UiO : University of Oslo

Mobility and the return to field of study

Snorre Skagseth

Masters thesis at the Department of Economics

Economics Master's two years

University of Oslo

November 2023

Acknowledgements

I would like to thank my supervisor, Edwin Leuven. I also want to thank Oda Elnan Lorentzen for putting up with me for the last few months. It has been a ride.

Abstract

The primary concern of the paper is whether self-selected migration is one of the factors driving the differences in the hourly wage premiums within fields of higher education, across regions in Norway.

To explore the question, I use Norwegian register data. Combining application data to tertiary education between 1998 to 2018 with earnings data from the Norwegian tax registry, the national education registry, and the Norwegian population registry.

To correct for self-selected migration I use a generalized Heckman correction model, using the multinomial logistic regression as the choice model, and implementing three separate control functions, to purge the bias from the coefficients of interest. The code is implemented in R.

The results show a few instances of bias in the OLS returns that rise to the level of statistical significance, with estimated biases ranging from 0.4 to -2.2 percentage points. Controlling for self-selected migration does not narrow the range of wage premiums within fields across regions.

The paper also documents the differing mobility patterns between different types of tertiary education and shows how migration propensity varies based on field of study and the region they resided in when growing up. Notably, the migration propensity into different regions varies greatly depending on the type of education individuals possess.

The paper also documents how even after controlling for age, gender, application scores, and parental education there is substantial variation in the returns to education across different counties, particularly pronounced in fields typical of the private sector.

Overall, this thesis sheds light on the intricate relationship between mobility patterns, education returns, and self-selected migration, contributing new insights to the existing literature, particularly in the context of Norway.

Keywords: return to field of study, migration, selection bias
JEL codes: D04; H43; I23; I28; J24

Contents

List of Figures

List of Tables

1 Introduction

When looking at the hourly wage returns to different fields of study across regions in Norway using wage data from 2018, it is striking to see the within-field differences. Examples are how civil engineering in Rogaland commands a 15% hourly wage premium compared to Trøndelag or how having studied medicine and working Sogn og Fjordane is associated with a 20% higher hourly wage rate compared to Oslo, after controlling for gender, experience, ability, and socioeconomic background characteristics. Another fact is that 43% of full-time working adults between the ages of 28 and 43 do not live in the county where they grew up.

With the returns to the field of study varying greatly and the Norwegian population being fairly mobile, it begs the question of why the returns to the type of education have not equalized across counties. One potential explanation is self-selected migration. If the types of workers that choose to move are different in the wage-earning potential to the population that stays, it could lead to upward or downward bias in the ordinary least squares estimates.

Often, when economists estimate equations of interest where self-selection can be a concern, they try to correct for the bias in various ways. One way is to use methods like instrumental variables, where the researcher models the dependent variable of interest directly. By using an exogenous source of variation that determines the dependent variable of interest but does not affect the outcome variable other than through the induced changing of the dependent variable, they purge the bias from the result. This would lead to estimates that could be interpreted causally with different interpretations depending on whether one assumes homogeneous or heterogeneous treatment effects.

An alternative approach pioneered by Heckman (1979) is to model the selection process directly and specify how it correlates with the error term. The intuition behind the Heckman correction arises from scenarios where the data available for analysis is not randomly sampled, and certain observations are systematically excluded or included based on some criteria.

The problem of potential self-selection migration causing differences in returns across regions is conceptually closer to a Heckman-type model where the worker makes a choice of where to work and live, and the choice can be dependent on wage-driving background characteristics. The difference is that instead of the choice margin being between two choices (e.g., work or not), the selection is over a large number of exclusive choices. There are different ways to deal with this, but in the current setting, I will mobilize a multinomial logistic regression to model the self-selected migration and use the estimated selection probabilities in the outcome equation to correct for the self-selection.

In this paper, I am mainly concerned with how self-selected migration affects wages across regions. There is however evidence from Norway by Kirkeboen et al. (2016) that the returns to different fields of higher education are in line with individuals choosing fields where they have a comparative advantage. This indicates that there is meaningful self-

selection on what field people choose to study. To get truly causal estimates on the returns to education across regions, one would need to deal with both types of self-selection at once, which is somewhat of a daunting task.

I, therefore, throughout the thesis, assume that the conditional independence assumption holds in such a way that when conditioning on the full list of covariates, the coefficients of interest capture the causal effect of obtaining a type of education with any remaining bias in the returns stemming from the self-selected migration.

Borjas et al. (1992) represents some of the earlier work on internal migration and the returns to the level of education. They develop a theoretical Roy (1951) model with regional differences in returns to skill. The model predicts that regions that offer higher returns to skill attract more skilled labor. In the American setting, their empirical findings indicate that regions with the highest returns to skill see the biggest inflow of skilled labor, with the mismatch between a worker's skills and the returns to the type of skill as one of the biggest determinants for them moving away.

One attempt at correcting potential self-selection bias in the returns to the level of education is Dahl (2002). The author develops a semi-parametric method where he estimates the migration probabilities by using cell means and includes them in the earnings equation using polynomial expansion. He applies his method to the 1980 US census data and finds that the ordinary least squares (OLS) method tends to overestimate earnings for people with higher education who move to states where the returns to their education are higher. However, the study does not help to narrow the range of returns to college education across states.

A more recent addition to the literature is Ransom (2021). He investigates the returns to college majors in the context of the US, also using a control function approach. Building on Dahl's, he uses the conditional inference recursive partitioning method developed Hothorn et al. (2006) to estimate the selection probabilities into occupation and where to live. His findings suggest that the OLS estimates for STEM and business majors are biased upwards compared to education majors by 15% at the median.

In the Scandinavian context Korpi and Clark (2015) uses Swedish panel data between 2001-2009, to assess the distribution of migration income change. Using matching and quantile regression, they find that the largest returns to education are captured by the highly educated, the ones lowest in the income distribution, and the ones moving to the largest metropolitan areas.

The focus of my thesis is threefold. Firstly, I try to document some of the overarching mobility patterns among full-time working individuals with tertiary education between the ages of 28 and 43 and how they differ depending on the field they have studied. Secondly, I document how the returns to the type of higher education differ within fields across regions. Thirdly, I investigate the effect self-selected migration has on biasing the regional wage premiums.

In many ways, my paper is closely related to Dahl (2002), and I will use his model as a starting point, though I will use parametric techniques to estimate selection probabilities. What my paper adds to the literature is primarily to explore the degree to what degree self-selected migration drives wage differences between regions in a new setting, that is Norway. In line with Ransom (2021) I only look at the returns for people with higher education but having a greater variety of fields being represented. One of the great features of my dataset is that I have access to a fairly good measure of ability (individuals application scores). This means that I, to some extent, can partial out the effect of ability before controlling for self-selected migration.

The mobility data show that people with different types of higher education experience stark differences in the propensity to live in a different county from where they grew up, with age and what county one grew up in being important factors. The data also show that migration propensity into different regions strongly varies depending on the type of education people have.

Using an OLS regression where I control for age as dummies interacted with gender, application score, and parental education shows a great degree of variation in the returns to the field of education across different counties, with the variation being more pronounced in typical private sector fields.

Correcting for self-selected migration, using the county where one grew up as an exogenous source of variation in the selection equation, shows statistically significant bias in the uncorrected returns within some field county combinations, the point estimates of the bias varying between 0.4 and -2.2 percentage points, compared to the reference group. That said, bias significance at the 5% level is only present in 3 out of 18 counties, with little evidence that correcting for self-selected migration narrows the within-field differences in returns to education across regions.

The remainder of the paper is structured as follows. Section 2 outlines a model of mobility and earnings and shows how self-selected migration can lead to bias in the OLS estimates. In section 3 I give institutional context on education, the labor market, as well as the population patterns in Norway and administrative regions. Section 4 discusses the sources of the data, sample restriction, provides summary statistics, and shows the general migration patterns in the data. Section 5 shows how I implement the empirical design, section 6 show the results while section 7, before I conclude in section 8.

2 A model of mobility and earnings

To fixate ideas I will adapt a model by Dahl (2002) to show how self-selected migration potentially can lead to bias in the ordinary least squares estimates. Considers a country with N distinct regions where people live for two periods. In the first period, they do not work but obtain a type of higher education. In the second period, they choose in what region to work

and live. At birth, people are randomly assigned to different regions, making it reasonable to think that skill and other wage-determining characteristics are equally distributed across regions, but that self-selected migration potentially changes the distribution across regions.

Consider the ex-post perspective where people have obtained an education within a $field \in \{1, ..., F\}$ and started working. Assume their earnings can be characterized by the following constant effects equation.

$$y_{ik} = x_i'\delta_k + \beta_{1k} + \sum_{f=2}^{F} \beta_{fk} field_{if} + u_{ik} \text{ for (k = 1, ...,N)} \tag{1}$$

Where y_{ik} is the natural logarithm of earnings for individual i in region k. $x_i'\delta_k$ is a vector of observables for individual i and the loading factor for their payoff in region k. β_{1k} is the payoff for the reference field in region k. $field_{if} \equiv \mathbb{1}_{[field=f]}$ is an indicator variable that takes the value 1 if individual i obtained an education in field f and zero otherwise. u_{ik} is the region-specific error for individual i. For notational simplicity, let c_i be a vector of all covariates, both for the field dummies and the observable personal characteristics. If we assume $E[u_{ik}|c_i] = 0$ the equation could be estimated by OLS. However, under self-selected migration, this might not be the case, as I will demonstrate below.

Throughout the paper, I assume people are utility-maximizing agents and that mobility across regions is the outcome of such a maximization. Assume that V_{ijk} is the utility person i would obtain from moving from region j to k, and that this is a function of the region-specific income y_{ik} and tastes t_{ijk}. For expositional purposes, assume that the utility function is linearly separable of the form.

$$V_{ijk} = y_{ik} + t_{ijk} \tag{2}$$

The individual taste factor t_{ijk} captures the non-wage part of the utility function. Including everything from the cost of moving from j to k , to the difference in amenities, tax rates, and so on.

An alternative formulation of the above equation is to split it up into

$$y_{ik} - E[y_{ik}|c_i] = u_{ik} \tag{3}$$

Where u_{ik} is the individual deviation in income if they were to live in region k compared to if an average person with the same characteristics were to live in region k. And

$$t_{ijk} - E[t_{ijk}|d_i] = w_{ijk} \tag{4}$$

Where d_i is a vector of observable characteristics and w_{ijk} is the individual deviation in taste from the average taste for people to live in k. The important difference between u_{ik} and w_{ijk} is that the earnings error is assumed not to depend on where you are from, while

the taste error is allowed to vary depending on the sending and receiving region.

We can use the two above equations to reformulate the utility function as a random utility model,

$$V_{ijk} = V_{jk} + e_{ijk} \tag{5}$$

With V_{jk} being the average utility of moving from j to k and e_{ijk} is the individual deviation from the average stemming from their income and taste deviations. For utility-maximizing individuals, let the indicator function $M_{ijk} = 1$ if and only if $V_{ijk} = max(V_{ij1}, ..., V_{ijN})$ and zero otherwise[1].

