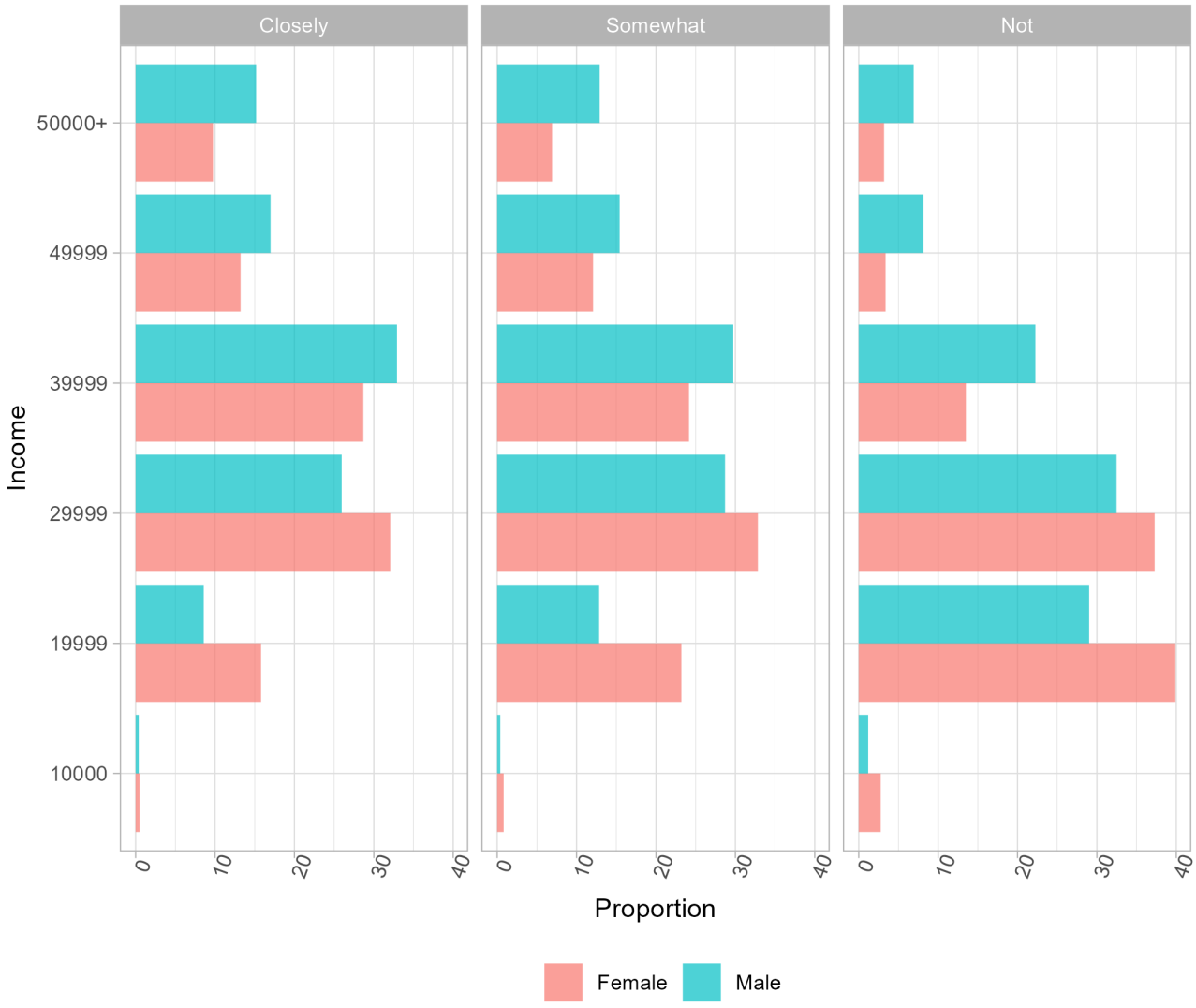
Notes: This figure shows the distribution of wages by workers' feeling of qualification. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C8: Wage distribution by qualification: 2018



