MASTERS THESIS

THE USE OF POST-DOUBLE LASSO METHOD IN ESTIMATING GENDER PAY GAP IN ESTONIA

Student:
Togzhan Khaval
Yolandah Chinyani

Advisor:
Jaan Masso
Associate Professor

Quantitative Economics

May 18, 2023

We have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced.

Abstract

This paper explores the gender pay gap in Estonia over ten years, from 2010 to 2020, and investigates the potential contribution of machine learning methods to the study of the gender pay gap. The paper discusses the history of gender inequality in wages in Estonia, provides an overview of relevant studies and data, and compares the use of traditional econometrics methods to the post-double LASSO machine learning method in estimating the gender pay gap. The results from both methods are used to decompose the gender pay gap into explained and unexplained using the Oaxaca Blinder decomposition method. This master thesis emphasizes the relevance of applying machine learning methods, such as the post-double LASSO used in this study for variable and model selection in wage gap estimation. Our findings showed that a significant portion of the gender pay gap remains unexplained both in the case of traditional and machine learning methods, potentially due to discrimination, gender biases, or other unobservable characteristics. Comparing results from the post-double LASSO method to the conventional model in the literature, we find that the size of the adjusted pay gap differs depending on the approach used.

Keywords: Gender Pay Gap, Oaxaca-Blinder decomposition, machine learning, Post-double LASSO

JEL classification: C52, C55, J16, J31, J71

Contents

Introduction

The difference in wages between men and women has been a concern for decades around the world, with several factors attempting to explain the reasons for the inequalities in wages. Cintas-Pena & Garcia Sanjuan (2019)[1] conducted a study on gender inequality in Neolithic societies using an archeological approach. Although the study's results did not show acute gender inequality, they did indicate signs of an increase in male predominance. Gender inequality dates back centuries throughout the evolution of women's rights. According to a study by Bryson et al. (2020)[2] on gender wage gap trends in Britain using data for those born in 1958 and 1970 through to the second decade of the 21st Century. The results indicated that women were paid approximately 50% less than men in the early 20th century.

Blau and Khan (1994)[3] conducted a study on trends in women's and men's hourly wages in the USA from the period between 1975 and 1987 using Michigan Panel Study of Income Dynamics (PSID) data. During the studied period, a higher percentage of women were employed in lower-paying jobs than men. However, the study's results indicated that the average log-wage gap fell from 0.504 to 0.359.

According to traditional economic theory, wages should reflect a worker's productivity, and workers with identical productivity should have identical wages. Unfortunately, it is almost impossible to accurately measure a worker's real productivity for most employers. In addition to productivity, wages are understood to result from a worker's skills, education, and labor market experience. Ideally, any difference between wages should result from variations in these factors. However, the observable wage determinants do not capture all the factors affecting wage differentials. Some other important determinants of wages are not easy to observe and not included in traditional data sources, such as social capital, favoritism, and personality traits. In particular, the gender pay gap – the difference in wages indicating that women are underpaid compared to men - is still important in promoting gender equality.

The economic status of women has changed over the years. Goldin(2014) examined the progress of gender convergence in the labour market in the United Sates over the last century. The progress theoratically applies to most parts of the world. The study highlighted that education at all levels has increased for women relative to men and job experience for women has also increased due to the increase in labour force participation. For this study, we will focus on the gender pay gap in Estonia over 10 years from 2010 to 2020. Eurostat data shows that Estonia has had one of the highest unadjusted gender pay gaps in the Euroarea during the last decade. In 2012 according to Eurostat data, the unadjusted gender pay gap in Estonia was

29.9%. The gap has decreased over time, with 2021 statistics showing that the wages and salaries gap between men and women was 14.9%. The gap has decreased, but it is still quite concerning.

Studying gender pay gaps is relevant to the economy as, in the long run, gender pay gaps potentially contribute to higher poverty rates for women during their working life and after retirement. According to Eurostat data, in 2018, women in the EU aged over 65 received pensions that were, on average, 30% lower than male pensions, and they are 3.5% more likely to be at risk of poverty than males. Clark (2016)[4] studied the economic impact of equal pay in the United States, and the results showed that if women over the age of 18 received what same wages as men, the poverty rate among working women would fall from 8.2% to 4.0%.

Using the 2014 data from the Structure of Earnings Survey, Leythienne and Ronkowski (2018)[5] performed a study to measure the impact of average characteristics between men and women on the gender pay gap. They used the Oaxaca-Blinder decomposition method to distinguish between the explained and unexplained portion of the wage differences. The results showed that the overall unadjusted gender pay gap for all countries was 16.6%, with the unexplained gender pay gap being 11.5%. In Estonia, the unadjusted gender pay gap was 28.1, and the unexplained gender pay gap was 20.1% which is a significant portion of the wage gap. It is quite concerning that women in Estonia earned 20.1% less than men, regardless of any differences in observable characteristics.

Despite the profound examination and effort to explain the differences in wages between men and women using conventional methods, there is very limited empirical research aimed at using machine learning approaches in gender inequality. Most traditional empirical gender pay gap estimations have used standard regression models even though the potential benefits of machine learning, such as its ability to handle large and complex data sets, have been highlighted recently (Athey 2018)[6]. Therefore, in this study, we aim to understand the contribution of machine learning methods to the study of the gender pay gap in comparison to the traditionally used regression models.

We estimate the gender pay gap in Estonia conditional on a set of covariates selected by the post-Least Absolute Shrinkage and Selection Operator (LASSO) machine learning approach. We will then compare the results with the estimation results from using a conventional model with covariates proposed by literature, such as education, occupation, economic sector, and age.

Machine learning methods such as the post-LASSO are used for variable and

model selection. It helps in understanding which variables are to be fit into the model to reduce overfitting and increase the accuracy of the results. The recent study by Briel and Topfer(2020)[7] applied the post-double LASSO procedure to estimate the gender pay gap using data from the German Socio-Economic Panel. The study compared the results of the model obtained by using the machine learning approach compared to results from conventional methods using the method of Oster (2019)[8]. The results showed that machine learning methods are relevant for model selection, and the effect of the variables differs along the wage distribution.

The second section of the paper is devoted to the in-depth analysis of existing literature on the current topic. The third and fourth sections provide the research methodology and descriptive data analysis, explaining the data and empirical model used for the analysis. The results of the empirical analysis are presented, and the discussion of the empirical analysis results is provided in section five.

Literature Review

In this chapter of the thesis, we examine academic studies that focus on our topic - gender wage inequality and the methods of its calculation. Firstly, we briefly talk about the gender pay gap and its importance. Secondly, we bring out papers examining gender income inequality using traditional and machine learning methods. Finally, we introduce Estonia, its economic history, and its current situation regarding wage inequality.

2.1 Traditional approach to the estimation of gender pay gap

Most studies on gender inequality depict strong convergence of the economic status of males and females over time. Despite this convergence, a gap exists between men's and women's earnings. Since wages are the largest and most accessible component of income, the gender wage gap has become the most widely studied topic in terms of gender inequality (Gupta, Smith and Ronald L. Oaxaca, 2003)[9]. The economics of discrimination equipped labour economists with the necessary tools for studying the gender wage gap, resulting in numerous research papers trying to quantify variables that contribute to the difference. Starting from the 1970s, many studies tried to explain the factors contributing to the wage differentials between men and women. Most of these studies aim to distinguish between the explained and unexplained portion of the gender pay gap. The explained portion of the gap can be explained by observable characteristics such as education and work experience while the unexplained portion of the wage gap may be a result of several factors such as variable selection, coefficient estimates, and intercepts, as well as other societal or cultural forms of discrimination that may not be considered in the models used.,(Goldin.C, 2014)[10].

The standard gender wage gap decomposition tool emerged from seminal studies of Oaxaca (1973) and Blinder (1973)[11]. The main idea of this approach is to write the gap as the sum of two parts: structure (unexplained) and composition (explained) parts. Over time, several modifications and extensions of this decomposition method have been developed. Juhn et al.(1991)[12] extended the method to study changes over time in the unexplained gap; Albrecht et al. (2003)[13] and Machado and Mata (2005[14]) integrated quantile analysis; Fairlie (2005) extended the model to treat dichotomous outcomes; Bauer and Sinning (2008)[15] modified

the model for censored outcomes, and Ñopo, (2008, a,b)[16][17] developed the model for non-parametric setups. Throughout time, the modeling framework advanced by including other distributional characteristics even more, and some methods of studying the entire distribution have been developed, DiNardo et al.(1996))[18].

The procedure known as the Oaxaca-Blinder decomposition has been the widely used method when studying wage differentials and distinguishing between the explained and unexplained portions of the wage differences. In studying the gender pay gap, the explained portion can be attributed to differences in education levels, years of experience, marital status, age, etc. However, to tackle the issue of gender discrimination, it is important to understand and reduce the unexplained portion of the gaps, as it may make policy implementation difficult if the root of the wage gap is not clear.

Weichselbaumer & Winter-Ebmer(2005)[19] conducted a meta-analysis on the international comparison of the gender pay gap reviewing over 260 published studies on wage differences between men and women from the 1960s to 1990s data covering 63 countries. The results indicated that although it is still persistent, over time, the raw gender wage gaps worldwide have fallen substantially. Additionally, country data restrictions also contributes to the differences in the estimated gaps. According to the author, estimates using the Neuwark decomposition technique found higher wage gaps than results from the Oaxaca-Blinder method. This indicates that methodological choices do have an impact on the results of gender pay gap studies.

