Marius Hallin Ekman
Oskar Eikeseth

The educational wage premium

An empirical analysis based on Norwegian registry data.

Master's thesis in Economics
Supervisor: Hildegunn Ekroll Stokke
June 2023

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Marius Hallin Ekman
Oskar Eikeseth

The educational wage premium

An empirical analysis based on Norwegian registry data.

Master's thesis in Economics
Supervisor: Hildegunn Ekroll Stokke
June 2023

Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Norwegian University of
Science and Technology

Preface

It is with great pleasure and enthusiasm that we present this master’s thesis, the culmination of months of research and hard work. This work is the result of our passion for the field of economics, and our desire to contribute to its knowledge and advancement.

This master's thesis marks the end of a 5-year long academic journey in economics at NTNU in Trondheim. Throughout this journey, we have been fortunate enough to learn from some of the brightest minds in the field of economics. One of these great minds has been our mentor and guidance throughout the work with this thesis, Hildegunn Ekroll Stokke. Her feedback and support have been instrumental in shaping our thinking and shaping the direction of this research. We are deeply grateful for her contributions.

This thesis is not only the result of our individual efforts but are also the product of the countless individuals who have contributed to its development. We would like to express our sincere gratitude to all those who have supported and encouraged us throughout this journey. We also want to give a huge thanks to family and friends for being good emotional support for us. A special dedication goes to Marius’ mom, who always believed in him and supported him, but sadly passed away due to cancer on April 1, 2023. She was truly an inspiring person with a warm heart and a special ability to take care of those around her.

Finally, we would like to give each other a sincere thank you for the companionship throughout these five years. The cooperation and support in several of our common subjects have been important for both our academic results, but also for our motivation and wellbeing. We have been able to develop a unique professional relationship, but also a unique friendship which will last for a lifetime. So, for that we will always be grateful for each other.

We hope that this thesis will contribute to the ongoing discourse and understanding of wage gaps, and that it will inspire future researchers to continue to push the boundaries of knowledge in the field of economics.

Trondheim, June 15, 2023

Marius Hallin Ekman Oskar Eikeseth

Abstract

The purpose of this thesis is to analyse the educational wage premium, or educational wage gap, amongst the population of Norway in 2014. Low-educated individuals are defined by having only completed secondary education, while high-educated individuals are defined by holding a college degree. The main analysis aims to confirm the existence of a wage premium, measure its size, and find possible heterogeneities in the wage premium between gender, immigration background, and location of residence. An econometric approach is used as the main method to estimate the wage premium, which includes both OLS regressions and IV regressions. The task relies on registry data from microdata.no, which provides access to data on the entire Norwegian population. The result is a dataset of more than 1.2 million individuals, along with the ability to control for various variables that also affect one's wages. This allows for the controlling of 30 control variables and isolates their effect on wages, thus obtaining a more precise estimate of the educational level's effect on wages. At the same time, the use of microdata.no has limited the access to other variables that could have been interesting to control for, resulting in other implications that are discussed in detail throughout the thesis.

The analysis finds a wage premium relative to individuals with high school education as their highest finished degree, of 18.16% for those with a bachelor's degree, 35.59% for those with a master's degree, and 47.51% for those with a doctorate's degree. The heterogeneity test finds that women have a slightly higher wage premium than men, immigrants from high-income countries have a slightly higher wage premium than natives, and those who live in the labour market region of Oslo have a slightly higher wage premium than those who live outside the labour market regions of Oslo, Trondheim, Stavanger, or Bergen. Furthermore, the test does not conclude whether there is a difference in the wage premium for immigrants from low-income countries and natives, or whether there is a difference for those who live in the labour market regions of Trondheim, Bergen, or Stavanger compared to those who live outside Oslo, Trondheim, Bergen, or Stavanger.

A pervasive problem throughout the thesis is that the effect of skills on wags has not been resolved, which affects the results of the wage premium for education. This is also discussed throughout the task and summarized towards the end.

Sammendrag

Formålet med denne oppgaven er å analysere lønnspremien ved å ta høyere utdanning, eller lønnsforskjellen mellom høyt og lavt utdannede, på Norges befolkning i 2014. Lavt utdannede er definert ved å bare ha fullført videregåendeutdanning, mens høyt utdannede er definert ved å inneha en grad fra universitet eller høyskole. Hovedanalysen går ut på å bekrefte at det finnes en lønnspremie, måle størrelsen på denne, samt mulige heterogeniteter i lønnspremien mellom kjønn, innvandringsbakgrunn og bosted. Det blir brukt en økonometrisk tilnærming som metode for å estimere lønnspremien, der oppgaven inkluderer både OLS regresjoner og IV regresjoner. Oppgaven baserer seg på individdata fra microdata.no, som gir tilgang på data på hele Norges populasjon. Dette har gitt et datasett på mer enn 1,2 millioner individer, samt tilgang på å kontrollere for ulike variabler som også påvirker lønna. Dette gjør at det blir kontrollert for 30 kontrollvariabler i modellen som løser ut deres effekt på lønna, og dermed oppnås et mer presist estimat på utdanningsnivåets effekt på lønna. Samtidig har bruken av microdata.no begrenset tilgangen på andre variabler som kunne vært interessant å kontrollere for, noe som har resultert i andre implikasjoner som blir diskutert nærmere gjennom oppgaven.

Analysen finner en lønnspremie relativ til individer med videregående utdanning som deres høyeste fullførte utdanning på 18,16% for de med bachelorgrad, 35,59% for de med mastergrad, og 47,51% for de med doktorgrad. Heterogenitetstesten finner at kvinner har en noe høyere lønnspremie enn menn, innvandrere fra høyinntektsland har en noe høyere lønnspremie enn nordmenn, og at de som bor i arbeidsmarkedsregionen Oslo har en noe høyere lønnspremie enn de som bor utenfor arbeidsmarkedsregionene Oslo, Trondheim, Stavanger eller Bergen. Videre finner testen ikke fram til en konklusjon på om det er forskjell i lønnspremien for innvandrere fra lavinntektsland og nordmenn, eller om det er forskjell for de som bor i arbeidsmarkedsregionene Trondheim, Bergen eller Stavanger mot de som bor utenfor Oslo, Trondheim, Bergen eller Stavanger. Resultatene fra disse testene blir knyttet opp mot tidligere litteratur.

Et gjennomgående problem for oppgaven er at effekten ferdigheter har på lønnen ikke har blitt løst for, noe som påvirker resultatet på lønnspremien for utdanning. Dette er også noe som diskuteres gjennom hele oppgaven og blir oppsummert mot slutten.

Content

1 Introduction

Income is essential to meet a person's basic needs and achieve a certain standard of living. Most people would accept a high income if it was offered, and people generally enjoy having financial breathing room. One needs the income to purchase food, clothing, shelter, healthcare, education, and other basic necessities. Without income, it would be difficult to survive and maintain a decent quality of life. Moreover, income also enables people to achieve their goals and aspirations, such as buying a house, starting a business, or pursuing further education. It can also provide people with a sense of security, independence, and freedom of choice. There are several articles implying that the marginal utility of income is positive, but decreasing as income becomes large. See for instance *The Marginal Utility of Income* (Layard, Mayraz and Nickell, 2008) or *Diminishing Marginal Utility Revisitied* (Kimball *et al.*, 2015). This is an interesting finding, and implies that at a certain stage of wealth, increased income will not necessarily give a significant increase in utility. According to Finansforbundet, the amount where the marginal utility flattens out is estimated to be \$95.000 USD from a study done by Purdue University in Indiana, but varies around the world (Finansforbundet, 2020). Based on this study, one can assume that people who earns less than \$95.000 USD yearly, will have interest in increasing their income and have incentives to take measures that will ensure so. For most people, their monthly salary is their main source of income and is determined by their wage levels. It is therefore reasonable to believe that people with less than \$95.000 USD in annual income also are more concerned about a wage gap.

In this thesis, the main focus is the wage gap between high- and low-educated individuals and determining the payoff related to taking higher education by using the OLS regression method. While doing so, heterogeneities between worker groups are being controlled for and variables with explanatory power on wages are included to isolate the effect of education on wages at a larger degree.

1.1 What is a wage gap?

First off, it is important to distinguish between wages, salary, and income. These are terms which are usually used interchangeably yet have a slightly different meaning. A salary is a fixed amount of money that is regularly provided to an employee, typically monthly or yearly.

When an employee is hired, a salary is typically agreed upon and stated in an employment contract. The amount of a salary usually remains the same regardless of the number of hours worked. Wages, as opposed to salaries, are often paid by hourly rates. The payment paid to an employee usually depends on the number of hours they have worked, as well as on the sector, industry, job title, and geographic location. Income is a broader term that refers to any money that a person receives, whether it's from salary, wages, investments, rent, dividends, interest, capital gains, public social security schemes, or any other source.

In layman's terms, a wage gap is the disparity in wages between different groups or individuals, typically based on elements like education level, work experience, age, ethnicity, gender, and occupation. For instance, the gender wage gap is the distinction in pay between men and women, with males typically earning more than women for equivalent or similar work. Similar to this, racial wage gaps describe the disparities in wages among various racial and ethnic groupings. In this thesis, the main focus is on educational differences to explain the wage gap between higher and lower educated individuals but supplemented with different control variables to isolate the actual effect of education to a higher degree. At the same time, it is important to mention that there are endless many factors that decide the wage gap and the model can only include a limited number of variables. This is due to some factors being hard to represent with numerical data, and other factors just haven’t been observed enough so that there is representable data available to use.

1.2 Wage premiums as a return on investment

A wage gap can easily be seen as some sort of unfair distribution of wealth. Hence, the wage gap is also subject of discrimination. That being said, the main interest in this thesis is to explain the educational impact on wages, rather than a discriminatory impact. For this reason, it is important to understand why wage differences between higher and lower educated individuals are necessary.

The wage gap is not necessarily an unfair distribution, but rather a necessary premium to create an incentive for people to voluntarily choose to take further education after they finish high school. But why do people need to take higher education? The answer is quite simple. Several jobs nowadays require the competence achieved from higher education to execute the job. Medicine, engineering, and technology are only a few examples of industries which have become way more advanced only during the last few decades. Higher education is crucial to continue progressing the competence within these sorts of fields with new research and innovation. The higher education is required to acquire knowledge that is necessary to understand certain jobs and for employers to be capable of performing with high quality at work. This is a desired requirement from employees at all types of firms and for the society as a customer of these services.

To achieve the desired behaviour of a certain proportion of the society’s individuals wanting to attend higher education, it needs to follow some sort of reward for delaying income and putting in the effort. The existence of a wage premium is a way to reward those individuals choosing to invest in their own education. This dynamic is what is called the education wage premium. The educational wage premium is meant to compensate for the investment cost of education.

Enrolling in higher public education at university level is considered free in Norway, but there is a small admission fee that only costs about 600 NOK per semester (Direktoratet for høyere utdanning og kompetanse, 2022). Even though attending university is free based on a monetary point of view, there are other costs related to taking higher education. If the opportunity costs are considered, taking higher education can suddenly seem quite expensive. Instead of attending higher education, one could start working and make an income to finance living costs. By attending higher education, those living costs need to be financed by some other economic supports, which normally is a student loan. At least 60% of the loan becomes debt which needs to be repaid later. The additional debt combined with lost income possibilities is also two arguments on why education wage premium is necessary.

In 2021, the average student debt for students completing their higher education was 378 000 NOK (Lånekassen, 2022). For comparison, in 2011 this number was 246 000 NOK. That means there has been an increase of approximately 132 000 NOK during a decade. In 2022, the average debt increased to 410 000 NOK (Eilers, 2023). Even though some of this increase can be explained by inflation, the average inflation from 2011 to 2021 has only been 24.4%, which yields a debt of about 306 000 NOK in 2021 (Statistisk Sentralbyrå, 2023). There is still a real increase of 72 000 NOK during the decade. Also, only 22 000 NOK of the increase of 32 000 NOK from 2021 to 2022 can be explained by inflation (Statistisk Sentralbyrå, 2023). There is still a real increase of 10 000 NOK during this 1-year period. Another explanation of the study debt growth is the extension of study support from 10 to 11 months (Lånekassen, 2022). This extension results directly in a larger debt per year from an additional month of payments.

So, taking both lost income and study debt into account, the opportunity cost of higher education is quite high. A high opportunity cost favours a high educational wage premium, so that individuals have incentives to voluntarily attend higher education.

1.3 Research question and its background

Some parts of economics are closer tied together than others. Education economics and labour economics are two fields that depend on each other's fundamentals, where for instance education is a key determinant of labour market outcomes. At the same time will incentives to take education first appear when entering the labour market. Hence, both of these studies are central in this thesis where the wage premium given by educational attainment is in focus. A generous number of papers already acknowledge that a wage premium to education indeed exists. Both the papers from Walker and Zhu (2008), and James (2012) show that the wage premium to education exists and is dynamic over time.

The studies of education economics and labour economics is the key field for this thesis. The background of the research question is the interest of analysing the return of educational investments in Norway, based on registry data from the entire population of Norway in 2014. This data is presented and explained in chapter 3. The wage premium is the only considered return of education, even though one could argue that for instance improved work environment, increased knowledge, or being able to work within a field one finds interesting, also are forms of educational returns. These types of returns have a weak database since they are hard to empirically measure, so the analysis fails to include them. Hence, wages are considered as the only direct return of educational investment.

The limitations imposed by the available data result in the following formulation of the research question:

> *How large is the educational wage premium in Norway in 2014, and does it differ between heterogeneities in gender, residence and immigrational background?*

This is an interesting research question and analysis to make when considering if education pays off. One of the goals of the thesis is to help young people make the decision of whether or not to attend higher education, even though the wages should not be the sole motivation for occupational choices.

1.4 Hypotheses

This paper will test several hypotheses. Testing hypotheses is the main purpose of statistics as a field. When the analysis begins, it is in fact the null hypothesis that is being tested. This method is based on Karl Popper's falsification theory, which the majority of scientific publications still are using (Wilkinson, 2013).

The null hypothesis is an additional hypothesis to the scientific hypotheses, where the null hypothesis is refusing that such an effect presented in the scientific hypotheses exists. The scientific hypotheses, or also known as alternative hypotheses, are stated as the following:

> H_1: An educational wage premium does indeed exist, where education has a positive, significant effect on wages.
>
> H_2: The expected wage always increases as the education level increases. Hence, a PhD. graduate expects a higher wage level than a master graduate.
>
> H_3: The educational effect on wages is positive but decreasing.

The contradictory null hypothesis, H_0, is defined as "H_i is not true", where $i \in [1, 2, 3]$.

So statistically, what is done is a test of the results from the analysis and then an evaluation of either keeping or rejecting the alternative hypotheses. Meanwhile, it is important to note that keeping a hypothesis because it cannot be rejected, does not necessarily prove that the hypothesis surely is correct. It can only prove that in this specific analysis, there is not enough evidence to reject it so there is a chance for it to be true.

1.5 Disposition

The thesis is divided into a total of 6 chapters. Chapter 1 is the introduction and background for the research question. Chapter 2 presents the theoretical framework and briefly reviews a selection of previous literature. Chapter 3 reviews the data used in the analysis, with a presentation of the variables and their corresponding descriptive statistics. Chapter 4 presents he empirical strategy, choice of method, and challenges associated with these, before the results from the analysis are reviewed in Chapter 5. Chapter 5 also includes discussions regarding the results while comparing them to earlier literature. Finally, chapter 6 contains a brief summary and conclusion of the results found in the thesis.

2 Theory and literature

In this part of the paper, relevant literature and theoretical issues are discussed. This is important in order to better understand the results of the analysis and why the wage gap is an interesting topic. The first thing to look at, is relevant published literature which gives an insight into the educational wage premium. Further on, the Norwegian school system is presented in order to understand what kind of education the different levels actually contain. There will also be a discussion about ability, and lastly the non-pecuniary effects of education.

2.1 Relevant literature review

As mentioned in chapter 1.3, the academic topic of wage premiums already has a lot of relevant research published. In this section, some of the most interesting papers for this thesis are presented. Supplying the thesis with external literature within the same research field helps understand the background for the hypotheses and analysis, and supplements the findings in this thesis with a comparison to other results.

Multiple papers have already tried to calculate the size of educational wage premiums in earlier research. For instance, James (2012) looked at how the premium has changed from the 1970s to 2010 using data from the Bureau of the Census in the United States. One could see that in this time-period, the raw premium has increased from 40 percent to upwards of 70 percent for individuals with bachelor’s degree or higher. Individuals with some college education, but not a completed degree, had in the same time-period slightly or no growth in their premium (James, 2012). Earlier findings based on data in the UK, found that the college wage premium for males did not substantially change in the 1980s and mid-1990s. While women on the other hand, had an increase of 10 percentage points in their college wage premium (Walker and Zhu, 2008).

According to a study published in the American Sociological Review, the wage gap between college-educated and non-college-educated workers has increased significantly in recent decades. The study found that college-educated workers earn about 50% more on average than their non-college-educated counterparts (Goldin and Katz, 2007). Also, another research based on data from 1959 to 1996 found that the college premium for younger male workers in the United States and the United Kingdom has risen substantially, while the premium for older male workers is about the same today as it was in the mid-1970s (Card and Lemieux, 2001). The same paper also found that the college-high school wage gap in Canada has increased for younger male workers, but the gap for older Canadian men has declined. This trend has contributed to growing economic inequality and imposed challenges for workers with lower levels of education. For instance, it can undermine social mobility and make it more difficult for lower-educated individuals to move up the economic ladder (Council, 1990). Research has shown that lower-educated workers are more likely to be in poverty and to have inadequate access to healthcare and other important resources (Blank, 2009). Also, the wage gap can make it difficult for these workers to save for retirement, leading to increased economic insecurity in old age (Gustman, Steinmeier and Tabatabai, 2012). Additionally, the wage gap can also have negative effects on overall economic growth, as it can lead to a less efficient allocation of labour and can reduce the purchasing power of lower educated workers (Katz and Murphy, 1992).

It is reasonable to believe there are differences in the wage premiums between different educational majors, which the paper by James (2012) finds evidence to support. For instance, he finds that engineering majors have a college major premium of 125 percent, while psychology and social work majors have a college major premium of 40 percent. This yields a staggering 85 percentage points difference in the college major premium (James, 2012). Even though his paper relies on data from the U.S, it is reasonable to assume that differences in college premiums between majors occur in Norway as well. Unfortunately, it is too extensive for this thesis to compare the premiums between majors, so the analysis only covers the average wage premium for all majors. Further research is encouraged to study possible differences.

In earlier literature, there has been discussions about the importance of seeing the labour market as a market where supply and demand determines the equilibrium. The Nobel Prize-winning economist Jan Tinbergen was the first to note the persistently rising demand for educated labour in advanced economies. This is often referred to as an “education race” model, where the primary implication is that if the supply of educated labour does not keep pace with persistent outward shifts in demand for skills, the skill premium will rise. When the rising supply of educated labour began to slacken in the early 1980s, the economic consequence was an increase in the college skill premium (Autor, 2014).

The secular time trend has over a long time shown that more and more individuals attend higher education. From the start of the 20th century to the 21st, the global higher education students per 10 000 capita goes from around 1-2 in 1900 to more than 160 in 2000. This trend has been even higher in industrialized countries (Schofer and Meyer, 2005). By educational reforms, a country such as China increased its enrolment in the relevant age cohort from only 1.5% in 1978 to 27% in 2010 and further growth after that (Tan, 2013). The increase in enrolment does not seem to decrease the educational wage premium (Lindley and Machin, 2016). This result also corresponds to the results found in a paper from a study done with data from 1980 and 1990 in Norway, where the substantially higher level of educational attainment for more recent cohorts does not seem to have a negative effect on educational wage premiums for these younger cohorts (Hægeland, Klette and Salvanes, 1999). Also, even earlier research from Norway has found that the estimated returns to education are quite stable across Norwegian birth cohorts from 1942 to 1970 (Hægeland, 2001). It can be found quite interesting how a low supply of educated labour seems to increase the college skill premium, while a high supply does not seem to decrease the premium. A possible explanation for this is the importance of educated labour, as discussed in chapter 1.2.

While looking at how the wage gap can increase, it is also important to discuss some sources to decrease the wage gap. Wage gaps can have a wide range of sources, so by locating the sources of why the gap exists, one can also figure out how to increase or decrease the gap. Changes in the labour market, such as the increasing demand for skilled workers and the decline of unionization, can contribute to the wage gap (Autor, Katz and Kearney, 2008). It can be shown that increasing the minimum wage and providing support for workers to access quality education and training, can help to reduce the wage gap and improve economic security for lower-educated workers (Dube, Lester and Reich, 2010). Research has also shown that discrimination against women and racial minorities can play a significant role in the wage gap (Budig and England, 2001). Policies that aim to address discrimination, such as equal pay laws and anti-discrimination protections, can then help to reduce the wage gap (Crosby *et al.*, 2003).

To summarize the relevant literature included in this thesis, there are papers published which found evidence not only for the wage gap to exist, but also for it to be determined by educational level and the chosen major, and also for it to have increased over the last few decades. Papers have also found that the college wage premium is affected by the supply and demand for educated labour, and that there has been an enormous increase in individuals attending higher education over the past century. These findings are important to keep in mind during the discussion of the analysis, so some of them are mentioned again in chapter 5.

2.2 The Norwegian schooling system

Educational attainment is a crucial part of this thesis. In the analysis, it will show how different levels of education affects one's wages. An important prerequisite to interpret the results, is to understand how the Norwegian schooling system is structured. This subchapter explains the schooling system in depth and include a table with a timeline overview of the system.

The normal age to begin primary school in Norway is the year a child turns 6 years old. Primary school lasts for 7 years, from 1st to 7th grade. After primary school, the child will attend lower secondary school. This is usually the year the child turns 13 years old. Lower secondary school lasts for 3 years, from 8th to 10th grade. Both primary and lower secondary school is mandatory in Norway, which means the child both have the right, and the obligation to attend these (Utdanningsdirektoratet, 2022). The academic calendar year lasts from August to June, which results in some variation in age amongst the students depending on whether the student's birthday is before or after school starts in August.

