Kigawa-Blinder-Oaxaca Decomposition: The Dutch Gender Wage Gap

Animesh Gadre

This is a small, unserious exercise capturing the changes in the gender wage gap in Netherlands with the oaxaca package on R. I try to decompose the wage gap with the help of the Kitagawa-Blinder-Oaxaca method. I originally intended to do this exercise on Python but I came around to the conclusion that it's neater and more convenient on R. Be that as it may, the functionality of oaxaca package in R still somewhat limited (for instance, it can treat only one categorical variable at a time) and this decomposition is best executed on STATA.

The decomposition analysis of the gender wage gap builds on the wage equation of male and female workers, which are formed using the Mincerian human capital earnings function. The wage equations are then estimated using OLS and divided into an explained component which reflects the mean differences in the endowments of male and female workers and a residual or unexplained portion which reflects the gender differences in the price of market skills or labour market discrimination. However, models of such nature invariably run into endogeneity issues. We'll discuss these in brief later.

The wage equations of male and female workers can be summarised as:

$$lnW_m = X_m\beta_m + \epsilon_m \tag{1}$$
$$lnW_f = X_f\beta_m + \epsilon_f \tag{2}$$

where lnW is the natural logarithm of hourly wages of male workers (m) and female workers (f). The variable X is a vector of explanatory variables which captures human capital, life course variables and job characteristics, and the standardised residual (with mean zero and variance 1) is denoted by ϵ. The decomposition of the labour market wage is then given by:

$$lnW_m - lnW_f = \overline{X}_m\widehat{\beta}_m - \overline{X}_f\widehat{\beta}_f \tag{3}$$
$$lnW_m - lnW_f = \widehat{\beta}_m(\overline{X}_m - \overline{X}_f) + \overline{X}_f(\widehat{\beta}_m - \widehat{\beta}_f) \tag{4}$$

or

$$lnW_m - lnW_f = \widehat{\beta}_f(\overline{X}_m - \overline{X}_f) + \overline{X}_m(\widehat{\beta}_m - \widehat{\beta}_f) \tag{5}$$

where the first term on the right hand side of (4) and (5) shows the gender differences in the predictors weighted by the coefficients of the male and female workers respectively. This refers to the explained component of the wage decomposition (also known as the quantity effect). The second term refers to the unexplained component of the decomposition which is reflected by the difference in prices (or different market wage structures) weighted by the mean characteristics of male or female workers. In essence, the second term in (4) presents the difference between wages that women would have earned if they had the same

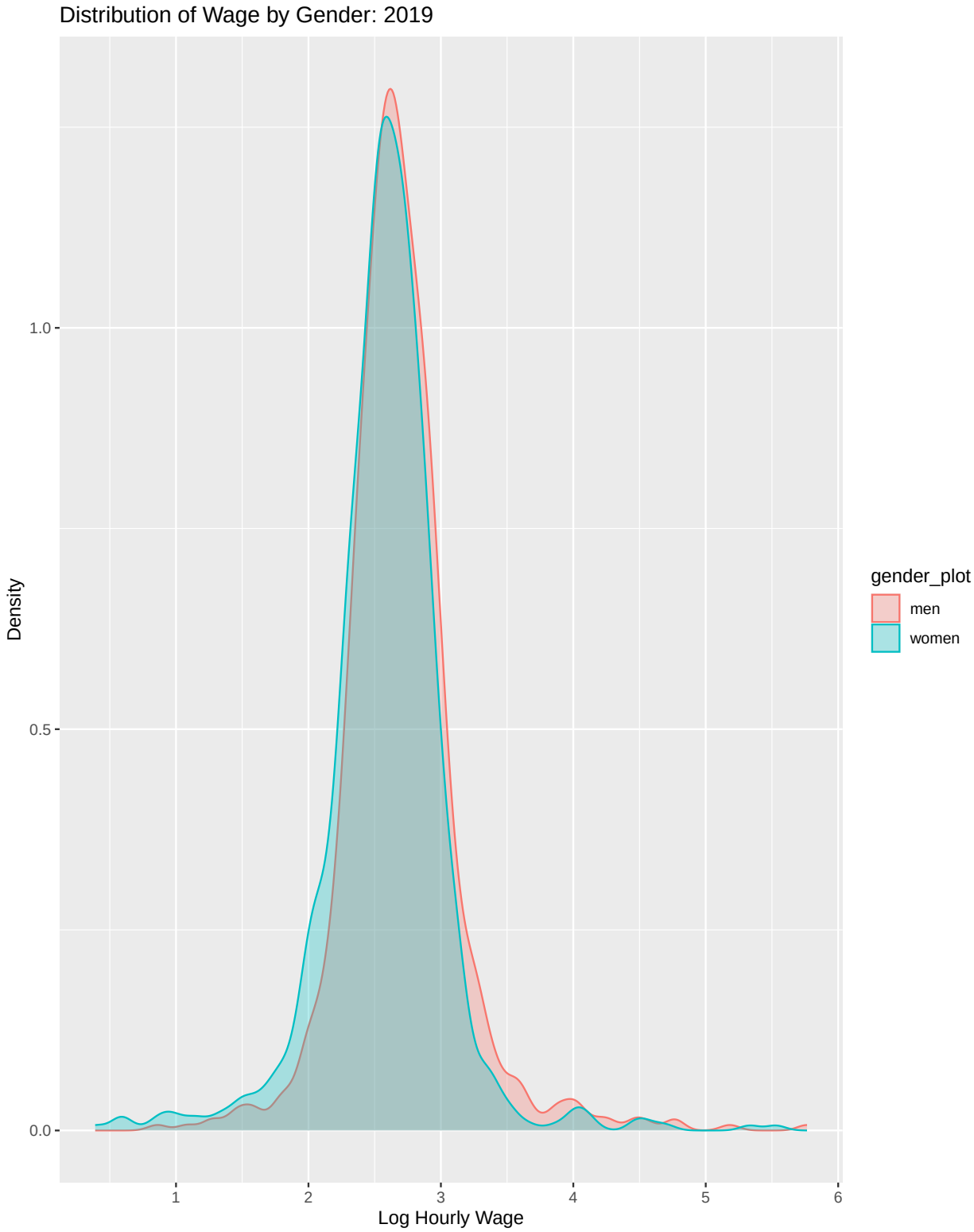
coefficients as men and the wage that they actually earned. This unexplained difference is often interpreted as an estimate of labour market discrimination faced by women.