2.2 Machine learning approach to the study of the gender pay gap

Machine learning and its applications have gained much attention worldwide. Despite economic literature widely discussing gender pay gaps and their reduction, very few studies have been concerned with the efficiency of estimating methods, especially the machine learning aspect. However, there have been several machine learning studies that have investigated wages. For example, researchers Voleti and Jana (2021)[20] have found that machine learning models can be effectively used to predict salaries of job postings considering certain features such as job title, company size, and industry. A similar study was conducted by Nandhini Shanmugam and Maheedhar Gunnam (2019)[21] and demonstrated the potential of using machine learning to predict salaries based on job postings and provides insights into the factors that influence salary levels in the job market. They applied several ma-

chine learning algorithms, including Random Forest, Gradient Boosting, and Neural Networks, to predict the salaries of these job postings. Other studies have used machine learning to identify patterns in wage data that can help to explain why some people earn more than others or identify factors contributing to wage inequality. Additionally, machine learning has been used to predict wage changes over time and identify factors that may be driving these changes. Big data and the ability to include a large variety of observable characteristics, including ones not traditionally considered, such as differences in job tasks, organizational characteristics, and social networks, could improve our understanding of why the wage gap between men and women persists.

Another paper that highlighted the potential of machine learning models discovered that the gender wage gap in Chile is primarily due to factors such as gender discrimination and occupational segregation rather than differences in education or experience, and the gap is larger for women who work in the private sector, and that it is particularly pronounced for women who work in administrative and support services (Kristjanpoller, Michell & Olson,2023)[22]. The authors find that tree-based models such as random forests and gradient boosting models outperform traditional linear regression models and suggest that policies aimed at reducing gender discrimination and promoting equal access to education and training opportunities could effectively reduce the gender wage gap in Chile.

One of these data-driven models introduced by Belloni et al.(2014)[23] is the post Least Absolute Shrinkage and Selection Operator (LASSO). This method is a way to refine the LASSO estimator further and is intended to improve the estimation of the coefficients by reducing the estimation error of the Lasso estimator. The technical part involves two auxiliary Lasso regressions.The first auxiliary Lasso regression involves regressing the outcome variable Y against the original set of covariates X. This step helps identify the covariates most strongly associated with the outcome variable and that should be retained in the final model. The second auxiliary LASSO regression involves regressing the treatment variable D against the covariates X. This step helps identify any covariates related to the treatment variable, which may be important for controlling confounding in the final model. The union of the selected controls from these two LASSO regressions gives us the full set of controls used in the final OLS regression.

Relying on the papers, we will use the post-double LASSO model to identify the most important factors that contribute to the gap and estimate their effects while controlling for other factors. We will first run a lasso for variable selection and

then an ordinary least suqares regression with variables with non-zero coefficients obtained by lasso. The size of the adjusted pay gap differs substantially depending on the approach used according to the literature review written by Stephanie Briel and Marina Topfer (2020)[7]. Using the US data, they compare two LASSO models (restricted and unrestricted) and three traditional models. The results detected that the machine learning-based model specifications are more flexible than conventional model specifications for estimating gender pay gaps (GPG). Machine learning models can help indicate which control variables are relevant at different along the wage distribution. By checking how robust the estimates are to the remaining selection on unobservables, they concluded that the LASSO models were more effective in accounting for observable selection factors and recommended using diverse control variables at different parts of the wage distribution.

The study by Böheim and Stöllinger (2021) underlies that OLS post-double LASSO approach is well-suited for decomposing the gender wage gap in case of a large number of explanatory variables. The results were obtained by comparing the OLS model and a post-double -LASSO specification.(Böheim, Stöllinger, 2021)[24]. From the investigation, they found that the explained gender pay gap is 1%-point greater than in the conventional OLS model and the estimated error variance of the post-double -LASSO specification was smaller than that of OLS.

One of the few studies focused on the impact of methodological choices was introduced by Strittmatter and Wunsch (2021)[25]. They examined the gender pay gap using a large dataset of individual earnings from Germany, testing different regression models. The study emphasizes the significance of methodological decisions and how they can significantly reduce the unexplained gender pay gap estimates up to 50%, even when relevant wage determinants are kept constant across the estimates.

Another term highlighted by the literature (Briel and Topfer 2020)[7] is glass ceiling and sticky floor. Generally, the glass ceiling and sticky floor can contribute to wage inequality and limited career opportunities for women and minority groups.(Christofide, Polycarpou, and Vrachimis, 2013)[26]. While the term "glass ceiling" is often used to describe the systemic discrimination that can limit opportunities for advancement and make it difficult for these groups to reach top leadership positions, the term "sticky floor" is often used to describe the opposite phenomenon, in which women and minorities are disproportionately represented in low-level, low-paying positions and have difficulty moving up in their careers (Arulampalam W. Booth, A. L. Bryan M. L. 2007)[27]. In their study, Stephanie Briel and Marina Topfer (2020) found that the conventional model specifications indicated significant

gender pay gaps (GPGs) at the top of the distribution, reflecting a glass ceiling effect. However, the results from machine learning showed a U-shaped GPG across the distribution, with high gaps observed at both the top (glass ceiling) and bottom (sticky floor) of the distribution. The results of this study motivate us to look whether the application of machine learning would change the results previously obtained with conventional methods.

2.3 Overview of the past studies on the gender pay gap in Estonia

Estonia became part of the Soviet Union in 1940 and had a centrally planned economy during that period. During Soviet times, gender equality was encouraged and enforced (Karu, 2011)[28]. Women were as involved in the labor market as men were. There was not much difference in the labor market in regard to the work itself. However, women were still expected to take over the traditional rules, such as taking care of the household. Over the years, regardless of the involvement of women in the labor market, Estonia has had one of the highest gender pay gaps in Europe (Anspal, 2025)[29] (Figure 2.1). The figure below gives an overview of the gender pay gap in Estonia compared to most countries in the Euro Area according to the data published by Eurostat.

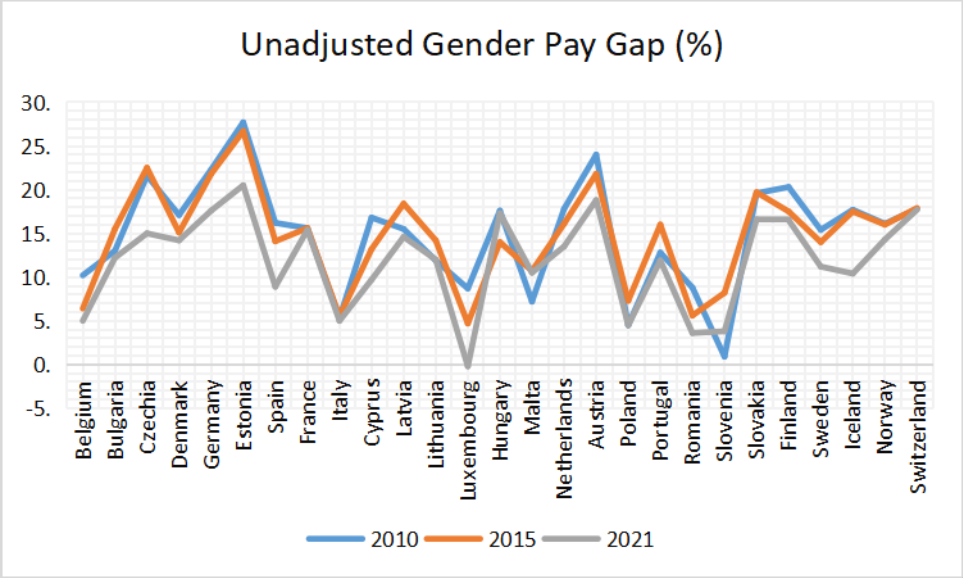

Figure 2.1: Unadjusted Gender Pay Gap
Source: Eurostat(online data code: EARN_GR_GPGR2)

Merikull and Tverdostup (2020)[30] studied the gender pay gap in Estonia over three decades, from 1989 to 2020, throughout the transition from communism to capitalism. The study showed that the position of women in the labour market has improved, and women now have better occupations. Additionally, it indicated that the gender pay gap declined due to a decline in wage inequality and women's higher returns on education, among other factors. Still, the unexplained portion of the gap has not declined much (Merikull and Tverdostup 2020)[30]. The Oaxaca-Blinder decomposition method was used for this study to understand the differences in mean wages between men and women through the transition. One of the main determinants of wage disparities was education level. Additionally, the authors argued that the unexplained gap could be explained by women's attitudes towards certain roles or other cultural factors which can be difficult to measure. During communism, women were encouraged to attain higher levels of education, the same as men, and to this day, in Estonia, women are enrolled in tertiary education more than men. According to Statistics Estonia census data from 2021, among the surveyed population, only 25.7% of males had attained higher education or specialized education after secondary education, while 37.3% of females had attained higher education or specialized education after secondary education.