The individual utility depends on the region of birth and where they live and work when they are grown up. The selection rule outlined above leads to the realization that one only observes y_{ik} in the utility-maximizing choice.

$$y_{ik} \text{ if and only if } M_{ijk} = 1 \tag{6}$$

The above equation describes a multi-market Roy (1951) model of earnings and mobility. The insight from the model is that the observed population living and working in k is not a random sample as.

$$E[u_{ik}|y_{ik} \text{ observed}] = E[u_{ik}|M_{ijk} = 1] \tag{7}$$

Where $E[u_{ik}|M_{ijk} = 1]$ is the potential selection bias for a given observation. If this conditional expectation is correlated with our field dummies, the OLS estimation of the returns to the field of study in a given region will be biased. It is not possible to determine apriori whether the bias is positive or negative as it depends on the correlation structure of both the earnings and the taste residuals.

3 Institutional context

3.1 Distribution of population and administrative regions

Norway is a well-developed industrial country with living standards and life expectancy among the highest internationally. Geographically, it is a long and narrow country with a long coastline. The surface area of the mainland is around 323 806 km², making it the 8th biggest in Europe with regards to area. With a population of only 5.5 million people (2023), it is a fairly thinly populated country. In terms of an American state, the surface area is close to that of New Mexico.

The country is partitioned into two regional levels of administration. With 18 counties and more than 400 municipalities at the time of the cross-section in 2018. There has been

[1] Assuming there is a singular maximum, the errors have finite moments, and are described by a finite number of parameters.

a general trend of municipalities and, to some extent, counties being merged, with the new sets of reforms starting in 2018 and ending in 2020, bringing the number of counties from 19 to 11 in 2020. The number of municipalities was also reduced to 365. Showing that neither the municipality nor the counties are truly stable objects.

Out of the total population in the country, around 40% reside in the counties surrounding Oslofjorden. This region is limited to the current counties of Viken and Oslo, as well as the former Vestfold, and covers approximately 8% of the mainland's area. Out of the total population, around 4/5 live in urban or suburban areas. Most people are living in the south and northern parts of the country have a relatively sparse population.

As of 2022, there were only six cities or metropolitan areas with more than 100,000 inhabitants. These urban areas are ranked in order of the number of people. Oslo (the biggest) is Norway's capital and the only city in Norway to be its own county. Inside the city borders, there reside around 600,000 people, but the greater metropolitan area bleeds into the neighboring regions like Akershus, with around 1 million. The remaining cities on the list are Bergen (Hordaland), Stavanger/Sandnes (Rogaland), Trondheim (Trøndelag), Fredriksta/Sarpsborg (Østfold), Drammen (Akershus).

3.2 Education

In Norway, the overall majority of people who obtain post-secondary education do so from public institutions. These can be grouped into universities and university colleges. The older universities (Oslo, Bergen, Trondheim, and Tromsø) generally offer a wider portfolio of fields like law, medicine, natural sciences, and civil engineering. In contrast, the university colleges offer more vocational training like nursing and engineering. Over the time of the sample, a few of the university colleges are merged together to form universities. Examples of these are the University of Agder (2007) and the University of Stavanger (2005), which both today offer civil engineering but not, for example, medicine or law. Obtaining a bachelor's degree usually takes three years, while a master's degree typically takes an additional two years.

The institutions are regulated and funded by the Ministry of Education and Research. This is also the case for private university colleges. For public universities and university colleges, there is just a low registration fee of less than 1000 NOK per semester, while private colleges can charge fees that are orders of magnitudes higher. Most students are eligible for financial support from the Norwegian State Educational Loan Fund. The support is issued as a loan partly converted into a grant when the students pass their exams. The students at private colleges are offered a bigger loan to cover the higher fees.

The admission process for higher education is centralized and handled by the Norwegian Universities and Colleges Admission Service, which handles the admissions for all uni-

versities and almost all colleges.[2] People apply for an institution/field combination (Technology at the Technical University in Trondheim). The number of spots for each combination is decided through the allocated funding from the Ministry of Education and Research. Ones grade point average (GPA) is the main determining factor for what spot a person gets allocated. In contrast, a minority of spots are allocated to certain people, f.e. people from the northernmost part of the country. Applicants can also get some additional points added to their GPA by taking certain subjects in high school, by having previous post-secondary education, by fulfilling military service, or by their age. There are also fields dominated by one gender where the minority gender receives a few extra points on their application score (such as women in certain STEM fields and men in psychology).

3.3 Labor market

The labor market in Norway is characterized by a big public sector that makes up around 30 % of the employed population, with around 20 % of the employees working for the municipalities and the remaining 10 % working at the state, either in the administration or at the county level according to numbers by Statistics Norway. The public sector is the main employer of both teachers and healthcare workers.

Another aspect of the Norwegian labor market is characterized by export lead two-tier wage bargaining as described by, for example, Bhuller et al. (2022). The idea of the system is that the general wage growth in the economy should not be higher than what the exporting sector can tolerate to still be competitive on the world market. To achieve this, the employers and workers' unions for the export-led sectors first negotiate what wage increases they can carry, which then set the pattern for the rest of the economy. This sets the wage floor for big parts of the economy. This is supplemented by local wage bargaining at the individual workplace. In the public sector, there is less room for local negotiations, meaning that in professions like teaching and healthcare, the wages are expected to be relatively equal across regions. In the private sector, on the other hand, there is more wage flexibility, leading to us expecting more variation in wages across regions. In a Roy-type model, one would, therefore, expect a higher propensity to migrate for private sector individuals.

4 Data

4.1 Data sources and sample restriction

I combine administrative data from Norway from various sources. I have access to application data for tertiary education for the years 1998 to 2018. This gives me data on their application score. I retain people's first observed application where they have no higher ed-

[2] The exception is the business school of Oslo.

ucation at the time of application. For the treatment variables, I have information on the educational outcomes of all applicants from 1998 - 2018[3]. I use the first degree that people finish to categorize them into ten broad fields [4]. The fields are science, engineering, technology, business, social science, humanities, law, health, and medicine based on Kirkeboen et al. (2016)s partition. See table 1 for more details. As my measure of returns to education, I use data from the Norwegian tax registry for 2018. The data contains the monthly wage for September 2018 and agreed-upon weekly working hours. To get a measure of the hourly wage rate, I compute hourly wage = monthly wage/(4 ∗ weekly hours)[5]. I link the data to the Norwegian population registry to obtain information about socioeconomic background characteristics, like whether the parents have higher education, as well as where the individual lived at age 16 and where they lived in 2018. I use the 18 counties that Norway was partitioned into in 2018 as my definition of the local labor market (LLM)[6].

I restrict the sample to people being 23 or younger when they apply and require them to be 28 or older in 2018. I drop all observations that are without a tertiary education. And all observations that do not belong to the 10 fields. I further drop all observations with log(Hourly wage) outside of the range 4.7 - 7.5 (an hourly wage between 109 - 1800 nok) and require the individuals to work more than 30 hours per week. This leaves me with a sample of 95425 individuals.

Table 1 shows the ten broad fields I with sub-field examples. Any attempt to classify types of education into fields glosses over within-group heterogeneity. One meaningful question about the returns to education is what type of job one is qualified for after obtaining the education. One could argue that there is an overlap between economics and business or information technology and computer science and that these fields should be combined in some other way. Another example is the difference between kindergarten teachers and high school teachers, where jobs and earnings are different from each other. With more data, it would be possible to aggregate the fields to a lesser extent, making the concern of in-field heterogeneity smaller.

4.2 Descriptive statistics

Table 2 reports summary statistics for the sample. In the sample, people work, on average, 37.67 hours per week. For context, the general full-time employment is considered to be

[3]This information comes from the national education register.

[4]One could argue that the last field one is observed studying is a better measure of the type of human capital a person has. In my sample, the change in definition does not affect the categorization of most people, but this alternative measure of human capital could be used as a robustness check.

[5]For people with more than one job I sum the income and number of weekly hours from all their jobs to get their total hours and monthly wage and perform the same calculation as above to get the wage rate

[6]Using county as the measure for LLM is potentially problematic as actual local labor markets might cross county borders. One alternative would be to use Bhuller (2009)'s partition into 46 local labor markets based on commuter data from 2000-2006 where the LLMs can cross county borders. I, however, disregard this option as this makes my data too thin.

Table 1. Classification of broad fields as in Kirkeboen et al. (2016) with examples of sub-fields

Science: biology; chemistry; computer science; mathematics; physics
Business; administration; accounting; business studies
Social science; sociology; political science; anthropology; economics; psychology
Teaching; kindergarten teacher; school teacher
Humanities: history; philosophy; languages; media
Health: nursing; social work; physical therapy
Engineering (BSc): electrical; construction; mechanical; computer
Technology (MSc): engineering; biotechnology; information technology
Law: law
Medicine: medicine; dentistry; pharmacology

37.5 hours per week, but the norm differs somewhat from field to field. In the sample, the log hourly wage rate is 5.77, which translates to approximately 320 NOK/hour. Women are somewhat over-represented in the sample, with 64 %. This, however, fits nicely with the fact that women in Norway have a higher propensity to acquire post-secondary education than their male counterparts. Considering that the youngest individual in the sample is 28 and the oldest is 43, the sample skews quite a bit, with a mean of only 33.16. This is a general pattern of the application data and is not caused by the sample restriction. That said, the main mass of the sample is in the younger age brackets, which is something to keep in mind when analyzing the results. Application score is re-scaled to be mean zero and has a standard deviation of 1. 43% of the people do not live in the county where they grew up, conveying that the sample is fairly mobile. Comparing this to Dahl (2002), who uses data from the 1990 US census, looking at white males between 25-34, 31% lived in their birth state. One apparent reason for the higher migration propensity in this sample is that the whole of Norway is the size of a US state, such that much smaller moves are required to be considered an internal migrant under my categorization[7]. The table also conveys some indicators for socioeconomic background characteristics, such as whether the parents had higher education when the individual was 16 and the father's average yearly earnings in 1000 NOK when the child was 16-18. The father's wage is corrected for inflation using 2015 as the base year.

Table 3 breaks the data down by field, and report means with standard deviation in parentheses. The three fields with the most observations are health, commerce, and teaching, all with lower-than-average mean application scores to get in. This tracks well with the fact that the Ministry of Education and Research provides more spots within this field. This is compared to the more selective fields, like law and medicine, where the application score is much higher. Taking application as a measure of ability, it is clear that the ability distribu-

[7]Migration propensity is calculated as the fraction of people living in a different county in 2018 from what they did at age 16.

Table 2. Descriptive statistics of the full estimation sample

	Mean	Std. Dev.
Age	33.16	4.17
Female	0.64	0.48
Application Score	0.00	1.00
Mother Higher Education	0.46	0.50
Father Higher Education	0.43	0.50
Fathers income (1000 NOK)	550.58	655.68
Propensity to migrate	0.43	0.50
ln(Hourly Wage)	5.77	0.29
Working hours	37.67	3.78

Notes: 95,542 Observations. Hourly wage is measured as the monthly wage in September divided by 4 times the agreed-upon number of working hours for the month. Application score is standardized to being mean zero and standard deviation of 1. Propensity to migrate is calculated as the fraction of people who live in a different county than when they were 16. Mother and father higher education is a dummy variable that equals 1 if the parent had at least a bachelor's degree when the individual was 16 and father income is the mean yearly income when the individual was aged 16 to 18 reported in thousands of NOK inflation-adjusted using 2015 as the base year.

tion differs between fields and is a potential confounding variable in the wage equation.

There seems to be a positive association between the hourly wage rate and the number of hours worked. For instance, medicine and civil engineering have the two highest hourly wage rates and are among the top three professions with the longest working hours. On the other hand, health and teaching have the lowest hourly wage rates and are among the professions with the shortest working hours. One possible explanation for this could be what Claudia Goldin describes as the phenomenon of "greedy work", where employers tend to pay disproportionately more for longer hours (see Goldin (2020) Goldin (2021)).