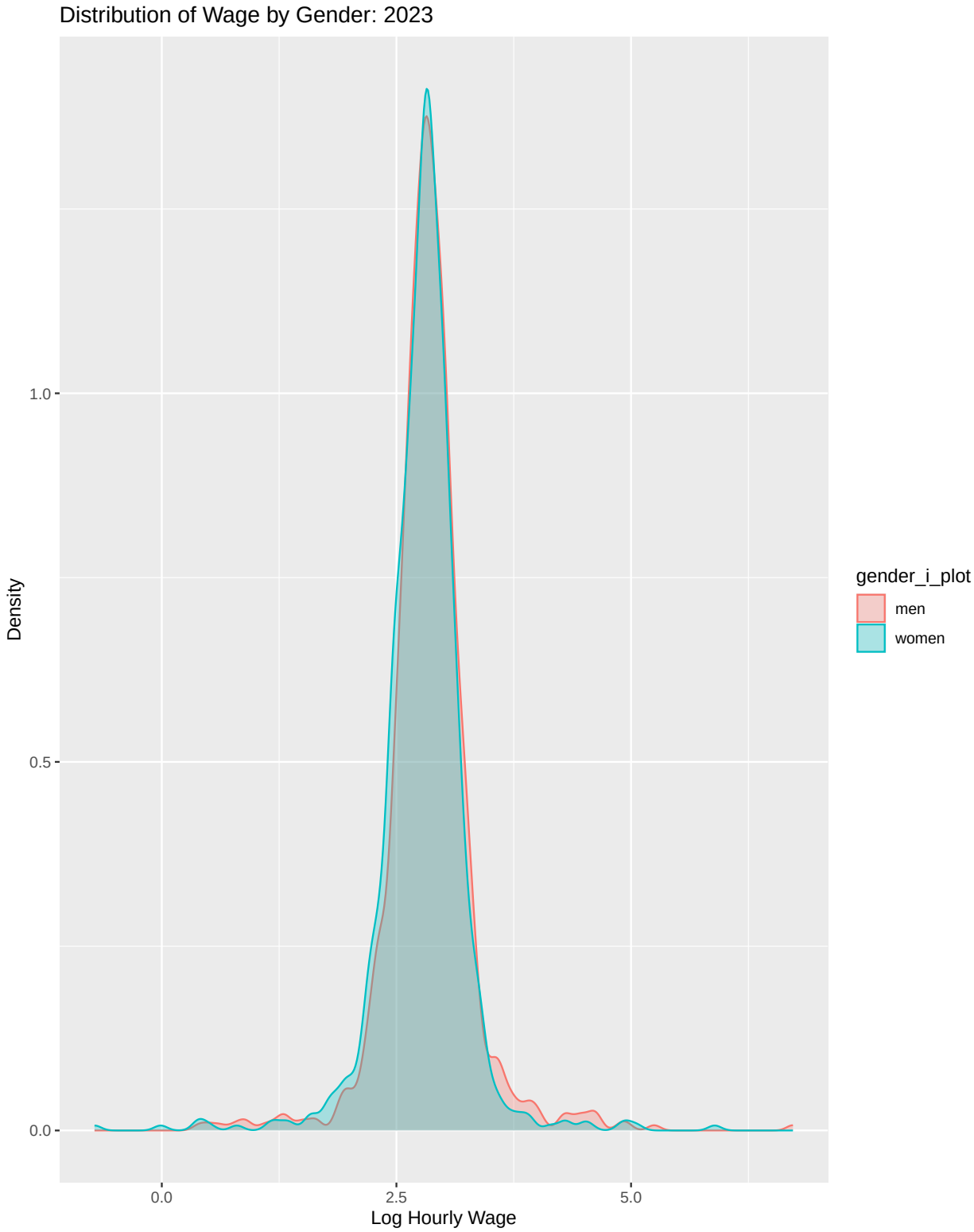
In the case of (4), the wage equation for male workers represents the nondiscriminatory wage structure, which is then used to compare the wage equation of female workers and determine the explained and unexplained differential. In the same manner, the equation (5) assumes the wage equation of female workers to be the nondiscriminatory wage structure. It has been widely debated as to which of the two equations should be used for the decomposition analysis. The wage decomposition can often be sensitive to the choice of the reference group and so it is of importance to understand the consequences of selecting either of the reference groups. The wage decomposition equation can be more generally written as:

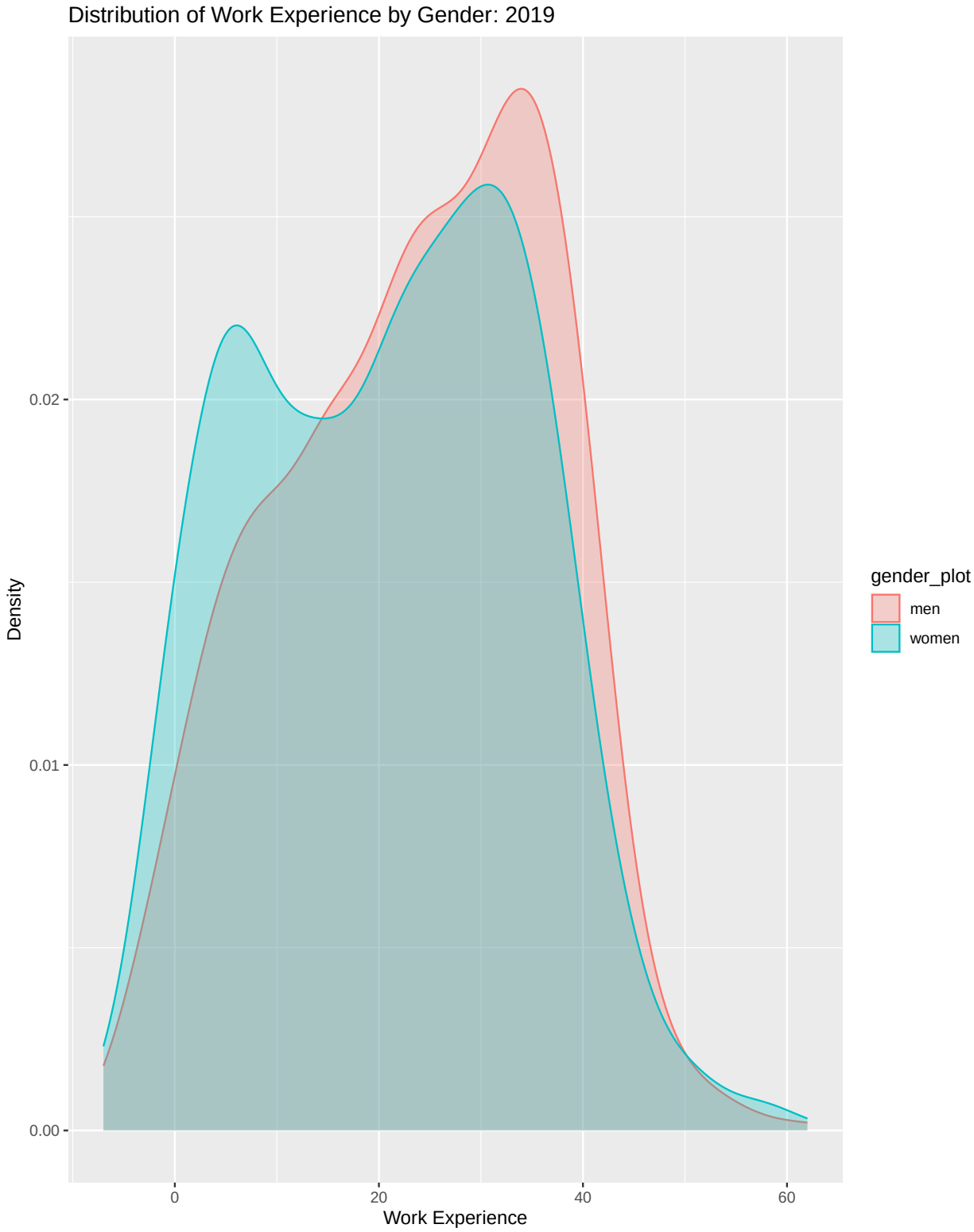
$$lnW_m - lnW_f = \widehat{\beta}^*(\overline{X}_m - \overline{X}_f) + [\overline{X}_m(\widehat{\beta}_m - \widehat{\beta}^*) + \overline{X}_f(\widehat{\beta}^* - \widehat{\beta}_f)] \tag{6}$$