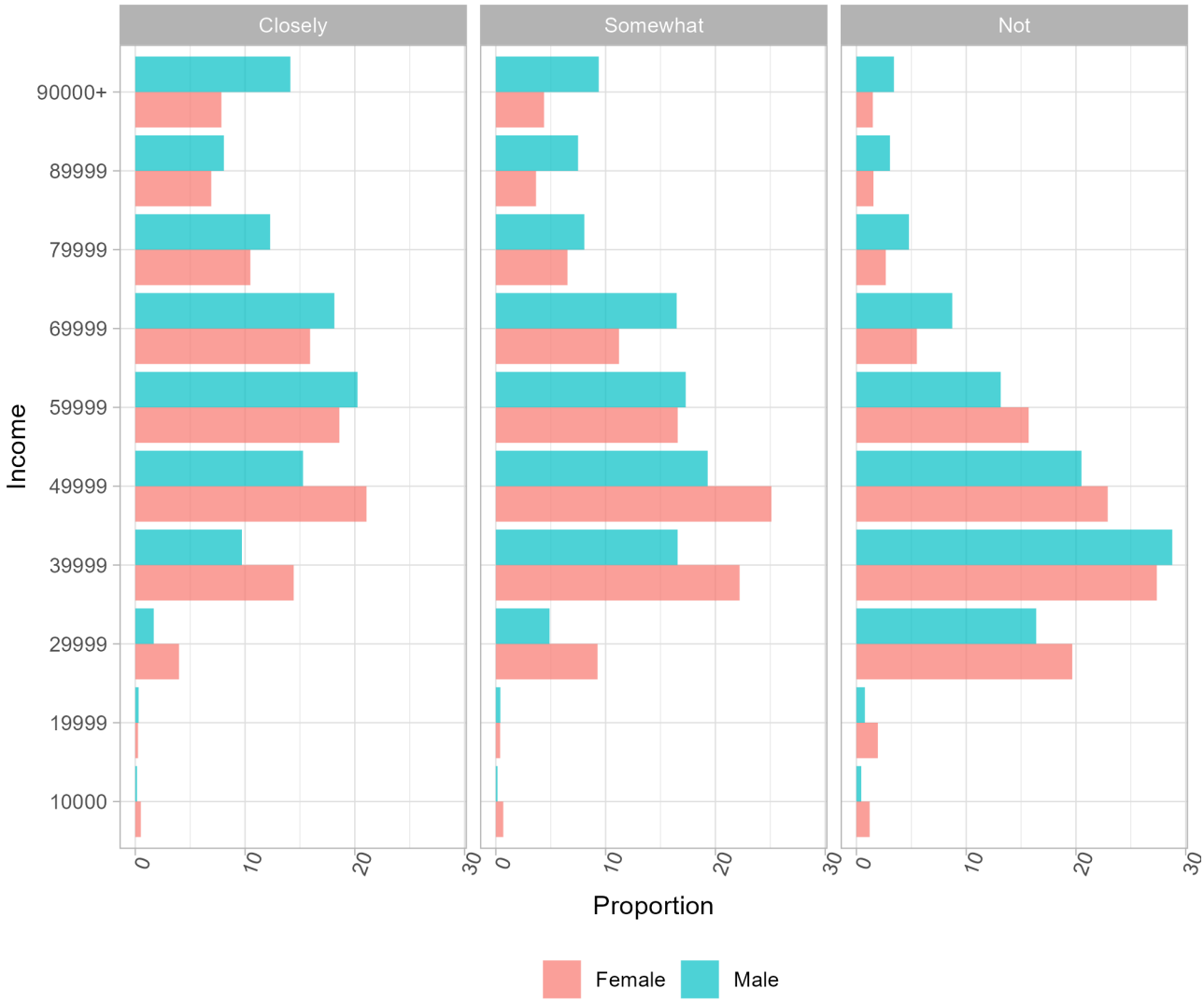
Notes: This figure shows the distribution of wages by workers' feeling of qualification. Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C9: Wage distribution by job relatedness: 1998