After lower secondary school, most students enrol in upper secondary school from the year they turn 16 years old. Upper secondary school is voluntary, and the student stands free to choose between several programs. The programs can be categorized into two main directions. The first direction is programs for general studies. These programs provide the students general or specific university and college admission certification. These kinds of programs have a duration of 3 academic years (Utdanningsdirektoratet, 2022). The other direction is vocational education programs. This direction contains a range of different occupations, for example building and construction workers, hairdressers, florists, healthcare workers and so on. The student attains normal classroom education for a duration of 2-3 years before attaining an apprenticeship within a firm. Another possibility instead of an apprenticeship is to enrol in a one-year supplementary program for general university and college admission certification (Utdanningsdirektoratet, 2022). Usually, by attending a program for general or specific university and college admission certificate, one can expect to finish at 19 years old. For the vocational education programs, the expected completion age with a profession certificate after completing an apprenticeship is 20 years old.

From the structure of the schooling system, it follows that the youngest age a student can enter university or college is in general the year they turn 19. In university, there is a range of different degrees that are defined as college degrees (Moody, 2021). The standard duration for a bachelor's, master's or doctoral degree, is respectively three, five and eight years (Studenttorget, 2016). There are mainly two ways to accomplish a master's degree. One can complete a bachelor's degree first, which qualifies for a 2-year master's program. After the bachelor's degree is completed, the student then has to apply for the 2-year master's program, where usually an average grade of C or better is necessary to be able to apply. The alternative way is to apply directly on a 5-year integrated master's program, which is called an undergraduate master's degree (Bennet, 2022). With an undergraduate master's degree, the student is guaranteed to end up with a master's degree after five years, without having to worry about obtaining certain grades. Either way, the total duration of a master's degree is 5 years and regardless of whether a student attends an undergraduate master's or not, it will not affect the results of the analysis. In the table below, there is an overview of the different levels, grades and duration for the Norwegian school system.

Table 1: Overview of the Norwegian school system.

Grade	Name	Student's age*	Duration	Level	Comment
Below school duty	Preschool	0 – 6	Varies, depending on parents' situation	0	Normally kindergarten and day-care
1st – 7th	Primary school	6 – 13	7 years	1	Mandatory education
8th – 10th	Lower secondary school	13 – 15	3 years	2	Mandatory education
11th – 14th	Upper secondary school or Highschool	16 – 20	3 years	3 – 5	Upper secondary school is voluntary in Norway, but most pupils choose to attend it. Students can choose majors, where the duration typically is 3 years. For vocational subjects, the duration is 2 years schooling + 2 years apprentice.
14th – 17th	University bachelor	19 – 22	3 years	6	180 study points. Also known as undergraduate academic degree.
18th – 19th	University master	22 – 24	2 years	7	120 study points (300 in total). Also known as postgraduate academic degree.

20th +	University doctoral	24 – 27	3 years	8	The highest academic degree one can achieve. Includes both Ph.D. and DPhil.

*Since a calendar year lasts January to December, while a school year lasts August to June and spans over two calendar years, the age of students in the same class may differ by one year. The age given in the table is therefore the age the student will turn in the current calendar year.

Since there are some differences between how the school system is structured in different countries, knowledge about the Norwegian school system specifically is important to be able to understand the interpretation of the different variables used in the data. That is because some of the variables are defined by assumption made using the knowledge of the school system. For instance, at what age students are able to attend certain degrees and at what age they possibly can start working.

2.3 Ability vs. disability

In the literature of educational wage premium, there is an ongoing debate about ability and skills. The main challenge to achieve a completely unbiased estimate of the wage return of education, is to make sure no unobserved variable with correlation to the measurement of education is left out of the model. Since there are so many factors that contributes to determine one’s choices in life, it is unlikely that all of them are able to be captured in a single model. An example is how one’s skills or ability affects the choice of whether to attend higher education or not. This causes a violation of the Gauss-Markov assumption of zero conditional mean, and the implications of this is addressed in chapter 4.

It is important to clarify how this paper defines skill and ability. These are two terms which in literature often are used interchangeably, but which are used for different matters in this paper. Hence, it is important to make a clear distinction between the two terms, so that confusion is avoided when discussing the results of the analysis. The skills debate will follow in chapter 4, when discussing omitted variable bias.

When referring to ability, it is thought to as a binary state of an individual either being disabled or able-bodied. Discussions about ability and disability has been going on since the time of Aristotle and Kant (Reynolds, 2019). Lack of ability does not necessarily mean lack of skills. For example, if an individual who works as an accountant lacks the ability to walk, it does not impact his skills as an accountant. He might just have the need to use a wheelchair but can still do his job equally as good as, or even better than, someone who possess the ability to walk. Hence, an individual’s ability is not necessarily correlated specifically with his tasks at work. But this obviously depends on the line of work and the interests of the disabled individual. If the individual who lacks the ability to walk wants to become a mailman instead of an accountant, the disability to walk can probably prevent him or her to obtain the desired job position. Research has found that some disabled people experience barriers to securing and maintaining employment, where different disabilities bring different barriers. This also depends on the severity of the disability (Lindsay, 2011). Another research found that people without hearing disabilities have a higher wage premium to education than people suffering from hearing disability (Benito, Glassman and Hiedemann, 2016). While this is not necessarily representative for all types of disabilities, Benito et al. (2016) also finds that people with severe eyesight disability and wheelchair users have lower employment rate than people with severe hearing disability. Statistics Norway (SSB) has reported a positive effect of educational level on employment, by people with mobility disability (Karlsen, 2022). Based on the findings in the research mentioned, it would be interesting to control for ability in this analysis. The educational wage premium is expected to be either under- or overestimated as a result of omitting ability. Unfortunately, due to lack of data in microdata.no, there is no good proxy available to use as a control variable. Therefore, it will not be controlled for, but it is still worth a discussion around the implications of omitting relevant variables. Chapter 4 contains a discussion regarding skills and further addresses the consequences of the omitted variable bias. Hopefully, future research can control for ability if microdata.no eventually obtains a good proxy for it.

2.4 Non-pecuniary effects of education

While this paper is looking into how education affects wages, education can also have a range of other effects on people's lives that are not necessarily directly economically related. The question is whether these secondary effects can indirectly also affect one's wages.

Oreopoulos and Salvanes (2011) finds a wide range of non-pecuniary effects both within and outside of the labour market. The first one to notice is that higher educated individuals are less unemployed than lower educated individuals. Higher educated individuals also have a higher satisfaction rate at work as well. Another earlier study also found a strong positive effect on intrinsic work value for both genders, in addition to higher satisfaction at work (Mottaz, 1984).

The second part is returns appearing outside the labour market. A larger share of higher educated individuals reports to have “very good health”. High-educated individuals believe to a higher degree that people can be trusted and are less in favour of spanking to discipline their child's (Oreopoulos and Salvanes, 2011). The same study shows that lower educated individuals are more likely to ever try smoking, get arrested and have their first child in their teenage years.

All these different variables could be interesting to control for in the analysis, but two main issues make this infeasible for the thesis. The first is based on the limitations given by microdata.no. Variables such as smoking and self-reported health perceptions do not currently exist in microdata.no. The second reason is based on an evaluation of whether the effects are spurious rather than causal. Sometimes, two phenomena can appear to correlate but in reality, does not have any causality (Kenton, 2021). Which means it can be hard to tell if an existing relationship between smoking and wages is just associated, rather than caused by a causal effect from the education level.

Research has found that smoking has a significant negative effect on wages, even when controlling for education (Bondzie, 2016). Interestingly, the study further reveals that this smoking wage penalty is only found by males. Combining the two findings, it can be seen as an additional effect of education.

3 Database

The database used for empirical analyses is a crucial part to determine the results found in the analyses. This chapter is dedicated to give an insight in the database, the analytical program that is used, and to explain and define the variables. The basis of the analysis is registry data from microdata.no, and a cross-section dataset is used on a sample of about 1 250 000 individuals from Norway in 2014. Since the educational level of a full-time worker is considered to not change over time, it is adequate to use cross-sectional data from a given year in the analysis rather than panel data over several years. After an individual has taken higher education, this will be constant over all following years to come, unlike variables that change from year to year. The first subchapter will start off by explaining how the platform microdata.no has been used and some implications and limitations it has brought along. Following, will also all the variables used in the analysis and associated descriptive statistics be presented.

3.1 Microdata.no

The database used for this thesis is the platform microdata.no. Microdata.no is a relatively new research platform first published in 2018, that is operated and developed through a collaboration between the Norwegian Centre for Research Data (NSD) and Statistics Norway (SSB).

Microdata.no provides raw data for 414 different variables, some of which dating back as far as 1964. The platform gives users access to annual demographic and socioeconomic microdata on almost 11 million people, including income, education, employment, and social security benefits on individual level. All people in the population who have ever had a permanent or temporary Norwegian personal identity number are included in the database, although direct personal identification has been eliminated to protect people's privacy (Johansen, 2020).

Macro- and microeconomics are frequently distinguished in economics, and the similar distinction may be made for economic data. While macro-level data examines the overall picture on a national scale, micro-level data looks more intently on an individual level (Bagdasarian, 2018). One can measure the same subject on both levels. For instance, a dummy-variable that is equal to 1 if a person is employed or 0 if unemployed might be used to quantify employment at the micro-level. The (un)employment rate of a particular area, such as a municipality, region, or country, could be a macro-level measure describing the same problem. Microdata.no provide access to a special direct online dissemination of micro-level data. Although the data is collected on an individual basis, it appears as metadata to the users.

The data used in the analysis consists of 43 variables for 1 268 018 individuals. The following chapter will address all the variables used and how they are defined. To ensure the privacy of individuals, the platform is as mentioned above metadata-driven, where raw data cannot be directly accessed or viewed. Only through numerous functionalities, including descriptive statistics and regression analysis, are descriptions of the data attainable. Privacy and confidentiality are further maintained through several different measures. It is not possible, for example, to define a population size of fewer than 1000 individuals, and a 2% winsorization is applied to the raw data. This means that the highest values are set equal to the value of the 99th percentile, while the lowest values are set equal to the 1st percentile. This results in the most extreme observations having no impact on the descriptive results. The distribution becomes less skewed, but consequently, the mean and standard deviation will be somewhat incorrectly estimated, typically underestimated. The winsorization will not affect the regression analyses, which only use the underlying non-winsorized data because the results from these analyses are not considered personally identifiable information. In addition to winsorization, the descriptive data is noisified by ±5 individuals, and this noise is constant and stochastic with an expected value of zero. Due to the analysis's large number of observations, this will not have a substantial impact.

3.2 Data limitations and challenges

This section will address some of the challenges this thesis has been facing due to the use of microdata.no. As mentioned, it is not possible to view or browse the actual data in the program, and one cannot download the data, which means that all analyses must be done directly in the program. The results from these analyses are then exported to Excel for further manual processing.

The selection of relevant variables available in the program for this analysis is also somewhat limited. There are currently 414 available variables in microdata.no, but not all of them are interesting to include in a wage premium analysis. The variables vary in validity period, where some goes as far back as to 1964 and ranges to 2021, while others only date back to

2015. This resulted in a trade-off between relevant variables to use, where the year that had the most variables overlapping in validity time was selected. This ended up being the year 2014.

There are some additional variables that would be desirable to include in the analysis but are not available. For example, it would likely have been rewarding to control for different firms within the same industry. Especially in the private sector, wage determination can vary between firms. A firm's profitability and wealth can contribute to determining policies regarding wages, which is uncorrelated with the individual factors controlled for in the analysis. Microdata.no actually has a variable for organization number, so an attempt was made to control for this in the analysis. But, because of privacy policy limitation this variable is a pseudonym variable which cannot be used for statistical operations, so it had to be left out.

It is also not possible to link external data to the program yet, and this, combined with the limited availability of relevant variables, makes it virtually impossible to find good proxy and instrumental variables for the analysis. This is one of the biggest challenges the thesis is facing. As a result, it is not possible to account for actual work experience, IQ, health (such as whether an individual is a smoker, heavily obese, etc.), and more.

The last notable weakness of the program is that there are limited functionalities. For example, it is not possible to extract pre-made tables like the ones in Stata using the command "outreg2" or "asdoc". One can only extract tables and results to Excel, so all tables have to be made manually. It is also not possible to save regression coefficients or residuals as a new variable, as one can do in Stata by the command "predict".

3.3 Delimitation of the sample

Initially, the population was 2 979 272 observations when the variables were imported. When removing the observations with missing data on education, 90 902 observations were removed from the sample. By implementing the restriction of only including the observations with more than the estimated minimum wage of 286 200 NOK annually, another 1 045 032 observations were removed (KarriereStart.no, 2014). Most of these are assumed to be children and pensioners. Additionally, by restricting the age to the interval 18 to 62 years, i.e. to legal individuals younger than retirement age, 97 717 observations were removed. Then, by limiting the data to only include individuals with full time contracts, 180 606 observations were also removed. Furthermore, everyone with a lower educational level than high school were removed, yielding 206 726 observations. Lastly, every individual with higher education but no completed degree has also been removed. This resulted in 89 848 observations being removed from the sample.

With all of these restrictions implied, a total of 1 710 831 have been removed from the sample, leaving a total of 1 268 018 observations in the dataset. Every restriction implied has been for a reason. Most of them to simplify the interpretation and making the analysis cleaner. The analysis can now compare individuals with a degree from higher education with individuals with high school as their highest completed education, with a dataset that only includes individuals with full time contracts, between the age of 18 and 62 who earns more than the minimum wage.

3.4 Variables and descriptive statistics

In this part, a table with full descriptive statistics are shown for all the variables used in the models. As mentioned in the previous chapter, the dataset contains a total of 43 variables, whereas 32 are control variables, one is the explained variable, and 8 are explanatory variables, and 2 are instrumental variables. Some of which are self-generated variables such as dummies, logarithmic variables, and interaction terms, so it might be useful to look at the attached script in appendix A1 for insight in the coding. There will also be a table showing every dummy variable’s distribution in appendix A21. In the following part, there is a description and interpretation of the values for each of the 43 variables.

Table 2: Descriptive statistics for all variables, plus age, for the full sample.

VARIABLES	OBS.	AVERAGE	STD. DEV.	1%	50%	99%
WAGE	1 268 018	586 368	246 353	298 229	518 122	1 693 375
LNWAGE	1 268 018	13.2138	0.3507	12.61	13.16	14.34
UTDNIV_HIGH	1 268 018	0.47764	0.4995	0	0	1
UTDNIV_VGS	1 268 018	0.52236	0.4995	0	1	1
UTDNIV_VGS_2Y	1 268 018	0.07165	0.2579	0	0	1
UTDNIV_VGS_3Y	1 268 018	0.40022	0.4899	0	0	1
UTDNIV_VGS_4Y	1 268 018	0.05050	0.2190	0	0	1
UTDNIV_BACHELOR	1 268 018	0.33739	0.4728	0	0	1
UTDNIV_MASTER	1 268 018	0.12770	0.3338	0	0	1
UTDNIV_PHD	1 268 018	0.01255	0.1113	0	0	1
MOTHEDUC	1 268 018	0.15918	0.3658	0	0	1
FATHEDUC	1 268 018	0.19443	0.3958	0	0	1
MALE	1 268 018	0.57149	0.4949	0	1	1

IMMIGRANTS_HIGHINC	1 268 018	0.06420	0.2451	0	0	1
IMMIGRANTS_LOWINC	1 268 018	0.04154	0.1995	0	0	1
MSTAT_UNMARRIED	1 268 018	0.39588	0.4890	0	0	1
MSTAT_MARRIED	1 268 018	0.49302	0.5000	0	0	1
MSTAT_DIVORCED	1 268 018	0.11110	0.3143	0	0	1
SECTOR_PRIVATE	1 268 018	0.58856	0.4921	0	1	1
OSLO	1 268 018	0.31849	0.4659	0	0	1
BIG_CITY	1 268 018	0.24027	0.4272	0	0	1
AGE	1 267 988	42.5	10.7	22	43	62
POT_EXP	1 267 988	21.8	11.0	2	22	43
POT_EXP_SQ	1 267 988	596	500	4	484	1 849
INDU_REF	1 268 018	0.10679	0.3088	0	0	1
INDU_1	1 268 018	0.00000	0.0000	0	0	0
INDU_2	1 268 018	0.03820	0.1917	0	0	1
INDU_3	1 268 018	0.00000	0.0000	0	0	0
INDU_4	1 268 018	0.00000	0.0000	0	0	0
INDU_5	1 268 018	0.08849	0.2840	0	0	1
INDU_6	1 268 018	0.10603	0.3079	0	0	1
INDU_7	1 268 018	0.04945	0.2168	0	0	1
INDU_8	1 268 018	0.01390	0.1171	0	0	1
INDU_9	1 268 018	0.04492	0.2071	0	0	1
INDU_10	1 268 018	0.02645	0.1605	0	0	1
INDU_11	1 268 018	0.00000	0.0000	0	0	0
INDU_12	1 268 018	0.06229	0.2417	0	0	1
INDU_13	1 268 018	0.03608	0.1865	0	0	1
INDU_14	1 268 018	0.08223	0.2747	0	0	1
INDU_15	1 268 018	0.09454	0.2926	0	0	1
INDU_16	1 268 018	0.17780	0.3823	0	0	1
INDU_17	1 268 018	0.01000	0.0995	0	0	1
INDU_18	1 268 018	0.01523	0.1225	0	0	1
INDU_19	1 268 018	0.00000	0.0000	0	0	0
INDU_20	1 268 018	0.00000	0.0000	0	0	0

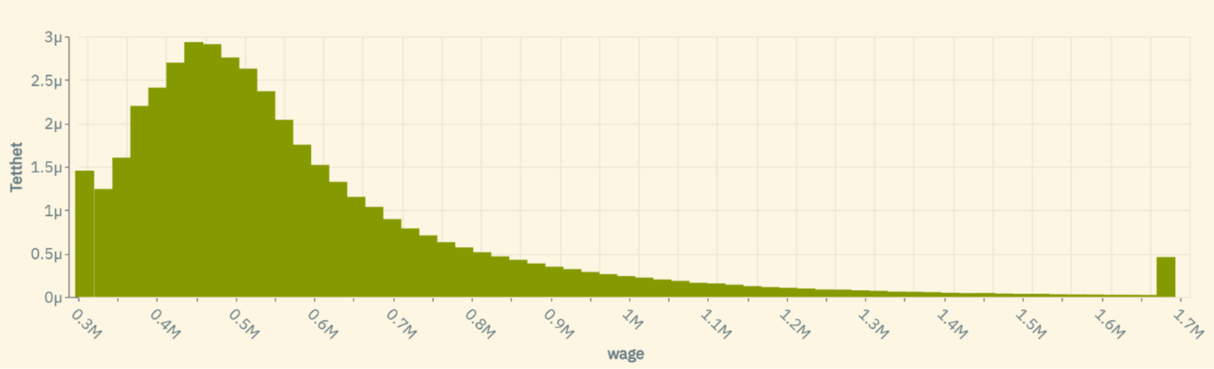
Two similar tables can be found in appendix A2 and A3 with descriptive statistics separating higher educated individuals and lower educated individuals.

3.4.1 The explained variable

Wage

lnwage is the explained variable, which is the individual's wage on logarithmic form. This makes the interpretation in the analysis easier, as one looks at percentage changes in the wage. The wage variable is numeric and includes cash salary, taxable fringe benefits, and sickness and maternity benefits during the calendar year. The average wage in the entire dataset is 586 368 NOK. Meanwhile the average wage amongst higher educated individuals is 634 467 NOK and amongst lower educated individuals is 543 523 NOK. As mentioned in chapter 3.2, every individual with less than minimum wage of 286 200 NOK according to KarriereStart.no or without a full-time contract has been removed from the dataset. Below is an illustration of the wage distribution for the entire dataset.

Figure 1: Wage distribution for the entire sample.



The wage distribution for only higher educated individuals and for only lower educated individuals can be found in appendix A4 and A5, respectively.

3.4.2 Explanatory variables

The explanatory variable when looking at an educational wage premium is obviously education. Higher and lower educated individuals have been separated by defining high school (upper secondary school) as lower educated, and university/college as higher educated. Everyone with less than high school education as their highest finished education, have been removed from the dataset.

High school

High school is the reference category for education in the main models and is defined as lower educational level in this thesis. This is a dummy named *utdniv_vgs*, which is equal to 1 if the individuals have high school education as their highest finished education, and 0 if not. The dummy has also been separated into three new dummy variables: one for two years, one for three years and one for four years of high school education. The separation is done to show how individuals are distributed between the different durations of high school. When higher education is included in the model, the interpretation of the education coefficient is the percentage difference in wages amongst higher educated individuals compared to lower educated individuals. Because of how high school works in Norway, as explained in chapter 2.2, there are three different variables for high school. *utdniv_vgs_2y* is a dummy equal to 1 for everyone who has one or two years of high school as their highest finished educational degree, and 0 if not. *utdniv_vgs_3y* is a dummy equal to 1 for everyone who has three years of high school as their highest finished educational degree, and 0 if not. *utdniv_vgs_4y* is a dummy equal to 1 for everyone who has four years of high school as their highest finished educational degree, and 0 if not. In the dataset, 52.24% of the individuals has high school as their highest finished educational level. Amongst them, 13.72% has 2 years of high school, 76.61% has 3 years of high school and 9.67% has 4 years of high school. In the total dataset this yields *utdniv_vgs_2y* to include 7.17% of the individuals, *utdniv_vgs_3y* to include 40.02% of the individuals and *utdniv_vgs_4y* to include 5.05% of the individuals.

University/college

There is one general college variable which is a dummy equal to 1 called *utdniv_high* for everyone with either a completed bachelor's, master's or doctorate's degree, and 0 if not. This covers 47.76% of the individuals in the dataset. The variable is used as the explanatory variable for some of the models, but it is also separated into an own variable for each of the different degrees. *utdniv_bachelor* is a dummy equal to 1 for everyone who have completed a bachelor's degree as their highest completed education, and 0 if not. *utdniv_master* is a dummy equal to 1 for everyone who have completed a master's degree as their highest completed education, and 0 if not. *utdniv_phd* is a dummy equal to 1 for everyone who have completed a doctorate's degree as their highest completed education, and 0 if not. Amongst the higher educated individuals in the dataset, 70.64% has a bachelor's degree, 26.73% has a master's degree and 2.63% has a doctorate degree as their highest finished educational level. In the total dataset, this yields *utdniv_bachelor* to include 33.74% of the individuals, *utdniv_master* to include 12.77% of the individuals, and *utdniv_phd* to include 1.26% of the individuals.