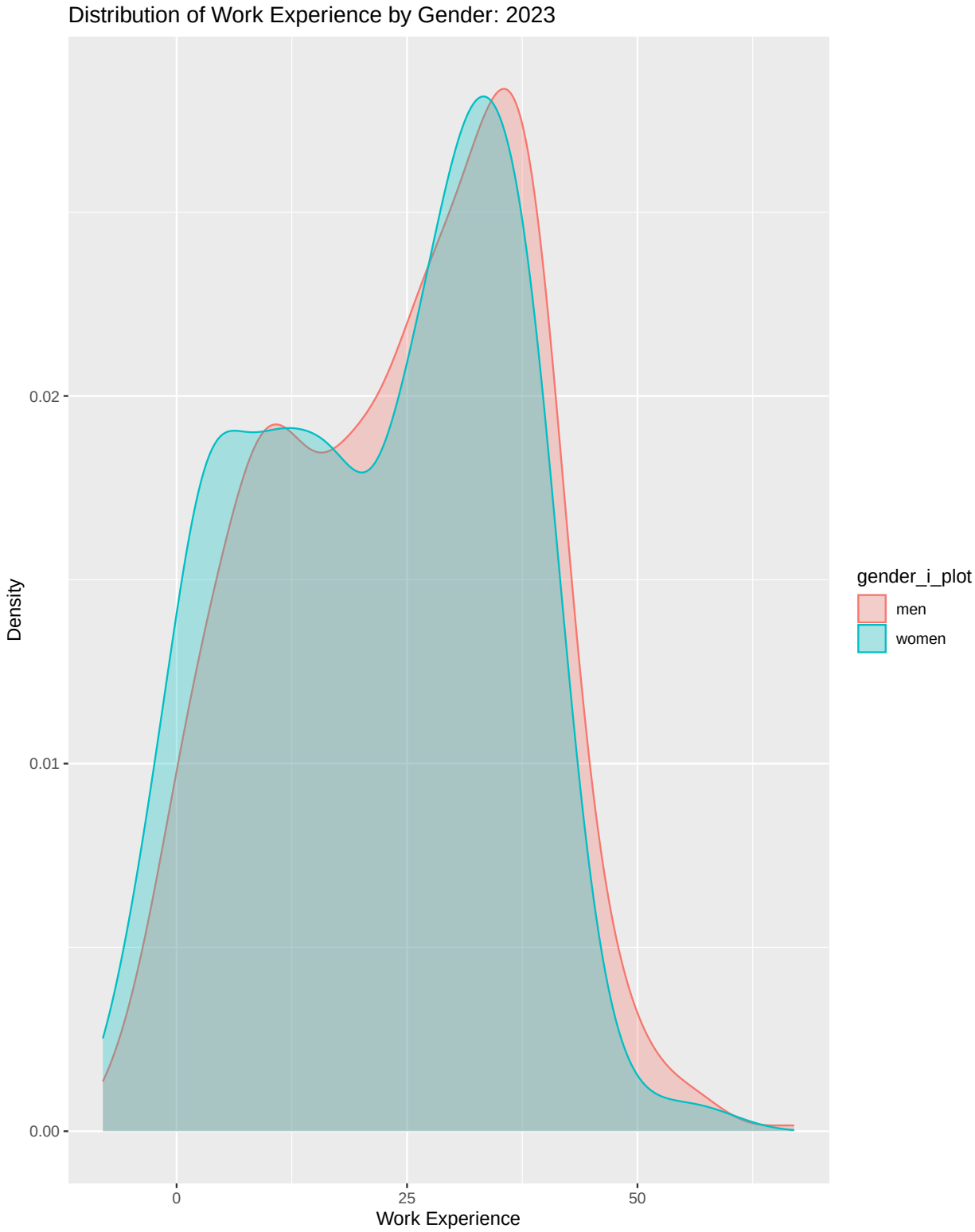
where, as before, the first term refers to the component explained by gender differences in characteristics and the second component refers to the discrimination residual. Here, $\widehat{\beta}^*$ represents the nondiscrimination wage structure. Thus, if the male wage equation is assumed to be nondiscriminatory, then $\widehat{\beta}^* = \widehat{\beta}_m$ and the resultant equation resembles (4). Conversely, if the female wage equation is assumed to be nondiscriminatory, then $\widehat{\beta}^* = \widehat{\beta}_f$ and the resultant equation resembles (5). Assume first that there is discrimination against women but no positive discrimination or nepotism towards men. In this case, when discrimination is eliminated, the wages of women will increase but men's wages will remain unchanged. Thus, $\widehat{\beta}^* = \widehat{\beta}_m$ can be assumed to be true in such a scenario. In the second case, assume that there is no discrimination against women but only positive discrimination or nepotism towards men. When discrimination is eliminated, the wages of men decrease but women's wages remain unchanged. In this scenario, $\widehat{\beta}^* = \widehat{\beta}_f$ holds true. Therefore, the choice of nondiscriminatory wage structure depends on the nature of discrimination prevailing in the economy, which can be hard to ascertain. This issue can be overcome by using a pooled model of workers which assumes that discrimination can work in either direction: against women or in favour of men. In this manner, the requirement to assume the direction of discrimination is eliminated and a nondiscriminatory wage structure is developed.

I use the rich demographic information available in the LISS Panel survey for the years 2019 and 2023. The LISS panel is a representative sample of Dutch individuals who participate in monthly Internet surveys. The panel is based on a true probability sample of households drawn from the population register. The observations with complete information are used for analysis, which includes 810 men and 879 women in 2019 and 832 men and 846 women in 2023. The dependent variable is the logarithm of hourly wage which is constructed using the data on monthly income and working hours. The LISS panel survey does not contain a variable which directly measures whether the respondent engages in part-time work or full-time work. Therefore, I construct a binary variable by using a condition on the number of hours worked per week. We can see the distribution of the log of hourly wage and work experience for men and women in the plots.


Distribution of Wage by Gender: 2019
Density
1.0
0.5
0.0
1
2
3
4
5
6
Log Hourly Wage
gender_plot
men
women


Distribution of Wage by Gender: 2023
Density
1.0
0.5
0.0
0.0
2.5
5.0
Log Hourly Wage
gender_i_plot
men
women


Distribution of Work Experience by Gender: 2019
Density
0.02
0.01
0.00
0
20
40
60
Work Experience
gender_plot
men
women



The traditional predictors in the wage equation are education and work experience. Since there is no reliable information available with respect to work experience in the survey data, I construct a rather crude measure by using years spent in education and the current age of the respondent. The model also takes into account the number of children of the respondent, civil status (co-habitation or marriage), full time work, urban

region, type of contract, occupation and type of organisation. Since the oaxaca package is unable to handle more than one categorical variable at a time, I estimate a model separately for each category.