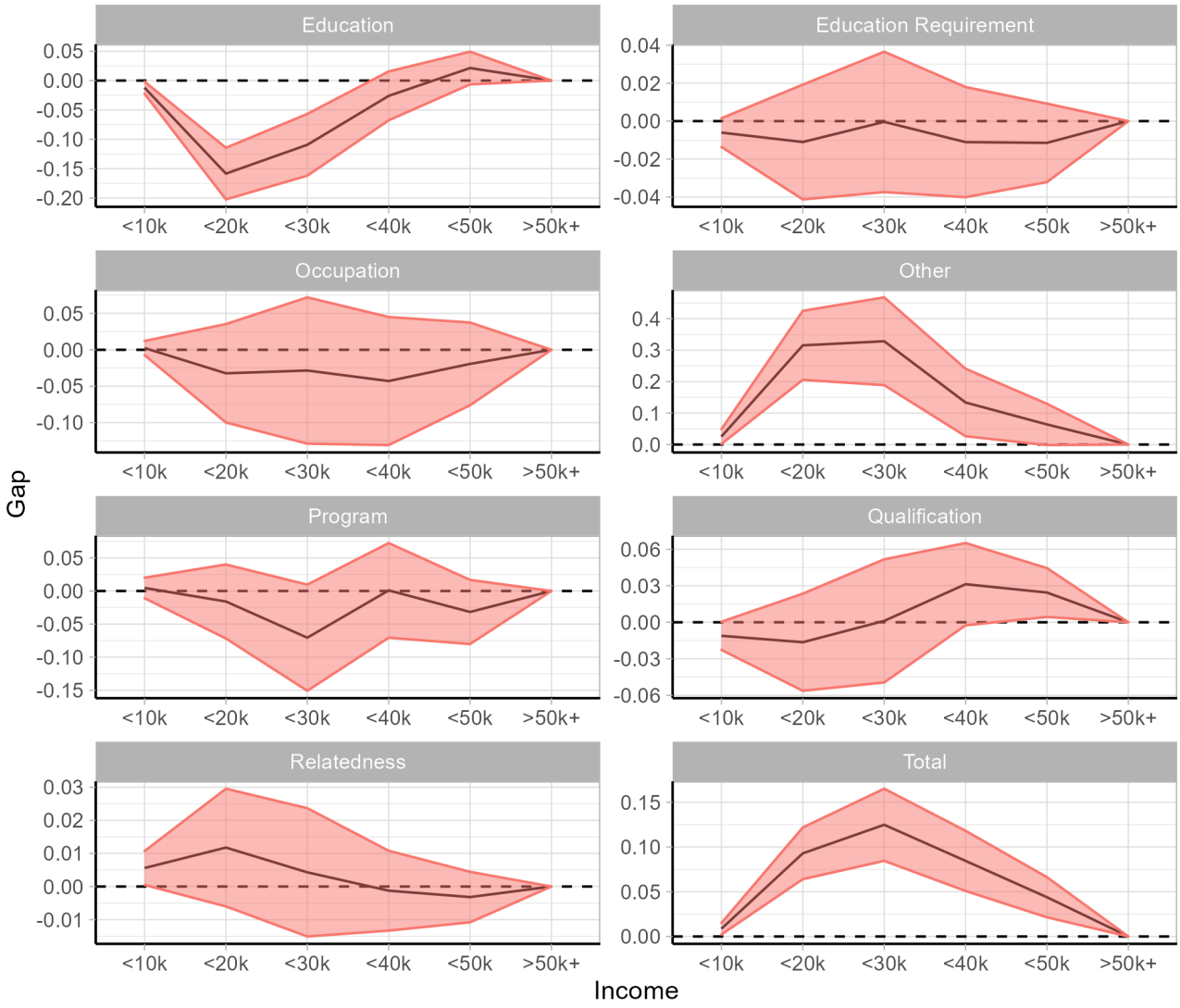
Notes: This figure shows the distribution of wages by workers' feeling of job relevance in relation to their education and training Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.) are represented on the vertical y-axis.