Alas and Kaarelson (2008) conducted a study in Estonian private companies to understand the attitude towards gender equality in Estonia. A survey covering 301 Estonian private sector companies was conducted to understand the attitude and preferences of Estonian executives (Alas and Kaarelson 2008)[31]. The results showed that men and women do not hold the same position in the labour force as men are mostly preferred by employers and usually earn more than women. For the gap to be narrowed, employers must be aware of the issue and place a high value on promoting gender equality. However, the results from this study showed that three out of four companies do not believe that it is necessary to compare wages by gender, 49% of the managers were not aware of the gender equality act and 65% of the respondents agreed to the statements that gender inequality is a far-fetched problem and not one for the employers in the private sector (Alas and Kaarelson 2008)[31].

The Centre for Applied Social Sciences (CASS) of the University of Tartu conducted a study to understand the state of gender equality in Estonian research. Regarding the wages portion of the study, they used data from Estonian universities between 2013 and 2019. The results showed that the share of women in research has increased, and the gender wage gap has decreased over time. However, regardless of

the increases in women's participation and wages, women still earn less than men in most universities. In 2019, the study showed that women earned slightly more than men at the University of Tartu and the Estonian Academy of Arts. This could indicate the progress being made in promoting gender equality and narrowing the gender pay gap.

Several studies have been conducted in Estonia to understand gender wage disparities and the contributing factors. Most studies estimated the gender pay gap using the traditionally used econometric methods. They decomposed the results using methods such as the Oaxaca-Blinder method to understand the explained and unexplained portion of the gap. The results have not indicated a significant decrease in the unexplained potion over the years. Therefore, in this study, we will use the conventionally used regression models and take a machine learning approach using the post-double-LASSO machine learning method for variable selection and compare the results.

Data and Descriptive Statistics

In this paper, we will use Estonian Labour Force Survey(LFS) data from 2010 to 2020. According to Statistics Estonia, this survey methodology follows the International Labour Organization(ILO) guidelines and contains data on the size and characteristics of the labor force, including different variables that provide valuable insights about employment trends, wage rates, and workforce demographics, which can help to make decisions related to labor market policies and workforce planning.

From 2000, each individual is surveyed two quarters, then not observed consequent two quarters, and thereafter again surveyed for two quarters [30]. The survey includes permanent residents of Estonia between the ages of 15 and 74. To reduce the impact of outliers, we excluded observations with net wages that are lower than minimum Estonian wage and more than 4000 euros per month. Initial data contains 221,317 observations and 1644 variables, including 105,983 men and 115,334 women. The dependent variable of our analysis is the natural logarithm of hourly net wages.

Table 3.1: Data Description

	Men			Female		
Year	Wage	Age	Percentage	Wage	Age	Percentage
2010	611.51	42.40	44.28	445.30	44.63	55.72
2011	662.68	42.04	45.89	467.83	44.57	54.11
2012	716.93	42.68	44.95	491.91	44.77	55.05
2013	768.07	43.28	43.76	529.56	45.19	56.24
2014	809.24	43.27	44.39	582.02	45.11	55.61
2015	843.65	43.54	47.68	590.04	45.72	52.32
2016	911.70	43.35	44.29	656.27	45.59	55.71
2017	975.48	43.54	44.82	717.57	45.74	55.18
2018	1100.03	44.02	44.88	850.84	46.39	55.12
2019	1191.61	44.31	44.78	909.82	46.76	55.22
2020	1242.89	44.20	45.15	1098.29	46.67	54.85

Source: Estonian LFS data, authors' calculations.

Table 3.1 presents the percentages of the sample for each survey year that are men and women, and their average income and age by gender. The data for net wages are widely available for most of the years; therefore, we will use the wages in net terms in this paper. The wages are converted from Estonian kroon to euro for the period before Estonia entered the Eurozone (2010). The gender pay gap, which

Table 3.2: Explanatory variables

Variables	Description
Wage	Income from primary job occupation, monthly. This is a numerical value and it is the response variable in the models
Gender	This is our treatment variable for the study.
Demographic	Age, nationality, Marital status, Location(four regions)
Education	Level(primary, secondary and tertiary), Field of education
Occupation	Consists of 9 main categories
Employer characteristics	Firm size, Firm ownership
Economic sector	Primary, Secondary, and Services sector

is the differences in wages between men and women, is calculated by subtracting the wages of women from the wages of men and dividing the result by the wages of men. Functionally, this is expressed as

$$PayGap = (MeanEarnings_{male} - MeanEarnings_{female})/MeanEarnings_{male} \quad (3.1)$$

Estonia has had one of the highest gender pay gaps, and the raw data from LFS does show a significant difference in wages, as depicted in Figure 4.1.

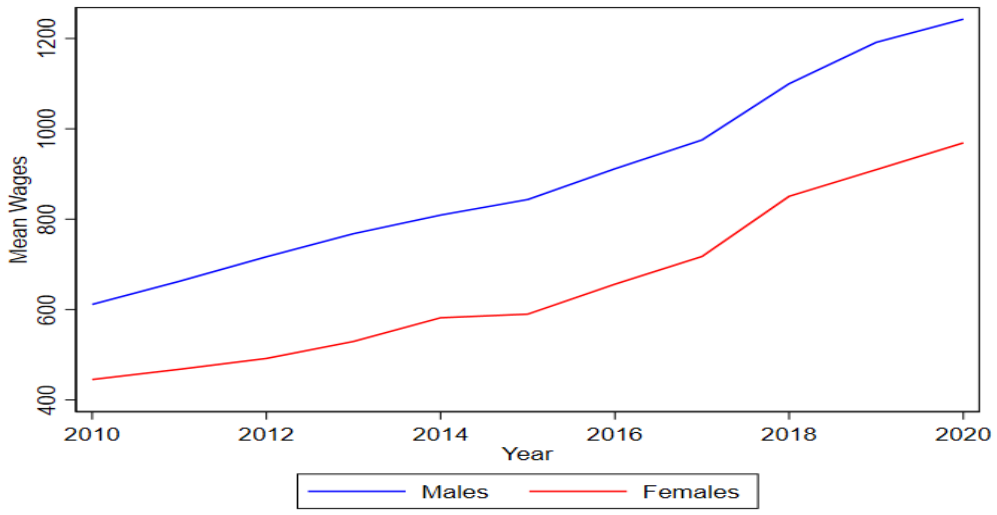

Figure 3.1: Yearly Mean wages by Gender

Source: Estonian LFS data, authors' calculations.

The LFS data contains information about the personal and job-specific observ-

able characteristics that are widely used in most studies to estimate the gender pay gap. For this study, we consider the following variables described in Table 4.2. These are considered in the analysis since they have been commonly identified as significant in the past research literature, (Danquah, Iddrisu,Boakye, Owusu, 2021)[32] More details on other categorical variables can be found in the Appendix. While the data are quite rich with regard to available wage determinants, there are several important variables that we do not observe. One example is actual work experience. Potential experience can be captured by age and education, assuming that time spent in school might reduce the time in the labour market.

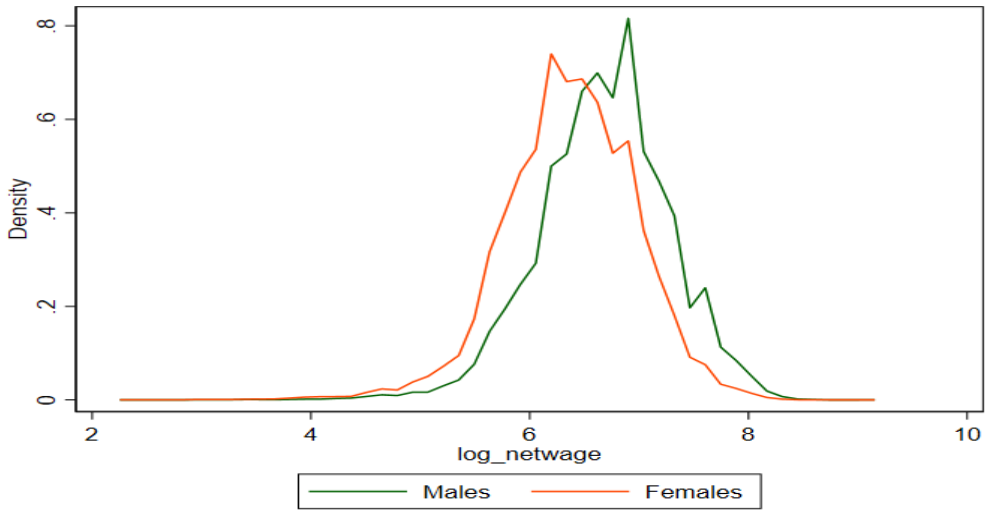

Figure 3.2: Kernel Density estimate: Distribution of Log monthly Net wages
Source: Estonian LFS data, authors' calculations.

Overall, men have higher wages, as evident from the Kernel density plot (see Figure 4.2) and the descriptive data table. The graph in green, which represents males, is positioned on the left side of the graph in orange, representing males. This indicates that women are more inclined to be present in lower wage brackets. Additionally, the highest point of the female wage distribution seems to be located to the left of the highest point of the male wage distribution.

Methodology

In this section, we outline the different estimation methods applied. First, we start by defining the parameter of interest of our research. Secondly, we describe the OLS wage regression and the machine learning approach used for model selection. Finally, we will describe the decomposition method used for this study.