The first model treats type of contract as a category. It includes 4 categories: (1) permanent employment; (2) employee in temporary employment; (3) on-call employee; and (4) temp-staffer.

```
## $y.A
## [1] 2.692524
##
## $y.B
## [1] 2.582551
##
## $y.diff
## [1] 0.1099728
```

The mean predictions of the log of hourly wages in 2019 come out to be 2.69 log points for men and 2.58 log points for women. The raw gender wage gap was roughly 11%. The decomposition results are presented below and the Neumark model (pooled coefficients) is indicated by group weight (-1).

```
##      group.weight coef(explained) se(explained) coef(unexplained)
## [1,]    0.0000000     0.018838279    0.01693939        0.09113452
## [2,]    1.0000000    -0.085991002    0.02575430        0.19596380
## [3,]    0.5000000    -0.033576361    0.01541207        0.14354916
## [4,]    0.4795737    -0.031435088    0.01567213        0.14140788
## [5,]   -1.0000000     0.009823636    0.01208265        0.10014916
## [6,]   -2.0000000    -0.018490265    0.01431341        0.12846306
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
## [1,]      0.02397628        9.113452e-02      2.397628e-02          0.00000000
## [2,]      0.03626789        0.000000e+00      0.000000e+00          0.19596380
## [3,]      0.02659942        4.556726e-02      1.198814e-02          0.09798190
## [4,]      0.02688962        4.742880e-02      1.149840e-02          0.09397909
## [5,]      0.02006631        5.212026e-02      1.049159e-02          0.04802890
## [6,]      0.02580256       -4.208005e-15      4.631045e-15          0.12846306
##      se(unexplained B)
## [1,]       0.000000000
## [2,]       0.036267887
## [3,]       0.018133944
## [4,]       0.018874762
## [5,]       0.009714268
## [6,]       0.025802558
```

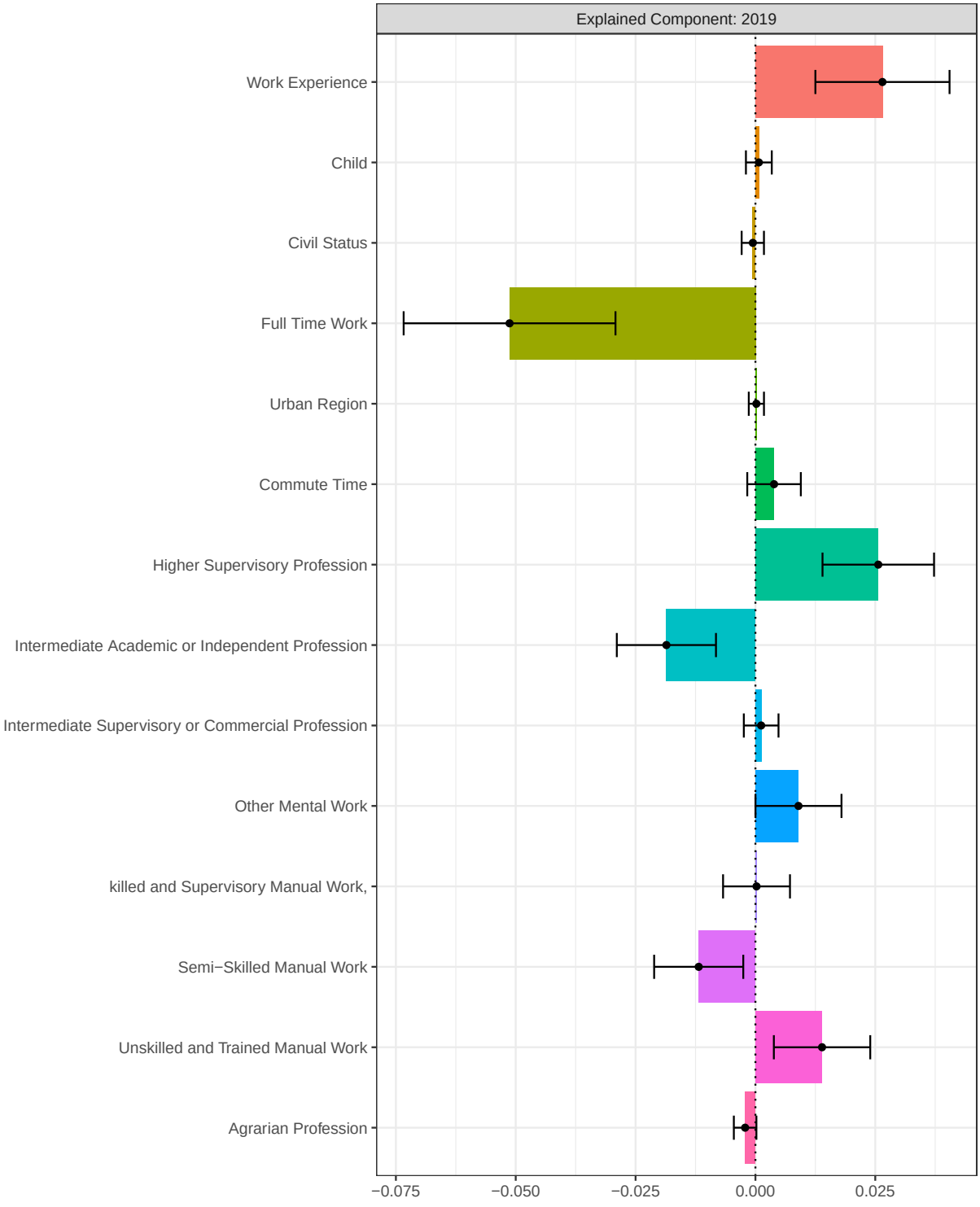
The decomposition suggests that roughly 1% out of 11% of the gender wage gap is explained by differences in human capital characteristics and job attributes. Let's look at the next model with occupation as the category.