Figure C10: Wage distribution by job relatedness: 2018



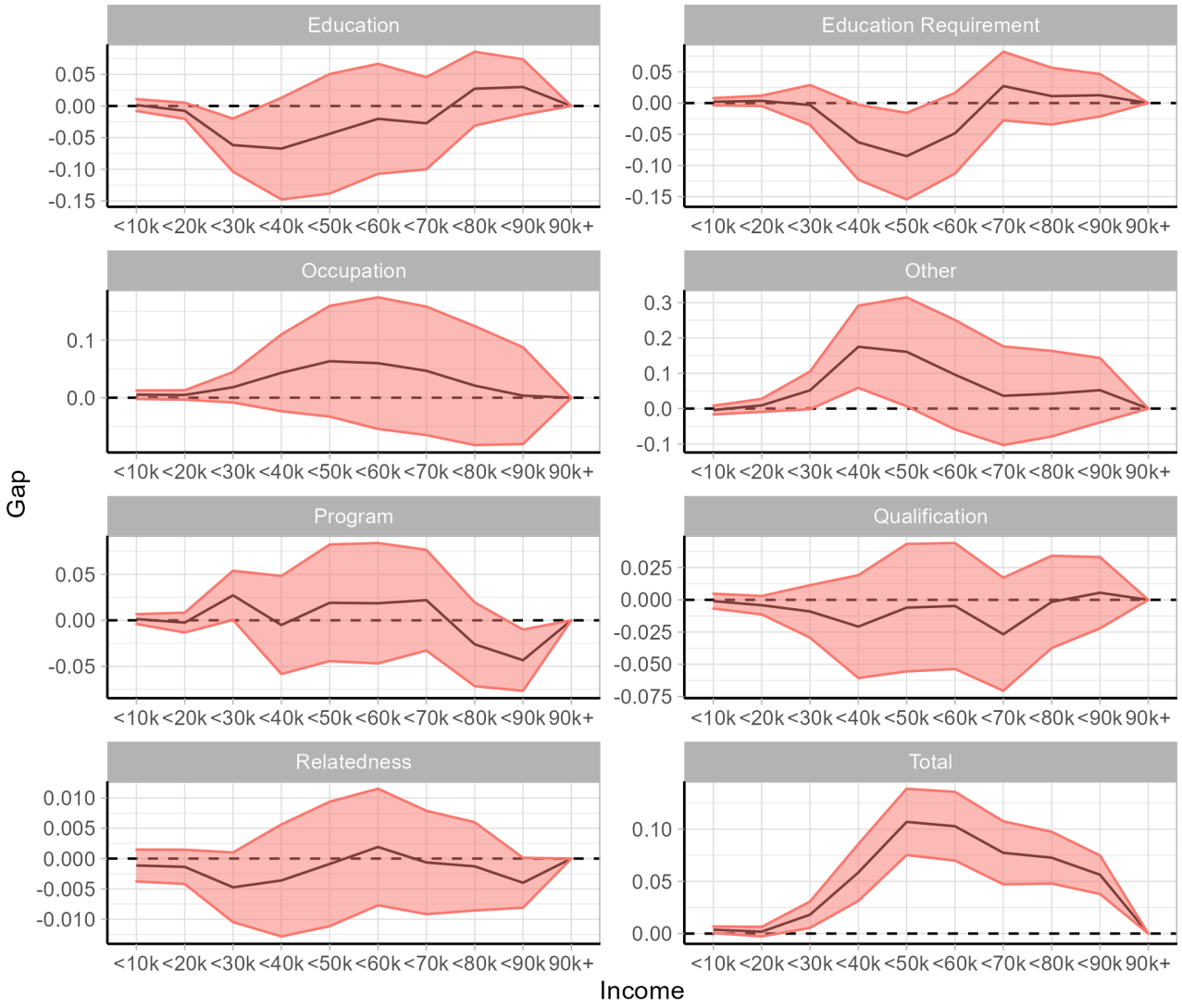
Notes: This figure shows the distribution of wages by workers' feeling of job relevance in relation to their education and training Data used in this figure comes from the National Graduate Survey (NGS). Income brackets ('<10k', '10-20k', etc.