4.1 Parameter of interest

In this study the parameter of interest that authors are more interested in understanding and explaining is the raw gender pay gap in Estonia over the 10-year period. First,we start by describing dummy values that will be used in this analysis. Dummy values are used in this analysis for categorical features of the data. In addition to the gender dummy variable, other categorical variables will be included in the analysis, such as education level, occupation, economic sector, firm size, and age. The outcome variable is denoted by Y_i which indicates the standardized monthly wage. It represents the actual wage earned by the individuals in the study and log-transformed to correct for the skewed distribution of wages and to interpret the results in terms of percentage changes. Then the raw gender pay gap that represents the difference in average earnings between men and women in a particular workforce will be shown as below:

$$\Delta = E[Y_i|G_i = 1] - E[Y_i|G_i = 0] \tag{4.1}$$

The raw gender pay gap can be decomposed into an explained and unexplained part. The explained part is the difference in average earnings between men and women due to differences in observable characteristics. In contrast, the unexplained part is the difference in average earnings after accounting for these factors. The vector X_i contains observed characteristics of employees or the predictors and a constant. The predicted wage of employed men, would they have the same observed characteristics as employed women are $E_{X|G=1}[\mu_0(x)]$, with $\mu_0(x) = E[Y_i|G_i = 0, X_i = x]$ Adding and subtracting $E_{X|G=1}[\mu_0(x)]$ in (3.1) gives,

$$\Delta = E[Y_i|G_i = 1] - E_{X|G=1}[\mu_0(x)] + E_{X|G=1}[\mu_0(x)] - E[Y_i|G_i = 0] \tag{4.2}$$

The second difference on the right side of (3.2) is the part of the raw gender pay

gap that can be explained by gender differences in the observed wage determinants X_i . The first difference on the right side of (3.2) is the gender pay gap for employed women, which gender differences in the observed wage determinants cannot explain. It shows us the expected difference in wages or salaries of employed women compared to observationally identical employed men, which we denote by

$$\delta = E[Y_i|G_i = 1] - E_{X|G=1}[\mu_0(x)] \tag{4.3}$$

4.2 The Standard OLS wage model

For our analysis, we will start by estimating the gender pay gap with the standard wage regression earnings function shown in equation (3.4)[33]. This equation is the fundamental part of empirical research on earnings determination (Danquah, Iddrisu,Boakye, Owusu,2021)[32]. Gender is included in the model as an explanatory variable of the wage rate to account for the potential differences between the log monthly wages of men and women. In the estimated equation (3.4), we rely on ordinary least squares(OLS). OLS estimates are based only on the sample of employed workers for whom wage is observed [34]. The OLS estimates are obtained by minimizing the residual error. Hence, this simple approach compares individuals at the mean of the distribution, i.e., the wage of the "average" man compared to that of the "average" woman, given their characteristics. The model will be specified as follows:

$$ln(Y_i) = \alpha + G_i\beta_1 + \beta_i H_i + \epsilon \tag{4.4}$$

where the log of wages will be regressed over the vector of observable personal, demographic, and labour market characteristics H_i contributing to the gender pay gap such as gender, age, nationality(Estonian or Non-Estonian), education, tenure at employer, employment sector(private or public), firm ownership (foreign, state or private domestically owned companies) and occupation, etc. The G_i is a gender dummy, and the coefficient on G_i measures the adjusted GPG. The term ϵ is an error term capturing all influences on the gender pay gap not captured by the observable variables, i.e., the unexplained part of the gender pay gap. Using the prepared data and the specified model, OLS regression analysis will be conducted to estimate the coefficients of the explanatory variables and interpreted later.

However, it's crucial to acknowledge the limitations of this OLS approach.Firstly, OLS may introduce an omitted variable bias if it fails to account for all the relevant

factors that could influence wages, such as unobserved abilities or career interruptions. Second, OLS operates under the assumption of a linear relationship between the dependent and independent variables, which may oversimplify the actual relationship between gender and wages. Lastly, OLS can produce biased estimates if there is endogeneity, for instance, if unobserved factors affect both gender and wages. The Mincerian wage equation, which includes education and experience (and the square of experience) as determinants of wages, can partially address some of these limitations. However, even the Mincerian equation might not fully capture all the complexities of the gender pay gap. Factors such as occupational segregation, differences in negotiation skills, and gender biases can also play a significant role in the gender pay gap, and these might not be fully captured by either the OLS or the Mincerian equation (Blau & Kahn, 2017)[35]. More sophisticated methods, like post-double LASSO and Oaxaca-Blinder decomposition, can help account for the influence of a larger set of variables and better isolate the effect of gender on wages (Chernozhukov et al., 2018; Jann.B, 2008)[36] [37].

4.3 Post-double LASSO Model

Following Briel and Töpfer (2020)[7], we will then use the post-double LASSO method to investigate variables contributing to the gender pay gap in Estonia. Selecting the right variables and controlling for the necessary covariates is essential in the analysis as it determines the accuracy of the model used (Gelback JB, 2016)[38]. In this study, we will use the Estonian Labour Force Survey data, which contains observable characteristics about each worker included; therefore, selecting the appropriate set of variables that correctly explain the gender wage differentials is important. We are interested in using the post-double LASSO estimator for variable and model selection for this study. The post-double LASSO method combines the LASSO method with a second stage of variable selection using ordinary least squares (OLS) regression. The LASSO method for variable selection was proposed by Tibshirani(1996)[39]. OLS estimates have low bias but large variance causing low prediction accuracy. This can be sometimes improved by shrinking or setting to zero some coefficients. Also with large number of predictors we would like to find a smaller subset that exhibit the strongest effects. The LASSO method selects variables by shrinking the parameter coefficient in linear regression, thus giving it a penalty called L1 regularization (Tibshirani, 1996). Hastie, Tibshirani, and Friedman (2009)[40] define the LASSO estimate of model coefficients as

$$\hat{\beta^{lasso}} = arg_{\beta}min1/2\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{4.5}$$

where N is the total number of observations in the data set. Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero. However, in the case of the LASSO, the penalty term has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large. Hence, the LASSO method shrinks the coefficients of the control variables towards zero, effectively selecting the most important wage predictors. (Hastie, Tibshirani and Friedman, 2009)[40].

The LASSO method prevents overfitting and improves the interpretability of the model by identifying which variables should enter the wage model and in what form.(Mullainathan, S., Spiess, J, 2017)[41]. Similarly to ordinary least squares, LASSO is an optimization problem that seeks to minimize the sum of squared residuals. The difference is the addition of a constraint. As stated by Belloni, Chen, Chernozhukov, and Hansen (2012)[42]. Before applying regularization methods, James et al. (2013)[43] recommend standardising the predictors using the formula:

$$\tilde{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{i,j} - \overline{x}_j)^2}} \tag{4.6}$$

where $x_{i,j}$ is the ith value of the jth predictor. The aim of this is for all the predictors to have unit variance and all be on the same scale. Relying on the equation (4.5) first we run LASSO for treatment variable (gender):

$$\hat{\beta^{lasso}} = arg_{\beta}min1/2\sum_{i=1}^{N}(g_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{4.7}$$

where g_i is the gender variable $x_{i,j}$ represents the control variables β_j are the coefficients to be estimated λ_1 is the penalty term for the first LASSO. In the following step, we run LASSO for outcome variable (log of wages):

$$\hat{\beta^{lasso}} = arg_{\beta}min1/2\sum_{i=1}^{N}(ln(Y_i) - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{4.8}$$

where $ln(Y_i)$ is the log wage variable $x_{i,j}$ represents the control variables β_j are the coefficients to be estimated λ_2 is the penalty term for the second LASSO.

The choice of the regularization parameter is critical to the success of the LASSO method – too large of a value and potentially important variables will be lost, and too small of a λ will not perform adequate variable selection. λ was selected through cross validation for each regression. The cross validation procedure works by taking a set of λ's, generating a model for each one, and choosing the λ that results in the model with the most predictive power. We determine the tuning parameter with a 10-fold cross-validation procedure. Cross-validation procedure selects the λ that minimizes the mean-squared-error. Optimal λ parameter that was selected by cross-validation in final adaptive step is 0.007569, i.e. the 90th value on the grid of 45 values.

. In the second stage, the post-double LASSO method further selects the best subset of the remaining predictors and identifies variables that may be correlated with gender. We will use the selected variables to perform Ordinary Least Squares regression.

We consider the following model:

$$Y_i = \tau d_i + \beta_i H_i + \epsilon \tag{4.9}$$

Where d_i is the gender wage gap at individual level i. τ and β_i are unknown parameters. In our case, we do not exactly know which of the set $H_1....H_i$ of controls is important for Y. Also, given large i, there are concerns with issues of overfitting since some of the H could be significant by chance. Following on from equation (3.5), we can specify Post-double LASSO model in equation (3.6) :

$$Y_i = \beta_0 + \sum \beta_i H_i + \sum \theta_j Z_j + \sum \phi_k W_k + \epsilon \tag{4.10}$$

The model includes control variables H_i, post-double LASSO selected confounding variables Z_j, and the post-double LASSO selected variables that may be correlated with gender W_k.

The combination of controls from both stages will then generate a full set of controls for the final OLS regression. This will also ensure that variables with large effects are included.