3.4.3 Control variables

The analysis includes a set of control variables to minimize issues with omitted relevant variables that lead to endogeneity if they affect individuals' wages while also affecting the included explanatory variables.

Gender

The gender wage gap has been analysed in an increasing number of longitudinal studies, where for instance Kunze (2005) analyses the male-female wage differential during the early career covering the period 1975 to 1990, while Blau and Kahn (2017) examines the gender pay gap in the United States in the period 1980 to 2010. Considering the wage differences that earlier literature has found, it can be argued that gender needs to be controlled for in order to obtain a more precise estimate of the educational wage premium in the analysis. *male* is a dummy equal to 1 if the individual is male and 0 if the individual is female. In the dataset there is 57.15% males and 42.85% females. Research has shown that in Norway, a relatively high share of females works part-time jobs compared to males (Bø, 2004). Since the sample is limited to only include individuals with full-time contracts, it is expected to be a majority of males in the total sample. Amongst only the higher educated individuals, there is 45.35% males and 54.65% females, and amongst the lower educated individuals there is 67.93% males and 32.06% females.

Marital status

Marital status is the status in relation to marriage legislation and has been separated into three dummy variables. *mstat_unmarried* is a dummy equal to 1 if the individual’s marital status is unmarried, and 0 if not. This is the reference category for marital status in the models. *mstat_married* is a dummy equal to 1 if the individual's marital status is married, widow, registered partner or alive partner, and 0 if not. *mstat_divorced* is a dummy equal to 1 if the individual's marital status is divorced, divorced partner, separated or separated partner, and 0 if not. In the dataset, 39.59% of the individuals are unmarried. 49.3% of the individuals are either married, widow, registered partner or alive partner, and 11.11% of the individuals are divorced, separated, separated partner or divorced partner. Amongst the higher educated individuals, 38.68% are unmarried, 51.61% are married, widow, registered partner or alive partner, and 9.72% are divorced, separated, separated partner or divorced partner. Amongst the lower educated individuals, 40.42% are unmarried, 47.20% are married, widow, registered partner or alive partner, and 12.38% are divorced, separated, separated partner or divorced partner.

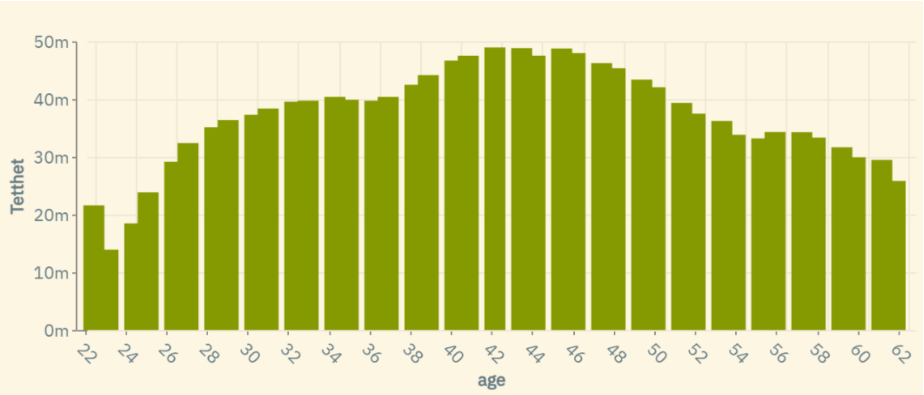
Immigrants

Immigrants have been defined according to the categorization in microdata.no, where the categories “immigrants” and “foreign-born with one Norwegian-born parent” are defined as immigrants in this thesis. If the individual is within the categories “Born in Norway with two Norwegian-born parents”, “Norwegian-born with immigrant parents”, “Norwegian-born with one foreign-born parent”, or “Foreign-born with two Norwegian-born parents”, the individual is considered a native. When controlling for immigrants, two variables have been used combined. The first one is a variable that shows different combinations of own or parents' country of birth. This one has been used to determine who should be categorized as immigrants, and who should be categorized as natives. The other variable is a variable that shows which country the individual is born in, if they are born abroad. By combining these two variables, it is possible to separate between where the immigrants immigrate from. By using an article from the *World Economic Situation and Prospects (WESP)* published in 2014 to classify high-income and low-income countries, immigrants have been divided into two dummy variables. *immigrants_highinc* is a dummy equal to 1 if the individual is an immigrant born in a high-income country, and 0 if not. *immigrants_lowinc* is a dummy equal to 1 if the individual is an immigrant born in a low-income country, and 0 if not. The reason immigrants are separated between high-income and low-income immigrants, is because earlier literature has found differences between these groups, and especially for those immigrating from one high-income country to another (Barstad, 2013). Since Norway is classified as a high-income country according to the United Nations, it is reasonable to assume these differences may occur in this research as well (Nations, 2014). The definition of a high-income versus low-income country in this thesis, is the United Nations classification of developed versus underdeveloped countries (Nations, 2014). The term "high-income country" is used for the developed countries, whereas "low-income countries" are the rest of the world, which includes underdeveloped countries, developing countries, and countries in transition. In the dataset there is 10.57% immigrants, whereas 6.42% are from high-income countries and 4.15% are from low-income countries. Amongst higher educated individuals, there are 11.07% immigrants, whereas 6.32% are from high-income countries and 4.75% are from low-income countries. Amongst lower educated individuals, there are 10.12% immigrants, whereas 6.51% are from high-income countries and 3.61% are from low-income countries.

Potential experience

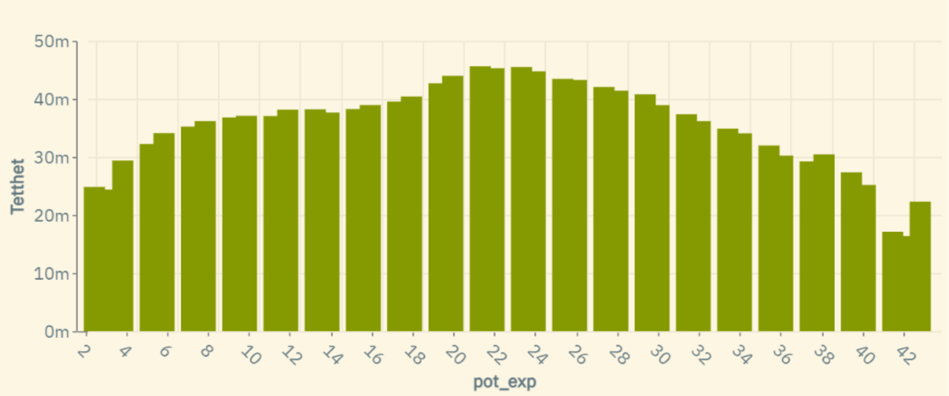
Potential experience, *pot_exp*, is a self-generated variable using the individual's age and an assumption that every individual finished their education continuously without breaks. The inclusion of this variable is necessary considering how actual work experience is crucial for determining wages. Meanwhile, microdata.no lacks variables related to actual work experience, so a proxy is the closest the analysis gets. If the individual has 2 years of high school as their highest finished education, it is expected that they finish school when they are 18 years old. This is the youngest age they can graduate, and they can then potentially start working and gain work experience. So, the formula for potential experience is the individual's age subtracted the expected/youngest age they can normally finish their education. For 3 or 4 years of high school, the expected age to graduate is respectively 19 and 20 years old. For higher educated individuals, the expected age to graduate is 22 years for bachelor's degree, 24 years for master's degree and 27 years for doctorate's degree. The potential experience for the individuals in the dataset ranges from 0 to 44 years, with an average of 21.79 years. The average age for the individuals in the dataset is 42.53 years. Potential experience squared, *pot_exp_sq*, has also been included to see if the return of potential experience is decreasing. Below follows some illustrations of both the age distribution as well as the distribution of potential experience in the entire sample, as well as separated for higher and lower educated individuals.

Figure 2: Distribution of age for the full sample.



The dataset is limited to individuals between 18-62 years, whereas there are so few full-time workers in the age range 18-21 years that the graph seems to start at 22 years of age. Still, there are 50 individuals that are 19 years, 3125 that are 20 years and 6205 that are 21 years in the dataset.

Figure 3: Distribution of potential experience for the full sample.



The distribution of potential experience amongst only lower and only higher educated individuals, can be found in appendix A15 to A18.

Place of residence/work region

Research has found evidence of a city wage premium to exist, meaning there are differences in wages between individuals' resident in urban and rural areas. For instance, Yankow (2006) finds that two-thirds of the premium can be explained by cities attracting workers of higher unmeasured skills and ability, while the remaining wage premium is shown to consist of both level and growth elements. The wage level effect is consistent with a productivity advantage for firms located in cities, while the wage growth effect is shown to relate in part to a cumulative advantage in the returns to job mobility for urban workers. Also Gould (2007) and Carlsen, Rattsø and Stokke (2016) finds evidence of an existing urban wage premium. Hence, it can be argued that residential area needs to be controlled for in the analysis based on the earlier literature. To control for this, there are two dummy variables included, according to labour market region classification from Statistics Norway (Bhuller, 2009). The work on the division of labour market regions is mainly based on commuting statistics for Norwegian municipalities for the period 2000-2006. The first dummy, *oslo*, is for the labour market region of Oslo and equals 1 if the individual is inhabitant in this region, and 0 if not. This region does not only include the municipality of Oslo, but a total of 51 municipalities including places like Bærum, Asker, Drammen and more. The second dummy, *big_city*, is a dummy equal to 1 if the individual is inhabitant in the labour market region of either Bergen, Trondheim, or Stavanger, and 0 if not. These does also not only include the cities themselves, but a total of 72 municipalities within these labour market regions. In the sample, 31.85% lives within the labour market region of Oslo, 24.03% lives within the labour market region of either Bergen, Trondheim or Stavanger, and 44.12% lives in other regions of Norway. For higher educated individuals, 37.77% lives in the labour market region of Oslo, 24.00% in the labour market region of either Trondheim, Bergen or Stavanger, and 38.24% in other regions of Norway. For lower educated individuals, 26.44% lives in the labour market region of Oslo, 24.05% in the labour market region of either Trondheim, Bergen or Stavanger, and 49.51% in other regions of Norway.

Sector

Sector is included as a dummy named *sector_private* which is equal to 1 if the individual works in private sector, and 0 if the individual works in public sector. The public sector is defined as all state or municipal enterprises or companies. The variable indicates the institutional sector of the enterprise where the individual has their main employment. This sector grouping is based on systems developed by the UN and the EU. In the sample, 58.86% of the individuals work in private sector, while 41.14% work in public sector. Amongst higher educated people, 43.17% work in private sector, and 56.83% work in public sector. Amongst lower educated people, 73.20% work in private sector, and 26.80% work in public sector. There are often differences in wages between private and public sector. Research based on microdata from Italy, France and Great Britain has found that public sector pays more to low skilled workers with respect to private sector, while private sector pays more to high skilled workers with respect to public sector (Lucifora and Meurs, 2006). Meanwhile, research from Norway found the existence of a wage premium in private sector for both lower educated and higher educated individuals (Rattsø and Stokke, 2020). The wage difference between sectors makes it an interesting and important variable to control for.

Industry

As research has shown, the college wage premium varies between different college majors (James, 2012). To control for this, 20 industry dummies have been introduced to capture wage differences between different industries according to classification from Statistics Norway. The reference category, *indu_ref*, is a dummy equal to 1 for if the individual works in manufacturing, and 0 if not. This yields 10.68% of the individuals in the sample. Below is a table showing all the industry dummies and their description.

Table 3: Overview of industry dummies and their description.

VARIABLE NAME	INDUSTRY DESCRIPTION	SHARE OF DATASET
indu_ref	Manufacturing and other industries.	10.68%
indu_1	Agriculture, forestry, and fishing.	0.61%
indu_2	Mining and quarrying.	3.82%
indu_3	Electricity, gas, steam, and hot water supply.	0.89%
indu_4	Water supply, sewerage, and waste management.	0.59%
indu_5	Construction.	8.85%
indu_6	Wholesale and retail trade, and repair of motor vehicles.	10.60%

indu_7	Transportation and storage.	4.95%
indu_8	Accommodation and food service activities.	1.39%
indu_9	Information and communication.	4.49%
indu_10	Financial and insurance activities.	2.65%
indu_11	Real estate activities.	0.83%
indu_12	Professional, scientific, and technical activities.	6.23%
indu_13	Administrative and support service activities.	3.61%
indu_14	Public administration and defence, and compulsory social security.	8.22%
indu_15	Education.	9.45%
indu_16	Human health and social work activities.	17.78%
indu_17	Arts, entertainment, and recreation.	1.00%
indu_18	Other service activities.	1.52%
indu_19	Activities of households as employers.	0.0029%
indu_20	Activities of extraterritorial organizations and bodies.	none
Total		**≈ 100% (98.1629%)**

A similar table can be found in appendix A6 and A7, which separates the industry distribution between higher and lower educated individuals. There are some rounding errors which makes the total only yield only 98.16%. It is important to note that when the industry dummies are included, sector is also implicitly controlled for. So, for the models where these are included, *sector_private* is intentionally left out.

3.4.4 Instrumental variables

When testing for robustness in chapter 5, some instrumental variables are included to remove non-random measurement error of an individual's educational level. Research suspects a correlation between education and the error term, due to parents' influence on a child's educational attainment. This makes parent's education level seem like a good instrument for education and an IV regression is conducted in chapter 5.4 using this as instrument for education.

Parent’s education

Parent’s education is separated into two instrumental variables, *motheduc* and *fatheduc*. *motheduc* is the induvidual’s mother’s highest finished education level when the individual was 16 years old. *fatheduc* is the individual’s father’s highest finished education level when the individual was 16 years old. Both of these instruments are dummies equal to 1 if the parent have higher education, and equal to 0 if not. In the sample, 26.03% have some parent with higher education. 6.59% of the individuals have only a mother with higher education, 10.12% of the individuals have only a father with higher education, and 9.33% of the individuals have both parents with higher education. Amongst the higher educated individuals, 39.07% have some parent with higher education. 8.62% have only a mother with higher education, 14.19% have only a father with higher education, and 16.25% have both parents with higher education. Amongst lower educated individuals, 14.12% have some parent with higher education. 4.74% have only a mother with higher education, 6.39% have only a father with higher education, and 2.99% have both parents with higher education.

4 Methodology

This part of the thesis presents the model specification and methodological framework used in the analysis. Knowledge about the method is useful as a reader to better understand the limitations and pitfalls of the analysis. Some econometric challenges are also briefly discussed towards the end of the chapter.

4.1 Ordinary least squares

The ordinary least squares (OLS) estimation method is the most common of the three main approaches to linear least squares (LLS) models and is the main analytical tool for this thesis. Linear least squares is one of two categories within the estimation method of least squares, whereas the other category is nonlinear least squares. The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems by minimizing the sum of the squares of the residuals made in the results of each individual equation (Lumivero, 2017).

By using OLS, it allows to quantify a relationship between two variables and determine if it results from a causal relationship (Bailey, 2020). If OLS is broken down to its simplest form, one can say it consists of one dependent variable and some independent variables that explains the dependent variables value. A linear regression is based on algorithms that provides a linear relationship between the dependent and independent variables (Kanade, 2022).

There are generally two categories within OLS, single linear regression (SLR) and multiple linear regression (MLR). The difference only depends on whether there is only a single explanatory variable (SLR) or if there are multiple explanatory variables (MLR), as the respective names reveal. The single linear regression model is also called a bivariate model, and can be written on mathematical form as following:

$$(1.1)\ y = \beta_0 + \beta_1 x + u$$

Where y is the dependent variable, x is the independent variable, β_0 is the intercept, β_1 is the independent variable's coefficient, and u is the residual. This is called the population regression function. If more than one independent variable are added, the model is expanded to a multiple linear regression model, with its population regression function looking like:

(1.2) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$

Where *k* denotes the total number of independent variables.

There are a set of assumptions that need to hold for the OLS models to provide unbiased results, called the classical linear model (CLM) assumptions. The CLM assumptions contain all of the Gauss-Markov assumptions, which is MLR.1 through MLR.5, plus the assumption of a normally distributed error term, MLR.6 (Wooldridge, 2019).

MLR.1 Linearity in the model.

MLR.2 Random sampling.

MLR.3 No perfect collinearity, meaning enough variation in *x*, the explanatory variable.

MLR.4 Zero conditional mean, meaning there should be no correlation between the regressor and the error term. Mathematically: $E(u_i|x_i) = 0$ or $Cov(u_i, x_i) = 0$.

MLR.5 Homoscedasticity, meaning the variance in the error term stays constant. I.e. as *x* increases, the error term stays the same. Mathematically: $V(u_i|x_i) = \sigma^2$, which is a constant.

MLR.6 Normality. The population error is independent of the explanatory variables and is normally distributed with zero mean and variance of σ^2. Mathematically: $u \sim \text{Normal}(0, \sigma^2)$

Note that if assumption 6 holds, 4 and 5 implicitly holds as well since it would be impossible to have a normally distributed error term if the error term were correlated with any of the *x*'s, either in the error terms mean value or in the error terms variance. If these assumptions hold, the model can be considered BLUE (Best Linear Unbiased Estimator). Yet, assumption 4 rarely holds, including in this thesis. Because of limited available data, there will always be some omitted variables (i.e. included in the residual) which correlates with one or more of the explanatory variables. This is further discussed in chapter 4.3.2.

4.2 Model specification

In chapter 4.1, the general population regression functions for the SLR and MLR model were introduced. In this chapter, the sample regression functions used in the analysis of this thesis are specified. The sample regression function looks somewhat the same as the population regression function, except it is specified specifically with the variables used in the dataset.

The models used in the analysis, are based on these sample regression functions. Generally, for all the functions will subscript *i* indicate individual *i* because the data is cross-sectional on individual level. Furthermore, the explained variable will always be wage on logarithmic form, the variable *lnwage*.

For model (1), a single linear regression is run with the dummy *utdniv_high* as explanatory variable and high school graduates as reference category. The sample regression function looks the following:

(1.3) $lnwage_i = \beta_0 + \beta_1 utdniv\_high_i + u_i$

This is the simplest model that is used in the analysis, and it excludes all control variables and does not categorize higher educated individuals by degree. Model (2) will add the control variables to model (1), and the sample regression function looks the following:

(1.4) $lnwage_i = \beta_0 + \beta_1 utdniv\_high_i + \delta_j \boldsymbol{x}_i + \gamma_n \boldsymbol{b}_i + u_i$

where $\boldsymbol{x}$ and $\boldsymbol{b}$ are two vectors with control variables to simplify the equation, and *j* and *n* indicates the different elements within the vector. The vectors are defined as following:

$\boldsymbol{x}$ = (*male*, *oslo*, *big_city*, *pot_exp*, *pot_exp_sq*, *immigrants_highinc*, *immigrants_lowinc*, *mstat_married*, *mstat_divorced*)

$\boldsymbol{b}$ = (*indu_1*, *indu_2*, *indu_3*, *indu_4*, *indu_5*, *indu_6*, *indu_7*, *indu_8*, *indu_9*, *indu_10*, *indu_11*, *indu_12*, *indu_13*, *indu_14*, *indu_15*, *indu_16*, *indu_17*, *indu_18*, *indu_19*, *indu_20*)

Model 3 will not include any control variables, similarly to model (1), but in this model the different college degrees amongst higher educated individuals are separated.

(1.5) $lnwage_i = \beta_0 + \beta_1 utdniv\_bachelor_i + \beta_2 utdniv\_master_i + \beta_3 utdniv\_phd_i + u_i$

One can see here how the educational wage premium differs between college degrees. This model is still too simple due to lack of control variables. So, the control variables are added to achieve the main model, with the corresponding sample regression function:

(1.6) $lnwage_i = \beta_0 + \beta_1 utdniv\_bachelor_i + \beta_2 utdniv\_master_i + \beta_3 utdniv\_phd_i + \delta_j \boldsymbol{x}_i + \gamma_n \boldsymbol{b}_i + u_i$

Because there are so many industry dummies, and a sector dummy (*sector_private*) is used for some of the models in the analysis, the industry dummies got its own vector to keep the variables organized. The sector dummy is also left out of the vector $\boldsymbol{x}$ controls, as it is not included simultaneous with the industry dummies.

The main focus of this thesis is to estimate the educational wage premium from 2014, which means to identify the coefficients β_1, β_2 and β_3 from the sample regression function (1.6) corresponding to model (4). According to hypothesis H_1 and H_2 from chapter 1.4, the wage premium is expected to be positive and increase with increased educational level, i.e. $\beta_3 > \beta_2 > \beta_1 > 0$. According to hypothesis H_3, the wage return from education is also expected to be decreasing, i.e.
$\beta_3 - \beta_2 < \beta_2 - \beta_1 < \beta_1 - 0$.

4.3 Econometric challenges

In this part, the methodological challenges related to empirical econometric analyses are addressed and their relevance towards the analysis of this thesis is discussed. The consequence of not solving these challenges will also be addressed, where the main issue often is that the models give biased results and no longer are BLUE.

4.3.1 Missing data, non-random sample and outlying observations

The issue of missing data, non-random samples and outlying observations is not something worth spending much time discussing, as it is somewhat irrelevant for this thesis. As mentioned in chapter 3.1, a 2% winsorization is applied to the raw data, which means the highest values are set equal to the value of the 99th percentile, while the lowest values are set equal to the 1st percentile. This solves the outlying observation problem. Microdata.no removes observations with missing values or values 0, which also solves the missing data issue. Lastly, because of the large number of observations and strict privacy policy of microdata.no, the issue with non-random sampling is not much of a concern.

4.3.2 Measurement error

If there is a discrepancy between the observed and true values of a variable, what is called a measurement error is present. This can be caused by using an imprecise measure of an economic variable in the regression model. OLS will be consistent under certain assumptions, but if these are violated it may cause the model to be inconsistent. In some of these cases, the size of the asymptotic bias can be derived. The measurement error can either be in the dependent variable, in the independent variable, or both.