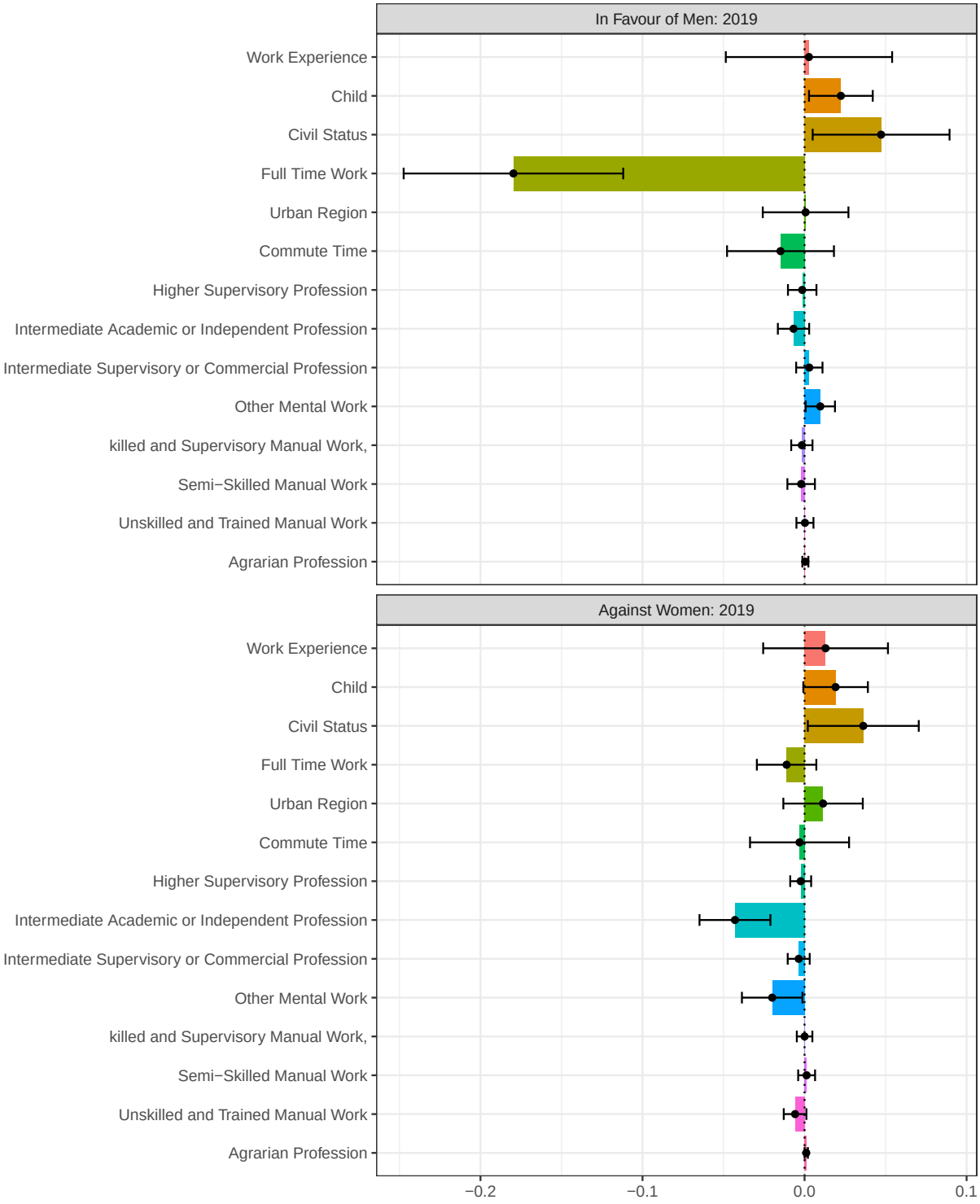
```
##      group.weight coef(explained) se(explained) coef(unexplained)
## [1,]    0.0000000    -0.002737304    0.03073976        0.11271010
## [2,]    1.0000000    -0.093037446    0.02460177        0.20301024
## [3,]    0.5000000    -0.047887375    0.02162417        0.15786017
## [4,]    0.4795737    -0.046042878    0.02147509        0.15601568
## [5,]   -1.0000000     0.014463531    0.01571042        0.09550927
## [6,]   -2.0000000    -0.024030035    0.01770878        0.13400283
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
```

```
## [1,]      0.03250075        1.127101e-01      3.250075e-02          0.00000000
## [2,]      0.03235373        0.000000e+00      0.000000e+00          0.20301024
## [3,]      0.02727721        5.635505e-02      1.625037e-02          0.10150512
## [4,]      0.02728304        5.865730e-02      1.558651e-02          0.09735838
## [5,]      0.01726893        4.970553e-02      8.771445e-03          0.04580373
## [6,]      0.02355651       -3.642269e-15      5.334046e-15          0.13400283
##      se(unexplained B)
## [1,]       0.000000000
## [2,]       0.032353727
## [3,]       0.016176864
## [4,]       0.016837730
## [5,]       0.008448364
## [6,]       0.023556509
```

With the inclusion of occupation, the model is able to explain roughly 1.5% of the wage gap. We can look at how the individual components of the occupational choice contribute to the wage gap. Next we can look at the organisation type model and plot the results of the explained and the unexplained component of the occupation model.

```
##      group.weight coef(explained) se(explained) coef(unexplained)
## [1,]    0.0000000    -0.016975221    0.01894790         0.1269480
## [2,]    1.0000000    -0.094985499    0.02281336         0.2049583
## [3,]    0.5000000    -0.055980360    0.01430533         0.1659532
## [4,]    0.4795737    -0.054386900    0.01443372         0.1643597
## [5,]   -1.0000000    -0.003200805    0.01161553         0.1131736
## [6,]   -2.0000000    -0.038268068    0.01313938         0.1482409
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
## [1,]      0.02223071        1.269480e-01      2.223071e-02          0.00000000
## [2,]      0.03598096        0.000000e+00      0.000000e+00          0.20495830
## [3,]      0.02567729        6.347401e-02      1.111535e-02          0.10247915
## [4,]      0.02600125        6.606709e-02      1.066126e-02          0.09829261
## [5,]      0.01893832        5.889852e-02      9.697846e-03          0.05427508
## [6,]      0.02442218       -2.414735e-15      4.754611e-15          0.14824086
##      se(unexplained B)
## [1,]       0.000000000
## [2,]       0.035980960
## [3,]       0.017990480
## [4,]       0.018725438
## [5,]       0.009421775
## [6,]       0.024422183
```


Explained Component: 2019
Work Experience
Child
Civil Status
Full Time Work
Urban Region
Commute Time
Higher Supervisory Profession
Intermediate Academic or Independent Profession
Intermediate Supervisory or Commercial Profession
Other Mental Work
killed and Supervisory Manual Work,
Semi−Skilled Manual Work
Unskilled and Trained Manual Work
Agrarian Profession
−0.075
−0.050
−0.025
0.000
0.025