Table 3. Descriptive statistics broken down by field.

field	n	ln(Hourly Wage)	Working hours	Propensity to migrate	Female	Application Score	Mother Higher Edu	Father Higher Edu	Fathers income
Civil Engineering	7357	5.94 (0.28)	38.01 (2.7)	0.6 (0.49)	0.33 (0.47)	1.09 (0.77)	0.67 (0.47)	0.67 (0.47)	680.27 (571.14)
Commerce	15818	5.83 (0.33)	37.68 (3.03)	0.43 (0.5)	0.52 (0.5)	-0.02 (0.91)	0.44 (0.5)	0.44 (0.5)	627.00 (706.68)
Engineering	5152	5.89 (0.28)	37.87 (3.03)	0.39 (0.49)	0.22 (0.42)	0.05 (0.87)	0.44 (0.5)	0.44 (0.5)	540.98 (357.89)
Health	23738	5.69 (0.23)	37.13 (4.26)	0.35 (0.48)	0.85 (0.36)	-0.4 (0.86)	0.35 (0.48)	0.3 (0.46)	465.04 (312.06)
Humanities	7477	5.67 (0.27)	37.68 (3.8)	0.53 (0.5)	0.64 (0.48)	-0.02 (0.92)	0.52 (0.5)	0.48 (0.5)	514.91 (375.51)
Law	4404	5.85 (0.31)	38.07 (3.26)	0.52 (0.5)	0.64 (0.48)	0.44 (0.67)	0.59 (0.49)	0.57 (0.5)	718.00 (2159.65)
Medicine	4525	6.07 (0.33)	39.22 (5.95)	0.48 (0.5)	0.67 (0.47)	1.14 (0.86)	0.67 (0.47)	0.64 (0.48)	638.68 (514.41)
Science	4771	5.79 (0.28)	37.8 (3.19)	0.47 (0.5)	0.44 (0.5)	0.07 (0.98)	0.48 (0.5)	0.48 (0.5)	548.47 (496.59)
Social Science	8560	5.73 (0.26)	37.83 (3.57)	0.52 (0.5)	0.67 (0.47)	0.14 (0.86)	0.57 (0.5)	0.53 (0.5)	577.2 (431.99)
Teaching	13740	5.66 (0.2)	37.57 (3.72)	0.33 (0.47)	0.78 (0.42)	-0.5 (0.9)	0.38 (0.49)	0.32 (0.47)	465.45 (339.07)

Notes: The table breaks down the summary statistics by field. n shows the number of observations within field. The other columns show means with standard deviations in parenthesis. The values are calculated in the same way as discussed in the text and in table 2.

The propensity to migrate varies widely across fields, with three groups emerging. The low propensity group is health, teaching, engineering, and commerce, with migration propensities between 0.33 and 0.43. The group that stands out as very mobile is the civil engineers, with a migration propensity of 0.60. The rest of the fields can be grouped into a medium group with migration propensities between 0.47 and 0.53. The difference in migration propensity between the low and middle-propensity groups corresponds well with the migration pattern observed when stratifying only by the level of education. Individuals with a bachelor's degree have a migration propensity of 0.41 compared to those with a master's degree, who have a migration propensity of 0.54 [8]. Still, civil engineering is an outlier in this respect that warrants further investigation.

It is noticeable that certain professions are associated with particular indicators of socioeconomic status. When classifying into low, medium, and high socioeconomic status, it is reasonable to include teaching and health professions in the low group, whereas civil engineering, medicine, and law belong to the high socioeconomic status group. The groupings are somewhat similar to the one previously discussed and may have a potential correlation with wages. If these variables are not included in the wage equation, omitted variable bias may occur.

4.3 General migration pattern

Figure 1 shows the general migration patterns in the data. The left panel compares the number of people living in a county when they were 16 to where they lived in 2018. The black line is the 45-degree line, and the scales are log-transformed. If a point is below the line, it indicates that people, on average, move away from the region, and above means that people are moving into the region. The smallest county, Finnmark, had 634 observations in 2018, while Oslo was the biggest in 2018, with 27070 observations.

Based on where the sample lived at the age of 16, the counties in Norway can be divided into three categories. Small, middle-sized, and big. Finnmark is the smallest county, while Oslo, Hordaland, Trøndelag, Rogaland, and Akershus are the biggest. The remaining counties fall under the mid-sized category. People tend to leave Finnmark, but this trend is not as significant as it is for the mid-sized counties. Although there is some variation within this group, with Troms having fewer people leaving, on average, people move away from mid-sized counties[9]. Among the larger cities, Hordaland, Trøndelag, and Rogaland are able to maintain their population, while Akershus is losing some people. However, Oslo is a clear outlier, with a significant influx of people.

The heat map displayed in figure 1 shows the probability of people migrating from one region to another, conditional on moving. The rows, therefore, sum to one, and the diagonal

[8] The migration propensity for people with higher than masters level education is 0.54.

[9] When plotting the data using Bhuller (2009)'s reg46 partition the pattern is the same, but Tromsø (the biggest city and labor market in Troms) had net immigration.

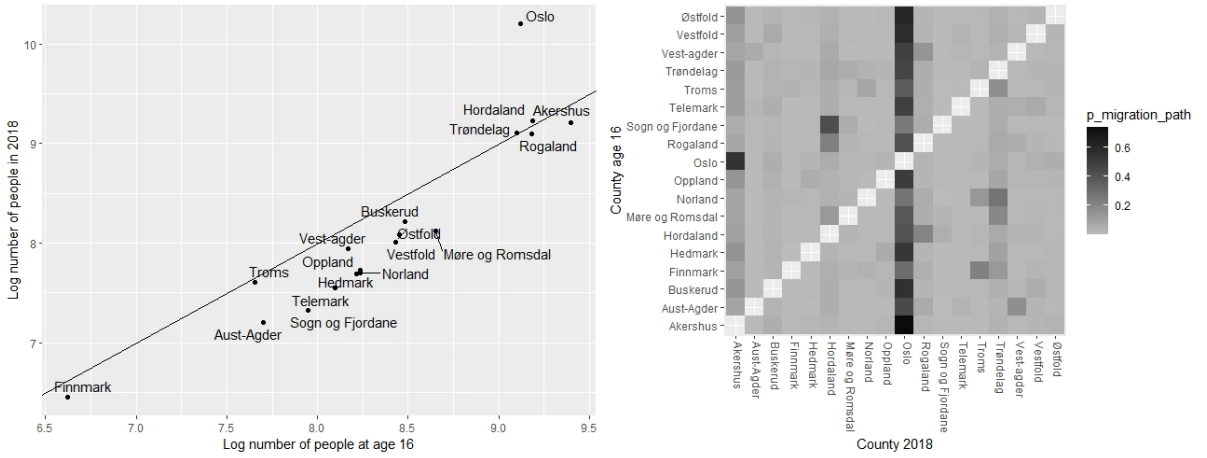

Figure 1. The left panel shows the relationship between the number of people growing up and living in a given county. The axes are log-scaled. The right panel shows a heat map of migration probabilities conditional on people moving. The probabilities are calculated as the fraction of people moving from, say, Rogaland to Oslo divided by the total number of people moving from Rogaland. The darker the color, the higher the probability.

elements are missing. The probabilities are calculated as the fraction of movers from, say, Rogaland to Oslo divided by everyone moving away from Rogaland. The y-axis represents the regions where people are moving from, and the x-axis represents the regions where they are moving to. The darker the color on the map, the higher the probability for the given migration path. By scanning horizontally, you can see where people are moving from, and by scanning vertically, you can see where they are moving to.

Oslo is the most popular destination for people moving within Norway, with high levels of in-migration from all counties. However, counties that are closer geographically are more likely to move to Oslo. Other popular destinations with noticeable levels of in-migration are the highly populated regions mentioned above. For these counties, the migration pattern shows that people tend to move short distances, such as from Sogn og Fjordane to Hordaland, Norland to Trøndelag, and Oslo to Akershus. This plot illustrates the expected migration patterns of a gravity-type model. Although migration is primarily unidirectional, all migration paths have been observed, although not by a significant proportion of people.

5 Empirical design

5.1 Ordinary least squares

When estimating the returns to education across regions, a natural place to start is by using OLS. The most naive model would be to regress the categorical variable field, county by county, using log hourly wage as the dependent variable. For simplicity, let's abstract from

the individual index.

$$y_k = \beta_{1k} + \sum_{f=2}^{F} \beta_{fk} field_f + u_k \text{ for (k = 1, ..., N)} \tag{8}$$

With y being log hourly wage, β_1 being the log hourly wage in the reference category, and $field$ is an indicator variable that takes the value 1 if an individual belongs to the field and 0 otherwise. Multiplying the β_f's by 100 they can approximately be interpreted as the percentage hourly wage premium in field f compared to the reference category, and u is an error term. The subscript k refers to the county where people resided in 2018.

A causal interpretation of this equation would be misguided as there would be reasonable suspicion of omitted variable bias. This happens when parts of the error term are a determinant of y_k and are correlated with the regressors. The list of potential confounding variables is long.

One place to start is to include labor market experience and gender as explanatory variables, as they are likely determinants of y_k and probably correlate with field. I use age in 2018 as my measure of experience [10]. As argued by, for example, Bertrand et al. (2010) and Kleven et al. (2019), women and men experience different rates of wage growth, both due to occupational sorting and the effect of having children on labor market decisions. My sample comprises individuals between 28 and 43, the prime age for having children. Lacking data on whether individuals have children, I try to control for the differing labor market decisions between the genders by using age as dummies interacted with gender.

Ability is another variable that probably determines y_k looks to be correlated with field. As discussed by many in econometrics, ability is a potential cofounder when performing returns education regressions Griliches (1977)Blackburn and Neumark (1993). To mitigate the problem, I include application score as a measure of ability.

Socioeconomic status is another potential determinant of y_k and is likely to correlate with the field. With potential channels being parental education explains both types of education and earning as shown by Erola et al. (2016), or that parents' education relates to the children's academic achievements through parental beliefs and behavior Davis-Kean (2005). I use whether the mother and father had higher education when the individual was 16 as a measure of socioeconomic status [11].

If we assume for a second that we have added all determinants of y_k that correlate with the regressors of interest, we could estimate the equation below and give a causal interpre-

[10] Alternative measures of experience could be the number of years since application in 2018 or the number of years since last observed in higher education.

[11] I do also have information on the parent's average yearly wage when the individual was between 16 and 19. As wages often follow more of a log-normal distribution, it would be necessary to transform the parental wage, thus restricting the sample even more due to parents having no observed income. Another way of keeping the wage information would be to make it into a categorical variable with some bin size, but this approach will not be pursued here.

tation of the returns to field coefficients in a given region.

$$y_k = x'\delta_k + \beta_{1k} + \sum_{f=2}^{F} \beta_{fk} field_f + u_k \text{ for (k = 1, ...,N)} \tag{9}$$

As discussed in section 2 the above regression equation can potentially suffer from self-selection problems, as the population residing in a given county is likely not random. If the people who are moving are different from the people who are staying with regard to earnings potential, this would lead to bias in the coefficient estimates of interest.