If the measurement error is in the dependent variable, then it is not really a big concern for the analysis. It is reasonable to assume that the measurement error has zero mean and is uncorrelated with the each of the explanatory variables in the model. If this is true, then the OLS estimators are unbiased and consistent and the usual OLS inference procedures are valid. If it is correlated, on the other hand, then the estimator of the intercept, β_0, becomes biased and this is rarely a cause for concern. If the measurement error is uncorrelated with the error term, as is usually assumed, then the variance of the error term is overestimated. This results in larger variances of the OLS estimators as well. This is to be expected and there is nothing one can do about it (except collect better data). As long as the measurement error is uncorrelated with the independent variables, then OLS estimation has good properties. So, measurement error in the dependent variable is not really much of a concern in this thesis, as it typically does not lead to bias in the estimator as long as the error is uncorrelated with the independent variables, even though it can result in higher standard deviation and variance (Wooldridge, 2019, pp. 308-310).

Measurement error in an explanatory variable, on the other hand, has been considered a much more important problem than measurement error in the dependent variable. This depends on whether the measurement error is dependent on the observed values or not. If the measurement error is independent of the observed value, the estimator will still be unbiased, similar to the measurement error in the dependent variable, but with higher variance. However, if the measurement error is correlated with the observed values, the regression results will be biased, where the estimator will be inconsistent and biased towards zero (Wooldridge, 2019, pp. 310-313).

To summarize with respect to this thesis, there can be a measurement error in the dependent variable in the analysis if there is a discrepancy between reported annual salary and actual annual salary. Individuals can work undeclared, or there may be other factors that cause the salary to deviate from what is stated in The Norwegian Tax Administration. It is considered reasonable to assume that this will not significantly affect the analysis as long as the discrepancies are not systematic, and even then, it will only lead to higher variance. Systematic measurement errors in some of the independent variables are, as mentioned, more problematic, and the most worrying for this analysis is possible measurement errors in educational levels. Since the other variables are just included as control variables, there is not much of a worry about the coefficients of these. The coefficients of interest are the ones for education, as they determine the educational wage premium found in the analysis.

Since it is Statistics Norway have collected the data on the individual's wages as well as their educational level, the concern that there is a measurement error present in the data is quite low. Statistics Norway endeavours to produce and disseminate statistics and analyses of a high quality (Statistisk Sentralbyrå, n.d.). More can be read about the quality of their statistics at their homepage found (link can be found in the reference list). If the wages or educational level were self-reported, the concern that a measurement error is present would be much larger. Statistics Norway, on the other hand, has access to data from The Norwegian Tax Administration (Statistisk Sentralbyrå, 2022) as well as the National Education Database (Statistisk Sentralbyrå, 2020), so the reported data is of high quality.

4.3.3 Omitted variables

By omitting relevant explanatory variables from the model, it can lead to biased estimates, because the model is underspecified. But the estimates are only biased if the omitted variables are correlated with some of the included explanatory variables. This violation of the exogeneity condition leads to over- or underestimation of the estimates, depending on the directions of the correlations. If the correlation is zero, i.e. the omitted variables and included variables are uncorrelated, the estimates for the included variables are unbiased. This is obviously also the case if the true coefficient for the omitted variable is zero, which means it does not appear in the true model (Wooldridge, 2019, pp. 84-85).

The limitation of variables from microdata.no means that it cannot be controlled for a lot of individual characteristics, such as skills and ability, and it also makes it difficult to control for actual experience. These variables certainly contribute to determine the individual's wages, which means that the omission leads to a bias. The most debated omitted variable in the wage premium literature, is skills. As mentioned in chapter 2.3, skills and ability are often used interchangeably in literature, while this paper clearly distinguished between the terms. Ability is defined as a binary state of an individual either being disabled or able-bodied, while skills refer to an individual's skills specifically linked to work-related tasks. These can be inherent skills or skills gained from different situations or experiences. It is important to note that skills are regarded as "noncognitive" or "soft skills", meaning they are not gained from education or caught in schooling measurement. Skill is often an issue to measure when doing these sorts of analyses, and the term "omitted ability bias" or "omitted skill bias" is often used in literature when addressing the issue. The main theory of omitted skill bias is that there is some correlation between the inherent skills of an individual and the level of education. Several economic theories suggest such a relationship (Blackburn and Neumark, 1993). The pitfall of omitted skill bias, such as with other omitted variable biases, is that the problem cannot be solved by increasing the sample size or repeating the study multiple times (Jargowsky, 2005).

Two different variables have been attempted to control for an individual's skills. First, the National Tests from 8th grade, also known as national assessments or standardized tests, which are tests that are administered to students across a country to assess their academic performance in specific subjects such as English, Calculus and Reading. The tests are designed to measure students' learning outcome in a standardized and objective way. The results of these tests can be used by schools, education departments, and policymakers to identify areas of strength and weakness in the education system, to track trends in student performance over time, and to make informed decisions about how to improve the quality of education for all students. The first time the National Tests were executed was in the spring of 2004, after the Norwegian Parliament adopted a new system for quality assessment in 2002-2003 (Utdanningsdirektoratet, 2022). Due to National Tests being a relatively new concept, it resulted in a lack of data on the variable in the sample. A first stage regression result can be found in appendix A20 and descriptive statistics for National Test scores in A24, both showing that missing data on the National Tests variables from 8th grade ruins the regression results.

The second measure of inherent skills used, is parents' educational level. Looking at parents` education and children's noncognitive skills is well documented in the literature. It has been established that the effect is more or less linear (Ganzach, 2000). The effect of parental education on educational attainment seems to be stable between cohorts (De Graaf and Huinink, 1992). The analysis will include this variable as an instrument for education in a 2SLS regression to control for skills, and a discussion around the validity of the results.

An attempt has been made to test for potential misspecification of the model for the analysis through a RESET test, which is considered a general test for investigating this. The test finds that the model is correctly specified within a 1% significance level, attached in appendix A8. However, this should only be interpreted as the functional form of the model being correctly specified, and there will still likely be problems with omitted relevant variables.

4.3.4 Simultaneity

The simultaneous causal relationship between education and wages is a well-known source of endogeneity. There is a mutual influence between the dependent and explanatory variables of interest, resulting in the explanatory variables being correlated with the error term and violating the exogeneity condition. How education affects wages has already been addressed. But the endogeneity resulting from a two-way causal relationship can be explained by individuals expecting higher wages in the future if they attain higher education, such that the choice of whether to attain education or not depends on the expected future wage rather than their actual wage at the time they make the choice. Hence, there is a two-way causal relationship between the two. This can mathematically be shown by a simplified population regression function without constant terms for both wages and education.

$$(2.1)\ \textit{wage} = \alpha_1 \textit{education} + \beta_1 z_1 + u_1$$

$$(2.2)\ \textit{education} = \alpha_2 \textit{wage} + \beta_2 z_2 + u_2$$

education will not be exogenous it depends on the individual's wages ("reverse causation"). *education* is correlated with the error term u_1, because u_1 is indirectly a part of *education*. This can be shown by inserting the *wage* equation into the *education* equation:

$$(2.3)\ \textit{education} = \frac{\alpha_2\beta_1}{1-\alpha_2\alpha_1} z_1 + \frac{\beta_2}{1-\alpha_2\alpha_1} z_2 + \frac{\alpha_2 u_1 + u_2}{1-\alpha_2\alpha_1}$$

Which gives the reduced form equation for education:

$$(2.4)\ \textit{education} = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

The full calculation can be found in appendix A9. It is clear from the reduced form for education that *education* and u_1 in the first equation is correlated. Thus, OLS applied to this equation will violate the zero conditional mean assumption and be inconsistent. Similarly, OLS is an inconsistent estimator for the parameters in the education equation as well.

To solve the problem in this thesis, a 2SLS or instrumental variables (IV) regression can be used, where an instrumental variable from the education equation is used to estimate the wage equation. The instruments need to be correlated with the endogenous independent variables but not with the error term in the regression equation, i.e. the education equation need to contain at least one exogeneous variable (with a nonzero coefficient) that is excluded from the wage equation. Then the rank condition is fulfilled, which is a necessary and sufficient condition to have a valid instrument for the education variable in the wage equation. The instrument of choice in this thesis, is parental education. The parent's educational level for an individual, is not expected to directly affect the individual's wage level. But it has been proven to affect the individual's choice of attaining higher education. Hence, it is a valid instrument for education since it is an exogeneous variable in the education equation and is excluded from the wage equation.

4.3.5 Multicollinearity

Multicollinearity means that there is a high, but not perfect, correlation between two or more independent variables (Wooldridge, 2019, p. 90). Even though if the analysis is affected by multicollinearity, it will not lead to biases in the estimates. It will only lead to higher variance, which is also not desirable because it can lead to less statistically significant estimates. Yet, all estimates in the full model turns out to be statistically significant. The correlation matrix attached in appendix A10 gives the impression that there are no correlations between the included explanatory variables in the analysis that are worrying high. To confirm this, a VIF test has also been run, which can be found in appendix A11. It reports an average VIF value of 2.34, which is well below the somewhat controversial threshold of 10. All in all, it is reasonable to assume that multicollinearity will not be an issue for this analysis.

4.3.6 Heteroscedasticity

In order to have an unbiased and consistent estimator, it is also important to not violate the conditions of homoscedasticity in the error term. As mentioned in chapter 4.1 under the CLM assumptions, homoscedasticity means that the error term has a constant variance, $V(u_i|x_i) = \sigma^2$. Heteroscedasticity means that this is not the case, i.e. $V(u_i|x_i) \neq \sigma^2$. Violation of this condition lead to incorrect standard deviations, and statistical inference based on them can therefore give erroneous results (Hayes, 2022). It is unrealistic that the assumption of homoscedasticity is fulfilled for this analysis. Cross-sectional studies often have a wide range in values, and especially studies of income (Frost, 2017). Even though there is a 2% winsorization that removes outliers, the range in values can still be quite wide. A Breusch-Pagan test, which is attached in appendix A12, confirms that there is heteroscedasticity within all significance levels. To solve this problem, the entire analysis is conducted using cluster-robust standard errors.

5 Results and analysis

In this chapter, the results from the analyses are presented and interpreted. The results of the raw versus adjusted educational wage gap will first be presented and discussed in chapter 5.1. Chapter 5.2 will include heterogeneity testing to find differences in the educational wage gap between different worker groups. Chapter 5.3 and 5.4 will include some robustness testing to the results, in which chapter 5.3 will present variations in the reference category and chapter 5.4 will present an instrumental variable regression.

5.1 Educational wage gap

Table 4: Regression results for model (1), (2), (3) and (4).

	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	OLS
VARIABLES	lnwage	lnwage	lnwage	lnwage
utdniv_high	0.13799***	0.22928***		
	(0.00064)	(0.00068)		
utdniv_bachelor			0.07056***	0.18159***
			(0.00069)	(0.00070)
utdniv_master			0.28941***	0.35593***
			(0.00108)	(0.00103)
utdniv_phd			0.41027***	0.47511***
			(0.00302)	(0.00269)
male		0.19609***		0.18854***
		(0.00061)		(0.00059)
oslo		0.08356***		0.07146***
		(0.00064)		(0.00063)
big_city		0.04226***		0.03509***
		(0.00066)		(0.00064)
pot_exp		0.02489***		0.02610***
		(0.00010)		(0.00009)
pot_exp_sq		-0.00042***		-0.00044***
		(0.00000)		(0.00000)
mstat_married		0.05222***		0.04474***
		(0.00062)		(0.00061)
mstat_divorced		0.04110***		0.03671***
		(0.00095)		(0.00093)
immigrants_highinc		-0.14570***		-0.14063***
		(0.00121)		(0.00118)
imigraints_lowinc		-0.14402***		-0.14261***
		(0.00136)		(0.00131)
Constant	13.15096***	12.67811***	13.15096***	12.67223***
	(0.00041)	(0.00132)	(0.00041)	(0.00130)
Industry dummies	No	Yes	No	Yes
Observations	1 268 018	1 267 992	1 268 018	1 267 992
R-squared	0.036	0.347	0.076	0.373

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

The regression results for the four main models are presented in table 4 above, showing the difference in the raw educational wage gap versus the gap adjusted for observable worker characteristics, both for all higher educated individuals and separate for the three college degrees. In model (2) and (4) have the industry dummies, i.e. vector ***b*** dummies *indu_1* through *indu_20*, been included in the regression but are not reported in the table. The coefficients of control variables are not interesting for this analysis, so the 20 dummies for industries have simply been left out of the table to save space. Instead, a single line in the bottom of the table can be found stating whether or not the dummies are included in the regression. The full model (4) reporting all coefficients, including the industry dummies, can be found in appendix A19.

The first model that is presented is model (1), which is the SLR. This model estimates the raw educational wage gap between higher and lower educated individuals. The model finds a statistically significant wage premium of 13.80% for higher educated individuals. The R-squared of the model is only 0.036, meaning that education only has an explanatory power of 3.6% of the variation in wages.

The next model is model (2), where model (1) is extended by including control variables. This adjusts the educational wage premium by observable worker characteristics, which provides a more precise estimate of the educational wage premium between higher and lower educated individuals. The model estimates the educational wage premium for higher educated individuals to 22.93%, all else equal. The estimate is statistically significant at a 1% significance level, and the model has an explanatory power of 34.7% of the variation in wages.

Model (3) is similar to model (1), where the raw wage gap is estimated. The only difference is that education has been separated into three categories, providing the raw wage gap between individuals with the respective degree and individuals with lower education. The model estimates the raw wage gap to be 7.06% for bachelor’s degree graduates, 28.94% for master’s degree graduates, and 41.03% for doctorate’s degree graduates. All estimates are statistically significant at a 1% significance level, and the explanatory power of education is 7.6% in this model.

In model (4), education has been separated between the different college degrees as in model (3), and all control variables are included in the model. This is the main model in the thesis, as it separates the educational wage gap between the different college degrees, as well as adjusts the wage gap with regard to observable worker characteristics. The model estimates a wage premium of 18.16% for bachelor's degree graduates, 35.59% for master's degree graduates, and 47.51% for doctorate's degree graduates, all else equal. The explanatory power is 37.3% in this mode, which is the highest of the four models in the table. All estimates are statistically significant at a 1% significance level.

From model (2) and (4), the coefficients for the ***x*** vector control variables are reported. There are only some small differences in the estimates of these between model (2) and model (4), so the interpretation is the same for both models. From the gender dummy, model (4) estimates males to have 18.86% higher wages than females, all else equal. The estimate is statistically significant at a 1% significance level. This is in line with the expectations of the model as well as earlier literature, where the existence of a gender wage gap between observable equal workers is well documented (Carlsen, Rattsø and Stokke, 2016). From the Oslo dummy, the model estimates individuals located within the labour market region of Oslo to have 7.15% higher wages than individuals located in rural areas. Meanwhile, from the big-city dummy, the model estimates individuals located within the labour market region of Trondheim, Bergen or Stavanger to have 3.51% higher wages than individuals located in rural areas. These results are in line with the expectations, as larger cities tend to have higher wages than smaller cities and districts (Baum-Snow and Pavan, 2012). The interpretation of potential experience is not as straightforward as for the dummy variables, since it is a numeric variable. The effect of one additional year of experience can be found by differentiating the wage model with regards to potential experience. It is then found that:

$$(3.1)\ \frac{dlnwage}{dpot\_exp} = 0.0261 - (2 \times 0.00044\text{pot\_exp})$$

This shows that the effect on wages by increasing experience is positive but decreasing. This is in line with earlier literature, where for instance Carlsen, Rattsø and Stokke (2016) found that experience matters for wage determination, and that the effect is non-linear. Furthermore, they found that wages increase with experience for the first 20 years, and that one extra year of experience adds 1% to wages calculated at average experience of 8.1 years. Whereas model (4) in this thesis finds that wages increase with experience for the first 29 years, and one extra year of experience adds 2.5% to wages calculated at average experience of 21.8 years. For marital status the model estimates married individuals to have 4.47% higher wages than unmarried individuals, and divorced individuals to have 3.67% higher wages than unmarried individuals. This is in line with earlier literature which also found that both categories of marital status had a positive effect on wages with unmarried as reference category. The earlier literature also found that being married had a larger effect than being divorced, statistically significant at a 1% significance level (Hill, 1979). Lastly, the model estimates immigrational background to have a negative effect on wages regardless of whether the individual immigrated from a high-income country or low-income country. This is in line with earlier literature, which also found that ethnic minority individuals tend to have lower wages than natives (Dustmann, Frattini and Theodoropoulos, 2011, p. 220). Model (4) estimates that immigrants from high-income countries have 14.06% lower wages than native Norwegians, and immigrants from low-income countries have 14.26% lower wages than native Norwegians, all else equal. Both estimates are statistically significant at a 1% significance level.

To show how the R-squared, i.e. explanatory power of the variables develop when the model expands, a table is presented with models gradually expanding the number of control variables included. There are 6 models in the table, (i) – (vi). Model (i) is similar to the SLR in model (1), except it has added *male* as a sole control variable. Model (ii) is similar to model (i), except it also controls for work region location through the dummies *oslo* and *big_city*. Model (iii) also includes potential experience through the variables *pot_exp* and *pot_exp_sq*, model (iv) includes marital status through the dummies *mstat_married* and *mstat_divorced*, model (v) includes immigrational background through the dummies *immigrants_highinc* and *immigrants_lowinc*, and model (vi) includes sector through the dummy *sector_private*. All these models have *utdniv_high* as explanatory variable and all the coefficients in every model are statistically significant at a 1% significance level. If the sector dummy in model (vi) is replaced with the industry dummies in vector ***b***, it will result in model (2) from table 4. By looking at the R-squared from every model in the table, one can see how it gradually increases as more control variables are added in the model.

Table 5: Regression results from model (i) – (vi).

	(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**
	OLS	**OLS**	**OLS**	**OLS**	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**
utdniv_high	0.19452***	0.18605***	0.21497***	0.20999***	0.20780***	0.23370***
	(0.00063)	(0.00063)	(0.00063)	(0.00063)	(0.00062)	(0.00066)
male	0.25032***	0.25071***	0.25612***	0.25656***	0.26018***	0.24023***
	(0.00061)	(0.00061)	(0.00059)	(0.00059)	(0.00058)	(0.00059)
oslo		0.07594***	0.08457***	0.08499***	0.09705***	0.08941***
		(0.00070)	(0.00068)	(0.00067)	(0.00067)	(0.00066)
big_city		0.06557***	0.07839***	0.07770***	0.08122***	0.07628***
		(0.00076)	(0.00073)	(0.00073)	(0.00072)	(0.00072)
pot_exp			0.02589***	0.02417***	0.02430***	0.02458***
			(0.00010)	(0.00010)	(0.00010)	(0.00010)
pot_exp_sq			-0.00042***	-0.00040***	-0.00042***	-0.00041***
			(0.00000)	(0.00000)	(0.00000)	(0.00000)
mstat_married				0.04319***	0.05566***	0.05494***
				(0.00067)	(0.00067)	(0.00066)
mstat_divorced				0.03118***	0.04169***	0.04144***
				(0.00103)	(0.00102)	(0.00101)
immigrants_highinc					-0.15964***	-0.16975***
					(0.00127)	(0.00128)
imigrannts_lowinc					-0.16515***	-0.16348***
					(0.00141)	(0.00142)
sector_private						0.08924***
						(0.00064)
Constant	12.98091***	12.94479***	12.60975***	12.61279***	12.62281***	12.56463***
	(0.00052)	(0.00059)	(0.00115)	(0.00115)	(0.00114)	(0.00125)
Industry controls	No	No	No	No	No	No
Observations	1 268 018	1 268 018	1 267 992	1 267 992	1 267 992	1 267 992
R-squared	0.145	0.154	0.221	0.224	0.242	0.254

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

When studying the educational wage premium, or wage gap between higher and lower educated individuals, it might be interesting to explore whether there are some major differences in the control variables between the two groups of individuals. These differences can help better understand the difference between the raw wage gap and the adjusted wage gap, as well as explain the contributions from different variables. Two tables with descriptive statistics separating higher and lower educated individuals can be found in appendix A2 and A3, where a few key differences can be found.

The first difference is the gender distribution. Amongst the individuals with lower education, 67.9% are males and 32.1% are females. Amongst the individuals with higher education, only 45.4% are males and 54.6% are females. This means that there is an overweight of women amongst higher educated individuals, while there is an overweight of men amongst the lower educated individuals. This uneven balance of the gender distribution will result in an underestimation of the educational wage gap unless gender is controlled for in the model, because females in average have lower wages than males. This is confirmed by the estimates from model (i) and (1). Without controlling for gender, the educational wage gap is estimated to be 13.80% in model (1), while it increases to 19.45% in model (i) by controlling for gender.

Next, some differences in place of residence can be found between the two groups. Amongst higher educated individuals, 61.77% live in the labour market region of either Oslo, Trondheim, Stavanger or Bergen. Amongst lower educated individuals, the same share is 50.49%. This uneven distribution of residence will cause an overestimate of the educational wage premium unless it is controlled for, because the urban wage premium effect will be included in the educational wage premium. This is confirmed in table 5, where the educational wage premium drops from 19.45% in model (i) to 18.61% in model (ii) when controlling for location.

Another, a bit larger difference, is the distribution between sectors. Amongst higher educated individuals, 43.2% work in private sector while 56.8% work in public. Amongst lower educated individuals, 73.2% work in private sector while only 26.8% work in public. Another perspective of this is to look at the distribution of higher educated individuals versus lower educated individuals within each sector. It is then seen that within private sector, 35.03% have higher education while 64.96% have lower education. Within public sector, 65.97% have higher education while 34.03% have lower education. Analyses based on Norwegian data shows the existence of a wage premium in private sector for both lower educated and higher educated individuals (Rattsø and Stokke, 2020). This is consistent with the positive coefficients for *sector_private* in the models found in appendix A27, where two regressions have been run with higher and lower educated individuals separated. The payoff of working in private sector is estimated to be 5.03% for lower educated individuals and 12.17% for higher educated individuals. This result contradicts the findings in the article from Lucifora and Meurs (2006) on the other hand, where it is found that low-skilled workers earn more in public sector than private, while high-skilled workers earn more in private sector than public sector (Lucifora and Meurs, 2006). It is reasonable to assume that there is a positive correlation between skills and educational attainment, as research has proven that non-cognitive skills play a critical role in student academic achievement (Lipnevich and Roberts, 2012). Hence, the *sector_private* coefficient would be expected to be negative in the model for lower educated individuals, according to the results from Lucifora and Meurs (2006). As the results in this thesis is in line with earlier literature from Norway, one can expect an underestimate of the educational wage premium unless sector is controlled for, regardless of the distribution between private and public sector in the sample. Yet, the distribution determines the size of the possible underestimate, as there is a significant difference in the payoff from working in private sector between higher and lower educated individuals. The expected underestimate of the educational wage premium is confirmed by table 5, where it increases from 20.78% in model (v) to 23.37% in model (vi) where the sector dummy is included.