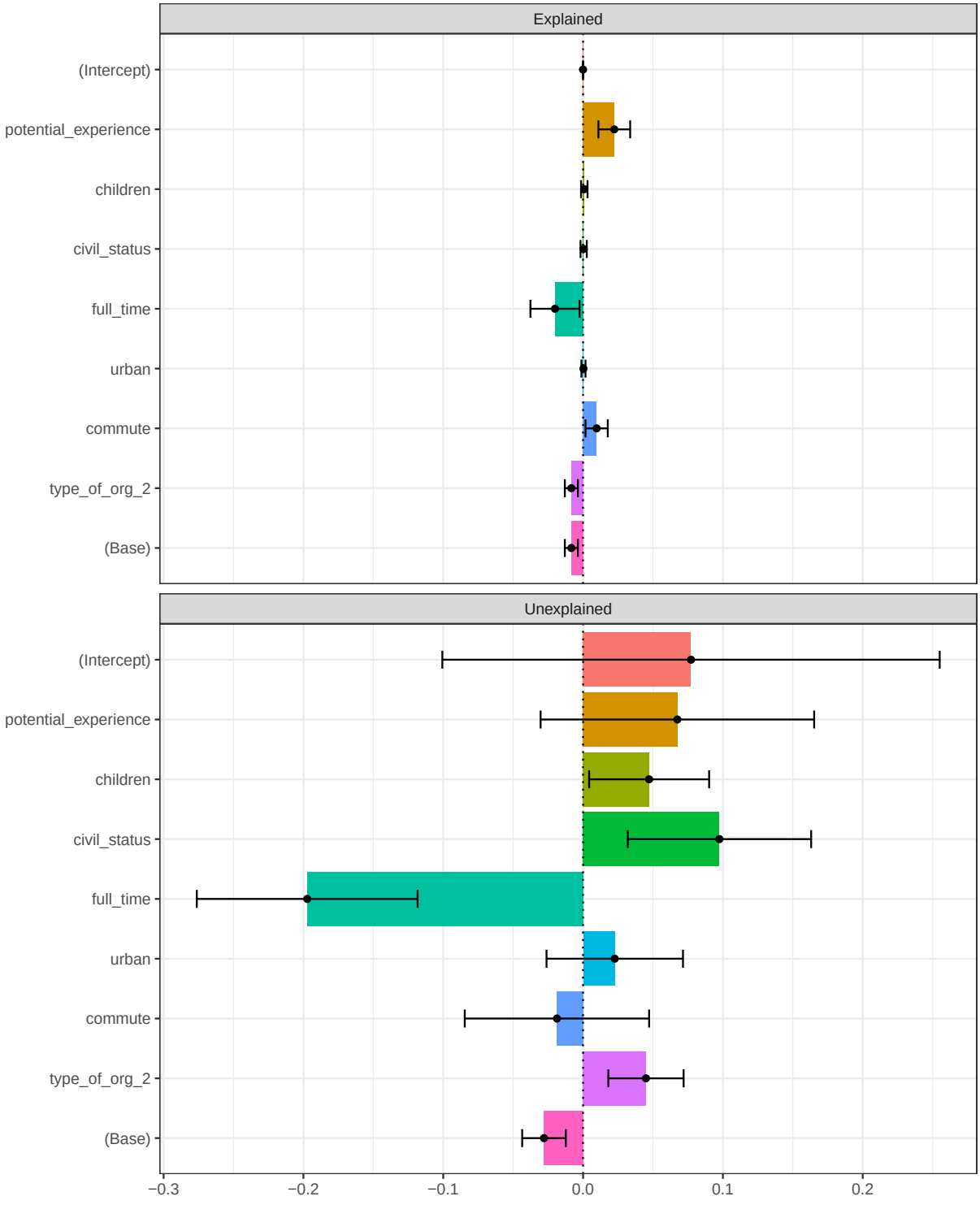


Next up is the model with type of organisation as a category. It indicates whether the respondent works in a public or a private company.

```
##       group.weight coef(explained) se(explained) coef(unexplained)
```

```
## [1,]    0.0000000    -0.016975221    0.01894790         0.1269480
## [2,]    1.0000000    -0.094985499    0.02281336         0.2049583
## [3,]    0.5000000    -0.055980360    0.01430533         0.1659532
## [4,]    0.4795737    -0.054386900    0.01443372         0.1643597
## [5,]   -1.0000000    -0.003200805    0.01161553         0.1131736
## [6,]   -2.0000000    -0.038268068    0.01313938         0.1482409
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
## [1,]      0.02223071        1.269480e-01      2.223071e-02          0.00000000
## [2,]      0.03598096        0.000000e+00      0.000000e+00          0.20495830
## [3,]      0.02567729        6.347401e-02      1.111535e-02          0.10247915
## [4,]      0.02600125        6.606709e-02      1.066126e-02          0.09829261
## [5,]      0.01893832        5.889852e-02      9.697846e-03          0.05427508
## [6,]      0.02442218       -2.414735e-15      4.754611e-15          0.14824086
##      se(unexplained B)
## [1,]       0.000000000
## [2,]       0.035980960
## [3,]       0.017990480
## [4,]       0.018725438
## [5,]       0.009421775
## [6,]       0.024422183
```

```
##                      group.weight coef(explained) coef(unexplained A)
## potential_experience           -1    0.0223260532         0.022644359
## children                       -1    0.0008776618         0.026089001
## civil_status                   -1    0.0004267434         0.051029699
## full_time                      -1   -0.0200733118        -0.192165654
## urban                          -1    0.0002726122         0.003800902
## commute                        -1    0.0097193041        -0.013180599
## type_of_org_2                  -1   -0.0083749342         0.023403850
##                      coef(unexplained B)
## potential_experience         0.044807039
## children                     0.021160144
## civil_status                 0.046498663
## full_time                   -0.005113114
## urban                        0.018862666
## commute                     -0.005500915
## type_of_org_2                0.021574043
```



The coefficients reflect the mean effect on women's wages if they had the same human capital as men. The inclusion of the type of organisation reduces the explained component of the wage gap. Women's mean hourly wage would increase by 2.23% if they had the same work experience as men.

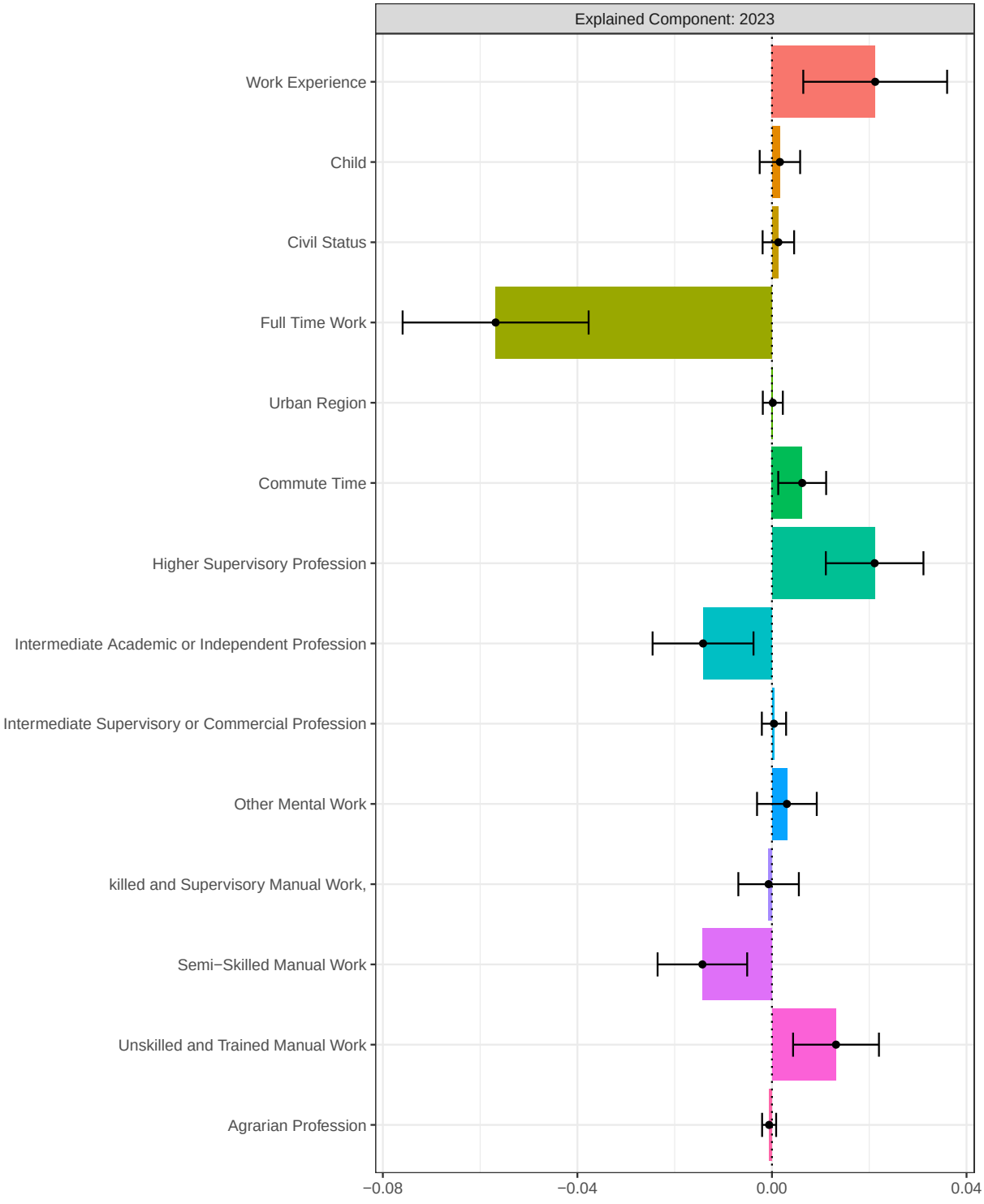
The next models estimate the gender wage gap for the year 2023. The information obtained through the