4.4 Decomposition

Once we run OLS regression and obtain the set of selected controls by the post-double LASSO procedure, we carry out the Blinder-Oaxaca decomposition on the difference in the mean wages of males and females. This procedure decomposes the difference in mean wages into explained and unexplained parts. (Ben Jann, 2008) [37]The procedure divides the wage differential between males and females into two parts: one that is "explained" by group differences in productivity characteristics, such as education or work experience, and a residual part (the "unexplained" part) that cannot be accounted for by such differences in wage determinants. This "unexplained" part is often used to measure discrimination, but it also includes the effects of group differences in unobserved predictors [37]. The decomposition we are interested in yields the following equation,

$$\overline{Y}_M - \overline{Y}_F = (\overline{X}_M - \overline{X}_F)\hat{\beta_M} + \overline{X}_F(\hat{\beta_M} - \hat{\beta_F}) \tag{4.11}$$

Whereby $\overline{Y}_M$ and $\overline{Y}_F$ are the observed averages of log wages of women and men, respectively. The $\overline{X}_M$ and $\overline{X}_F$ are the averages of individual characteristics and $\hat{\beta_M}$ and $\hat{\beta_F}$ are the regression coefficients of the model explaining monthly wages, estimated separately for men and women. The left part of the equation is the difference between the mean wages for men and the mean wages for women representing the raw gap. The first part of the right side in the equation shows the explained portion of the gap that can be attributed to the differences in observable characteristics. The last term represents the unexplained portion that can be attributed to the differences in coefficients and other unobserved characteristics. The results provided will show a detailed description of the contribution of each observable characteristic used to the differences in mean wages between men and women.

Results and Discussions

In this section, we present the results of our analysis consisting of 3 steps. We start by comparing two model specifications and their performance. Next, we present the results of the post-double lasso procedure used for variable selection and the selected variables. Finally, the gender pay gap is decomposed using the Oaxaca-Blinder decomposition method.

5.1 Regression outcome

In this section apart from estimating GPG we want to learn more about the performance of OLS and post-double LASSO regressions. Hence, the results of OLS regression is compared with the outcome of post-double LASSO estimator. First we look in the following set of controls selected by post-double LASSO. After having selected variables that are related to wages we end up with 64 raw potential controls. Out of 64 high-dimensional controls, the post-regularization with LASSO selected 45 controls as important determinants of wages and the rest of them that had the same value for all observations and collinearity with another variables in our dataset were omitted.

The results of wage regressions performed by OLS and post-double LASSO for male and female are presented in Table 5.1.

Table 5.1: Comparison of Regression Models

	OLS	post-double LASSO
# observations	81,933	81,595
# coefficients	43	45
$\hat{\sigma}_{MPE}$	.42132	.41268
adj.R^2	.5253	.5444

Source: Estonian LFS data, authors' calculations.

OLS is based on the OLS specification and used a sample of 81,933 observations, and a total of 43 variables was included in the model.The variable selection in conventional model was performed relying on previous literature. Post-double LASSO is a re-estimation by OLS-regression of the wage regressions including only the explanatory variables selected by the LASSO-estimator according to the one standard error rule.The table shows number of non-zero coefficients generated by

different models, the error variance estimated based on the mean squared prediction error generated by cross-validation, and the adjusted coefficient of determination for different models by gender. The adjusted R-squared for the conventional model estimated by ordinary least squares (OLS) is 0.5253, and the adjusted R-squared for the model estimated by OLS after the post-double LASSO selection is 0.5444. This means that the model estimated by OLS after the post-double LASSO selection explains about 54.44% of the variability in the dependent variable (log of netwage), while the conventional model explains about 52.53%. Given these results, the model estimated by OLS after the post-double LASSO selection performs slightly better than the conventional model in terms of explaining the variability in the dependent variable. This could be due to the post double lasso selection procedure having selected a set of control variables that are more relevant for explaining the dependent variable. The estimated error variance for the conventional model estimated by ordinary least squares (OLS) is 0.42132, and the estimated error variance for the model estimated by OLS after the post double lasso selection is 0.41268. The post-double LASSO specification has a slightly smaller estimated error variance compared to the conventional OLS model. This suggests that, on average, the model fitted after the post-double LASSO selection procedure makes slightly more accurate predictions than the conventional OLS model. As for estimating the gender pay gap, it is necessary to look at the estimated coefficient on the gender variable in each model. This coefficient represents the estimated gender pay gap, controlling for the other variables in the model. If the coefficient on the gender variable is similar in both models, it suggests that the post-double LASSO selection procedure did not substantially change the estimated gender pay gap. If the coefficient is significantly different, this implies that the control variables selected by the post-double LASSO procedure are important in estimating the gender pay gap. Recall that our conventional model controls for age (4 categories), education level (2 categories), region (4 categories), nationality, marital status (3 categories), occupation(8 categories), industry sector, firm size (5 categories), field of education (4 categories). We further extended the set of potential controls in post-double LASSO specification.

Since our dataset is missing the values of the parameter which identifies individuals who have kids for 2014 and 2015 we have excluded it from the analysis. But it is undeniable that so-called "motherhood penalty" has significant and persistent impact on the gender pay gap. According to past literature Budig and England (2001) each child reduced women's earnings by about 7% even after controlling for factors such as experience, education, and hours worked. [44]

In the following, we discuss estimation results obtained from the conventional model and the variables that were selected by post-double LASSO. In Table 5.2 described the list estimated coefficients of selected controls for both specification that represents the effect of each variable on the wages, controlling for other variables. Starting with the gender coefficient, in both models, it is negative, meaning women earn less than men on average. The pay gap is slightly reduced in the post-double LASSO regression (-.241943) compared to OLS (-.2493766). In terms of education, individuals with tertiary education earn more than those with primary education in both models. The impact of tertiary education appears slightly higher in the post-double LASSO model. For the region, post-double LASSO selection dropped two variables. Being in Northern Estonia increases wages in both models, but the effect is stronger in the post-double LASSO model. Being in Southern Estonia reduces wages in both models, but the effect is less in the post-double LASSO model. For firm size, wages increase with the size of the firm in both models, although the effect is slightly less in the post-double LASSO model.

For the field of education, the impact varies depending on the field. In most cases, the effects are less in magnitude in the post-double LASSO model. For instance, having an education in 'Education' has a negative impact on wages, but the effect is less negative in the post-double LASSO model. Interestingly, the coefficient for 'Humanities and Arts' flips sign, suggesting individuals with an education in this field earn more in the post-double LASSO model.For firm ownership, being in a foreign firm or state-owned firm has a negative impact on wages in the OLS model, but the coefficients are not estimated in the post-double LASSO model, suggesting these variables are not important predictors in the LASSO model.The coefficients for occupations vary in both models, coefficients for several occupations are not estimated in the post-double LASSO model, suggesting these occupations do not significantly contribute to explaining wage differences.For economic sectors, only the coefficient for the private sector is estimated in the post-double LASSO model, suggesting that being in the private sector slightly increases wages.For age groups, the coefficients are relatively stable across both models, suggesting age impacts wages similarly in both models.For marital status, being married increases wages in both models, but the effect is less in the post-double LASSO model. Data-driven approach selected potential variables such as industries and other job characteristics that are shown in the table and it's notable that 'monthly working hours' has a negative effect, while 'weekly working hours' has a positive effect in the post-double LASSO model.

Table 5.2: Regression results (by Gender)

	OLS	Post-double LASSO
Gender	-.2493766	-.241943
Education level		
Primary Education	- .1518857	-.1511169
Tertiary education	.2080003	.2195759
Region		
Northern Estonia	.1166772	.165750
Central Estonia	-.052303	-
Western Estonia	-. 1280443	-
Southern Estonia	-. 1600532	-.0815408
Firm		
Firm size (1-10)	-.0249211	-.053957
Firm size (11-49)	.095463	.068488
Firm size (50-199)	.169567	.141609
Firm size (200-499)	.242727	.211355
Firm size (above 500)	.248485	.222471
Estonian nationality	.163732	.137488
Field of education		
Education	-.349035	-.149784
Humanities and Arts	-.152508	.0178939
Social sciences and business	-.178274	-.1532568
Science, mathematics and computing	.089189	0.960862
Engineering, manufacturing and construction	- .228270	-.1967562
Agriculture and veterinary	-.082725	-.072149
Health and welfare	.068418	.075933
Services	-.117315	-.098825
Firm ownership		
Foreign firm	-.201174	.063945
Domestic private owners	-.269021	-
State owned	-.308121	-
Occupations		
Managers	.076733	.0823744
Tech and Associate professors	-.099373	.148637
Clerical Support	- .254390	-
Services and sales	-.347343	-
Skilled agricultural, forestry and fishery	-.317087	.0539579
Craft and related trades	-.212975	–.0065508
Plant and machinery operators and assemblers	- .271103	-
Elementary occupations	-.498187	-.192242
Economic Sector		
Primary sector	.133319	-
Private sector	-	0.0274029
Secondary sector	.064224	-
Services sector	.094633	-
Age group		
Age (1-24)	-.003347	.0022681
Age (25-35)	.147863	.147893
Age (36-45)	.177819	.176573
Age (46-59)	.113642	.113713

Table 5.3: Continued from previous page

	OLS	Post-double LASSO
Marital Status		
Single	-.031654	-.038337
Married	.0900956	.0674767
Cohabiting	-.0687058	-.0505302
Intercept Term	6.060098	-35.8268
Industries		
Hotels	-	.0646227
Transport	-	.0.144895
Finance	-	.1202535
Salestrade	-	.0260642
Other job characteristics		
Working full-time	-	0.0534646
Undereducated	-	0.0534646
Overeducated	-	-.1391707
Monthly Working hours	-	-.0420637
Weekly Working hours	-	1.216655
Upper Class	-	0.2391685

Source: Estonian LFS data, authors' calculations.