5.2 Self-selection with multiple unordered choices

Wishing to correct the returns to education for self-selected migration in Norway across an array of regions, a general model of earnings could be an alteration of equation 5.1.

$$y_k = x'\delta_k + \beta_{1k} + \sum_{f=2}^{F} \beta_{fk} field_f + \lambda(\cdot) + u_k for (k = 1, ..., N) \tag{10}$$

This alteration incorporates the control function $\lambda(\cdot)$, an unknown function that absorbs the correlation between the error term and the parameters of interest such that when conditioning on, it solves the problem of self-selection.

Following Bourguignon et al. (2007) for simplicity, only look at the wage in county 1. And let all covariates be condensed such that.

$$y_1 = x\beta + w_1 \tag{11}$$

$$y_k^* = z\gamma_k + \eta_k, \text{ for k = 1...N} \tag{12}$$

Where w_1 is a disturbance term that is not parametrically identified and verifies that $E(w_1|x,z) = 0$ and $V(w_1|x,z) = \sigma^2$. k is a categorical variable that describes the individual choice between N different regions in which to work based on the "utility" of y_k^*. z is a vector that represents the maximum set of explanatory variables for all the alternatives, and x contains all determinants of coefficients of interest. Further, assume that the model is non-parametrically identified from exclusion from at least one of the variables in z from x. Lets focus on the case where region 1 is chosen.

$$y_1^* > \max_{k \neq 1}(y_j^*) \tag{13}$$

Define

$$\varepsilon_1 = \max_{k \neq 1}(y_k^* - y_1^*) = max_{k \neq 1}(z\gamma_1 - z\gamma_k + \eta_k - \eta_1) \tag{14}$$

Under these conditions $\varepsilon_1 < 0$. Assuming that the (η_k)s are independent and identically

Gumbel distributed leads to the choice equation being a multinomial logistic model as shown by McFadden et al. (1973).

$$P(\varepsilon_1 < 0|z) = \frac{exp(z\gamma_1)}{\sum_{k=1}^{N} exp(z\gamma_k)} \tag{15}$$

Where the (γ_k)s are vectors of parameters that can be estimated by maximum likelihood.

defining Γ

$$\Gamma = \{z\gamma_1, z\gamma_2, ..., z\gamma_N\} \tag{16}$$

The generalization of the Heckman (1979) bias correction will be the conditional mean of w_k.

$$E(w_1|\varepsilon_1 < 0, \Gamma) = \int \int_{-\infty}^{0} \frac{w_1 f(w_1, \varepsilon_1|\Gamma)}{P(\varepsilon_1 < 0|\Gamma)} d\varepsilon_1 dw_1 = \lambda(\Gamma) \tag{17}$$

with $w_1 f(w_1, \varepsilon_1|\Gamma)$ being the conditional joint density function over w_1 and ε_1. Define P_k as the probability that region k is preferred:

$$P_k = \frac{exp(z\gamma_k)}{\sum_{k=1}^{N} exp(z\gamma_k)} \tag{18}$$

Given the N components of Γ and the corresponding N probabilities are an invertible function, we get.

$$E(u_1|\varepsilon_1 < 0, \Gamma) = \mu(P_1, ..., P_N) \tag{19}$$

Going back to the regression equation of interest, one can get consistent estimates of the coefficients on the β coefficients based on either

$$y_1 = x'\delta_1 + \beta_{11} + \sum_{f=2}^{F} \beta_{f1} field_f + \lambda(\Gamma) + u_1 \tag{20}$$

or

$$y_1 = x'\delta_1 + \beta_{11} + \sum_{f=2}^{F} \beta_{f1} field_f + \mu(P_1, ..., P_N) + u_1 \tag{21}$$

Where u_1 is an error term that is mean-independent of the regressors.

Both λ and μ are still unknown functions and need further assumptions to bring them to the data. In the literature, there are several different ways of implementing the control function, and I will consider three. The first one is simply assuming that the form of the control function is a first-degree polynomial expansion of the estimated choice probabilities.

Simply,

$$\mu(P_1, ..., P_N) \cong \mu_1 \hat{P_1} + ... + \mu_N \hat{P_N} \tag{22}$$

Where $\mu_1, ..., \mu_N$ are parameters and $\hat{P_1}, ..., \hat{P_N}$ are the estimated probabilities from a multinomial logistic regression.

The second control function takes the form.

$$\mu(P_1,...,P_N) \cong \begin{cases} -\gamma_1 \log(P_1), & \text{if county 1 is chosen} \\ \gamma_k \log(P_k)\frac{P_k}{1-P_k} & \text{for all } k \neq 1 \end{cases} \tag{23}$$

The third control function takes the form proposed by Dubin and McFadden (1984).

$$\mu(P_1,...,P_N) \cong \sum_{k=2}^{N} \gamma_k \left(\frac{P_k \log(P_k)}{1-P_k} + \log(P_1) \right) \tag{24}$$

With the γ's being parameters. The three control functions enter equations 21. As the model is linear in parameters, it can be estimated by OLS, and given a correct specification of the choice model, correction function, and covariates in the regression, be a consistent estimator of the coefficients of interest.

5.3 Identification

As mentioned above, we need an excludable variable in the choice equation for the coefficients of interest to be identified. For clarification, I will use the term instrument to refer to such variables. One proposed instrument by Dahl (2002) is to use where a person is born as an exogenous source of variation in the selection equation. The assumption is that where you are born can be viewed as a random cost shifter, changing the probability of selection. If we further assume that two people from different counties with the same observable characteristics and the same type of education will earn the same on average in $county_k$ including the correction function, will purge the bias from the coefficients. In other words, it means that "unproductive" characteristics like region of birth are not allowed to affect wage potential in any of the k regions. In my implementation, I use where the individual lived at age 16, but the logic is the same.

If the region one lived in at 16 carries additional information not picked up in the other regressors, the exclusion restriction would be violated, leading to biased estimates. One such reason could be that the quality of schooling differs between regions. Another would be if the ability is inherited (genetically or socially) and the ability distribution of parents is unequal across regions. A third reason the exclusion restriction would be violated is if there is prejudice in the local labor market against individuals from certain regions[12].

The setup also assumes that the required selection correction is the same for all individuals within a region. One could imagine having different correction functions for each field, but this quickly becomes intractable as the curse of dimensionality kicks in.

[12] One example of discrimination in Norway was for people from the north in general and for the Sami-people in specific See Midtbøen (2015) and Hellstad (2010). If this discrimination also affected what type of jobs one would be able to acquire the exclusion restriction would be violated.

5.4 Common critiques of the method

A common critique of using the multinomial logistic regression as the choice model is the "independence of irrelevant alternatives" assumption. In the current context, it means that the odds of preferring one region over another one do not depend on the presence or absence of "irrelevant" regions. One potential violation in the current context could be that the perceived difference between the two northernmost counties is irrelevant, leading to an inflated joint probability of choosing to live in one of the two. As we are not interested in the estimated probabilities per se, but rather purging the selection bias from the earnings equation Bourguignon et al. (2007), shows Monte Carlo evidence that the multinomial logit model can perform fairly well, even when the IIA hypothesis is violated.

That said, one way to circumvent the problem of the IIA hypothesis being violated is to use a multinational probit model as the choice model, allowing for correlations between regions. This solution, however, becomes intractable as the number of distinct choices increases. Another alternative is to use a nested logistic regression and specify a tree structure with correlated alternatives in each node. It is, however, not straightforward to define the structure. For example, should choices be categorized by, say living in the north or south of Norway, or should it be specified by population density where people choose between living in the cities or the countryside?

Another approach put forward by Dahl (2002) building on Lee (1983) maximum order statistic approach is to use only a subset of probabilities and their higher order polynomial expansions as control functions in the earnings equation. By estimating the probabilities non-parametrically using cell means, Dahl sidesteps the IIA assumption. He assumes that the joint density between the errors in the selection equation and the errors in the outcome equation can be expressed as the bivariate distribution between the error term in the earnings equation and the error for the utility-maximizing choice conditional on a subset of probabilities $\vec{p}$. In essence, he assumes conditioning on $\vec{p}$ one describes all information in the sub-utility differences that affect the joint density in the outcome equation. Though this method is relatively easy to implement, it requires more data than I have available to estimate the selection probabilities non-parametrically. A further critique against this approach is that it is not testable whether one has corrected for the relevant probabilities to purge the outcome equation of bias.

5.5 Estimation

5.5.1 Two step estimation

The first step to implementing the control function is to estimate the selection probabilities using a multinomial logistic regression outlined above. The outcome variable is the county the individual lived in in 2018. As predictors, I include gender, age as dummies, the applica-

tion score, what field they have an education in, parental education, and what county they lived in at age 16. All variables enter linearly [13].

The second step is to extract the fitted values from the multinomial model. This yields a vector of probabilities for each individual $\hat{P}_i = (\hat{p_{i1}}, ..., \hat{p_{iN}})$ where $\hat{p_{ik}}$ is the estimated probability that individual i lived in region k in 2018. I use the estimated probabilities to implement the three control functions outlined above and estimate them by OLS. As the three approaches yield similar results, I will only report corrected results using equation 22 (the first-degree polynomial expansion).

5.5.2 Bootstrapping

Given that the model is correctly specified, the estimates will be consistent, but as noted by, among others Murphy and Topel (2002), the naive standard errors in two-step econometric models are likely to be incorrect. There are several ways to account for this. One way would be to derive some feasible estimator of the asymptotic variance-covariance matrix. Another is to use resampling methods. In the current application, I use Bootstrap with replacement. For each bootstrap iteration, I draw n random samples with replacements from the dataset. n is the total number of samples in the dataset. For each iteration, I fit the multinomial model and implement the correction as outlined above. For each iteration, I store the second-stage regression coefficients. I also run an OLS without the correction terms on the resamples each time. There is different advice on how many times one resample is necessary. I do 1000 re-samples[14] and show bootstrapped standard errors for both the corrected and uncorrected model[15].

6 Mobility and the return to field of study

6.1 Migration patterns

As mentioned in section 4.2, civil engineering is somewhat of an outliner with regard to migration propensity, a fact that warrants some further investigation. Figure 2 contrast the migration patterns of civil engineering and health. The plot is constructed in the same way as the left panel in figure 1. With regards to where people grow up, we see a similar pattern both for civil engineering and health, with regards to the size of the region they are from. The difference comes down to what extent people are moving away from the smaller regions.

[13] I use the multinom function from the nnet package in R written by Ripley et al. (2016) to fit the the multinomial logistic regression. It fits the coefficients using a neural net and not maximum likelihood. There is a package called mlogit by Croissant et al. (2012) that uses maximum likelihood to fit the models, but it is to slow for my needs.

[14] Efron and Tibshirani (1994) recommends at least 200 for standard errors on regression coefficients. As I want to graphically show the distribution of the coefficients 200 resamples is too low.