Figure C11: Unexplained wage gap by income quantiles, reversed order: NGS 1998



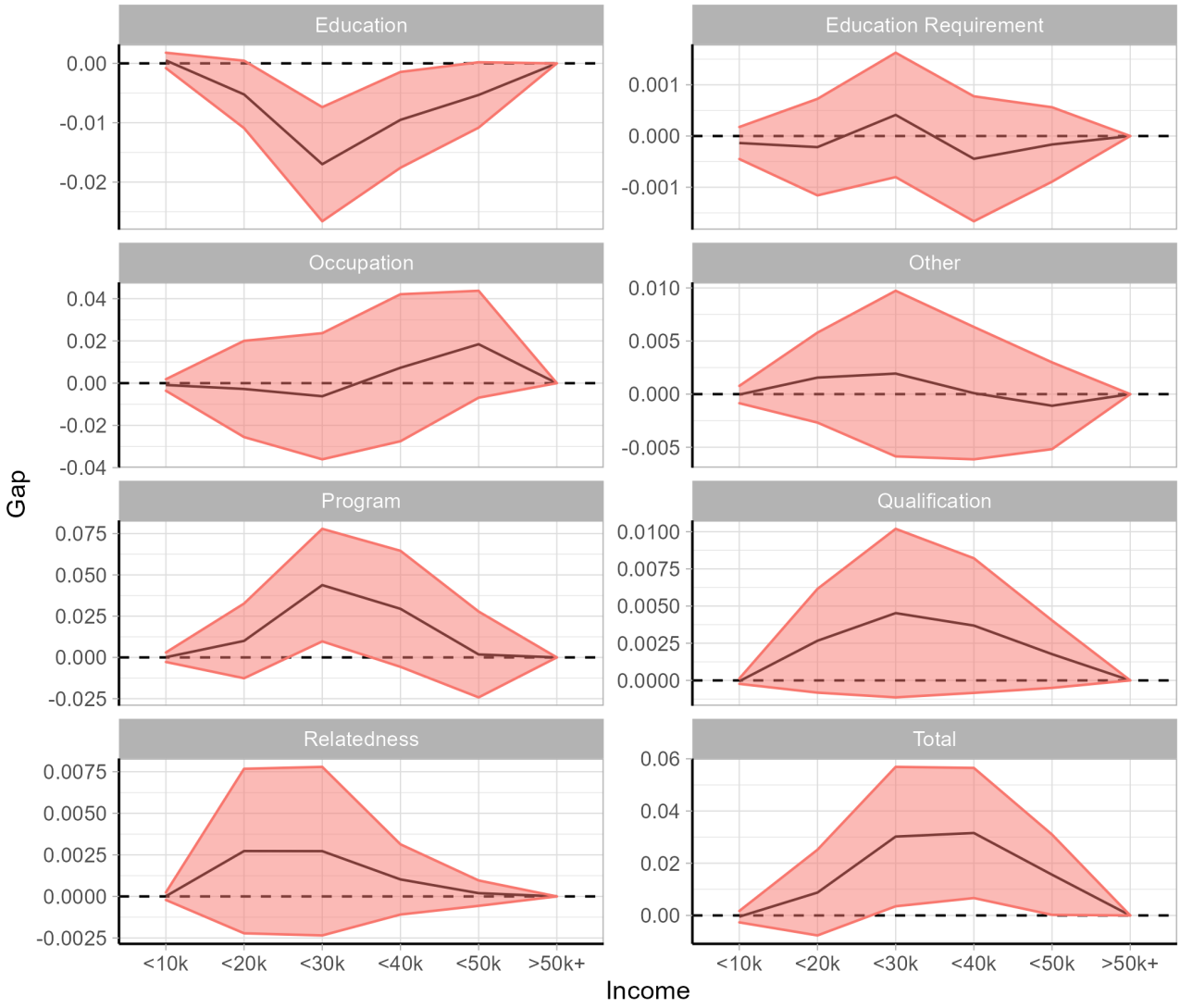
Notes: This figure shows the unexplained wage gap calculated using National Graduate Survey (NGS) data. Order of decomposition has been reversed compared to the main results. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure C12: Unexplained wage gap by income quantiles, reversed order: NGS 2018