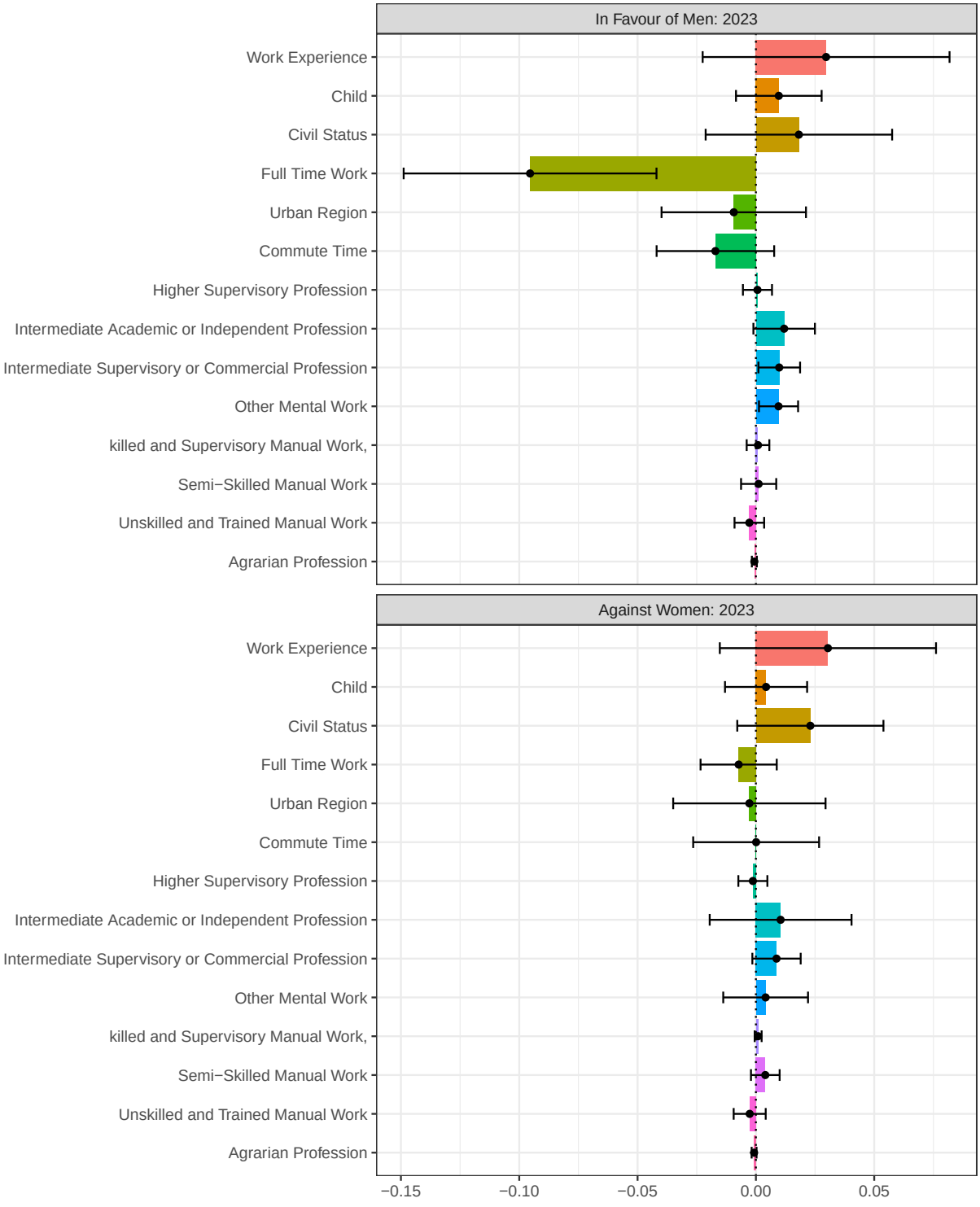
surveys for 2023 reflects the economic conditions for the year 2022, by which the pandemic induced lockdown had been for the large part lifted.

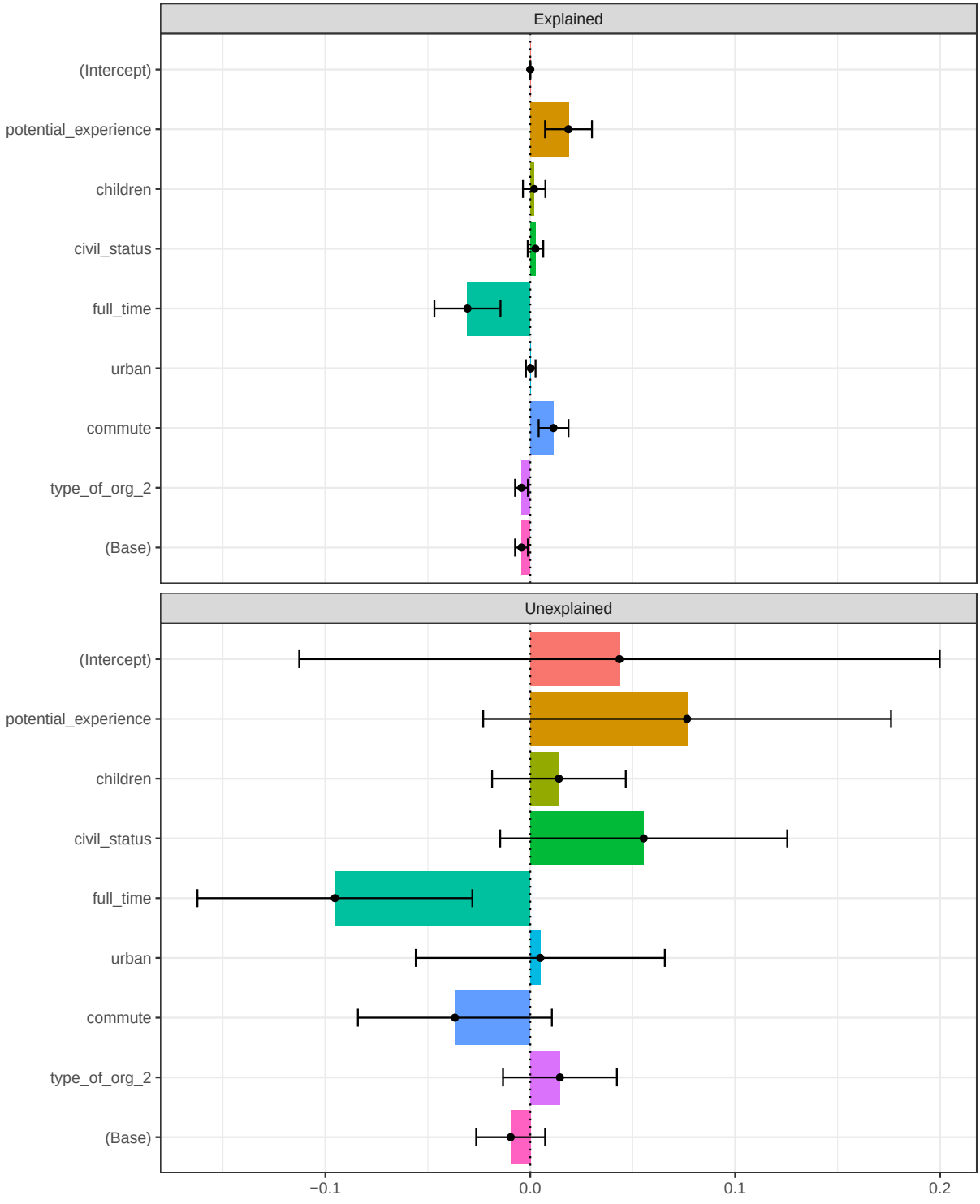
```
## $y.A
## [1] 2.845889
##
## $y.B
## [1] 2.783493
##
## $y.diff
## [1] 0.06239555
```