[15] The bootstrap standard errors are calculated as $sd(coef)$

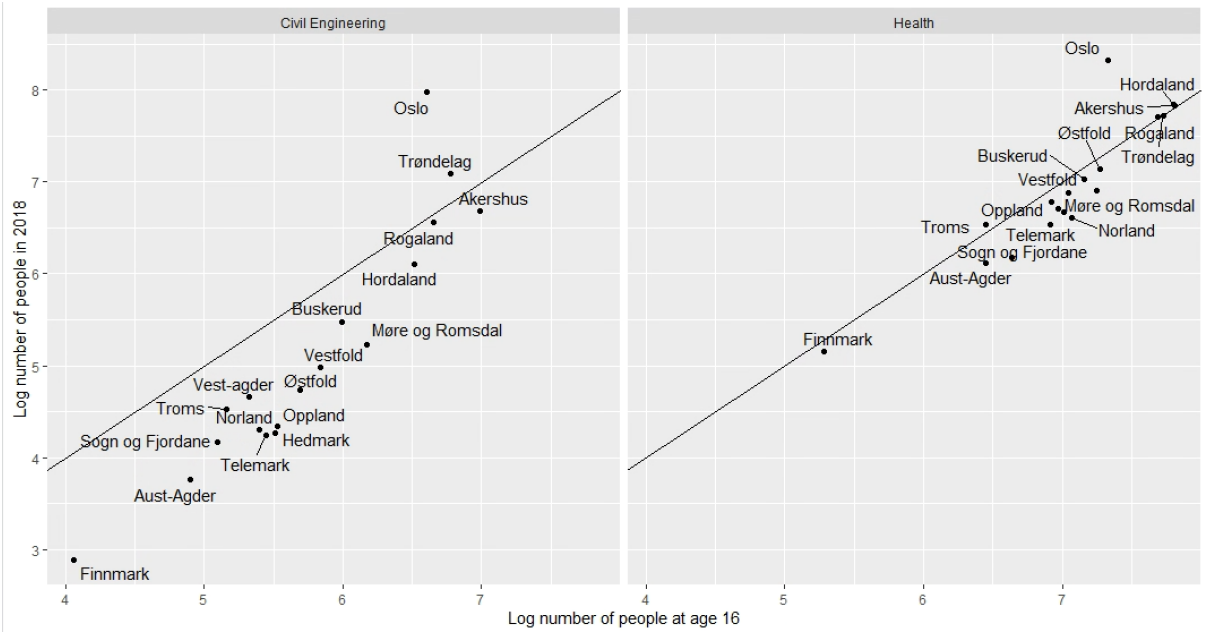

Figure 2. This plot contrasts the mobility patterns for civil engineering to health. The x-axis represents the number of people who lived in a region at age 16. The y-axis represents the number of people living in the region in 2018. Both axes are log-scaled, and the black line is the 45-degree line.

As is evident for the health field, though some people are moving away from the smaller counties, health is quit tightly clustered around the 45-degree line. In contrast, are the civil engineers moving away from the less populated regions and to the bigger ones. At least two potential reasons spring to mind. The potential difference between the groups can be preference-based. Another can be that it is easier to find relevant work within the health field in the smaller regions compared to civil engineering.

Figure 3 shows the dependence between living in one's home region and age. The data is grouped into three categories: one for people aged 30 or younger, another for people aged 31-35, and the last category is for people older than 35. Looking at the youngest age group, we see that the pattern of low, medium, and high propensity to stay between the fields becomes more evident. The people in engineering, teaching, and health are more likely to stay in their home region, whereas those in civil engineering are the least likely. On the other hand, the older people within the same fields are more likely to live in their home county. This pattern is not as evident in the first two age brackets but becomes more prominent in the last bracket. It is interesting to note that humanities and social sciences have a dip in the propensity to stay in the middle bracket. In contrast, engineering and medicine have a falling propensity to stay.

The declining pattern in medicine can potentially be explained by the LIS (medical doctors undergoing specialization) program, where one first works 12 months in a hospital and then 6 months in the local health service after having completed studying. Without delays in studying medicine, it takes 6 years to complete. This means that medicine students be-

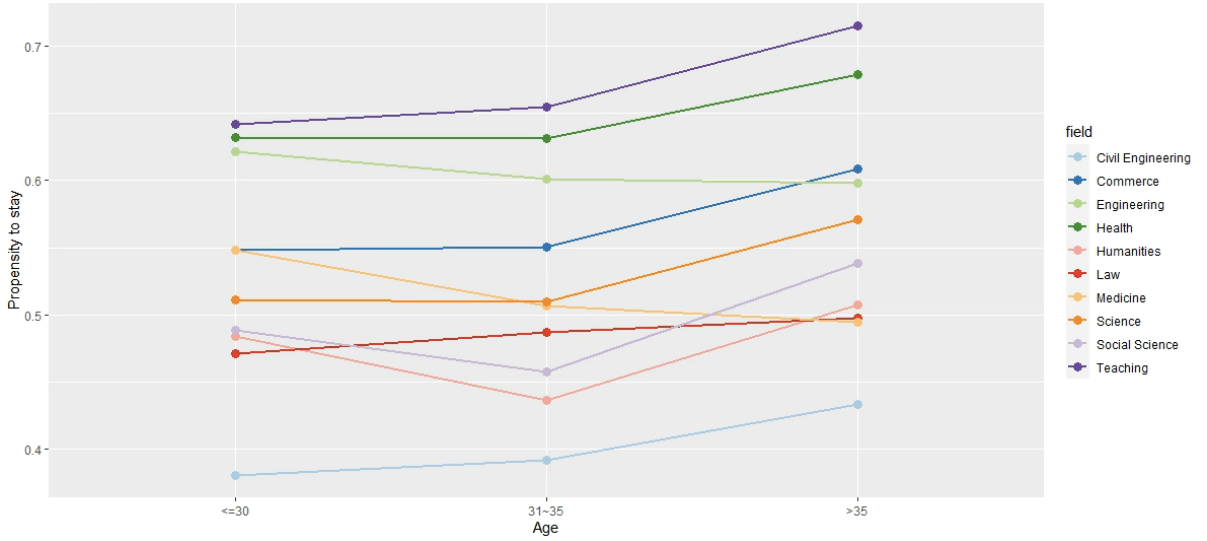

Figure 3. This figure shows the dependence between living in one's home region and age. The data is grouped into three categories: one for people aged 30 or younger, another for people aged 31-35, and the last category is for people older than 35, and the propensity is calculated as the fraction that lives in the county they grew up in.

tween the ages of 31-35 are undergoing or have just undergone specialization. The nature of the LIS program makes it difficult to choose where the individual undergoes specialization, forcing people to move. The author doesn't have any explanation as to why engineers move away. Still, comparing the change in the propensity of engineering to medicine, the drop is around 2.5 percentage points compared to 5.

One crucial aspect that the plot reveals is that the two-period model fails to capture the nature of mobility accurately. A more accurate model would potentially be one in which a person is randomly born in period one. In period two, the person decides where to study and work in their early career, and in period three, they decide where to work in their later career. Although this complication will not be taken into account directly, age enters the selection equation, which mitigates this problem to some extent.

Figure 4 showed the difference between the in-migrants and stayers fields distributions for six regions, as showing all 18 would take up to much space. Five of the regions are chosen as they are the ones with the most observations in my sample. Troms is added as an example of one of the smaller regions. The y-axis is independent for each cell and shows probabilities. The stayer probabilities are calculated as the fraction of people who in 2018 lived in the region where they grew up. The in-migrants probabilities are calculated as the fraction of all migrants who end up in a local labor market in 2018.

One of the most notable patterns can be seen in Oslo. The fields of study with the highest probability of being stayers are the same fields in which in-migrants are likely to come from. Conversely, fields with the highest likelihood of moving away are the same fields (such as engineering, health, medicine, and teaching) in which in-migrants are least likely to live in

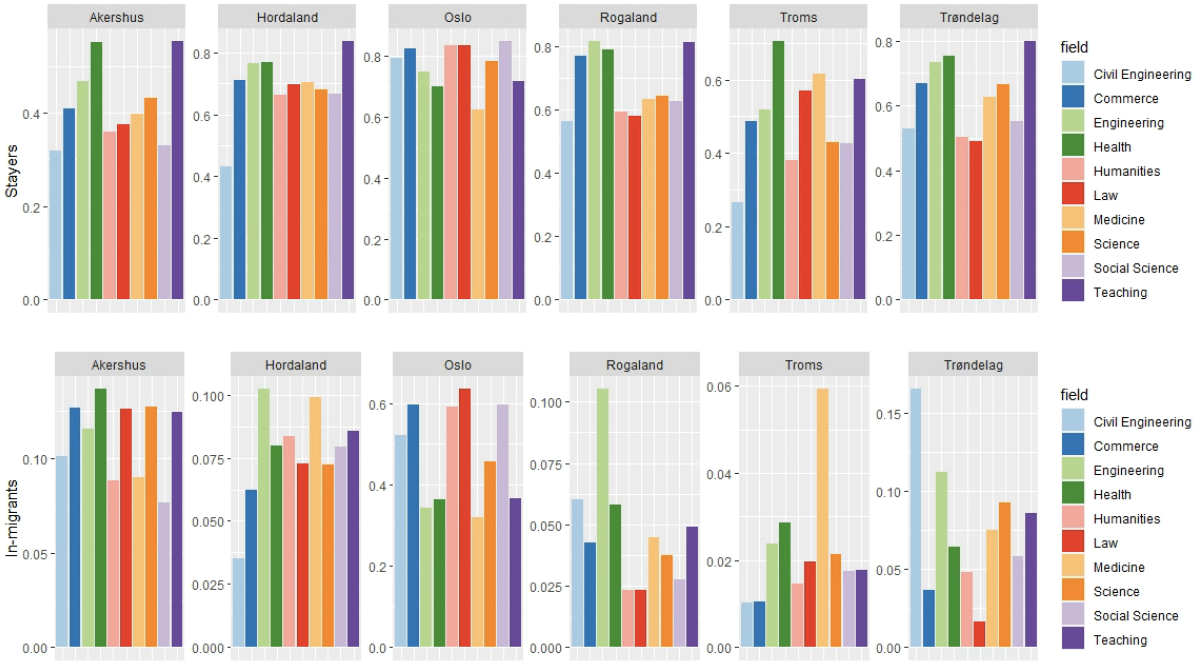

Figure 4. This plot shows the difference between the in-migrants and stayers fields distributions for six regions. The y-axis shows probabilities. The stayer probabilities are calculated as the fraction of people who live in the region where they grew up. The in-migrants probabilities are calculated as the fraction of all migrants who end up in a local labor market in 2018.

Oslo. The difference in probability between the likely and unlikely fields is greater among the in-migrants. One interpretation is that the general pattern shows what fields are in demand, affecting in-migrants to a greater extent, as stayers in low-demand fields have a stronger "taste" for Oslo as they grew up there. A similar pattern can be observed in Akershus and Hordaland, although it is not as clear as in Oslo.

Rogaland, Troms, and Trøndelag do not show the same degree of similarity between the two distributions. Each region has a dominant field of in-migration. Until 2005, Rogaland only had a university college, which typically offers more vocational training like engineering rather than science and civil engineering. This could be an explanation for the prevalence of engineering in-migrants. Rogaland is also home to the oil industry, with technical fields being in demand. In Troms, medicine is the popular field of in-migration, as it is one of only four regions that offer this field of study and has the lowest application score threshold. Trøndelag is home to the Norwegian Technical University (NTNU), which correlates well with the three most probable fields among in-migrants being civil engineering, engineering, and science.