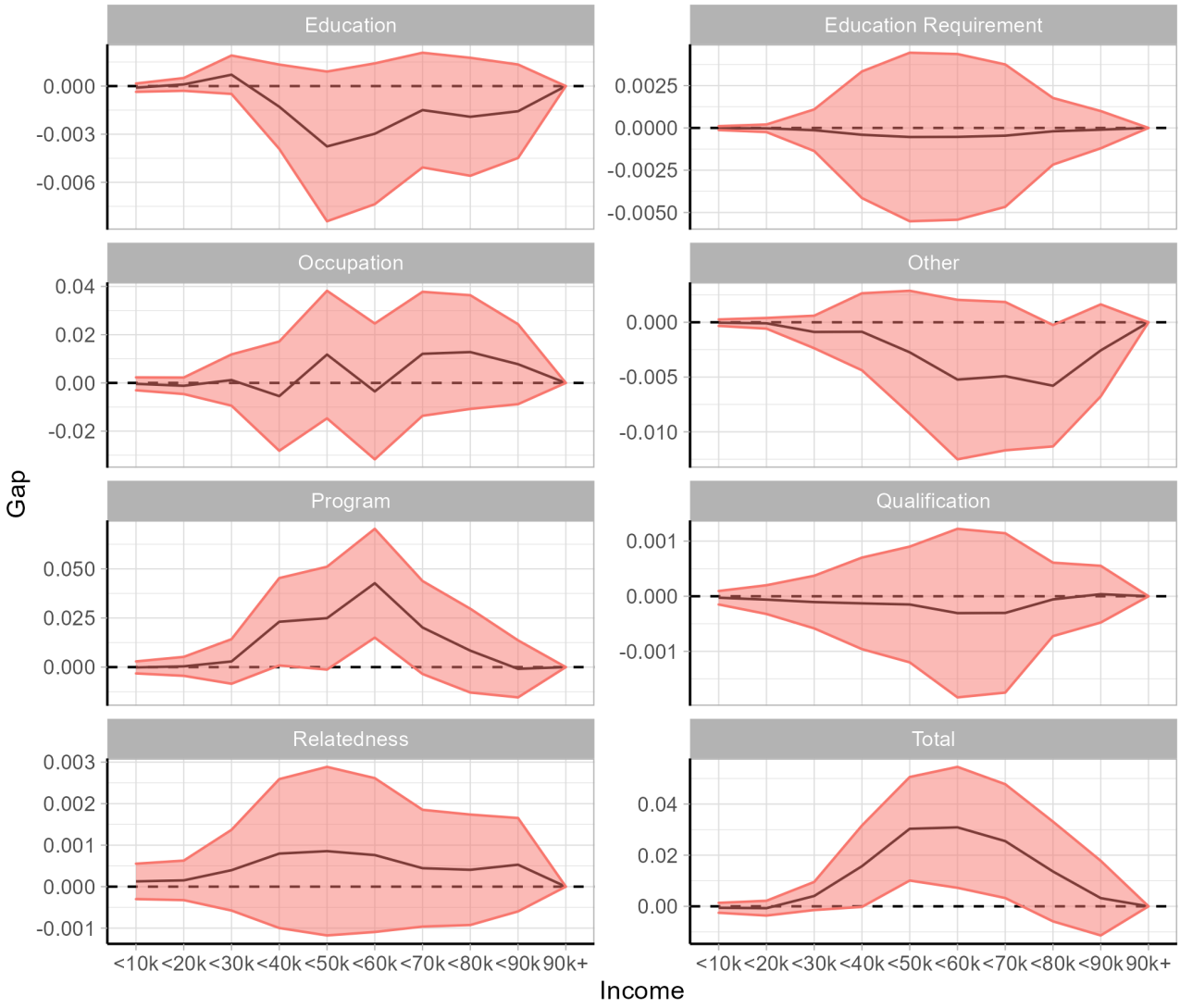
Notes: This figure shows the unexplained wage gap calculated using National Graduate Survey (NGS) data. Order of decomposition has been reversed compared to the main results. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total unexplained gap.