The final difference to address, is the difference in average age between higher and lower educated individuals. Lower educated individuals are on average 1.4 years older than higher educated individuals. According to Lindley and Machin (2016), it has been a continuing rise in the 2000s among the stock of adults who are college graduates. Hence, it makes sense that the cohort of higher educated individuals have a lower average age than the lower educated individuals. This affects potential experience, as it is a function of the individual’s age and expected graduation age. The difference in potential experience is larger, with an average of 24.2 years amongst lower educated individuals and 19.1 years amongst higher educated individuals. The difference is expected, since attaining higher education comes with a cost of giving up some years of potential experience caused by entering the labour market at a later point in time. Because experience is such an important determinant for wages, and higher education results in lower potential experience, one can expect the model to underestimate the educational wage premium unless it is controlled for. When this is controlled for, the educational wage premium increases from 18.61% in model (ii) to 21.50% in model (iii), both found in table 5.

Now that some of the statistical differences in the control variables between individuals with and without higher education has been addressed, it is easier to understand why the coefficient differs between the models. The first notable result is how all three of the premiums in model (4), where all the control variables are included, are higher than the average premium given in model (1) without any control variables. This can also easily be seen by comparing premium for premium between model (3) and model (4), where a significant increase in the premium is seen when including control variables. This result suggests that the joint effect of leaving out the control variables from the model, underestimates the educational wage premium. From model (i) to (vi) one can see how the educational wage premium develops when adding more control variables one by one. In model (i) one can see that just by adding *male*, the premium increases from 13.80% in model (1) to 19.45% in model (i). This does not come as a surprise, after seeing how the gender distribution differs between higher and lower educated individuals in the sample. There are more females attending higher education than males, but at the same time can it be seen from the *male* coefficient that males earn significantly more than females, all else equal.

In all models run with the joint higher education variable, *utdniv_high*, the wage premiums vary from 13.80% to 23.37%. In the model with the highest premium, every control variable is included, except the industry dummies. Instead, it includes *sector_private* which is a dummy equal to 1 if the individual works in the private sector and 0 if in the public sector. Between model (i) and (vi), the premium varies less than 5 percentage points. The impact from adding more control variables does not seem to change the premium drastically. The lowest educational wage premium in the models including control variables, is model (ii) which includes the controls *male*, *oslo* and *big_city*. In this model the premium is estimated to 18.61%. The reduction in the premium from model (i) to (ii), can be explained by the differences in residence between higher and lower educated individuals, as explained above. This leads to an overestimation of the premium in model (i). Gender seems to be the control variable that impacts the educational wage premium the most. This can also be seen from the table in appendix A13. When only including one explanatory variable, the educational wage premium changes from 0.81 percentage points at the lowest including only residence, to 5.65 percentage point at the highest including only gender.

There are also the models where higher education are divided into three subgroups, one for each college degree attainable. A table presenting the results by only including one control variable at a time, can be found in the appendix A14 for these models as well. This table shows how the different control variables affect the educational wage premium for each of the degrees, when they are the only control variable included compared to model (3) without any control variables. The bachelor's degree premium differs with 0.21 percentage points at the lowest including only immigration dummies, to 6.45 percentage points at the highest including all the industry dummies. The master's degree premium differs with 0.11 percentage points at the lowest including only marital status, to 4.70 percentage points at the highest including only potential experience. The doctorate's degree premium differs with 0.96 percentage points at the lowest including only immigration dummies, to 5.71 percentage points at the highest including only the sector dummy. The full model with all control variables, model (4), also have drastically higher premiums than model (3) without control variables. The bachelor's degree premium increases with 11.10 percentage points, the master's degree premium increases with 6.65 percentage points, and the doctorate's degree premium increases with 6.48 percentage points. These results implies that the educational wage premium is underestimated unless the control variables are included. The R-squared is also an important factor to evaluate when addressing this. In model (3) the R-squared is only 0.076, which means that the model only explains 7.6% of the variation in wages while there are other variables left out of the model that explains 92.4% of the variation. By including statistically significant control variables, the explanatory power of the model, i.e. the R-squared, gradually increases. This can easily be seen from table 5. In model (4) the R-squared is 0.373, meaning the model explains 37.3% of the variation in wages. This also implies that 62.7% of the variation is explained by other variables which are left out of the model. As long as the model has a R-squared of less than 1.00, i.e. the model explains less than 100% of the variation in the dependent variable, the coefficients are assumed to be under-/overestimated to some degree. By adding control variables in the model, the R-squared increases from 0.076 to 0.373, and the premiums became significantly larger. Model (4) is therefore a much more precise estimate of the educational wage premium than model (3), but one can still not conclude that the estimates are completely correct. If the model was able to include more relevant control variables, for instance skills, ability, actual experience, firms, etc. a higher R-squared would be obtained for the model, and it would have an even more precise estimate of the premiums.

The results found in this thesis show an educational wage premium of 18.16% for bachelor's graduates, 35.59% for master's graduates, and 47.51% for doctorate's graduates, i.e. the educational wage premium depends on the level of education. The average educational wage premium is 33.75%, whereas the average educational wage premium weighted for the distribution of individuals who obtain the different levels of higher education, is estimated to be 22.93% using the variable *utdniv_high*. Interestingly, earlier papers like James (2012) found a way higher educational wage premium than the results in this thesis. James (2012) reported a premium of 40% to 70% for individuals with bachelor's degree. Another article by Goldin & Katz (2008) found a premium for all college educated employees of 50%, which corresponds to the variable *utdniv_high*. I.e. Goldin & Katz found a premium that is 30 percentage points higher than the findings in this thesis.

One of the most incidental reasons that possibly could cause these differences, is what country the data used is collected from. While both of the aforementioned articles, James (2012) and Goldin & Katz (2008), are based on data from the United States of America, this paper is using data collected from Norway in 2014. While the data in this thesis also is more recent, James (2012) used data from 1977 to 2010 and Golding & Katz used data from 1915 to 2005. Because the analyses vary across countries and time, there may be major social, economic, and political differences that affects the premium. Norway has a higher density of workers owning a membership in a trade union, than the US has. This could possibly increase the wages for the one with the lowest initial wages, which most likely are groups of lower educated workers. A rise in minimum wage, for instance, will benefit these groups without increasing higher educated individual's wages (assuming these are higher initially). As a consequence, the wage gap will decrease, hence also the educational wage premium.

5.2 Heterogeneity with respect to gender, ethnicity, and location

There has also been run some regressions with five interaction terms to see if there are differences in the educational wage premium between different worker groups. The interaction terms for these heterogeneities are respectively *male_high*, *immhinc_high*, *immlinc_high*, *oslo_high* and *bigcity_high*. The interaction terms all interact with the variable *utdniv_high*, i.e. whether the individual has higher education or not, and the respective heterogeneity. These heterogeneities are gender, immigration, and residence. Immigration has been separated into two interaction terms, one for high-income countries (*immhinc_high*) and one for low-income countries (*immlinc_high*). Residence has also been separated into two interaction terms, one for residence in the labour market region of Oslo (*oslo_high*), and one for residence in the labour market region of one of the three big cities in Norway except Oslo (*bigcity_high*). These three big cities are Trondheim, Stavanger and Bergen, as elaborated in chapter 3.4.3.

For the interaction terms, there are a total of 8 regressions, model (I) – (VIII). All models include the ***x*** vector control variables, while model (I) – (IV) excludes the ***b*** vector industry dummies. Instead, these models include the sector dummy, *sector_private*. Model (I) includes the interaction term for gender. Model (II) includes the interaction term for immigrants. Model (III) includes the interaction term for residence. Model (IV) includes all interaction terms in the same model.

Model (V) – (VIII) excludes the sector dummy and instead includes the industry dummies. Model (V) includes the interaction term for gender. Model (VI) includes the interaction term for immigrants. Model (VII) includes the interaction term for residence. Model (VIII) includes all the interaction terms in the same model.

Apart from the interaction terms, model (I) – (IV) looks the same as model (vi) in table 5, while model (V) – (VIII) looks the same as model (2) in table 4. The results are presented in table 6 and 7 below.

Table 6: Regression results including interaction terms and control variables but excluding industry dummies.

	(I)	**(II)**	**(III)**	**(IV)**
	OLS	**OLS**	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**	**lnwage**
utdniv_high	0.23469*** (0.00078)	0.22541*** (0.00069)	0.20697*** (0.00086)	0.20310*** (0.00097)
male	0.24111*** (0.00075)			0.24126*** (0.00075)
male_high	-0.00171 (0.00114)			-0.00336*** (0.00114)
immigrants_highinc		-0.21958*** (0.00151)		-0.21547*** (0.00151)
immhinc_high		0.10543*** (0.00255)		0.09956*** (0.00256)
immigrants_lowinc		-0.17656*** (0.00177)		-0.16879*** (0.00178)

immlinc_high		0.02421*** (0.00274)		0.01186*** (0.00276)
oslo			0.05709*** (0.00090)	0.05955*** (0.00090)
oslo_high			0.06423*** (0.00131)	0.06060*** (0.00132)
big_city			0.06724*** (0.00094)	0.06808*** (0.00094)
bigcity_high			0.02272*** (0.00144)	0.02138*** (0.00144)
sector_private	0.08925*** (0.00064)	0.08867*** (0.00064)	0.08755*** (0.00063)	0.08720*** (0.00063)
Constant	12.56393*** (0.00126)	12.56738*** (0.00125)	12.57510*** (0.00125)	12.57555*** (0.00127)
Vector *x* controls	Yes	Yes	Yes	Yes
Vector *b* controls	No	No	No	No
Observations	1 267 992	1 267 992	1 267 992	1 267 992
R-squared	0.254	0.255	0.255	0.256

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

Table 7: Regression results including interaction terms and all control variables from vector ***x*** and vector ***b***.

	(V)	**(VI)**	**(VII)**	**(VIII)**
	OLS	**OLS**	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**	**lnwage**
utdniv_high	0.23934*** (0.00076)	0.22396*** (0.00071)	0.22091*** (0.00087)	0.22758*** (0.00095)
male	0.20572*** (0.00077)			0.20654*** (0.00078)
male_high	-0.01827*** (0.00110)			-0.01956*** (0.00111)
immigrants_highinc		-0.18354*** (0.00144)		-0.18251*** (0.00144)
immhinc_high		0.07985*** (0.00242)		0.07892*** (0.00242)
immigrants_lowinc		-0.14116*** (0.00171)		-0.13680*** (0.00172)
immlinc_high		-0.00504* (0.00262)		-0.01091*** (0.00264)

oslo			0.06845*** (0.00087)	0.07017*** (0.00087)
oslo_high			0.02884*** (0.00126)	0.02662*** (0.00127)
big_city			0.04524*** (0.00086)	0.04587*** (0.00086)
bigcity_high			-0.00499*** (0.00131)	-0.00551*** (0.00131)
Constant	12.67041*** (0.00136)	12.67953*** (0.00132)	12.67989*** (0.00133)	12.67276*** (0.00136)
Vector x controls	Yes	Yes	Yes	Yes
Vector b controls	Yes	Yes	Yes	Yes
Observations	1 267 992	1 267 992	1 267 992	1 267 992
R-squared	0.348	0.348	0.348	0.349

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

Remember from chapter 4.2 how the two sets of control variables were defined.

> ***x*** = (*male*, *oslo*, *big_city*, *pot_exp*, *pot_exp_sq*, *immigrants_highinc*, *immigrants_lowinc*, *mstat_married*, *mstat_divorced*)
>
> ***b*** = (*indu_1*, *indu_2*, *indu_3*, *indu_4*, *indu_5*, *indu_6*, *indu_7*, *indu_8*, *indu_9*, *indu_10*, *indu_11*, *indu_12*, *indu_13*, *indu_14*, *indu_15*, *indu_16*, *indu_17*, *indu_18*, *indu_19*, *indu_20*)

In the regression results above, the vector ***x*** controls are included, in addition to *sector_private*. The coefficients for *male*, *oslo*, *big_city*, *immigrants_highinc* and *immigrants_lowinc* have also been reported to see how the respective premium varies between higher and lower educated individuals.

The interaction term *male_high* measures the difference in education wage premiums for males compared to females. In model (I), the result is insignificant even at a 10% significance level. But in model (IV), (V) and (VIII), the coefficient is negative and significant at a 1% significance level. In model (IV) the wage premium seems to be 0.34 percentage points lower for males than females, while in model (V) and (VIII) the difference is a bit larger at respectively 1.83 and 1.96 percentage points. This can also be interpreted as the male wage premium, or gender wage difference, being lower for higher educated individuals than for lower educated individuals. This finding contradicts what earlier literature has found within the field of gender wage gap. A paper by Stokke (2021) found that the increase in the male wage premium is almost twice as large for workers with up to four years of college education, as for postgraduates (Stokke, 2021). The paper by Stokke (2021) also reproduces the findings of Barth, Kerr and Olivetti (2021) and Goldin *et al.* (2017), where the increase in the male wage premium is largest for the most educated workers. An explanation of the finding in this thesis, i.e. that the male wage premium is lower for higher educated individuals, is that qualifications and knowledge is more important for high-educated jobs than for low-educated jobs. Hence, personal characteristics such as gender becomes less important as the educational level increases, and one can expect the male wage premium to shrink.

Immigrants have been separated into two different interaction terms, the first being *immhinc_high*. This interaction term measures the difference in educational wage premiums for immigrants born in high-income countries, compared to Norwegian natives. The other interaction term is *immlinc_high*, which measures the difference in educational wage premiums for immigrants born in low-income countries, compared to Norwegian natives. Model (II) estimates the educational wage premium to be 10.54 percentage points higher for immigrants born in high-income countries and 2.42 percentage points higher for immigrants born in low-income countries, compared to Norwegian natives. Both coefficients are statistically significant at a 1% significance level. Model (IV) estimates the educational wage premium to be 9.96 percentage points higher for immigrants born in high-income countries and 1.19 percentage points higher for immigrants born in low-income countries, compared to Norwegian natives. Both coefficients are statistically significant at a 1% significance level. Model (VI) estimates the educational wage premium to be 7.99 percentage points higher for immigrants born in high-income countries compared to Norwegian natives, significant at a 1% significance level. Meanwhile, the same model estimates the educational wage premium to be 0.50 percentage points lower for immigrants born in low-income countries compared to Norwegian natives, only significant at a 10% significance level. Finally, model (VIII) estimates the educational wage premium to be 7.89 percentage points higher for immigrants born in high-income countries and 1.09 percentage points lower for immigrants born in low-income countries, compared to Norwegian natives. Both of these coefficients are significant at a 1% significance level. It seems that immigrants from high-income countries have a significantly higher return of education than native Norwegians, regardless of which model one looks at. An explanation for this is that attaining higher education reduces some of the initial wage gap between immigrants from high-income countries and natives. From model (4) it is seen that immigrants from high-income countries are estimated to have 14.06% lower wages than natives. By running the same regression for only high-educated individuals, it can be seen from the model in appendix A26 that immigrants from high-income countries are estimated to have 9.65% lower wages than natives. This finding is also in line with earlier literature, where a paper by Smith and Fernandez (2017) found that immigrant workers make significantly less than their native peers, but that the wage differential in the United States disappears after accounting for education and cognitive skills. Meanwhile, the return for immigrants from low-income countries seems to be positive when controlling for sector instead of industries, and negative when controlling for industries. The results are also small, the highest being a 2.4% return. It is therefore difficult to conclude whether the additional educational return is present or not, and if the return is higher or lower for immigrants from low-income countries compared to the native Norwegians.

The last two interaction terms control for residence as a heterogeneity. The first one is *oslo_high*, which compares the educational wage premium amongst individual's resident in the labour market region of Oslo, to individual's resident outside the labour market region of Oslo, Trondheim, Stavanger or Bergen. The other one is *bigcity_high*, which compares the educational wage premium amongst individual's resident in the labour market region of Trondheim, Stavanger or Bergen, to individual's resident outside the labour market region of Oslo, Trondheim, Stavanger or Bergen. As mentioned in chapter 3.4.3, 31.85% of the individuals in the sample lives within the labour market region of Oslo, 24.03% lives within the labour market region of either Bergen, Trondheim or Stavanger, and 44.12% lives in other regions of Norway. In model (III) the educational wage premium is estimated to be 6.42 percentage points higher in the labour market region of Oslo and 2.27 percentage points higher in the labour market region of Trondheim, Bergen or Stavanger, both compared to other labour market regions of Norway. Both estimates being significant at a 1% significance level. Model (IV) estimates the educational wage premium to be 6.06 percentage points higher in the labour market region of Oslo and 2.14 percentage points higher in the labour market region of Trondheim, Bergen or Stavanger, compared to other labour market regions of Norway. Both estimates being significant at a 1% significance level. Model (VII) estimates the educational wage premium to be 2.88 percentage points higher in the labour market region of Oslo and 0.50 percentage points lower in the labour market region of Trondheim, Bergen or Stavanger, compared to other labour market regions of Norway. Both estimates being significant at a 1% significance level. And finally, model (VIII) estimates the educational wage premium to be 2.66 percentage points higher in the labour market region of Oslo and

0.55 percentage points lower in the labour market region of Trondheim, Bergen or Stavanger, compared to other labour market regions of Norway. Both estimates being significant at a 1% significance level. A higher educational wage premium in Oslo than in the outskirts is consistent with the agglomeration literature, which finds that the urban wage premium is higher for high-educated individuals than for low-educated individuals. See for example the paper *Education, experience, and urban wage premium* (2016) in Regional Science and Urban Economics by Carlsen, Rattsø and Stokke. These are some interesting results, where a possible explanation is a supply and demand effect causing these differences in the return of education. This supply and demand effect is a result of many the distribution of industries which requires higher education inside and outside the respective labour market region. If many industries requiring higher education are being located in the labour market region, whereas not so many are being located outside the region, it will increase the demand for these positions inside the labour market region. This creates an overweight of higher educated individuals also living in the large labour market region, since it is close to their job. This overweight was mentioned earlier, where 61.77% of higher educated individuals live in the labour market region of either Oslo, Trondheim, Bergen or Stavanger, while only 50.49% of lower educated individuals live there. This uneven distribution between higher and lower educated individuals in the labour market region will impact the estimate from the return of education.

To summarize this chapter, it seems that the educational wage premium differs between males and females, where females have a slightly higher return than males. This contradicts earlier literature within the gender wage gap, but there is a possible explanation which seems logical. Immigrants from high-income countries also have a higher return than Norwegian natives, while the effect is ambiguous for immigrants from low-income countries, so it is difficult to conclude anything for this group. This finding is in line with the findings from a paper by Smith and Fernandez (2017). The premium seems to be a little higher in the labour market region of Oslo, while the result is ambiguous for Trondheim, Stavanger and Bergen. This finding is also in line with earlier literature by Carlsen, Rattsø and Stokke (2016).

5.3 Robustness: Variations in the reference category

To test the robustness of the results, the reference category for education has been changed to see if there are any significant difference between the degrees of higher education. Robustness testing can have several different goals, whereas one official goal is to see what happens to the analysis when the assumptions changes (Gelman, 2017). According to Gelman (2017), the most common reason for testing robustness is to see if the initial analysis is valid. This can sometimes cause problems in cases where the author uses robustness tests only to defend the models used instead of objectively trying to find the best fitted model.

So far, lower educational level, i.e. high school has been the reference category. In table 8 are three new models presented, where the reference category has been changed to bachelor's graduates in model (4-1), master's graduates in model (4-2) and doctorate's graduates in model (4-3).

Table 8: OLS regression results with new reference categories.

	(4-1)	**(4-2)**	**(4-3)**
	OLS	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**
utdniv_vgs	-0.18159***	-0.35593***	-0.47511***
	(0.00070)	(0.00103)	(0.00269)
utdniv_bachelor		-0.17433***	-0.29352***
		(0.00097)	(0.00266)
utdniv_master	0.17433***		-0.11919***
	(0.00097)		(0.00273)
utdniv_phd	0.29352***	0.11919***	
	(0.00266)	(0.00273)	
Constant	12.85382***	13.02815***	13.14734***
	(0.00135)	(0.00150)	(0.00295)
Vector x controls	Yes	Yes	Yes
Vector b controls	Yes	Yes	Yes
Observations	1 267 992	1 267 992	1 267 992
R-squared	0.373	0.373	0.373

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

The results show that there are significant differences in premiums between the three levels of higher education, at a 1% significance level. The highest educational level also obtains the highest wage premium, which is in line with hypothesis H_2 from chapter 1.4. The educational wage premium by increasing the educational level from high school to bachelor's degree, is estimated to be 18.16%. The premium by increasing the educational level from bachelor's degree to master's degree is 17.43%. The premium by increasing the educational level from master's degree to doctorate's degree is 11.92%. This is also in line with hypothesis H_3, stating that the return from educational investment is expected to be positive but decreasing.

5.4 Instrumental variable regression

In this chapter, another robustness test of the results is done by running a Two-Stage Least Squares (2SLS), or Instrumental Variable (IV) model. The reason a 2SLS model is applicable, is because of the concern of a possible overestimation of the educational wage premium caused by correlation between the educational level and skills. Since there is no available variable to control for skills, the models are suspected to be biased due to omitted variables. This causes the estimated educational wage premium to consist of at least two effects, the actual educational wage premium itself and a skills premium. It is assumed that the skills premium is positive, as earlier literature has found the importance of cognitive skills for wage determination as increasing (Murnane, Willett and Levy, 1995). This leads to an overestimate of the educational wage premium, in the models run in the analysis. For the 2SLS model, the individual's parents' education is used as instruments for education. If controlling for the skills effect on wages through parents' education is a successful instrument, the estimates of the educational wage premium is expected to be lower than in the original analysis. A table with the 2SLS regression results is presented below.