Figure 5 is a transition matrix and provides a greater overview. Contrasting it with the previous figure, it also provides information on the sending regions. The rows indicate where someone lived when they were 16, and the columns where they lived in 2018. The diagonal elements are the fraction of people staying, broken down by field, and the off-

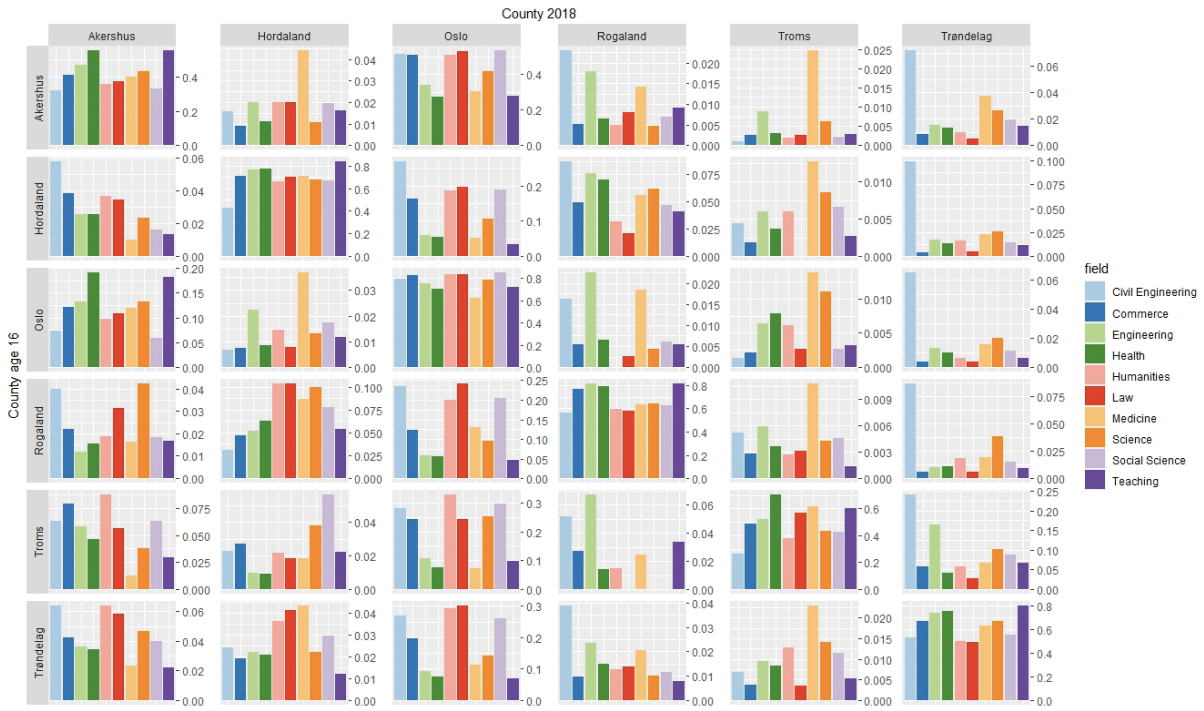

Figure 5. This figure shows a transition matrix for a subset of counties, with estimated migration probabilities broken down by field. The y-axis is county at 16, the x-axis is county in 2018. The retention probabilities are along the diagonal, and migration probabilities are on the off-diagonal. The probabilities are estimated by taking the number of people moving from, say, Rogaland to Oslo, divided by the total number of observations that lived in Rogaland at age 16. The probabilities are estimated separately for each field.

diagonal shows the fraction of people who move from one region to another.

Inspecting the mobility patterns along the diagonal shows that teaching and health are, in general, more likely to stay in the region where they grew up. The exception to this pattern is Oslo, where teachers and health are the most likely to leave. Scanning Oslo horizontally reveals that the majority of them move to Akershus, with the "diaspora" making up around 18% of the people who grew up in Oslo within these fields. Oslo is the most expensive real estate market in Norway, and teachers and health are among the lowest earning fields, making a potential explanation for the pattern that they move out of the city where property prices are lower.

We have identified civil engineers as the most mobile group, moving away from the smallest regions. Here, inspecting the pattern among the bigger regions (and troms), it is clear that civil engineers are the, or among, the fields with the highest propensity to move away. The exception is Oslo, where they are among the most likely to stay. Interestingly, civil engineering is also one of the fields with the highest propensity to move into Oslo. Contrasting this with Hordaland and Troms, which are the two regions (amongst the ones under inspection) where Civil Engineers clearly are the most likely to move away, there are very few in-migrants within the field. In the last three regions, the difference in staying propen-

sity between civil engineering and other fields is smaller, and we see that among the people who move to these regions, civil engineers are among the ones with the highest propensity.

The plot further highlights an interesting aspect - it reveals how sparsely the data is distributed when conditioned on the field. For instance, consider the migration path from Oslo and Troms to Rogaland, where no one has taken five out of the available migration paths. These migration paths belong to the bigger regions, which raises the question of how many empty or imprecisely measured cells would be observed if a non-parametric approach were used to estimate the selection probabilities, as proposed Dahl (2002).

6.2 Uncorrected returns

First, when I refer to corrected and uncorrected returns, this is in the same spirit as Dahl (2002) using it to refer to whether or not the estimates are corrected for self-selected migration, not to be misunderstood as adding covariates or having instruments that reasonably make field exogenous.

I will begin by demonstrating the importance of controlling for background characteristics in the OLS model, focusing on Oslo. In table 4, four specifications show the estimated returns to the field of study. The first column is the most basic, assuming no covariates are necessary. Column two adds age as dummies interacted with gender. Column three includes application score, and column four adds parental education. Health is the left-out reference category, and the table displays robust standard errors in parentheses.

After controlling for age and gender, the estimated returns on various fields decrease by 0.7 to 6.3 percentage points. Engineering experiencing the biggest drop, and social science experiencing the smallest. Accounting for ability by including application scores has an even more pronounced effect, with returns decreasing by 1.6 to 7 percentage points across all fields. Civil engineering and medicine, which have the highest average application scores, experience the largest decrease. Additionally, the sign on the returns for social science changes from positive to negative. Accounting for socioeconomic background with parental education results in a further decrease in returns across all fields and a marginal increase in the intercept.

The point of this exercise is not to claim that the estimated coefficients in column four are causal but rather to show that the added covariates should be included in the regression, changing the estimated hourly wage premium compared to the reference category by between 3.5 and 9.5 percentage points[16].

[16]Adding application as a second order polynomial changes estimates marginally and increases the adjusted r squared somewhat, with the biggest changes being in medicine and civil engineering which both decreases by 0.5 percentage points. A third-order polynomial of the application score has almost no effect on the variables of interest. In the further analysis, I will stick with including the application score linearly, mainly due to the corrected models only including the application score linearly, and a rerun of the model would take 25 hours.

Table 4. OLS Oslo

	Oslo, dep var log(hourly wage)			
	(1)	(2)	(3)	(4)
Intercept	5.72054	5.61847	5.37358	5.37405
	(0.00357)	(0.00778)	(0.01366)	(0.01365)
Humanities	-0.03099	-0.05030	-0.06647	-0.06801
	(0.00603)	(0.00591)	(0.00596)	(0.00596)
Social Science	0.01979	0.01907	-0.01103	-0.01303
	(0.00555)	(0.00540)	(0.00554)	(0.00556)
Engineering	0.18654	0.12180	0.10283	0.10206
	(0.01001)	(0.00950)	(0.00958)	(0.00958)
Commerce	0.16551	0.14084	0.11866	0.11761
	(0.00570)	(0.00543)	(0.00547)	(0.00548)
Teaching	-0.05109	-0.05426	-0.05159	-0.05211
	(0.00534)	(0.00526)	(0.00530)	(0.00531)
Civil Engineering	0.22778	0.20327	0.13614	0.13395
	(0.00620)	(0.00603)	(0.00669)	(0.00672)
Science	0.09858	0.06617	0.04199	0.04088
	(0.00816)	(0.00764)	(0.00779)	(0.00780)
Law	0.15969	0.14423	0.10318	0.10139
	(0.00787)	(0.00745)	(0.00763)	(0.00763)
Medicine	0.29632	0.28128	0.21152	0.20967
	(0.01003)	(0.00973)	(0.01025)	(0.01025)
Application Score			0.00580	0.00563
			(0.00026)	(0.00026)
Father higher edu				0.01013
				(0.00346)
Mother higher edu				0.00484
				(0.00346)
Gender X Age fixed effects	-	X	X	X
Num.Obs.	27109	27109	27109	27109
R2	0.123	0.230	0.245	0.246
R2 Adj.	0.123	0.229	0.244	0.244

Note: This table shows regression results for Oslo, with log(hourly wage) being the dependent variable and showing robust standard errors in parenthesis. The omitted field is health and. "Gender X Age fixed effects" refers to age as dummies interacted with gender.

Figure 6 shows uncorrected returns estimated from the same as equation column four in table 4. The left-out reference category is nursing[17], and the estimation is done region by region. To get a better feel for the wage difference between regions within fields, the estimates are grouped by field. The horizontal line indicates the wage of the reference category, and the error bars are 95% confidence intervals based on robust standard errors.

The graph highlights an interesting aspect - the difference between typical public and private sector fields. Teaching and health are two of the obvious public sector fields. Still, there is an argument to be made that humanities and social sciences are also heavily situated in the public sector. It can be argued that medicine is also a public sector field but is, in such, an outlier. I will get back to this later.

Categorizing the rest of the fields as private sector fields can help us understand wage differences across regions. In the public sector fields, the wage premium is relatively constant across regions. Humanities and teaching have a somewhat negative wage premium on average, and social science has close to no wage premium compared to health. On the other hand, in the private sector fields, there seems to be more variability in the point estimates, with a few regions standing out. Interestingly, these regions are somewhat in common across the fields. The regions are Oslo, Akershus, Rogaland, and, to some extent, Trøndelag.

In the private sector, two sensible groupings could be civil engineering, engineering, and science combined, commerce and law combined. Civil engineering, engineering, and science exhibit similar patterns in wage premiums across regions, though the fields are shifted by a constant compared to each other. In all three categories, the region that stands out with higher wage premiums is Rogaland. This could be due to the fact that Rogaland is home to Norway's biggest export sector, the oil industry, which is highly productive, requiring skilled professionals within the typical STEM (science, technology, engineering, and mathematics) fields. This fits nicely with the fact shown in figure 4 and 5 with the two most likely fields in-migrants having are engineering and civil engineering. Surprisingly, people in science are not that likely to move into Rogaland, though the uncorrected estimated wage premium is the highest there.

For commerce and law, the regions that stand out, with higher wage premiums, are Akershus, Oslo, and Rogaland, though the pattern is much more apparent in commerce. Oslo has a lot of big law firms and is also the finance and business hot spot in Norway.

The strong similarities between Oslo and Akershus highlights a potential problem with my definition of the local labor market. It is not uncommon to live in, for example, Akser and Bærum (Two places in Akershus) and commute to work in Oslo. This speaks to the

[17] It is not ideal to have the reference category being so overwhelmingly female. The reason why it was chosen is due to the wage being comparatively stable across regions with a span of 6 percentage points between points estimates between the highest earning region and the lowest. Teaching was relatively similar. Commerce, on the other hand, has a fairly balanced gender profile, but the difference in wage distribution was much bigger, and therefore not as suited as a reference category.