The mean predictions of the log of hourly wages in 2023 come out to be 2.85 log points for men and 2.78 log points for women, resulting in a wage gap of 6.2%.

```
##      group.weight coef(explained) se(explained) coef(unexplained)
## [1,]    0.0000000     0.011237558    0.01434481        0.05115799
## [2,]    1.0000000    -0.040782720    0.02079590        0.10317827
## [3,]    0.5000000    -0.014772581    0.01419253        0.07716813
## [4,]    0.4958284    -0.014555572    0.01422609        0.07695112
## [5,]   -1.0000000     0.007012846    0.01139802        0.05538270
## [6,]   -2.0000000    -0.006142701    0.01334621        0.06853825
##      se(unexplained) coef(unexplained A) se(unexplained A) coef(unexplained B)
## [1,]      0.02731354        5.115799e-02      2.731354e-02          0.00000000
## [2,]      0.03461145        0.000000e+00      0.000000e+00          0.10317827
## [3,]      0.02922836        2.557900e-02      1.365677e-02          0.05158913
## [4,]      0.02926073        2.579241e-02      1.354283e-02          0.05115871
## [5,]      0.02349390        2.792239e-02      1.174098e-02          0.02746031
## [6,]      0.02923801       -5.570522e-15      5.207511e-15          0.06853825
##      se(unexplained B)
## [1,]        0.00000000
## [2,]        0.03461145
## [3,]        0.01730572
## [4,]        0.01745011
## [5,]        0.01173418
## [6,]        0.02923801
```


Explained Component: 2023
Work Experience
Child
Civil Status
Full Time Work
Urban Region
Commute Time
Higher Supervisory Profession
Intermediate Academic or Independent Profession
Intermediate Supervisory or Commercial Profession
Other Mental Work
killed and Supervisory Manual Work,
Semi−Skilled Manual Work
Unskilled and Trained Manual Work
Agrarian Profession
−0.08
−0.04
0.00
0.04


In Favour of Men: 2023
Work Experience
Child
Civil Status
Full Time Work
Urban Region
Commute Time
Higher Supervisory Profession
Intermediate Academic or Independent Profession
Intermediate Supervisory or Commercial Profession
Other Mental Work
killed and Supervisory Manual Work,
Semi−Skilled Manual Work
Unskilled and Trained Manual Work
Agrarian Profession
Against Women: 2023
Work Experience
Child
Civil Status
Full Time Work
Urban Region
Commute Time
Higher Supervisory Profession
Intermediate Academic or Independent Profession
Intermediate Supervisory or Commercial Profession
Other Mental Work
killed and Supervisory Manual Work,
Semi−Skilled Manual Work
Unskilled and Trained Manual Work
Agrarian Profession
−0.15
−0.10
−0.05
0.00
0.05



It's important to note that the decomposition of the gender wage gap is generally subject to a lot of caveats due to the potential presence of endogeneity. Endogeneity entails that the predictors of the wage equation are correlated with the error term such that $E[\epsilon_{it}|X_i] \neq 0$ or $E[v_{it} + u_{it}|X_i] \neq 0$. Generally, there are three main sources of endogeneity in wage earnings models: (a) unobserved heterogeneity, (b) measurement error and (c) non-random sample selection. Unobserved heterogeneity may arise because of

the omission of relevant variables, which in the case of wage equations relate to any missing productivity or demographic characteristics. The problem of correlation between unobserved individual-specific effects and wage predictors may persist even with the inclusion of comprehensive measures of hu-man capital, job and demographic characteristics. The effect of the omission of a valid variable on the unexplained portion of the gender wage gap may be overstated or understated depending on the endowment levels of male and female workers in the unobserved predictors. For instance, if men (women) have a higher endowment in an unobserved predictor, then the unexplained portion of the decomposition analysis will be overstated (understated).

Measurement error in the wage predictors may cause the OLS estimates of the slope coefficients to be inconsistent. This most commonly occurs when the survey data does not contain information about a specific variable and thus a proxy variable is constructed in its place, such as potential work experience. The proxy for work experience returns inconsistent estimates because all the individuals do not work full-time or without any interruptions. In reality, men and women have vastly different working life cycles; women endure work interruptions primarily due to marriage and child birth while men do not.

Lastly, the non-random selection into work and consequently in the sample of study is another potential source of endogeneity in the model. The essential selection problem arises from the fact that men and women have different work periods; women tend to move in-and-out of the workforce intermittently while men generally work continuously until their retirement. Because there is no documentation of the wage offers made to people who are unemployed, the sample of workers only represents the self-selected group of workers. Thus, the wages observed in the sample are influenced by the individual's decision to accept orreject the wage offer. For women, the decision to accept a job offer not only depends on the wage but also on the other characteristics such as number of children, marital status and job flexibility. The non-random selection into work,hence, may bias the OLS estimates either positively or negatively. Although the problem of non-random selection into work is more acute for studies on trends of the gender wage gap than for cross-sectional studies such as this one, I make use of life course variables such as civil status and children to at least partially correct for selection into work.

The estimation of the gender wage gap after adjusting for human capital characteristics also misses an important detail: that these characteristics themselves may be affected by discrimination at various levels. There are deep seated norms which dictate the entry of women into specific educational fields, industries of work, occupation and their working hours. So if men and women do not have equal opportunities and fair choices to pursue their fields of interest, the model adjustments do not meaningfully convey the extent or the nature of discrimination. It is this line of thought which erroneously leads to people making an argument along the lines of–"There is no discrimination once we control for discrimination". Given these issues, the estimate of labour market discrimination is only a suggestive figure and not a conclusive one.