Table 9: 2SLS regression results.

VARIABLES	**(#1) FIRST STAGE utdniv_high**	**(#2) SECOND STAGE lnwage**	**(#3) FIRST STAGE utdniv_high**	**(#4) SECOND STAGE lnwage**
utdniv_high		0.46883*** (0.00389)		0.43878*** (0.00485)
fatheduc	0.20020*** (0.00095)			
motheduc			0.17000*** (0.00101)	
Constant	0.38900*** (0.00182)	12.57396*** (0.00219)	0.38060*** (0.00185)	12.58702*** (0.00251)
Vector x controls	Yes	Yes	Yes	Yes
Vector *b* controls	Yes	Yes	Yes	Yes
Observations	1 267 992	1 267 992	1 267 992	1 267 992
R-squared	0.336	0.273	0.327	0.291

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

The Two-Stage Least Squares model consists, as its name reveals, of two stages. The purpose of the estimation method is to instrument a variable that is suspected to be biased in the original model, to obtain a more valid estimate. Stage one consists of running a regression on the instrumented variable, which in this case is *utdniv_high,* and include all the other independent variables used in the original model, in addition to the instrumental variable. In this case that is a dummy on whether the father has higher education or not in model (#1) and whether the mother has higher education or not in model (#3). From the coefficients in model (#1) and (#3) it seems that parent's education is a valid instrument, as it increases the probability for the individual to attain higher education by respectively 20% and 17%, at a 1% significance level. In stage two, the results seem more surprising. In both models, the educational wage premium doubles from what the OLS model estimated, where model (#2) estimates a premium of 46.88% and model (#4) estimates a premium of 43.89%. Model (2) using OLS only estimated the educational wage premium to be 22.93%. This is the opposite result of what was expected. As mentioned, the goal with the 2SLS was to isolate the educational wage premium by getting rid of the skill premium effect that impacts the results. So, what can be an explanation for this result? Remember from chapter 4.3.4 that an important condition for an instrumental variable regression, is that the instrument fulfils the rank condition. This condition implies that the education equation need to contain at least one exogeneous variable that is excluded from the wage equation, which is parental education in this case. I.e. parental education should not be correlated to– or have a causal effect on wages. To control if this is the case, an OLS regression model with the instruments as explanatory variables on wages was run, also including the control variables. The results are presented in table 10 below.

Table 10: OLS regression results of mother's and father's education on the individual's wages.

VARIABLES	(X) OLS lnwage
utdniv_high	0.21981*** (0.00069)
fatheduc	0.04309*** (0.00080)
motheduc	0.01966*** (0.00085)
Constant	12.66610*** (0.00134)
Vector x controls	Yes
Vector b controls	Yes
Observations	1 267 992
R-squared	0.350

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

As seen from the results in model (X), both mother's and father's education have a statistically significant effect on the individual's wages, at a 1% significance level. This causes parents' education to not be validly excluded from the wage equation, which violates the rank condition. The premium is expected to decrease when instrumenting for parent's education, as it was hoped it would separate the skills wage premium from the educational wage premium. Instead, due to the violation of the rank condition for valid instruments, the educational wage premium doubled. By taking a look at model (X), one can see that the wage is estimated to increase by 4.31% if the father has higher education, all else equal. Meanwhile, the estimated effect of having a mother with higher education is 1.97%, all else equal. Even though both these results are statistically significant at a 1% significance level, it cannot be said for sure that the effect is causal and not spurious. It would not make sense for it to be causal, considering the structure and laws of wage determination in Norway in 2014.

An explanation for the significant effect can for instance be that higher educated parents provides their children with a network that gives them access to better paid jobs. For instance, if a father is a bank manager, it may provide his children with the possibility to work at the bank for him. In contrast, if a father is a grocery store manager, his children may be able to work at the grocery store. Typically, will a bank manager position require a higher education than a grocery store manager position, while also working at a bank is better paid than working at a grocery store. In this scenario, the father's educational level will have an impact on his children's wages through a networking effect. Another explanation could be that parents with higher education have more experience to share with their children regarding job applications and interview situations for better paid jobs.

There are probably other instruments that would work better than parental education for an individual's wages. Distance to the nearest educational institution is a typically used instrument in research papers (Pokropek, 2016). Another possibility could be to use an older siblings' educational level as instrument. As an older sibling, you are often a role model for the younger siblings, who may tend to follow their tracks. If the older sibling has chosen to attend higher education, it could have an effect on the individual's own choice of whether or not to attend, without directly affecting the wages. These two variables would assumably be validly excluded from the wage equation, and therefore work well as an instrument for education. Unfortunately, such variables are not available in microdata.no as of today.

6 Summary and concluding remarks

In this paper, the educational wage premium in Norway based on registry data from 2014 have been the subject of research. The purpose of the thesis has been to prove that the wage gap exists and try to estimate it using an econometric approach with rational underlying assumptions. By using registry data provided by microdata.no, the analysis has been able to study a large sample of more than 1.2 million full-time employees. The result from the analysis provides evidence that supports the theory that the educational wage premium indeed does exist. Further, does the educational wage premium increase as the educational level increases, but the marginal increase is decreasing. I.e. the doctorate's degree premium was found to be the largest, at 47.51%, the master's degree premium was the second largest at 35.59%, and the bachelor's degree premium was the lowest at 18.16%. These findings are in line with previous research done within the field of educational wage premiums. Even though the size of the educational wage premium varies between papers, it is not considered as an issue for this thesis. The findings in this paper are not assumed to be any less credible than the ones in the papers mentioned throughout this thesis, as there are numerous reasons for the premium to vary across country and time.

Due to deficient data for the last couple of years, the analysis has been done on data from 2014 while it would have been more interesting to analyse an even more recent year. This means the analysis are not able to capture a completely up-to-date estimate of the educational wage premium. As mentioned earlier, the secular trend shows an increase in individuals attending higher education. It would have been interesting to compare the results from 2014 with for instance 2019. Unfortunately, was 2014 the most recent year with adequate accessible data. Other effects that also may affect the results is some recent major events, such as the covid-19 pandemic. For individuals that were employed during the pandemic, there could possibly be a difference in those who got to have home-office and those who did not. There are usually a higher proportion of higher educated individuals working in offices, rather than out in the field, whereas lower educated individuals tend to work with more hands-on type of jobs. Therefore, higher educated individuals had better opportunities for working from home-offices, whereas many lower educated individuals got temporarily laid off. Because many industries with a majority of lower educated employees struggled financially during the pandemic, the educational wage premium assumably increased during this period of time.

Because microdata.no have access to data on individual level, the sample size is the actual true population for Norway in 2014. Even though restrictions that reduces the sample size are imposed, the analysis still manages to include more than 1.2 million observations. This is for sure an extremely important argument for the internal validity of the analysis. Increasing the sample size enhances the probability of detecting statistically significant findings, thereby allowing for the identification of even marginal effects as significant (Sumeracki, 2018). All variables in the main models are statistically significant at a 1% significance level. Still, there is a difference between the result being statistically significant and economically significant. The interpretation of the result might not necessarily be meaningful if the coefficient for instance is really small. Another example is if the sign changes between models, but the results are statistically significant in both models. This was seen happening with the interaction term for residence and higher education, where both models have significant results at a 1% significance level, but the result still seems ambiguous because the sign changes.

While the unique sample size for this data from Norway makes the internal validity seem high, the external validity focuses on to what degree the results can be applied to other groups or situations (Streefkerk, 2019). As mentioned, will differences between countries and dates have an important impact on the results, which affects the external validity. The existence of an educational wage premium has been proven to be true for this sample, whereas the question about external validity mostly will focus on the size of the premium. I.e. most countries, if not all, have an educational wage gap, but the size of it may differ.

In the analysis, it has not been taken into account the individuals having a vocational certificate. This is expected to influence the results found in the analysis, as they are expected to have a higher wage level than other individuals with high school as their highest finished education. Higher education has been defined as attending college and finishing a degree there. Everyone without college education, is defined as lower educated individuals, meaning the individuals with vocational certificates are included as lower educated individuals. Even though they don't attend any college or such, they are fully educated within their field of work. This is a contrast to the individuals attaining a regular 3-year program where they achieve general study competence. This is because an individual is considered “skilled” within his field of work with a vocational certificate, while the individual’s attaining a regular 3-year program are considered “unskilled”. By having a vocational certificate, it will provide job opportunities that pays better than the opportunities unskilled individuals have. Hence, the vocational certificate yields a wage premium itself. Unfortunately, microdata.no does not provide data on occupational certificates, so it is not possible to control for this as of today. This will possibly lead to an underestimate of the wage premiums in the models, since they cannot include a control for occupational certificates, which separates skilled versus unskilled workers amongst the lower educated individuals. The average wage is expected to be highest for higher educated individuals, while the average wage for individuals with occupational certificates is expected to be higher than for individuals considered unskilled without higher education or occupational certificates.

To give a further insight of the wage premium, the thesis has also controlled for heterogeneities between worker groups to see if the wage premium differs between certain individual characteristics. The chosen heterogeneities are gender, immigration, and residence, in which the analysis found evidence of females having a larger premium than males, immigrants from high-income countries having a larger premium than native Norwegians, and that the premium is larger in the labour market region of Oslo than in the districts. The heterogeneity analysis failed to provide evidence for a difference in the premium between immigrants from low-income countries and native Norwegians, and inhabitants in the labour market region of Trondheim, Stavanger or Bergen and in the districts, as the results of these were ambiguous.

Lastly, the thesis also provided two robustness tests of the results. The first test was to run the model with different variations in reference category for education. The second test was to construct a 2SLS model to isolate the effect of a skills wage premium from the educational wage premium. While the model with variations in the reference category was a success, the 2SLS failed to include valid instruments. Both National Tests and parents' educational level was attempted as instruments but failed due to insufficient data and a violation of the rank condition. The limitation of possible instrumental variables provided by microdata.no made it difficult to find a new instrumental variable, so the analysis failed to solve the issue of omitting skills from the model. The model still managed to include a sample of more than 1.2 million observations with 37% explanatory power on wages, where all the included variables were statistically significant at a 1% significance level. This is more than sufficient to prove that an educational wage premium exists and to obtain quite valid estimates, even though controlling for skills would strengthen the model even further. Hopefully will future research be able to detach the skills effect on wages from the educational effect, to push the boundaries of knowledge within the field of educational wage premiums.

References

Autor, D. H., Katz, L. F. and Kearney, M. S. (2008) TRENDS IN U.S. WAGE INEQUALITY: REVISING THE REVISIONISTS, *The Review of Economics and Statistics*, 90(2), pp. 300-323. Available at: https://direct.mit.edu/rest/issue/90/2 (Accessed: 17th of March, 2023).

Autor, D. H. (2014) Skills, education, and the rise of earnings inequality among the "other 99 percent", *Science*, 344(6186), pp. 843-851. doi: 10.1126/science.1251868.

Bagdasarian, B. (2018) Talking Data Part 2: Macro Data vs. Micro Data. Available at: https://blog.hubspot.com/customers/talking-data-part-2-macro-micro-data.

Bailey, M. A. (2020) *Real Econometrics - The Right Tools to Answer Important Question*. 2nd edn. Oxford University Press.

Barstad, A. (2013) Innvandring, innvandrere og livskvalitet, *Statistisk Sentralbyrå (SSB).*

Barth, E., Kerr, S. P. and Olivetti, C. (2021) The dynamics of gender earnings differentials: Evidence from establishment data, *European Economic Review*, 134, pp. 103713.

Baum-Snow, N. and Pavan, R. (2012) Understanding the city size wage gap, *The Review of economic studies*, 79(1), pp. 88-127.

Benito, S. G., Glassman, T. S. and Hiedemann, B. G. (2016) Disability and labor market earnings: Hearing earnings gaps in the United States, *Journal of Disability Policy Studies*, 27(3), pp. 178-188.

Bennet, M. (2022) Integrated Masters Degrees - A Guide. Find A Masters. Available at: https://www.findamasters.com/guides/integrated-masters-degrees-guide.

Bhuller, M. (2009) Inndeling av Norge i arbeidsmarkedsregioner, *Statistisk Sentralbyrå (SSB). Notater*, 24.

Blackburn, M. L. and Neumark, D. (1993) Omitted-Ability Biaa and the Increase in the Return to Schooling, *Journal of labor economics*, vol 11, nr 3. doi: https://doi.org/10.1086/298306.

Blank, R. M. (2009) *Social Protection vs. Economic Flexibility: Is There a Trade-Off?* Chicago, IL: University of Chicago Press.

Blau, F. D. and Kahn, L. M. (2017) The gender wage gap: Extent, trends, and explanations, *Journal of economic literature*, 55(3), pp. 789-865.

Bondzie, E. (2016) Effect of smoking and other economic variables on wages in the Euro Area, *Available at SSRN 2727228.* doi: http://dx.doi.org/10.2139/ssrn.2727228.

Budig, M. J. and England, P. (2001) The wage penalty for motherhood, *American sociological review*, pp. 204-225.

Bø, T. P. (2004) Høy yrkesdeltakelse blant kvinner i Norden, *Samfunnsspeilet*, 1/2004, pp. 12-17.

Card, D. and Lemieux, T. (2001) Can falling supply explain the rising return to college for younger men? A cohort-based analysis, *The Quarterly Journal of Economics*, 116(2), pp. 705-746.

Carlsen, F., Rattsø, J. and Stokke, H. E. (2016) Education, experience, and urban wage premium, *Regional Science and Urban Economics*, 60, pp. 39-49. doi: https://doi.org/10.1016/j.regsciurbeco.2016.06.006.

Council, N. R. (1990) *Inner-City Poverty in the United States*. Washington, DC: The National Academies Press.

Crosby, F. *et al.* (2003) Affirmative Action: Psychological Data and the Policy Debates, *The American psychologist*, 58, pp. 93-115. doi: 10.1037/0003-066X.58.2.93.

De Graaf, P. M. and Huinink, J. J. (1992) Trends in measured and unmeasured effects of family background on educational attainment and occupational status in the Federal Republic of Germany, *Social Science Research*, 21(1), pp. 84-112. doi: doi.org/10.1016/0049-089X(92)90019-D.

Direktoratet for høyere utdanning og kompetanse (2022) *Studere i Norge*. Available at: https://utdanning.no/tema/hjelp_og_veiledning/studere_i_norge (Accessed: 13th of March 2023).

Dube, A., Lester, T. W. and Reich, M. (2010) MINIMUM WAGE EFFECTS ACROSS STATE BORDERS: ESTIMATES USING CONTIGUOUS COUNTIES, *The Review of Economics and Statistics*, 92(4), pp. 945-964. Available at: https://direct.mit.edu/rest/issue/92/4 (Accessed: 17th of March, 2023).

Dustmann, C., Frattini, T. and Theodoropoulos, N. (2011) Ethnicity and second generation immigrants, *The Labour Market in Winter: the state of working Britain*, pp. 220-239.

Eilers, C. (2023) *Dette må du vite om studiegjeld*. Available at: https://www.dnb.no/dnbnyheter/no/din-okonomi/dette-bor-du-vite-om-studiegjeld (Accessed: 23rd of February 2023).

Finansforbundet (2020) *Så mye må du tjene for å bli lykkelig*. Available at: https://www.finansforbundet.no/folk-og-fag/lonn/sa-mye-ma-du-tjene-for-a-bli-lykkelig/ (Accessed: 22nd of February 2023).

Frost, J. (2017) *Heteroscedasticity in Regression Analysis*. Available at: https://statisticsbyjim.com/regression/heteroscedasticity-regression/ (Accessed: 27th April 2023).

Ganzach, Y. (2000) Parents' education, cognitive ability, educational expectations and educational attainment: Interactive effects, *British Journal of Educational Psychology*, 70(3), pp. 419-441. doi: doi.org/10.1348/000709900158218.

Gelman, A. (2017) *What`s the point of a robustness check?* Available at: https://statmodeling.stat.columbia.edu/2017/11/29/whats-point-robustness-check/ (Accessed: 27th April 2023).

Goldin, C. and Katz, L. F. (2007) The Race between Education and Technology: The Evolution of U.S. Educational Wage Differentials, 1890 to 2005, *National Bureau of Economic Research Working Paper Series*, No. 12984. doi: 10.3386/w12984.

Goldin, C. *et al.* (2017) The expanding gender earnings gap: Evidence from the LEHD-2000 Census, *American Economic Review*, 107(5), pp. 110-114.

Gould, E. D. (2007) Cities, workers, and wages: A structural analysis of the urban wage premium, *The Review of economic studies*, 74(2), pp. 477-506.

Gustman, A. L., Steinmeier, T. L. and Tabatabai, N. (2012) How Did the Recession of 2007-2009 Affect the Wealth and Retirement of the Near Retirement Age Population in the Health and Retirement Study?, *Social Security Bulletin*, 72, pp. 47-66. Available at: https://www.ssa.gov/policy/docs/ssb/v72n4/index.html (Accessed: 17th of March, 2023).

Hayes, A. (2022) *Heteroscedasticity Definition: Simple Meaning and Types Explained*. Available at: https://www.investopedia.com/terms/h/heteroskedasticity.asp (Accessed: 28th April 2023).

Hill, M. S. (1979) The wage effects of marital status and children, *Journal of Human Resources*, pp. 579-594.

Hægeland, T., Klette, T. J. and Salvanes, K. G. (1999) Declining returns to education in Norway? Comparing estimates across cohorts, sectors and over time, *Scandinavian Journal of Economics*, 101(4), pp. 555-576.

Hægeland, T. (2001) *Changing Returns to Education Across Cohorts: Selection, School System or Skills Obsolescence?* : Discussion Papers.

James, J. (2012) The college wage premium, *Economic Commentary*, (2012-10). doi: https://doi.org/10.26509/frbc-ec-201210.

Jargowsky, P. A. (2005) Omitted Variable Bias, i Dallas, U. o. T. a. (ed.). (Accessed: 22nd of February 2023).

Johansen, S. (2020) microdata.no – søknadsfri tilgang til registerdata om boforhold, *Tidsskrift for boligforskning*, 3(1), pp. 87-98. doi: https://doi.org/10.18261/issn.2535-5988-2020-01-06.

Kanade, V. (2022) *What Is Linear Regression? Types, Equation, Examples, and Best Practice for 2022*. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/ (Accessed: 10th of March 2023).

Karlsen, H. T. (2022) Bevegelseshemmedes sysselsetting øker med stigende utdanningsnivå, *SSB - Statistic Norway*. Available at: https://www.ssb.no/helse/funksjonsevne/statistikk/levekar-hos-personer-med-funksjonsnedsettelse/artikler/bevegelseshemmedes-sysselsetting-oker-med-stigende-utdanningsniva (Accessed: 22.02.2023).

KarriereStart.no (2014) *Statens lønnstrinn for 2014*. Available at: https://karrierestart.no/lonn-og-frynsegoder/783-lonnstrinn-2014-komplett-oversikt-over-alle-lonnstrinn-for-2014 (Accessed: 21st of April 2023).

Katz, L. F. and Murphy, K. M. (1992) Changes in Relative Wages, 1963-1987: Supply and Demand Factors, *The Quarterly Journal of Economics*, 107(1), pp. 35-78. doi: 10.2307/2118323.

Kenton, W. (2021) *Spurious Correlation: Definition, How It Works, and Examples*. Available at: https://www.investopedia.com/terms/s/spurious_correlation.asp (Accessed: 15th March 2023).

Kimball, M. S. *et al.* (2015) Diminishing marginal utility revisited, *Available at: https://ssrn.com/abstract=2592935*. doi: https://dx.doi.org/10.2139/ssrn.2592935.

Kunze, A. (2005) The evolution of the gender wage gap, *Labour Economics*, 12(1), pp. 73-97.

Layard, R., Mayraz, G. and Nickell, S. (2008) The marginal utility of income, *Journal of Public Economics*, 92(8), pp. 1846-1857. doi: https://doi.org/10.1016/j.jpubeco.2008.01.007.

Lindley, J. and Machin, S. (2016) The Rising Postgraduate Wage Premium, *Economica*, 83(330), pp. 281-306. Available at: http://www.jstor.org/stable/24751920 (Accessed: 2023/02/02/).

Lindsay, S. (2011) Discrimination and other barriers to employment for teens and young adults with disabilities, *Disability and rehabilitation*, 33(15-16), pp. 1340-1350. doi: doi.org/10.3109/09638288.2010.531372.

Lipnevich, A. A. and Roberts, R. D. (2012) Noncognitive skills in education: Emerging research and applications in a variety of international contexts, *Learning and Individual Differences*, 22(2), pp. 173-177.

Lucifora, C. and Meurs, D. (2006) THE PUBLIC SECTOR PAY GAP IN FRANCE, GREAT BRITAIN AND ITALY, *Review of Income and Wealth*, 52(1), pp. 43-59. doi: https://doi.org/10.1111/j.1475-4991.2006.00175.x.

Lumivero (2017) *ORDINARY LEAST SQUARES REGRESSION (OLS)*. Available at: https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols (Accessed: 26th of April 2023).

Lånekassen (2022) *Flere med over 1 million i studiegjeld*. Available at: https://lanekassen.no/nb-NO/presse-og-samfunnskontakt/nyheter/pressemelding_flere-far-1-million-i-studiegjeld/ (Accessed: 23rd of February 2023).

Moody, J. (2021) A Guide to Different Types of College Degrees, *US News*. Available at: https://www.usnews.com/education/best-colleges/articles/a-guide-to-different-types-of-college-degrees (Accessed: 20.02.2023).

Mottaz, C. (1984) Education and work satisfaction, *Human Relations*, 37(11), pp. 985-1004.

Murnane, R., Willett, J. B. and Levy, F. (1995) The growing importance of cognitive skills in wage determination: National Bureau of Economic Research Cambridge, Mass., USA.

Nations, T. U. (2014) Country Classification, *World Economic Situation and Prospects*. Available at: https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf.

Oreopoulos, P. and Salvanes, K. G. (2011) Priceless: The Nonpecuniary Benefits of Schooling, *Journal of Economic Perspectives*, 25(1), pp. 159-184. doi: 10.1257/jep.25.1.159.