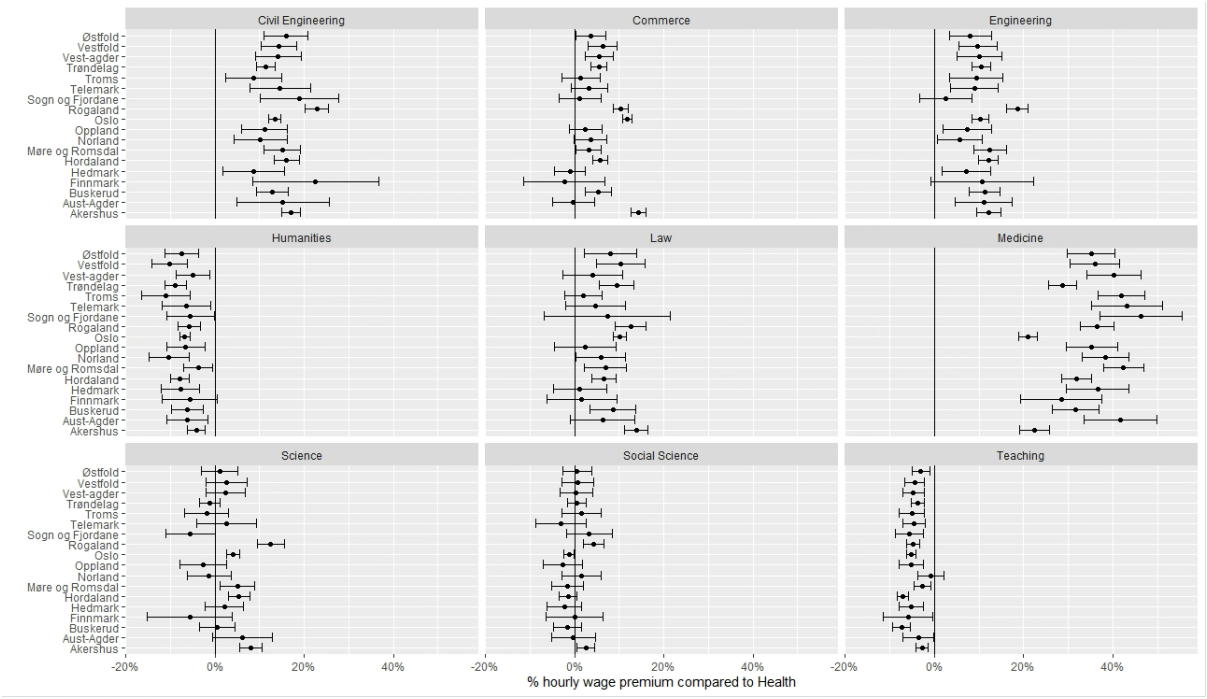

Figure 6. This figure shows the unadjusted % hourly wage premium to field of study for different regions. Health is set as the reference category. The points are spot estimates, and the error bars show 95% confidence intervals based on robust standard errors. The estimation is done region by region, controlling for application score, parental education, and age as dummies interacted with gender.

problem of using the county as a measure of the local labor market and not a classification like Bhuller (2009) based on commuting zones, making these two regions correlated as the labor markets are not distinct.

A potential fix would be to use where people work and not where they live to divide them into regions. This could be a potential robustness check that will not be pursued here. Interestingly, though both law and commerce have high uncorrected wage premiums in Rogaland, people with these types of education are not especially likely to move into the region.

Turning to the outlier among the public sector fields medicine commands the highest wage premium on average, with Oslo and Akershus and, to some extent, Trøndelag being notably lower than the rest.

One explanation could trace back to there having been reports of problems of getting people who practice medicine in the smaller regions. With less supply of labor, the wages must increase to attract talent. If this is the only explanation one would expect Hordaland and Rogaland also to have comparatively lower wage premiums than the rest of the regions, which is not the case. Another potential explanation is that there are differences in the type of employment between the regions, with more work at hospitals and less in private practice in the more populated regions.

Somewhat unexpectedly, science does not command much of a wage premium compared to health. The exceptions are Akershus, Hordaland, Møre og Romsdal, Oslo, and Ro-

galand, where science commands a statistically significant wage premium. One potential explanation could be that these regions are where it is easiest to find relevant work in the private sector. Many people with a science background end up taking a one-year course in pedagogy to begin work as teachers, potentially driving the result. One way to investigate this would be to classify if they work in a relevant field to their education or not, but this is outside the scope of this master's thesis.

It can be argued that a more sound comparison is to run the regressions separately by field and leave out one of the regions as the reference. The reason why this is not presented here is due to the nature of the self-selection correction that will be presented in the next part of the paper. That said, figure 11 shows the results of running the regressions field by field, using Oslo as the reference category, controlling for the same covariates. This plot is to be found in the appendix and it largely tells the same story.

6.3 Corrected returns

With the uncorrected returns, the four regions that stood out from the rest were Oslo, Akershus, Rogaland, and, to some extent, Trøndelag. I will, therefore, initially focus on the self-selection bias in these regions before turning to provide an overview of the rest of the regions.

6.3.1 First stage

As with instrumental variable approaches, it is important to show that the instruments (where people lived at age 16) have predictive power after conditioning on the other covariates of the second stage. In IV this can be done using the first stage F-statistic, where the rule of thumb is that the F-statistic should be 10 or higher Staiger and Stock (1994) for the instrument to be considered strong.

I am unaware of any similar method of testing the strength of the instruments in a control function setting. But to show that the instruments have predictive power, I have included table 7 in the appendix. The table shows the estimated coefficient from the full multinomial model as specified above, with age, application score, parental education, gender, field, and county at age 16 as predictors. I only show the estimated coefficients on the county at age 16. The rows are the estimated coefficients for the region people lived in at age 16, and the columns are the outcome variable (where people lived in 2018). Oslo is the left-out reference category among the outcomes, and Akershus is set as the intercept for where people lived at age 16.

The likelihood ratio test, comparing the null model to the fitted model, follows a $\chi^2(765)$ with a test statistic = 208185.09 with an associated p-value = 0, which leads us to reject H_0 of the null model in favor of the fitted model. The model has a McFadden pseudo-R-squared of 0.4482.

Unsurprisingly, whether one grew up in a region is a strong and significant predictor of whether they will live there in 2018. Let's call these the diagonal elements. For all outcomes, the diagonal coefficients are the largest among the region coefficients within each outcome and are all highly statistically significant.

Considering the off-diagonal element, the size and significance vary greatly. For example, growing up in Oslo lowers the log odds of all other regions being the outcome, compared to Oslo, with all the effects being statistically significant with a p-value of 0.05 or less. For Rogaland as the outcome county, all coefficients are significantly different from zero at the 5% level or below except for Østfold. Comparing all estimated coefficients to the intercept yields that only growing up in Hordaland or Vest-Agder increases the log-odds of living in Rogaland in 2018 compared to Oslo, all else equal. For Trøndelag growing up in Finnmark, Troms, Norland, and Møre og Romsdal increases the log-odds compared to Oslo, with all coefficients being significant at the 5% level except for Vestfold, Østfold, and Vestagder.

Comparing these highly populated regions to a smaller region that mainly has people moving away, like Hedemark, reveals a lack of precision in the estimated coefficients across many of small regions. For Hedemark, only 5 out of 18 region coefficients are individually significant at the 5% level, with none of the implied log odds being positive except for Hedemark. When many of the region's partial effects are noisily estimated, it is reasonable to assume that the estimated probabilities from the model will be noisy, biasing the corrected model towards zero.

To test whether the IIA assumption is appropriate, I try dropping the outcome Nordland and performing a Hausman and McFadden (1984) IIA test. Under the null, there is no difference in coefficient (IIA is not rejected). Under the H_1 the coefficients are different. The test follows a $\chi^2(704)$ and has a test statistic of 33.14 and an associated p-value = 1. The test is unable to reject H_0. That said, 16 out of 720 coefficients are dropped due to the differenced covariance matrix not being positive semidefinite.

6.3.2 Second stage

Table 5 shows the corrected and uncorrected returns to the field of study in four different regions with spot estimates and bootstrapped standard errors in parenthesis [18]. The dependent variable is log(hourly wage). The table also reports p-values and degrees of freedom for an F-test on the joint significance of the correction terms[19]. The correction terms are put in

[18] I opted for using bootstrapped standard errors both for the corrected and uncorrected, in favor of using robust standard errors for the uncorrected model. The difference between the two was small, but especially in Akershus, the implied precision in the estimates was smaller in the corrected estimates. Using the uncorrected ones, this is only the case for medicine in Akershus and engineering in Trøndelag. I also include 10 in the appendix to show the bootstrapped densities of coefficient estimates

[19] To perform the F-test, I update the diagonal elements of the variance-covariance matrix with the bootstrapped estimates. I use the estimated covariance matrix from the regression for the covariance between the elements of the correction function. Though there are ways to get bootstrapped estimates of the full covariance matrix as proposed by, for exampleMachado and Parente (2005) this will not be pursued here. An alternative

the table 8 in the appendix for the sake of space. As noted by Dahl (2002), the intercepts of the corrected models are not separately identified from the correction terms and should not be interpreted causally[20].

The correction terms in Oslo, Trøndelag, and Akershus are all jointly significant at the 5% level or below, while the correction terms on Rogaland are highly insignificant. The significance of the correction terms, or lack thereof, indicates whether they jointly change estimates in the regression equation.

The general trend for all regions is that the corrected returns are less efficiently estimated with bigger standard errors. The exceptions are engineering in Trøndelag and medicine in Akershus, where the standard errors are marginally smaller. This could indicate that including the selection probabilities increases the precision of the estimate, but more likely, it is an artifact of the bootstrapping.

Table 6 offers an easier way of interpreting the results. I calculated the bias in the uncorrected model as $\beta_{uncorrected} - \beta_{correctred}$. When this value is negative, the uncorrected model underestimates the returns compared to the reference group. To assess whether the changes in estimates are significant, I use a Hausman-type test and report relevant p-values in parenthesis [21]. As expected from the fact that the correction terms for Rogaland were jointly insignificant, it is somewhat surprising to see that the Hausman test deems the change in the coefficient of medicine to be highly significant. Looking at the standard errors for this coefficient, the ones for the corrected and uncorrected are different by 0.0001, meaning that the Hausman test statistic will be scaled up a lot. I will get back to this later.

way of performing the test would be to assume that the off-diagonal elements of the covariance matrix on the correction terms are all zero. When doing so, all F-tests are highly insignificant. The problem with this approach is that it does not take into account the correlation between terms in the correction function

[20] It is worth noting that including all the estimated probabilities will lead to the model being rank deficient. The regression package, therefore, automatically drops one of the correction terms.

[21] The test evaluates the consistency of a more efficient estimator (b_1) to a less efficient estimator that is assumed to be consistent b_0. Under H_0 it is assumed that both estimators are consistent, while under H_1 only the less efficient estimator is assumed to be consistent.

In the multivariate case, the test statistic is

$$H = (b_1 - b_0)^T (Var(b_0) - Var(b_1))(b_1 - b_0) \tag{25}$$

Following a chi-square distribution with $rank((Var(b_0) - Var(b_1)))$ degrees of freedom. In the univariate case, it reduces to

$$H = \frac{(b_1 - b_0)^2}{Var(b_0) - Var(b_1)} \tag{26}$$

Following a $\chi^2(1)$ under the null.