Overall, these results suggest that the gender pay gap is robust to different model specifications and that various individual, job, and firm characteristics contribute to explaining wage differences. The specific impacts of these variables vary slightly between the OLS and post-double LASSO models. The LASSO model effectively removes variables that do not contribute significantly to the prediction, providing a potentially more parsimonious and interpretable model. The key message remains that gender is a significant predictor of wage differences, even after controlling for various factors.

5.2 Oaxaca-Blinder Decomposition outcome

To present the differences in log net wages between two groups(in this case, Group 1: Male and Group 2: Female), we used Oaxaca-Blinder decomposition as mentioned in previous sections. The Oaxaca-blinder decomposes the gap at its mean into explained and the unexplained gender pay gap. The explained portion of the wage differences is due to gender differences in observable characteristics such as demographics and job characteristics. The unexplained portion is due to differences in coefficients and other non-observable characteristics not included in the model, such as discrimination. In Table 5.4, we present the estimated explained and unex-

plained gender pay gap with coefficients and standard error from our proposed full model with controls suggested by literature and controls selected by the post-double LASSO procedure.

Table 5.4: Oaxaca-Blinder decomposition of the gender pay gap

	Avg log monthly wages	
	Conventional	Post-double LASSO
Men	6.6483***	6.6455***
	(0.0030)	(0.0030)
Women	6.3590***	6.3588***
	(0.0028)	(0.0028)
Difference (raw wage gap)	0.2892***	0.2867***
	(0.0042)	(0.0042)
Explained	0.0398***	0.0447***
	(0.0034)	(0.0035)
Unexplained	0.2493***	0.2419***
	(0.0035)	(0.0035)

Source: Estonian LFS data, authors' calculations.
Note: *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively, and standard errors are given in parentheses.

The results showed an overall wage gap of about 0.289 percentage points between males and females. From our full model, of this difference, 0.0398 percentage points can be attributed to the differences in observable characteristics included in the model. This indicates that about 13.76% of the gender pay gap can be explained by the controls included in the model, and 86.24 remains unexplained. A significant portion of the wage gap is due to factors not captured by our model. Contrary to the conventionally used model, the post-double LASSO procedure provided similar results. Of the gap, 0.0447 percentage points can be attributed to differences in gender observable characteristics, and 0.2419 percentage points are due to other factors not captured by the model. A significant portion of the gender pay gap in Estonia may be due to discrimination, gender biases, or other unobservable characteristics. The post-double LASSO selected controls explain about 15.59% of the gap, while about 84.41% remains unexplained.

Table 5.5 shows the gender wage gap, the explained differential, and the unexplained differential calculated using the male-based Oaxaca-Blinder decomposition. The dependent variable is the logarithm of the monthly net wages. The table shows

the part of the gender wage gap attributed to categorical differences between men and women. From the table, gender differences in educational characteristics, occupation, and age signify a substantial portion of the wage gap. Table 7.2 shows the contribution of each variable included in our models. From both models, we observed that differences in education level have a significant impact on the gender wage gap Controlling for education level using conventional model, the wage gap is reduced by 0.0377 log points and using the post-double LASSO method, the wage gap reduced by 0.0212 log points. This suggests that a portion of the gender wage gap can be explained by variations in educational attainment between males and females. Furthermore, disparities in the choice of educational fields between men and women also contribute to the overall gender wage gap.

Regional differences were also found to have an impact on the wage gap. When considering variations in regions, the wage gap is increased by 0.0081 log points using conventional methods and by 0.0029 log points using the post-double LASSO method. This suggests that regional wage disparities contribute to the observed gender wage gap. It is also important to note that the conventional method included all four regions while the post-double LASSO model selected only the Northern and Sourthern regions. Additionally, the results indicated that differences in occupation contribute to an increase of 0.0013 log points in the gender wage gap using the conventional model and 0.0262 using the post-double LASSO model. This implies that disparities in occupational choices between males and females partially explain the observed wage gap. The result differences for each model are due to the differences in variables used in each model (Table 7.2 presents the full breakdown of the decomposition)

Other variables, such as nationality, age, firm ownership, firm size, marital status, and working full time, also influenced the gender wage gap.

Table 5.5: Oaxaca-Blinder decomposition of the gender pay gap- Grouped variables

	OLS	Post-double LASSO
Variable group	Log points	Log points
Education level	-0.0377	-0.0212
Field of education	0.0063	0.0029
Region	0.0081	0.0057
Nationality	-0.0027	-0.0025
Occupation	0.0013	0.0262
Economic sector	0.0009	-
Industry	-	0.0052
Age	0.0010	0.0009
Firm ownership	0.0078	0.0007
Firm size	0.0022	0.0016
Marital status	0.0006	0.0001
Working full time	0.0522	0.0511
Other job characteristics	-	-0.0162
Explained differential	0.0398	0.0447
Unexplained differential	0.2493	0.2419
Gender wage gap	0.2892	0.2867

Source: Estonian LFS data, authors' calculations.

Comparing the two models, the results suggest that we underestimated the explained gender pay gap using the conventionally used model. The post-double LASSO model selected a set of controls that are relevant predictors of wages. Using the post-double LASSO estimator, certain controls included in the conventional model were removed, and other controls not included were included. There were substantial differences in the variables included in the models. For demographics, the conventional method included age, nationality, and location. The location variable consisted of four geographical regions in Estonia; among the four, the post-double LASSO estimator selected only the North and Southern parts of Estonia. This may be due to the population distribution in Estonia as 2 of the major cities in Estonia are in the selected regions. Another significant difference in selected controls would be in the industry and sectoral variables. The conventional model did not consider specific industries but included sectors. Variable selection by the post-double lasso eliminated sectoral variables and selected four industries (salestrades, hotels, transport, and finance) out of ten industries as relevant predictors. Although

the differences in the decomposition results is relatively small, model selection has a substantial effect on the gender pay gap decomposition.

Overall, the Blinder decomposition analysis provided valuable insights into the factors contributing to the gender wage gap. By accounting for differences in various explanatory variables, we were able to estimate the individual contributions of each variable to the wage gap. These findings contribute to our understanding of the underlying causes of the gender wage gap and can inform policy interventions aimed at reducing gender-based wage disparities

Conclusion

Our main goal of this study was to investigate whether or not a systematic control variable selection using the post-double LASSO could contribute to the understanding of the gender pay gap. In our study, we applied a machine learning technique (post-double LASSO estimator) in order to find the appropriate set of control variables for the estimation of gender differences in pay. We then conducted Oaxaca-Blinder decomposition analysis on two different regression models, conventionally used models with variables suggested by literature and post-double LASSO. The Blinder-Oaxaca decomposition helps to understand the gender pay gap by breaking it down into explained and unexplained components. The paper aimed to compare the conventionally estimated adjusted GPGs to estimates of the latter based on the post-double-LASSO estimator proposed by Belloni et al.[42] A few studies have used machine learning techniques for the studying of wages. This study is the first to use machine learning techniques in studying the gender pay gap in Estonia. Compared to traditionally used econometrics methods, machine learning models such as the post-double LASSO used in this study accounts for variable selection bias, overfitting and improve predition accuracy by identifying the most important variables while avoiding spurious variables.

Using variables suggested by literature, the decomposition results show that the overall gender pay gap is 0.2892 percentage points with 0.0398 percentage points explained by the differences in characteristics between men and women, and 0.2493 percentage points remain unexplained. Some of the factors contributing to the explained portion of the pay gap include region (Northern, Central, Western, and Southern Estonia), occupation (managerial, clerical support, services and sales, and elementary occupations), education (primary, secondary, and tertiary), and age. Although the Estonian constitution has made improvements in promoting equality, according to a publication by the World Trade Press, the Estonian society has a patriarchal attitude and women still play a subordinate role. The society's has restricted the activities of women at home and work. This may in turn mean women spend more time taking care of the household and children which may affect their careers.

From the post-double LASSO selected variables, the decomposition results show a gender pay gap of 0.2867 percentage points, with 0.0447 percentage points explained and 0.2419 unexplained. Similar factors contribute to the explained portion of the pay gap, such as occupation, region, firm size, firm ownership, education level, Estonian nationality, and organizational sector. For the post-double LASSO proce-

dure, industry dummy variables and economic sector variables were included. It is important to note that the post-double LASSO estimator did not identitfy sectoral variables as important wage predictors but rather selected several industries such as finance, transport and hotels as significant predictors.

Overall, the analysis of the gender pay gap in Estonia using traditional regression models and the post-LASSO machine learning approach revealed a significant unexplained portion of the wage gap. While education, occupation, economic sector, age, and other factors were identified as significant predictors of wages, a large part of the wage gap remains unaccounted for, possibly resulting from discrimination, gender biases, or unobservable factors not included in the models. When using high-dimensional data sets in the study of wages, the post-double-lasso variable selection play a significant role in selecting only the most important variables while avoiding the inclusion of spurious variables.

This study contributes to the understanding of the gender pay gap in Estonia and highlights the potential benefits of incorporating machine learning methods in future research on wage inequality. Machine learning methods such as the post-double LASSO variable selection have the potential to provide more reliable and accurate estimates of the factors that contribute to the gender pay gap, which could inform policy decisions aimed at reducing this gap.