Pokropek, A. (2016) Introduction to instrumental variables and their application to large-scale assessment data, *Springer*. doi: 10.1186/s40536-016-0018-2.

Rattsø, J. and Stokke, H. E. (2020) Private-public wage gap and return to experience: Role of geography, gender and education, *Regional Science and Urban Economics*, 84, pp. 103571.

Reynolds, J. M. (2019) The meaning of ability and disability, *The Journal of Speculative Philosophy*, 33(3), pp. 434-447. doi: https://doi.org/10.5325/jspecphil.33.3.0434.

Schofer, E. and Meyer, J. W. (2005) The worldwide expansion of higher education in the twentieth century, *American sociological review*, 70(6), pp. 898-920. doi: 10.1177/000312240507000602.

Smith, W. C. and Fernandez, F. (2017) Education, skills, and wage gaps in Canada and the United States, *International Migration*, 55(3), pp. 57-73.

Statistisk Sentralbyrå (2020) Nasjonal utdanningsdatabase. Available at: https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/utdanning/nasjonal-utdanningsdatabase (Accessed: 28th of April, 2023).

Statistisk Sentralbyrå (2022) Inntekter, personlig næringsdrivende. Available at: https://www.ssb.no/inntekt-og-forbruk/inntekt-og-formue/statistikk/inntekter-personlig-naeringsdrivende (Accessed: 28th of April).

Statistisk Sentralbyrå (2023) *Konsumprisindeksen*. Available at: https://www.ssb.no/priser-og-prisindekser/konsumpriser/statistikk/konsumprisindeksen (Accessed: 8th of March 2023).

Statistisk Sentralbyrå (n.d.) *An institution that counts*. Available at: https://www.ssb.no/en/omssb/ssbs-virksomhet/tall-som-forteller (Accessed: 27th of April 2023).

Stokke, H. E. (2021) The gender wage gap and the early-career effect: the role of actual experience and education level. doi: https://doi.org/10.1111/labr.12191.

Streefkerk, R. (2019) *Internal vs. External Valudity | Understanding Differences and Threats*. Available at: https://www.scribbr.com/methodology/internal-vs-external-validity/ (Accessed: 5th of May 2023).

Studenttorget (2016) *De ulike veiene til en bachelorgrad*. Available at: https://studenttorget.no/index.php?artikkelid=8462 (Accessed: 20th of February 2023).

Sumeracki, M. (2018) Understanding Sample Sizes and the Word "Significant" (vol. 2023). Available at: https://www.learningscientists.org/blog/2018/11/1-1.

Tan, G. Y. (2013) Higher education reforms in China: For better or for worse?, *International Education*, 43(1), pp. 101. doi: https://www.proquest.com/scholarly-journals/higher-education-reforms-china-better-worse/docview/1467994525/se-2.

Utdanningsdirektoratet (2022) *Information for newly arrived parents and guardians: The education system in Norway*. Available at: https://www.udir.no/laring-og-trivsel/minoritetsspraklige-og-flyktninger/minoritetsspraklige/informasjon-til-nyankomne/information-for-newly-arrived/#a176785 (Accessed: 20th of February 2023).

Walker, I. and Zhu, Y. (2008) The College Wage Premium and the Expansion of Higher Education in the UK*, *The Scandinavian Journal of Economics*, 110(4), pp. 695-709. doi: 10.1111/j.1467-9442.2008.00557.x.

Wilkinson, M. (2013) Testing the null hypothesis: the forgotten legacy of Karl Popper?, *Journal of sports sciences*, 31(9), pp. 919-920. doi: doi.org/10.1080/02640414.2012.753636.

Wooldridge, J. M. (2019) *Introductory Econometrics: A Modern Approach*. 7th edn. Florence, AL: South-Western College Publishing.

Yankow, J. J. (2006) Why do cities pay more? An empirical examination of some competing theories of the urban wage premium, *Journal of Urban Economics*, 60(2), pp. 139-161.

Appendix

A1: Script with coding in microdata.no.

The script can be found in a separate document.

A2: Descriptive statistics for only higher educated individuals

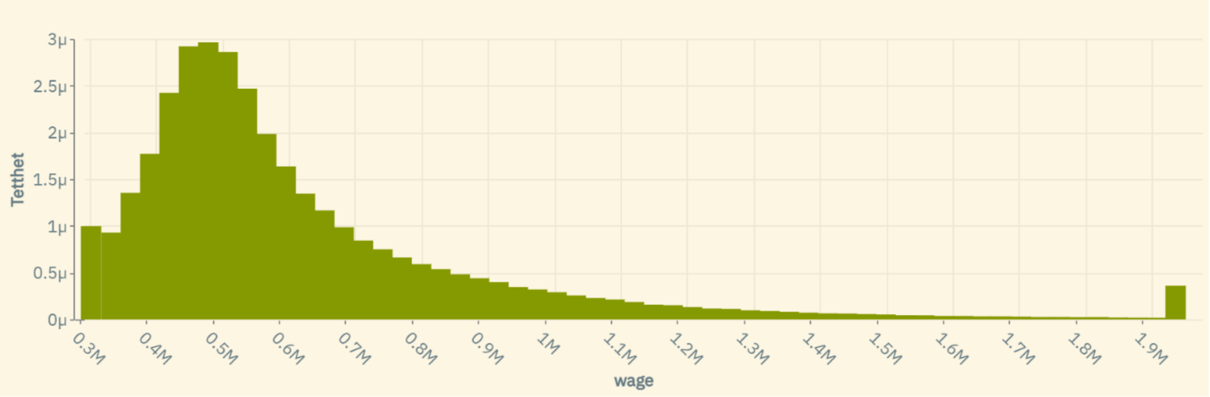
VARIABLE	OBS.	AVERAGE	STD. DEV.	1%	50%	99%
WAGE	605 649	634 467	285 131	304 000	547 000	1 960 000
LNWAGE	605 649	13.2855	0.3659	12.62	13.21	14.49
UTDNIV_HIGH	605 649	1.00	0.00	1	1	1
UTDNIV_VGS	605 649	0.00	0.00	0	0	0
UTDNIV_VGS_2Y	605 649	0.00	0.00	0	0	0
UTDNIV_VGS_3Y	605 649	0.00	0.00	0	0	0
UTDNIV_VGS_4Y	605 649	0.00	0.00	0	0	0
UTDNIV_BACHELOR	605 649	0.70637	0.4554	0	1	1
UTDNIV_MASTER	605 649	0.26735	0.4426	0	0	1
UTDNIV_PHD	605 649	0.02628	0.1600	0	0	1
MOTHEDUC	605 649	0.24869	0.4323	0	0	1
FATHEDUC	605 649	0.30445	0.4602	0	0	1
MALE	605 649	0.45354	0.4978	0	0	1
IMMIGRANTS_HIGHINC	605 649	0.06320	0.2433	0	0	1
IMMIGRANTS_LOWINC	605 649	0.04753	0.2128	0	0	1
MSTAT_UNMARRIED	605 649	0.38676	0.4870	0	0	1
MSTAT_MARRIED	605 649	0.51604	0.4997	0	1	1
MSTAT_DIVORCED	605 649	0.09719	0.2962	0	0	1
SECTOR_PRIVATE	605 649	0.43172	0.4953	0	0	1
OSLO	605 649	0.37764	0.4848	0	0	1
BIG_CITY	605 649	0.23997	0.4271	0	0	1
AGE	605 634	41.8	10.1	24	41	62
POT_EXP	605 634	19.1	10.1	2	19	39
POT_EXP_SQ	605 634	467	418	4	361	1 520
INDU_REF	605 649	0.05711	0.2320	0	0	1
INDU_1	605 649	0.00000	0.0000	0	0	0
INDU_2	605 649	0.03307	0.1788	0	0	1
INDU_3	605 649	0.00000	0.0000	0	0	0
INDU_4	605 649	0.00000	0.0000	0	0	0
INDU_5	605 649	0.02150	0.1450	0	0	1
INDU_6	605 649	0.05029	0.2186	0	0	1
INDU_7	605 649	0.01998	0.1399	0	0	1
INDU_8	605 649	0.00000	0.0000	0	0	0
INDU_9	605 649	0.06050	0.2384	0	0	1
INDU_10	605 649	0.03239	0.1770	0	0	1
INDU_11	605 649	0.00000	0.0000	0	0	0
INDU_12	605 649	0.09174	0.2887	0	0	1
INDU_13	605 649	0.02311	0.1503	0	0	1
INDU_14	605 649	0.11165	0.3149	0	0	1
INDU_15	605 649	0.17425	0.3793	0	0	1
INDU_16	605 649	0.24925	0.4326	0	0	1
INDU_17	605 649	0.01215	0.1095	0	0	1
INDU_18	605 649	0.01523	0.1225	0	0	1

INDU_19	605 649	0.00000	0.0000	0	0	0
INDU_20	605 649	0.00000	0.0000	0	0	0

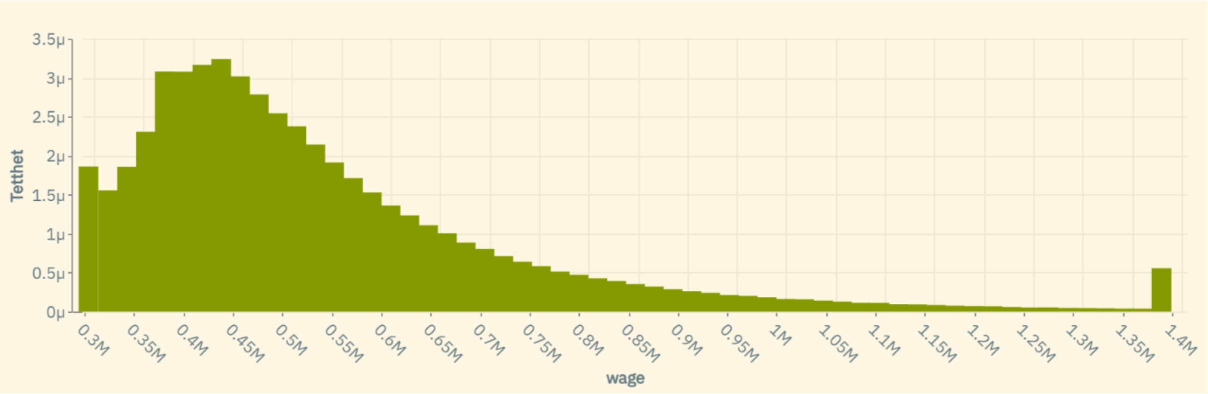
A3: Descriptive statistics for only lower educated individuals

VARIABLE	**OBS.**	**AVERAGE**	**STD. DEV.**	**1%**	**50%**	**99%**
WAGE	662 370	543 523	204 518	295 000	488 000	1 390 000
LNWAGE	662 370	13.14857	0.3248	12.60	13.10	14.15
UTDNIV_HIGH	662 370	0.00	0.00	0	0	0
UTDNIV_VGS	662 370	1.00	0.00	1	1	1
UTDNIV_VGS_2Y	662 370	0.13717	0.3440	0	0	1
UTDNIV_VGS_3Y	662 370	0.76616	0.4233	0	1	1
UTDNIV_VGS_4Y	662 370	0.09667	0.2955	0	0	1
UTDNIV_BACHELOR	662 370	0.00	0.00	0	0	0
UTDNIV_MASTER	662 370	0.00	0.00	0	0	0
UTDNIV_PHD	662 370	0.00	0.00	0	0	0
MOTHEDUC	662 370	0.07733	0.2671	0	0	1
FATHEDUC	662 370	0.09382	0.2916	0	0	1
MALE	662 370	0.67935	0.4667	0	1	1
IMMIGRANTS_HIGHINC	662 370	0.06512	0.2467	0	0	1
IMMIGRANTS_LOWINC	662 370	0.03606	0.1864	0	0	1
MSTAT_UNMARRIED	662 370	0.40422	0.4907	0	0	1
MSTAT_MARRIED	662 370	0.47196	0.4992	0	0	1
MSTAT_DIVORCED	662 370	0.12382	0.3294	0	0	1
SECTOR_PRIVATE	662 370	0.73198	0.4429	0	1	1
OSLO	662 370	0.26441	0.4410	0	0	1
BIG_CITY	662 370	0.24054	0.4274	0	0	1
AGE	662 354	43.2	11.1	21	44	62
POT_EXP	662 354	24.2	11.2	2	25	43
POT_EXP_SQ	662 354	714	536	4	625	1840
INDU_REF	662 370	0.15222	0.3592	0	0	1
INDU_1	662 370	0.00000	0.0000	0	0	0
INDU_2	662 370	0.04289	0.2026	0	0	1
INDU_3	662 370	0.01026	0.1007	0	0	1
INDU_4	662 370	0.00000	0.0000	0	0	0
INDU_5	662 370	0.14975	0.3568	0	0	1
INDU_6	662 370	0.15699	0.3638	0	0	1
INDU_7	662 370	0.07639	0.2656	0	0	1
INDU_8	662 370	0.01992	0.1397	0	0	1
INDU_9	662 370	0.03068	0.1724	0	0	1
INDU_10	662 370	0.02102	0.1435	0	0	1
INDU_11	662 370	0.00000	0.0000	0	0	0
INDU_12	662 370	0.03535	0.1847	0	0	1
INDU_13	662 370	0.04793	0.2136	0	0	1
INDU_14	662 370	0.05534	0.2286	0	0	1
INDU_15	662 370	0.02166	0.1456	0	0	1
INDU_16	662 370	0.11246	0.3159	0	0	1
INDU_17	662 370	0.00000	0.0000	0	0	0
INDU_18	662 370	0.01523	0.1224	0	0	1
INDU_19	662 370	0.00000	0.0000	0	0	0
INDU_20	662 370	0.00000	0.0000	0	0	0

A4: Wage distribution amongst higher educated individuals.



A5: Wage distribution amongst lower educated individuals.



A6: Overview of industry dummies, their description and distribution for only higher educated individuals.

VARIABLE NAME	INDUSTRY DESCRIPTION	SHARE OF DATASET
indu_ref	Manufacturing and other industries.	5.71%
indu_1	Agriculture. forestry. and fishing.	0.28%
indu_2	Mining and quarrying.	3.31%
indu_3	Electricity. gas. steam. and hot water supply.	0.74%
indu_4	Water supply. sewerage. and waste management.	0.28%
indu_5	Construction.	2.15%

indu_6	Wholesale and retail trade. and repair of motor vehicles.	5.03%
indu_7	Transportation and storage.	2.00%
indu_8	Accommodation and food service activities.	0.73%
indu_9	Information and communication.	6.05%
indu_10	Financial and insurance activities.	3.24%
indu_11	Real estate activities.	0.83%
indu_12	Professional. scientific. and technical activities.	9.17%
indu_13	Administrative and support service activities.	2.31%
indu_14	Public administration and defence. and compulsory social security.	11.16%
indu_15	Education.	17.42%
indu_16	Human health and social work activities.	24.93%
indu_17	Arts. entertainment. and recreation.	1.21%
indu_18	Other service activities.	1.52%
indu_19	Activities of households as employers.	none
indu_20	Activities of extraterritorial organizations and bodies.	none
Total		**≈ 100% (98.07%)***

*Rounding errors makes the total not sum up to exactly 100%.

A7: Overview of industry dummies, their description and distribution for only lower educated individuals.

VARIABLE NAME	**INDUSTRY DESCRIPTION**	**SHARE OF DATASET**
indu_ref	Manufacturing and other industries.	15.22%
indu_1	Agriculture. forestry. and fishing.	0.90%
indu_2	Mining and quarrying.	4.29%
indu_3	Electricity. gas. steam. and hot water supply.	1.03%
indu_4	Water supply. sewerage. and waste management.	0.88%
indu_5	Construction.	14.97%

indu_6	Wholesale and retail trade. and repair of motor vehicles.	15.70%
indu_7	Transportation and storage.	7.64%
indu_8	Accommodation and food service activities.	1.99%
indu_9	Information and communication.	3.07%
indu_10	Financial and insurance activities.	2.10%
indu_11	Real estate activities.	0.83%
indu_12	Professional. scientific. and technical activities.	3.54%
indu_13	Administrative and support service activities.	4.79%
indu_14	Public administration and defence. and compulsory social security.	5.53%
indu_15	Education.	2.17%
indu_16	Human health and social work activities.	11.25 %
indu_17	Arts. entertainment. and recreation.	0.80%
indu_18	Other service activities.	1.52%
indu_19	Activities of households as employers.	0.005%
indu_20	Activities of extraterritorial organizations and bodies.	none
Total		**≈ 100% (98.225%)***

*Rounding errors makes the total not sum up to exactly 100%.

A8: Ramseys RESET test for the full model, model (4).

Ramseys RESET test	
F(3, 1 267 952):	3581.358601
Prob > F:	0

A9: Simultaneity analytics

Taking a closer look at the simultaneity between wages and education within the two simplified regression models without constant terms. We have the equations:

(1) $wage = \alpha_1 education + \beta_1 z_1 + u_1$

(2) $education = \alpha_2 wage + \beta_2 z_2 + u_2$

We start off by inserting (1) into (2) and gets:

$$\textit{education} = \alpha_2[\alpha_1 \textit{education} + \beta_1 z_1 + u_1] + \beta_2 z_2 + u_2$$

Then we solve for *education*:

$$\textit{education} = \alpha_2\alpha_1 \textit{education} + \alpha_2\beta_1 z_1 + \alpha_2 u_1 + \beta_2 z_2 + u_2$$

$$\textit{education} - \alpha_2\alpha_1 \textit{education} = \alpha_2\beta_1 z_1 + \alpha_2 u_1 + \beta_2 z_2 + u_2$$

$$\textit{education}(1 - \alpha_2\alpha_1) = \alpha_2\beta_1 z_1 + \alpha_2 u_1 + \beta_2 z_2 + u_2$$

$$\textit{education} = \frac{\alpha_2\beta_1 z_1}{1-\alpha_2\alpha_1} + \frac{\alpha_2 u_1}{1-\alpha_2\alpha_1} + \frac{\beta_2 z_2}{1-\alpha_2\alpha_1} + \frac{u_2}{1-\alpha_2\alpha_1}$$

And finally, we simplify it a little by gathering the error terms together and leaving the variables outside of the fraction:

$$\textit{education} = \frac{\alpha_2\beta_1}{1-\alpha_2\alpha_1} z_1 + \frac{\beta_2}{1-\alpha_2\alpha_1} z_2 + \frac{\alpha_2 u_1 + u_2}{1-\alpha_2\alpha_1}$$

By defining the following:

$$\pi_{21} = \frac{\alpha_2\beta_1}{1-\alpha_2\alpha_1}$$

$$\pi_{22} = \frac{\beta_2}{1-\alpha_2\alpha_1}$$

$$v_2 = \frac{\alpha_2 u_1 + u_2}{1-\alpha_2\alpha_1}$$

We're left with the reduced form equation for education:

$$\textit{education} = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

A10: Correlation matrix.

The correlation matrix can be found in a separate document.

A11: Variation Inflation Factor (VIF) test

VARIANCE INFLATION FACTOR	VIF	1/VIF
utdniv_bachelor	1.468617	0.680913
utdniv_master	1.365457	0.732356
utdniv_phd	1.072011	0.932826
male	1.300503	0.768933
immigrants_highinc	1.037086	0.96424
immigrants_lowinc	1.042514	0.95922
mstat_married	1.463718	0.683192
mstat_divorced	1.344561	0.743737
indu_1	1.042785	0.95897
indu_2	1.286653	0.77721
indu_3	1.06436	0.939531
indu_4	1.041991	0.959701
indu_5	1.592311	0.628018
indu_6	1.681999	0.594531
indu_7	1.334029	0.749609
indu_8	1.112571	0.898819
indu_9	1.354885	0.73807
indu_10	1.21184	0.825191
indu_11	1.062399	0.941266
indu_12	1.496933	0.668033
indu_13	1.255927	0.796224
indu_14	1.616751	0.618525
indu_15	1.837932	0.54409
indu_16	2.338885	0.427554
indu_17	1.077844	0.927778
indu_18	1.122417	0.890934
indu_19	1.000331	0.999669
indu_20	1.000019	0.999981
pot_exp	18.328439	0.05456
pot_exp_sq	17.497523	0.057151
oslo	1.277455	0.782806
big_city	1.212909	0.824464
AVERAGE	2.341989	-

A12: Breusch-Pagan heterogeneity test for the full model, model (4).

Breusch-Pagan	
chi2(1):	105481.651379
Prob > chi2:	0

Breusch-Pagan, studentisert	
chi2(1):	33741.536758
Prob > chi2:	0

Breusch-Pagan, f-test	
F(1, 1 267 986):	34663.900203
Prob > F:	0

A13: Table showing the development of educational wage premium when only including one control variable.