Figure C13: Explained wage gap by income quantiles, reversed order: NGS 1998



Notes: This figure shows the explained wage gap calculated using National Graduate Survey (NGS) data. Order of decomposition has been reversed compared to the main results. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total Explained gap.

Figure C14: Explained wage gap by income quantiles, reversed order: NGS 2018



Notes: This figure shows the explained wage gap calculated using National Graduate Survey (NGS) data. Order of decomposition has been reversed compared to the main results. The x-axis denotes yearly income while the y-axis denotes differences in the CDF between men and women. The gap is interpreted as the difference in probability of earning less than the income threshold on the x-axis. The shaded area is the 95% confidence interval, calculated using bootstrap. Each panel denotes the contribution of key characteristics to the total Explained gap.

Figure C15: Unexplained counterfactual

$$
\begin{aligned}
F_{Y(1,0,0,0,0,0|0)} &= \int F_{Y(1,0,0,0,0,0|X_0}(y|x)dF_{X_0}x \\
F_{Y(1,1,0,0,0,0|0)} &= \int F_{Y(1,1,0,0,0,0|X_0}(y|x)dF_{X_0}x \\
F_{Y(1,1,1,0,0,0|0)} &= \int F_{Y(1,1,1,0,0,0|X_0}(y|x)dF_{X_0}x \\
F_{Y(1,1,1,1,0,0|0)} &= \int F_{Y(1,1,1,1,0,0|X_0}(y|x)dF_{X_0}x \\
F_{Y(1,1,1,1,1,0|0)} &= \int F_{Y(1,1,1,1,1,0|X_0}(y|x)dF_{X_0}x \\
F_{Y(1,1,1,1,1,1|0)} &= \int F_{Y(1,1,1,1,1,1|X_0}(y|x)dF_{X_0}x
\end{aligned} \tag{7.1}
$$

Figure C16: Explained counterfactual

$$
\begin{aligned}
F_{Y(1|1,0,0,0,0,0)} &= \int F_{Y(1|X_{(1,0,0,0,0,0)})}(y|x)dF_{X_{(1,0,0,0,0,0)}}x \\
F_{Y(1|1,1,0,0,0,0)} &= \int F_{Y(1|X_{(1,1,0,0,0,0)})}(y|x)dF_{X_{(1,1,0,0,0,0)}}x \\
F_{Y(1|1,1,1,0,0,0)} &= \int F_{Y(1|X_{(1,1,1,0,0,0)})}(y|x)dF_{X_{(1,1,1,0,0,0)}}x \\
F_{Y(1|1,1,1,1,0,0)} &= \int F_{Y(1|X_{(1,1,1,1,0,0)})}(y|x)dF_{X_{(1,1,1,1,0,0)}}x \\
F_{Y(1|1,1,1,1,1,0)} &= \int F_{Y(1|X_{(1,1,1,1,1,0)})}(y|x)dF_{X_{(1,1,1,1,1,0)}}x
\end{aligned} \tag{7.2}
$$