Bibliography

[1] M. Cintas-Pena and L. Garcia Sanjuan. Gender inequalities in neolithic iberia: A multi-proxy approach. *European Journal of Archaeology*, 22(4):499–522, 2019.

[2] A. Bryson, H. Joshi, B. Wielgoszewska, and D. Wilkinson. "A Short History of the Gender Wage Gap in Britain". 2020.

[3] F. D. Blau and L. M. Kahn. "Analyzing the gender pay gap". *Quarterly Review of Economics & Finance*, (39(3)):625, 1999.

[4] J Clark. "The Economic Impact of Equal Pay by State. Women in the States". 2016, February.

[5] Leythienne & Ronkowski. "Decomposition of the Gender Pay Gap Using a Sample of Establishments in the Structure of Earnings Survey". *Statistical Journal of the IAOS,2018*, 2018.

[6] Susan Athey. "The Impact of Machine Learning on Economics". *University of Chicago Press*, 2018.

[7] S. Briel and M. Topfer. "The Gender Pay Gap Revisited: Does Machine Learning offer New Insights?". 2020.

[8] E Oster. "Unobservable selection and coefficient stability: Theory and evidence". 2019.

[9] Nina Smith Nabanita Datta Gupta and Ronald L. Oaxaca. "Swimming Upstream, Floating Downstream: Comparing Women's Relative Wage Position in the U.S. And Denmark ". *Industrial and Labor Relations Review*, 2003.

[10] C. Goldin. "A grand gender convergence: Its last chapter.". *American Economic Review*, 104(4), 2014.

[11] A Blinder. "Wage discrimination: Reduced form and structural estimates". 1973.

[12] Kevin M. Murphy Juhn, Chinhui and P. Brooks. "Accounting for the slowdown in blackwhite wage convergence". *'Workers and their wages', AEI Press, Washington*, 1991.

[13] J. Albrecht, A. Björklund, and S. Vroman. "Is there a glass ceiling in Sweden?". *J. Labor Econ.*, 21(1), 2003.

[14] Jose A. F. Machado and M José. "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression" . *Journal of Applied Econometrics*, 20(4), 2005.

[15] T. K. Bauer and M. Sinning. "An extension to the Blinder-Oaxaca decomposition to nonlinear models". *Advances in Statistical Analysis*, 92.2, 2008.

[16] H Ñopo. "Matching as a tool to decompose wage gaps". *The Review of Economics and Statistics*, 90.2, a,2008.

[17] H Ñopo. "An extension of the Blinder-Oaxaca decomposition to a continuum of comparison groups". *Economic Letters*, 100.2, b,2008.

[18] J. N. Fortin DiNardo and T. Lmieux. “Labour market institutions and the distribution of wages, 1973-1992: a semi-parametric approach.”. *Econometrica*, 64.5, 1996.

[19] D. Weichselbaumer and R. Winter-Ebmer. “A meta-analysis of the international gender wage gap.”. *Journal of Economic Survey*, 19.3, 2005.

[20] Ritvik Voleti and Bappaditya Jana. “Predictive Analysis of HR Salary using Machine Learning Techniques ”. 2021.

[21] Nandhini Shanmugam and Maheedhar Gunnam. “Machine Learning Based Salary Prediction from Job Postings ”. 2019.

[22] Kevin Michell1 Werner Kristjanpoller and Josephine E. Olson. “Determining the gender wage gap through causal inference and machine learning models: evidence from Chile ”. 2023.

[23] A. Belloni, V. Chernozhukov, and C. Hansen. “Inference on treatment effects after selection among high-dimensional controls. ”. *Econ.Stud*, 81(2), 2014.

[24] R. Böheim and P. Stöllinger. "Decomposition of the gender wage gap using the LASSO estimator". *Applied Economics Letters,*, (28(10)):817–828, 2021.

[25] A. Strittmatter and C. Wunsch. "The Gender Pay Gap Revisited with Big Data: Do Methodological Choices Matter?". 2021.

[26] A. Polycarpou Christofides, L.N. and K. Vrachimis. “Gender Wage Gaps, “Sticky Floors” and “Glass Ceilings” in Europe”. *Labour of Economics*, 21, 2013.

[27] A. L. Bryan M. L. Arulampalam W. Booth. “Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution”. 60(2), 2007.

[28] Marre Karu and Kairi Kasearu. “Slow Steps towards Dual Earner/Dual Carer Family Model: Why Do Fathers Not Take Parental Leave ”. 3(1), 2011.

[29] S. Anspal. “TEssays on gender wage inequality in the Estonian labour market. PhD dissertation, Faculty of Economics and Business Administration, University of Tartu.”. 2015.

[30] J. Meriküll and M. Tverdostup. "The Gap that Survived the Transition: The Gender Wage Gap Over Three Decades in Estonia. SSRN Electronic Journal". 2020.

[31] Ruth Alas and Tõnu Kaarelson. “Gender equality in post-socialist country: case of Estonia” . 6(2), 2008.

[32] Ernest Owusu Boakye Michael Danquah, Abdul Malik Iddrisu and Solomon Owusu. “Do gender wage differences within households influence women’s empowerment and welfare? Evidence from Ghana”. 2021.

[33] J Mincer. “Investment in human capital and personal income distribution.”. *Journal of Political Econom*, 66, 1958.

[34] A. Belloni and C. Chernozhukov, V.and Hansen. “High-dimensional methods and inference on structural and treatment effects. ”. *Econ.Perspect*, 28(2), 2014.

[35] F. D. Blau and L. M. Kahn. "The gender wage gap: Extent, trends, and explanations". *Journal of Economic Literature*, 55(3), 2018.

[36] V Chernozhukov, D Chetverikov, M Demirer, E Duflo, C Hansen, W Newey, and J Robins. "Double/debiased machine learning for treatment and structural parameters". *The Econometrics Journal*, 21(1), 2018.

[37] Ben Jann. “The Blinder–Oaxaca decomposition for linear regression models”. *The Stata Journal*, (4), 2008.

[38] Gelback JB. “When do covariates matter? And which ones, and how much?”. *Labor Econ*, 34(2), 2016.

[39] R. Tibshirani. “Regression shrinkage and selection via the lasso.”. *Stat. Methodol*, 1996.

[40] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning: data mining, inference and prediction". 2009.

[41] S. Mullainathan and J. Spiess. “Machine learning: An applied econometric approach.”. *Econ. Perspect*, 31(2), 2017.

[42] A. Belloni, D. Chen, and C. Chernozhukov, V.and Hansen. “Sparse models and methods for optimal instruments with an application to eminent domain.”. *Econometrica*, 80(6), 2012.

[43] James, Gareth, Daniela Witten, and Trevor Hastieand Robert Tibshirani. “An Introduction to Statistical Learning.”. 2013.

[44] Michelle J. Budig and Paula England. Children and the gender wage gap. *Title of the Journal*, Number of the Volume:Range of Pages, 2001.

Resumee

MASINÕPPE MEETODITE KASUTAMINE SOOLISE PALGALÕHE HINDAMISEKS EESTIS

Käesolevas magistritöös uuritakse soolise palgalõhe muutumist Eestis kümne aasta jooksul, aastatel 2010-2020, rõhuasetusega masinõppe meetodite kasutamisvõimalustele soolise palgalõhe uurimisel. Töös käsitletakse soolise palgaerinevuse ajalugu Eestis, antakse ülevaade asjakohastest uuringutest ja andmetest ning võrreldakse traditsiooniliste ökonomeetriliste meetodite kasutamist topelt-LASSO-järgse (post-double LASSO) masinõppe meetodiga soolise palgalõhe hindamisel. Mõlema meetodi tulemusi kasutatakse soolise palgalõhe dekomponeerimiseks selgitatud ja selgitamata osadeks, kasutades Oaxaca-Blinderi dekomponeerimist. Magistritöös rõhutatakse masinõppe meetodite, nagu käesolevas uuringus kasutatud topelt-LASSO-järgse meetodi, kasutamise asjakohasust palkasid selgitavate muutujate ja mudeli kuju valikul palgalõhe hindamisel. Meie tulemused näitasid, et märkimisväärne osa soolisest palgalõhest jääb selgitamata nii traditsiooniliste kui ka masinõppe meetodite kasutamisel, mis võib olla tingitud diskrimineerimisest, soolisest eelarvamusest või muudest mittevaadeldavatest omadustest. Võrreldes topelt-LASSO-järgse meetodi tulemusi kirjanduses levinud tavapärase mudeliga, leidsime, et korrigeeritud palgalõhe suurus erineb sõltuvalt kasutatud lähenemisviisist.