	(*1)	**(*2)**	**(*3)**	**(*4)**	**(*5)**	**(*6)**	**(*7)**
	OLS	**OLS**	**OLS**	**OLS**	**OLS**	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**
utdniv_high	0.19452***	0.12987***	0.16288***	0.13546***	0.13943***	0.17573***	0.18579***
	(0.00063)	(0.00064)	(0.00064)	(0.00063)	(0.00064)	(0.00070)	(0.00072)
male	0.25032***						
	(0.00061)						
oslo		0.07209***					
		(0.00074)					
big_city		0.06888***					
		(0.00081)					
pot_exp			0.02564***				
			(0.00011)				
pot_exp_sq			-0.00043***				
			(0.00000)				
mstat_married				0.11223***			
				(0.00066)			
mstat_divorced				0.09075***			
				(0.00102)			
immigrants_highinc					-0.12405***		
					(0.00130)		
imigrannts_lowinc					-0.14606***		
					(0.00140)		
sector_private						0.12567***	
						(0.00068)	
Constant	12.98091***	13.11533***	12.83825***	13.08676***	13.16431***	13.05897***	13.16832***
	(0.00052)	(0.00050)	(0.00104)	(0.00053)	(0.00043)	(0.00062)	(0.00092)
Industry controls	No	No	No	No	No	No	Yes
Observations	1 268 018	1 268 018	1 267 992	1 268 018	1 268 018	1 268 018	1 268 018
R-squared	0.145	0.045	0.095	0.057	0.048	0.062	0.203

Standard deviation in parentheses

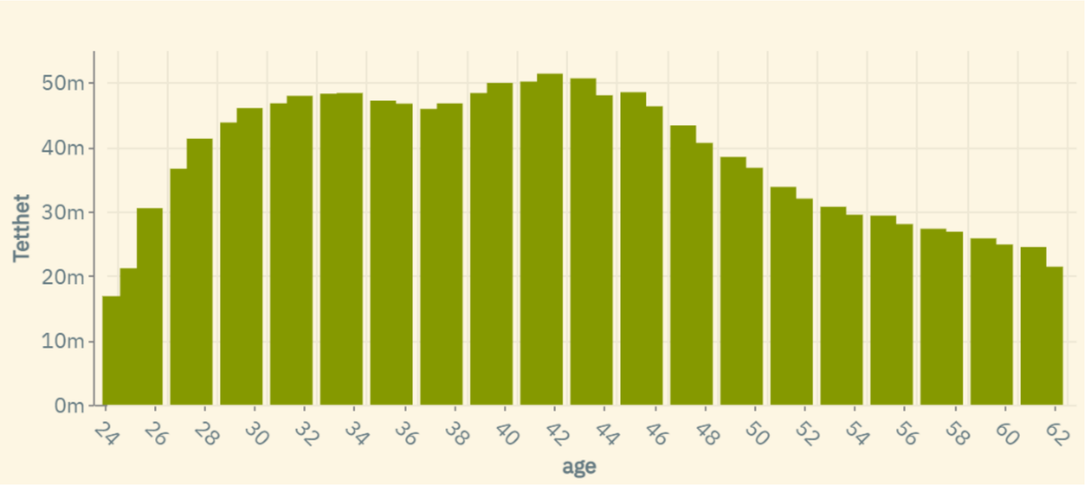
***p<0.01, **p<0.05, *p<0.1

A14 Table showing the development of educational wage premium for the subgroups of degrees when only including one control variable.

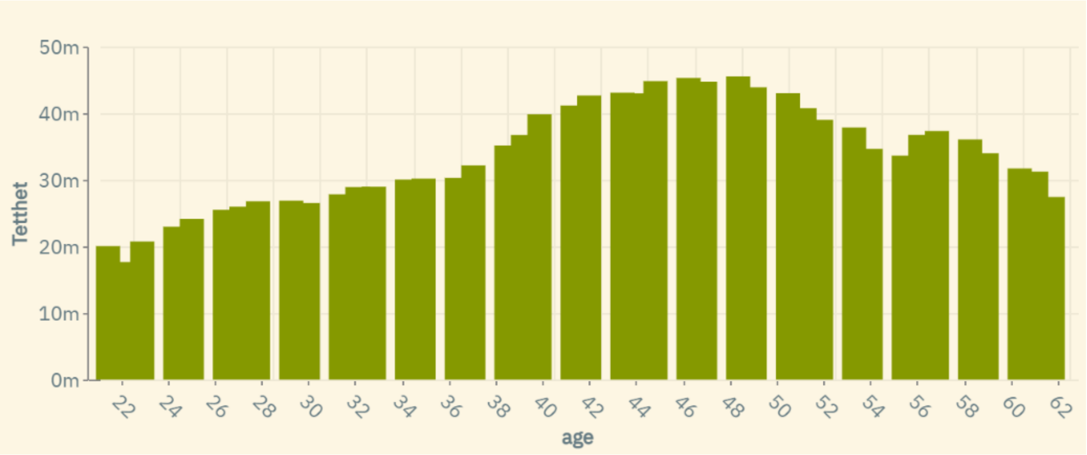
	(‘1)	**(‘2)**	**(‘3)**	**(‘4)**	**(‘5)**	**(‘6)**	**(‘7)**
	OLS	**OLS**	**OLS**	**OLS**	**OLS**	**OLS**	**OLS**
VARIABLES	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**	**lnwage**
utdniv_bachelor	0.13279*** (0.00067)	0.06648*** (0.00067)	0.09058*** (0.00067)	0.06836*** (0.00067)	0.07269*** (0.00067)	0.10864*** (0.00073)	0.13508*** (0.00075)
utdniv_master	0.32604*** (0.00104)	0.27773*** (0.00109)	0.33636*** (0.00104)	0.28828*** (0.00106)	0.28827*** (0.00108)	0.32584*** (0.00109)	0.30335*** (0.00111)
utdniv_phd	0.42789*** (0.00295)	0.39348*** (0.00303)	0.42974*** (0.00282)	0.39321*** (0.00297)	0.41988*** (0.00299)	0.46738*** (0.00301)	0.46630*** (0.00309)
male	0.24010*** (0.00060)						
oslo		0.05720*** (0.00073)					
big_city		0.05799*** (0.00079)					
pot_exp			0.02707*** (0.00010)				
pot_exp_sq			-0.00045*** (0.00000)				
mstat_married				0.11022*** (0.00065)			
mstat_divorced				0.09589*** (0.00100)			
immigrants_highinc					-0.11840*** (0.00128)		
imigrannts_lowinc					-0.14646*** (0.00135)		
sector_private						0.12703*** (0.00066)	
Constant	12.98785*** (0.00052)	13.12189*** (0.00050)	12.81805*** (0.00103)	13.08707*** (0.00052)	13.16395*** (0.00043)	13.05798*** (0.00061)	13.16735*** (0.00091)
Industry controls	No	No	No	No	No	No	Yes
Observations	1 268 018	1 268 018	1 267 992	1 268 018	1 268 018	1 268 018	1 268 018
R-squared	0.176	0.082	0.144	0.097	0.088	0.103	0.229

Standard deviation in parentheses, ***p<0.01, **p<0.05, *p<0.1

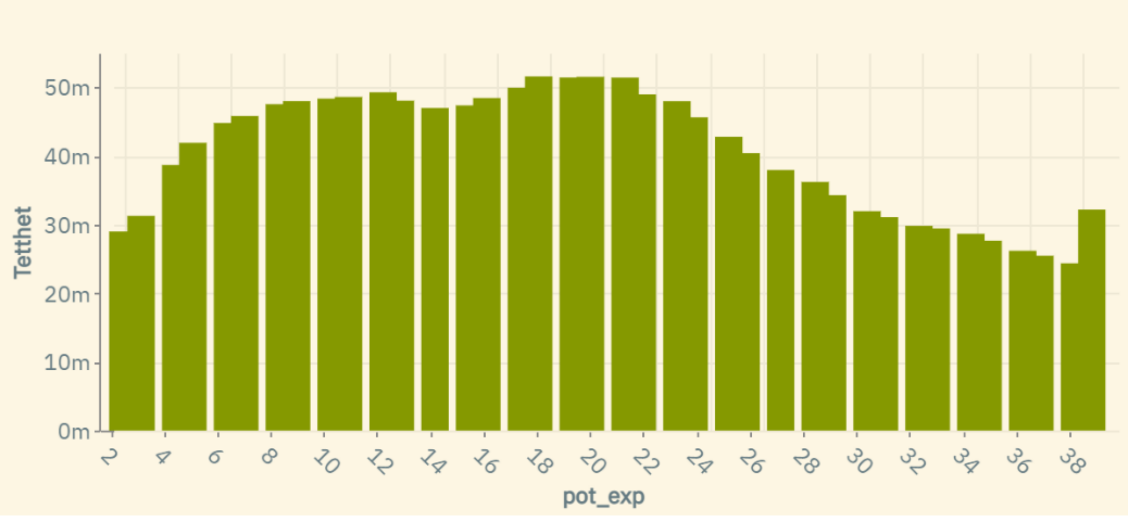
A15: Age distribution for higher educated individuals.



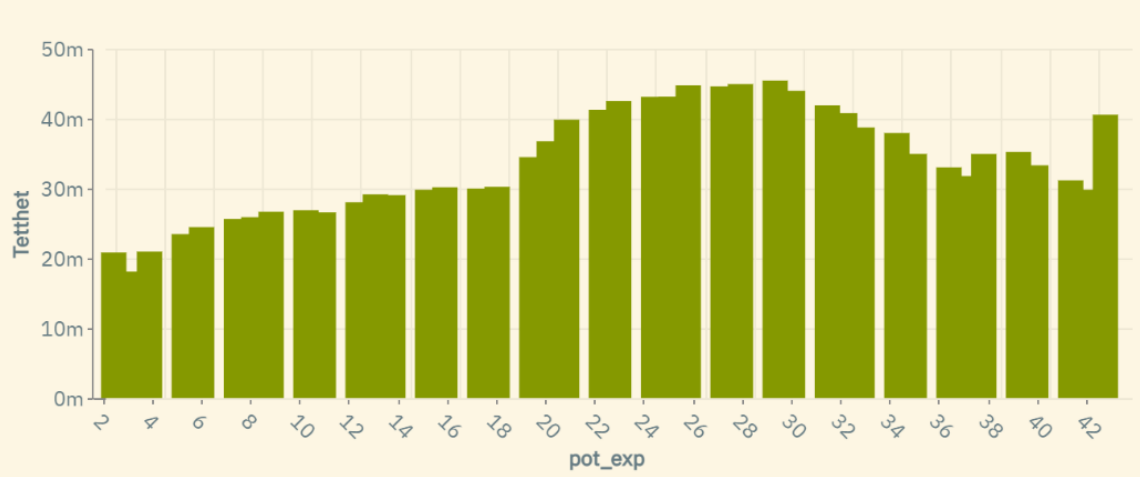
A16: Age distribution for lower educated individuals.



A17: Potential experience distribution for higher educated individuals.



A18: Potential experience distribution for lower educated individuals.



A19: Regression model (4) reporting all coefficients.

	(3)
	OLS
VARIABLES	**lnwage**
utdniv_bachelor	0.18159*** (0.00070)
utdniv_master	0.35593*** (0.00103)
utdniv_phd	0.47511*** (0.00269)
male	0.18854*** (0.00059)

oslo	0.07146*** (0.00063)
big_city	0.03509*** (0.00064)
pot_exp	0.02610*** (0.00009)
pot_exp_sq	-0.00044*** (0.00000)
mstat_married	0.04474*** (0.00061)
mstat_divorced	0.03671*** (0.00093)
immigrants_highinc	-0.14063*** (0.00118)
immigrants_lowinc	-0.14261*** (0.00131)
indu_1	-0.05197*** (0.00384)
indu_2	0.42334*** (0.00184)
indu_3	0.13692*** (0.00253)
indu_4	-0.06318*** (0.00300)
indu_5	-0.01006*** (0.00118)
indu_6	-0.03906*** (0.00118)
indu_7	0.02883*** (0.00160)
indu_8	-0.13625*** (0.00228)
indu_9	0.08987*** (0.00164)
indu_10	0.15579*** (0.00229)
indu_11	0.10652*** (0.00422)
indu_12	0.05571*** (0.00149)
indu_13	-0.07905*** (0.00175)
indu_14	-0.07075*** (0.00117)

indu_15	-0.17579*** (0.00113)
indu_16	-0.12261*** (0.00107)
indu_17	-0.14509*** (0.00267)
indu_18	-0.10070*** (0.00223)
indu_19	-0.12244*** (0.03833)
indu_20	-0.32208*** (0.00139)
Constant	12.67223*** (0.00130)
Observations	1 267 992
R-squared	0.373

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

A20: Regression results from first stage 2SLS attempting National Tests as instrument.

VARIABLES	**(^)** **OLS** **utdniv_high**	**(^^)** **OLS** **utdniv_high**	**(^^^)** **OLS** **utdniv_high**
ntest_eng	-0.02105** (0.00346)		
ntest_read		0 (0)	
ntest_math			-0.01784** (0.00322)
Constant	0.94210*** (0.09224)	0 (0)	0.92607** (0.11814)
Vector *x* controls	No	No	No
Vector *b* controls	No	No	No
Observations	0	5	0
R-squared	0.923	NaN	0.901

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

A21: Distribution of all dummy variables for full sample.

	Numeric		**Percentage**	
VARIABLES	**1**	**0**	**1**	**0**
utdniv_high	605 650	662 368	47.76 %	52.24 %
utdniv_vgs	662 368	605 650	52.24 %	47.76 %
utdniv_vgs_2y	90 855	1 177 157	7.17 %	92.83 %
utdniv_vgs_3y	507 480	760 532	40.02 %	59.98 %
utdniv_vgs_4y	64 034	1 203 981	5.05 %	94.95 %
utdniv_bachelor	427 811	840 205	33.74 %	66.26 %
utdniv_master	161 922	1 106 091	12.77 %	87.23 %
utdniv_phd	15 917	1 252 097	1.26 %	98.74 %
male	724 668	543 353	57.15 %	42.85 %
immigrants_highinc	81 405	1 186 609	6.42 %	93.58 %
immigrants_lowinc	52 664	1 215 351	4.15 %	95.85 %
mstat_unmarried	501 988	766 031	39.59 %	60.41 %
mstat_married	625 148	642 866	49.30 %	50.70 %
mstat_divorced	140 873	1 127 135	11.11 %	88.89 %
oslo	403 860	864 157	31.85 %	68.15 %
big_city	304 665	963 346	24.03 %	75.97 %
sector_private	746 304	521 704	58.86 %	41.14 %
indu_ref	135 410	1 132 610	10.68 %	89.32 %
indu_1	7 679	1 260 330	0.61 %	99.39 %
indu_2	48 434	1 219 581	3.82 %	96.18 %
indu_3	11 296	1 256 719	0.89 %	99.11 %
indu_4	7 515	1 260 500	0.59 %	99.41 %
indu_5	112 206	1 155 805	8.85 %	91.15 %
indu_6	134 451	1 133 565	10.60 %	89.40 %
indu_7	62 705	1 205 312	4.95 %	95.05 %
indu_8	17 631	1 250 385	1.39 %	98.61 %
indu_9	56 963	1 211 051	4.49 %	95.51 %
indu_10	33 543	1 234 472	2.65 %	97.35 %
indu_11	10 544	1 257 471	0.83 %	99.17 %
indu_12	78 978	1 189 034	6.23 %	93.77 %
indu_13	45 744	1 222 270	3.61 %	96.39 %
indu_14	104 272	1 163 739	8.22 %	91.78 %
indu_15	119 875	1 148 135	9.45 %	90.55 %
indu_16	225 449	1 042 564	17.78 %	82.22 %
indu_17	12 687	1 255 337	1.00 %	99.00 %
indu_18	19 309	1 248 707	1.52 %	98.48 %
indu_19	37	1 267 975	0.00 %	100.00 %
indu_20	0	1 268 018	0.00 %	100.00 %
motheduc	201 840	1 066 180	15.92 %	84.08 %
fatheduc	246 535	1 021 481	19.44 %	80.56 %

A22: Distribution of all dummy variables for higher educated individuals.

	Numeric		Percentage	
VARIABLES	**1**	**0**	**1**	**0**
utdniv_high	605 649	0	100.00 %	0.00 %
utdniv_vgs	0	605 649	0.00 %	100.00 %
utdniv_vgs_2y	0	605 649	0.00 %	100.00 %
utdniv_vgs_3y	0	605 649	0.00 %	100.00 %
utdniv_vgs_4y	0	605 649	0.00 %	100.00 %
utdniv_bachelor	427 813	177 835	70.64 %	29.36 %
utdniv_master	161 920	443 729	26.73 %	73.27 %
utdniv_phd	15 913	589 733	2.63 %	97.37 %
male	274 686	330 965	45.35 %	54.65 %
immigrants_highinc	38 280	567 371	6.32 %	93.68 %
immigrants_lowinc	28 784	576 860	4.75 %	95.25 %
mstat_unmarried	234 248	371 402	38.68 %	61.32 %
mstat_married	312 547	293 108	51.61 %	48.40 %
mstat_divorced	58 866	546 782	9.72 %	90.28 %
oslo	228 725	376 934	37.77 %	62.24 %
big_city	145 338	460 312	24.00 %	76.00 %
sector_private	261 469	344 181	43.17 %	56.83 %
indu_ref	34 589	571 062	5.71 %	94.29 %
indu_1	1 724	603 926	0.28 %	99.72 %
indu_2	20 022	585 625	3.31 %	96.69 %
indu_3	4 503	601 148	0.74 %	99.26 %
indu_4	1 699	603 953	0.28 %	99.72 %
indu_5	13 019	592 624	2.15 %	97.85 %
indu_6	30 465	575 192	5.03 %	94.97 %
indu_7	12 097	593 551	2.00 %	98.00 %
indu_8	4 430	601 219	0.73 %	99.27 %
indu_9	36 636	569 008	6.05 %	93.95 %
indu_10	19 614	568 037	3.24 %	93.79 %
indu_11	5 016	600 632	0.83 %	99.17 %
indu_12	55 568	550 085	9.17 %	90.83 %
indu_13	13 993	591 649	2.31 %	97.69 %
indu_14	67 615	538 029	11.16 %	88.84 %
indu_15	105 531	500 121	17.42 %	82.58 %
indu_16	150 962	454 690	24.93 %	75.07 %
indu_17	7 354	598 293	1.21 %	98.79 %
indu_18	9 220	596 429	1.52 %	98.48 %
indu_19	0	605 648	0.00 %	100.00 %
indu_20	0	605 648	0.00 %	100.00 %
motheduc	150 622	455 028	24.87 %	75.13 %
fatheduc	184 394	421 255	30.45 %	69.55 %

A23: Distribution of all dummy variables for lower educated individuals.

	Numeric		Percentage	
VARIABLES	**1**	**0**	**1**	**0**
utdniv_high	0	662 370	0.00 %	100.00 %
utdniv_vgs	662 370	0	100.00 %	0.00 %
utdniv_vgs_2y	90 854	571 509	13.72 %	86.28 %
utdniv_vgs_3y	507 474	154 886	76.61 %	23.38 %
utdniv_vgs_4y	64 028	598 339	9.67 %	90.33 %
utdniv_bachelor	0	662 370	0.00 %	100.00 %
utdniv_master	0	662 370	0.00 %	100.00 %
utdniv_phd	0	662 370	0.00 %	100.00 %
male	449 974	212 381	67.93 %	32.06 %
immigrants_highinc	43 137	619 230	6.51 %	93.49 %
immigrants_lowinc	23 888	638 477	3.61 %	96.39 %
mstat_unmarried	267 739	394 626	40.42 %	59.58 %
mstat_married	312 606	349 751	47.20 %	52.80 %
mstat_divorced	82 006	580 350	12.38 %	87.62 %
oslo	175 137	487 225	26.44 %	73.56 %
big_city	159 323	503 043	24.05 %	75.95 %
sector_private	484 840	177 523	73.20 %	26.80 %
indu_ref	100 820	561 544	15.22 %	84.78 %
indu_1	5 957	656 411	0.90 %	99.10 %
indu_2	28 408	633 962	4.29 %	95.71 %
indu_3	6 794	655 570	1.03 %	98.97 %
indu_4	5 822	656 551	0.88 %	99.12 %
indu_5	99 187	563 172	14.97 %	85.02 %
indu_6	103 986	558 381	15.70 %	84.30 %
indu_7	50 598	611 759	7.64 %	92.36 %
indu_8	13 192	649 170	1.99 %	98.01 %
indu_9	20 324	642 041	3.07 %	96.93 %
indu_10	13 930	648 443	2.10 %	97.90 %
indu_11	5 528	656 837	0.83 %	99.16 %
indu_12	23 421	638 950	3.54 %	96.46 %
indu_13	31 750	630 614	4.79 %	95.21 %
indu_14	36 659	625 706	5.53 %	94.46 %
indu_15	14 346	648 020	2.17 %	97.83 %
indu_16	74 493	587 876	11.25 %	88.75 %
indu_17	5 323	657 045	0.80 %	99.20 %
indu_18	10 089	652 275	1.52 %	98.48 %
indu_19	32	662 331	0.00 %	99.99 %
indu_20	0	662 370	0.00 %	100.00 %
motheduc	51 222	611 144	7.73 %	92.27 %
fatheduc	62 141	600 218	9.38 %	90.62 %

A24: Descriptive statistics for National Test scores.

VARIABLE	OBS.	AVERAGE	STD. DEV.	1%	50%	99%
NTEST_MATH	5	-	-	-	-	-
NTEST_ENG	5	-	-	-	-	-
NTEST_READ	5	-	-	-	-	-

A25: Descriptive statistics for interaction terms.

VARIABLE	OBS.	AVERAGE	STD. DEV.	1%	50%	99%
MALE_HIGH	1 268 018	0.21663	0.4119	0	0	1
IMMHINC_HIGH	1 268 018	0.03019	0.1711	0	0	1
IMMLINC_HIGH	1 268 018	0.02270	0.1490	0	0	1
OSLO_HIGH	1 268 018	0.18038	0.3845	0	0	1
BIGCITY_HIGH	1 268 018	0.11462	0.3186	0	0	1

A26: OLS model (4) for only higher educated workers.

VARIABLES	OLS lnwage
utdniv_bachelor	3.07931*** (0.00094)
utdniv_master	3.27032*** (0.00102)
utdniv_phd	3.38546*** (0.00204)
male	0.16967*** (0.00084)
oslo	0.07840*** (0.00091)
big_city	0.02699*** (0.00096)
pot_exp	0.03115*** (0.00015)
pot_exp_sq	-0.00050*** (0.00000)
mstat_married	0.04325*** (0.00087)
mstat_divorced	0.03071*** (0.00144)
immigrants_highinc	-0.09650*** (0.00187)
imigraints_lowinc	-0.14310*** (0.00194)
Constant	9.73509*** (0.00181)

Vector *b* controls	Yes
Observations	605 635
R-squared	0.402

Standard deviation in parentheses

***p<0.01. **p<0.05. *p<0.1

A27: Regressions with sector in interest, where higher and lower educated individuals are separated.

	(Lower)	**(Higher)**
	OLS	**OLS**
VARIABLES	**lnwage**	**lnwage**
utdniv_high	0.00000*** (0.00000)	6.35944*** (0.00079)
sector_private	0.05029*** (0.00083)	0.12173*** (0.00094)
Constant	12.65574*** (0.00159)	6.35944*** (0.00079)
Vector *x* controls	Yes	Yes
Vector *b* controls	No	No
Observations	662 347	605 635
R-squared	0.201	0.259

Standard deviation in parentheses

***p<0.01, **p<0.05, *p<0.1

NTNU

Norwegian University of
Science and Technology