Appendix

Table 7.1: Descriptive Statistics by Gender,Selected Variables

	Men		Women	
	Mean	Std.Dev	Mean	Std.Dev
Log Net Wage	6.691377	.5881755	6.37826	.601553
Education level				
Primary Education	.1323482	.3388728	.068918	0.253317
Tertiary education	.2406519	.4274843	.387465	0.487176
Region				
Northern Estonia	.5171853	.4997109	.498364	0.50002
Central Estonia	1281015	.334207	.124627	0.330299
Western Estonia	.094036	.291882	.109906	0.312776
Southern Estonia	.17048	.376058	.186847	0.389793
Firm				
Firm size (1-10)	.201094	.400823	0.215600	411242
Firm size (11-49)	.404121	.490726	.389045	0.487538
Firm size (50-199)	.249023	.432452	.245268	0.430250
Firm size (200-499)	.068984	.253430	.076369	0.265590
Firm size (above 500)	.069542	.252377	.070877	0.256622
Nationality				
Estonian nationality	.740842	.438176	.750890	0.432501
Field of education				
Education	.003626	.060412	- .04007	0.196135
Humanities and Arts	.0203	.141026	.047718	0.213170
Social sciences and business	.029956	.170469	.121799	0.327057
Science, mathematics and computing	.176562	.381302	.120129	0.325114
Engineering, manufacturing and construction	.203454	.172163	.076345	0.265551
Agriculture and veterinary	.037188	.189224	0.030630	0.172316
Health and welfare	.2106822	.407797	.123191	0.328659
Services	.092873	.290258	.057309	0.232435
Firm ownership				
Foreign firm	.218122	.412975	.186517	0.389527
Domestic private owners	.595711	.490759	.4316331	0.495308
State owned	.1855	.388706	.381478	0.485754
Occupations				
Managers	.0937104	.291428	.070278	0.255617
Tech and Associate professors	.117272	.321747	.146308	0.353418
Clerical Support	.035278	.184484	.081465	0.273551
Services and sales	.065964	.248222	.205213	0.403861
Skilled agricultural, forestry and fishery	.011029	.104443	.011261	0.105522
Craft and related trades	.252883	.434099	.030927	0.173123
Plant and machinery operators and assemblers	.21624	.411684	.075992	0.264988
Elementary occupations	.081371	.2734079	.120463	0.325506
Economic Sector				
Primary sector	.0725735	.259438	.027217	0.162716
Secondary sector	.2578056	.437431	.166456	0.372493
Services sector	.642539	.479057	.753951	0.430711
Age group				
Age (1-24)	.075427	.264082	.058719	0.235101
Age (25-35)	.239994	.428084	.171706	0.377129
Age (36-45)	.238728	.426311	.245621	0.430459
Age (46-59)	.312933	.463692	.369406	0.482648
Marital Status				
Single	.195081	.396267	.164452	0.370689
Married	.7593399	.4284889	.681948	0.465724
Cohabiting	.7593828	.427462	.682133	0.465651
Employment characteristics				
Working full-time	.7593828	.427462	.682133	0.465651

Source: Estonian LFS data, authors' calculations.

Table 7.2: Oaxaca-Blinder decomposition of the gender pay gap

	Conventional		Post-double LASSO	
		p > |z|		p > |z|
Male	6.6483 (0.0030)	0.000	6.6456 (0.0030)	0.000
Female	6.3591 (0.0028)	0.000	6.3589 (0.0028)	0.000
Difference	0.2892 0.0042	0.000	0.2867 (0.0041)	0.000
Explained	0.0398 (0.0034)	0.000	0.0447 (0.0035)	0.000
Unexplained	0.2494 (0.0035)	0.000	0.2419 (0.0035)	0.000
Explained				
Education level				
Primary education	-0.0092 (0.0005)	0.000	-0.0092 (0.0005)	0.000
Tertiary education	-0.0286 (0.0008)	0.000	-0.0304 (0.009)	0.000
Region				
Northern Estonia	-0.0001 (0.0001)	0.000	0.0042 (0.0006)	0.000
Central Estonia	-0.0001 (0.0001)	0.245	- -	-
Western Estonia	0.0022 (0.0004)	0.000	-	
Southern Estonia	0.0029 (0.0004)	0.000	0.0015 (0.0002)	0.000
Firm size				
Firm size (1-10)	0.0004 (0.0005)	0.398	0.0008 (0.0004)	0.000
Firm size (11-49)	0.0006 (0.0004)	0.113	0.0006 (0.0003)	0.000
Firm size (50-199)	0.0017 (0.0006)	0.003	0.0016 (0.0005)	0.000
Firm size (200-499)	-0.0012 (0.0005)	0.010	-0.0013 (0.0004)	0.000
Firm size (Above 500)	0.0007 (0.0005)	0.141	-0.0001 (0.000406)	0.000
Nationality	-0.0027 (0.0005)	0.000	-0.00249 (0.00043)	Value
Field of education				
Education	0.1499 (0.0005)	0.000	0.0135438 (0.0004781)	0.000
Humanities and arts	0.0044 (0.0003)	0.000	0.0041756 (0.000312)	0.000
Social sciences, business and law	0.0171 (0.0006)	0.000	0.0147742 (0.000624)	0.000
Science, mathematics and computing	-0.0061 (0.0006)	0.000	-0.0065724 (0.000456)	0.000
Engineering, manufacturing and construction	-.0251 (0.0007)	0.000	-0.0217673 (0.0007153)	0.000
Agriculture and veterinary	-0.0005 (0.0001)	0.000	-0.0004374 (0.0001086)	0.000
Health and welfare	0.0051 (0.0004)	0.000	0.0057936 (0.0004002)	Value

Table 7.2 – Continued from previous page

	Conventional		Post-double LASSO	
		p > |z|		p > |z|
Services	-0.0036 (0.0003)	0.000	-0.0029712 (0.0002575)	0.000
Firm ownership				
Foreign firm	-0.0018 (0.0008)	0.000	0.0006706 (0.0001828)	0.000
Domestic private owned	-0.0469 (0.0112)	- 0.000	- -	
State owned	0.0565 (0.0.0117)	- 0.000	- -	
Occupations				
Managers	(0.0004)	0.000	0.0026736 (0.0002662)	0.000
Technicians and associate professionals	0.0025 (0.0003)	0.000	-0.0036471 (0.0003943)	0.000
Clerical support	0.0116 (0.0005)	0.000		
Services and sales	0.0445 (0.0011)	0.000	0.0099633 (0.0008682)	0.000
Skilled agricultural, forestry and fishery	0.0003 (0.0002)	0.000		
Craft and related trades	-0.0426 (0.0014)	0.000	0.0109201 (0.0014306)	0.000
Plant and machinery operators and assemblers	-0.0362 (0.0011)	0.000	-0.0008817 (0.0009036)	0.000
Elementary occupations	0.0189 (0.0011)	0.000	0.00719 (0.0004816)	0.000
Economic sector				
Primary sector	0.0054 (0.0005)	- 0.000	- -	
Secondary sector	0.0065 (0.0009)	- 0.000	- -	
Services sector	-0.0111 (0.0010)	- 0.000	- -	
Age group				
Age(less than 25	-0.0000 (0.0001)	0.0.688	0.0000393 (0.0001453)	0.784
Age (25-35)	0.0099 (0.0006)	0.000	0.0096685 (0005601)	0.000
Age (36-45)	-0.0027 (0.0005)	0.000	-0.002898 (0.0005402)	0.000
Age (46-59)	-0.0060 (0.0004)	0.000	-0.0058759 (0.0004529)	0.000
Marital status				
Single	-0.0010 (0.0002)	0.000	-0.0011809 (0.0002188)	0.000
Married	0.0068 (0.0067)	0.312	0.0050858 (0.0059283)	0.391
Cohabiting	-0.0052 (0.0067)	0.441	-0.0038001 (0.0059208)	0.521
Other job characteristics				
Working full time	0.0522 (0.0014)	0.441	0.0511094 (0.0014083)	0.000
Upper class	- -	- -	-0.0242318 (0.0010238)	0.000
Calender monthly working hours	-	-	0.0005563	

Table 7.2 – Continued from previous page

	Conventional		Post-double LASSO	
		p > |z|		p > |z|
			(0.0002213)	0.012
Calender weekly working hours	-	-	0.0005945	
	-	-	(0.0005232)	0.256
Overeducated	-	-	0.0062464	
		-	(0.0003804)	0.000
Undereducated	-	-	0.0005833	
			(0.0001256)	0.000
Industry				
Salestrade	-	-	-0.0012694	
		-	(0.0002459)	0.000
Hotels	-	-	0.0044362	
	-	-	(0.0004225)	0.000
Transport	-	-	-0.0004905	
	-		(0.0002795)	0.079
Finance	-	-	0.0024877	
	-	-	(0.0002501)	0.000

Source: Estonian LFS data, authors' calculations.

Note: Standard errors are given in parentheses.

Non-exclusive licence to reproduce thesis and make thesis public

I, Togzhan Khaval
(date of birth: 15.05.1997)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
1.1. reproduce, for the purpose of preservation and making available to the public, including for
addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
1.2. make available to the public via the web environment of the University of Tartu, including
via the DSpace digital archives until expiry of the term of validity of the copyright,

THE USE OF POST-DOUBLE LASSO METHOD IN ESTIMATING GENDER PAY GAP IN ESTONIA

supervised by Jaan Masso,

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property
rights or rights arising from the Personal Data Protection Act.

Tartu, 18.05.2023

Non-exclusive licence to reproduce thesis and make thesis public

I, Yolandah Chinyani
(date of birth: 06.09.1996)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
1.1. reproduce, for the purpose of preservation and making available to the public, including for
addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
1.2. make available to the public via the web environment of the University of Tartu, including
via the DSpace digital archives until expiry of the term of validity of the copyright,

THE USE OF POST-DOUBLE LASSO METHOD IN ESTIMATING GENDER PAY GAP IN ESTONIA

supervised by Jaan Masso,

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property
rights or rights arising from the Personal Data Protection Act.

Tartu, 18.05.2023