Table 5. Corrected and uncorrected returns to field of study

	Oslo		Rogaland		Akershus		Trøndelag	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
Intercept	5.37405	5.27885	5.37492	5.40695	5.37876	5.30844	5.43021	5.32506
	(0.01378)	(0.06529)	(0.02498)	(0.15486)	(0.02307)	(0.11489)	(0.02131)	(0.14330)
Humanities	-0.06801	-0.06163	-0.05787	-0.05315	-0.04117	-0.03819	-0.08845	-0.07019
	(0.00595)	(0.00615)	(0.01270)	(0.01399)	(0.00992)	(0.01026)	(0.01203)	(0.01316)
Social Science	-0.01303	-0.00669	0.04222	0.04625	0.02542	0.02872	0.00447	0.02066
	(0.00555)	(0.00589)	(0.01192)	(0.01294)	(0.00986)	(0.01033)	(0.01034)	(0.01159)
Engineering	0.10206	0.10198	0.18645	0.18589	0.12188	0.12088	0.10573	0.10180
	(0.00938)	(0.00938)	(0.01226)	(0.01233)	(0.01417)	(0.01418)	(0.01083)	(0.01079)
Commerce	0.11761	0.12227	0.10332	0.10621	0.14214	0.14593	0.05409	0.06465
	(0.00562)	(0.00576)	(0.00857)	(0.00908)	(0.00861)	(0.00906)	(0.00903)	(0.01000)
Teaching	-0.05211	-0.05167	-0.04726	-0.04669	-0.02776	-0.02612	-0.03751	-0.03816
	(0.00534)	(0.00537)	(0.00739)	(0.00744)	(0.00679)	(0.00681)	(0.00707)	(0.00714)
Civil Engineering	0.13395	0.13590	0.22814	0.22796	0.17065	0.17179	0.11368	0.12460
	(0.00648)	(0.00663)	(0.01466)	(0.01513)	(0.01065)	(0.01081)	(0.01064)	(0.01123)
Science	0.04088	0.04247	0.12541	0.12651	0.07974	0.08167	-0.01090	-0.00513
	(0.00756)	(0.00760)	(0.01591)	(0.01614)	(0.01242)	(0.01249)	(0.01183)	(0.01202)
Law	0.10139	0.10798	0.12543	0.12913	0.13842	0.14255	0.09408	0.11613
	(0.00708)	(0.00744)	(0.01851)	(0.01950)	(0.01310)	(0.01335)	(0.01895)	(0.01921)
Medicine	0.20967	0.21051	0.36513	0.36252	0.22512	0.22283	0.28803	0.28287
	(0.01000)	(0.01005)	(0.01970)	(0.01971)	(0.01676)	(0.01673)	(0.01545)	(0.01565)
Application Score	0.00563	0.00578	0.00475	0.00482	0.00479	0.00487	0.00408	0.00442
	(0.00026)	(0.00026)	(0.00047)	(0.00049)	(0.00041)	(0.00043)	(0.00041)	(0.00042)
Father higher edu	0.01013	0.01247	0.00847	0.00966	0.02207	0.02393	-0.00667	-0.00193
	(0.00344)	(0.00351)	(0.00645)	(0.00669)	(0.00549)	(0.00561)	(0.00575)	(0.00605)
Mother higher edu	0.00484	0.00703	0.00627	0.00748	-0.00098	-0.00066	0.01173	0.01670
	(0.00326)	(0.00329)	(0.00583)	(0.00604)	(0.00553)	(0.00559)	(0.00559)	(0.00588)
Num.Obs.	27109	27109	8963	8963	9980	9980	9081	9081
RMSE	0.26	0.26	0.26	0.26	0.25	0.25	0.24	0.24
Gender X Age fixed effects	X	X	X	X	X	X	X	X
F test, p-value on correction terms	-	0.0066	-	0.8712	-	0.0033	-	0.0416
F test df	-	(17,27092)	-	(17,8946)	-	(17,9963)	-	(17,9064)

Note: The table shows corrected and uncorrected returns to field of study compared to health, with spot estimates and bootstrapped standard errors based on 1000 replications in brackets. The dependent variable is log(hourly wage). Estimation is done county by county. All models include age as dummies interacted with gender. The table also reports the p-value from an F-test for the joint significance of the correction terms and the associated degrees of freedom. RMSE is short for root mean squared error.

Table 6. OLS bias

term	Oslo	Rogaland	Akershus	Trøndelag
(Intercept)	0.0952 (0.1358)	-0.032 (0.834)	0.0703 (0.5322)	0.1052 (0.458)
Humanities	-0.0064 (0)	-0.0047 (0.4201)	-0.003 (0.2518)	-0.0183 (0.0006)
Social Science	-0.0063 (0.0015)	-0.004 (0.4214)	-0.0033 (0.2827)	-0.0162 (0.002)
Engineering	0.0001 (0.5861)	0.0006 (0.6806)	0.001 (0.1127)	0.0039 (-)
Commerce	-0.0047 (0.0004)	-0.0029 (0.3384)	-0.0038 (0.1772)	-0.0106 (0.0139)
Teaching	-0.0004 (0.4636)	-0.0006 (0.5292)	-0.0016 (0.003)	0.0007 (0.5327)
Civil Engineering	-0.0019 (0.1569)	0.0002 (0.9602)	-0.0011 (0.5379)	-0.0109 (0.0023)
Science	-0.0016 (0.0268)	-0.0011 (0.6877)	-0.0019 (0.1293)	-0.0058 (0.0072)
Law	-0.0066 (0.004)	-0.0037 (0.5457)	-0.0041 (0.1119)	-0.022 (0)
Medicine	-0.0008 (0.4231)	0.0026 (0.0001)	0.0023 (-)	0.0052 (0.0374)

Note: Shows the difference between $\beta_{uncorrected} - \beta_{corrected}$ for the different fields in four counties. P-values from Hausman-tests in brackets. When no number is in brackets, it indicates that the efficiency assumptions of the Hausman-test is violated. For more details, see footnote 21.

For Oslo, the estimates on humanities, social science, commerce, science, and law change significantly with p-values less than 0.05. The model suggests that fields are all underestimated in the uncorrected model compared to the reference category. The estimated size of the bias is only around 0.02-0.5 percentage points.

In Trøndelag, the Hausman assumption that the corrected model is less efficient is violated for engineering, thus not providing us with any inference about the significance of the change in parameters. All other fields except teaching experienced statistically significant changes at the 5% level of significance or lower. Medicine is the only field that has overestimated returns, while all the other significant fields are underestimated by between 0.5-2 percentage points.

For Akershus, the assumptions of the Hausman test are violated for medicine. Except for this, only the change in the coefficient on teaching rises to the level of statistical significance with p-value = 0.003 and a change in the coefficient of -0.16 percentage points.

One aspect that stands out is that the direction of the bias is the same within most fields

for the four regions. The two exceptions are medicine and civil engineering, which exhibit both positive and negative differences in spot estimates between the uncorrected and corrected models. Though this is the case, are the ranges of estimates decreased in most fields within the four regions.

One example is commerce, which in all four regions is underestimated. As discussed in 6.2 are Akershus, Rogaland, and Oslo, the three regions that clearly stand out, having a greater hourly wage premium, with Trøndelag lagging a fair bit behind. Though the direction of the bias is the same, the returns in Trøndelag are underestimated by around twice the size of the others, having the effect of decreasing the dispersion in wage premia across the regions. Among the four regions, the difference between the highest and lowest wage premia decreases in all fields except engineering and teaching, mainly being driven by the fact that Trøndelag closes some of the gaps to the top regions within each field.

An alternative way of performing inference on whether correcting for self-selection is changing the estimates is by treating the difference between the coefficients in the corrected and uncorrected returns equation as random variables and estimating the distribution of the difference by bootstrapping. For each bootstrap iteration, I fit the corrected and the uncorrected models on the resampled data before I take the difference in estimated coefficients. If the corrected coefficient is the same as the uncorrected, the difference will, on average, be zero, and one can test the hypothesis $H_0 : \beta_{uncorrected} - \beta_{corrected} = 0$ vs $H_1 : \beta_{uncorrected} - \beta_{corrected} \neq 0$ with the percentile method. This method should also be robust to situations where the corrected estimates are more efficient than the uncorrected ones, as is not the case for the Hausman test.

Figure 7 shows density plots of the estimated distribution of the difference between the uncorrected and corrected coefficients, using 1000 bootstraps. The colored regions show the probability density for different percentiles. In large, the figure leaves us drawing the same conclusions as table 6. This is, however, not the case for medicine in Rogaland, where the figure indicates only about one standard deviation difference between the corrected and uncorrected estimates, leading to us not rejecting the null of no difference. This is a different conclusion from the Hausman test on the same parameter.

The general tendency comparing the Hausman test to the bootstrapped test is that the indicated p-values are smaller in the latter. This could be exemplified with science in Oslo, which is only approaching the 5% level of significance in using the bootstraps compared to a p-value of 0.03 with the Hausman test, or teaching in Akershus approaching the 5% level of significance using the bootstrapped method and having a p-value of 0.003 using the Hausman test.

For the two instances, the corrected equation is more efficient than the figure provides us with inference. First, for engineering in Trøndelag, we reject the null of no difference at the 95% level of confidence, concluding that engineering is overestimated. For medicine in Akershus, we do not reject the null at either the 5 or 10% level of significance.

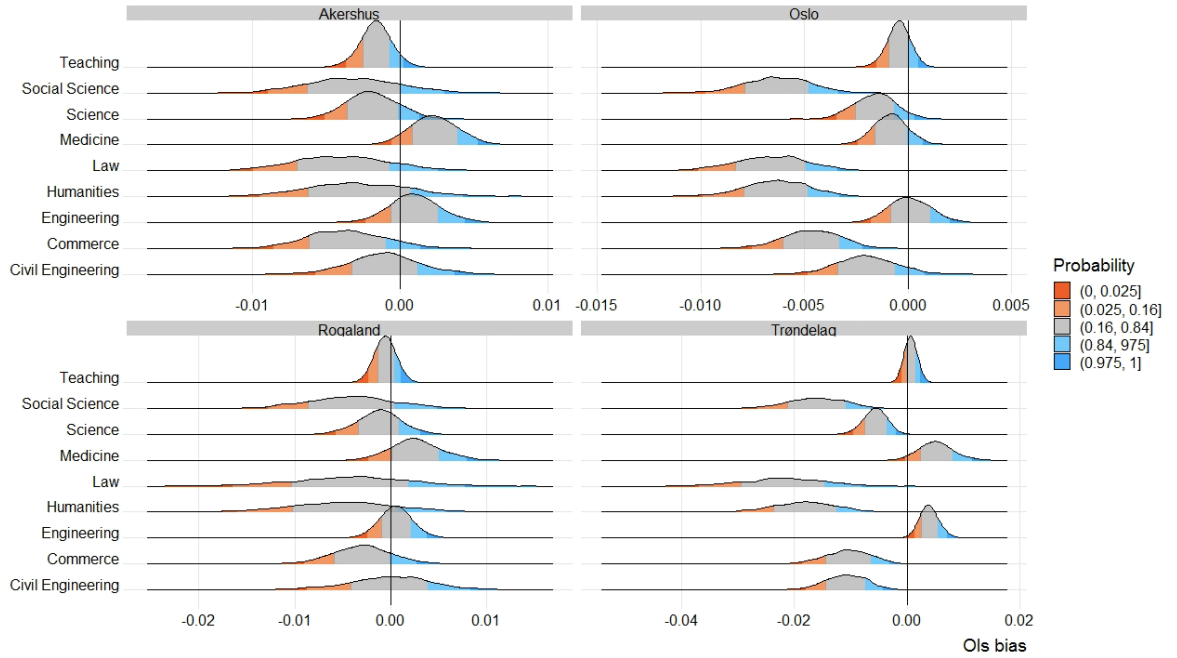

Figure 7. This plot shows the density of 1000 bootstrapped resamples where, for each iteration I take the difference $\beta_{uncorrected} - \beta_{corrected}$. The colored regions show the probability density for different percentiles.

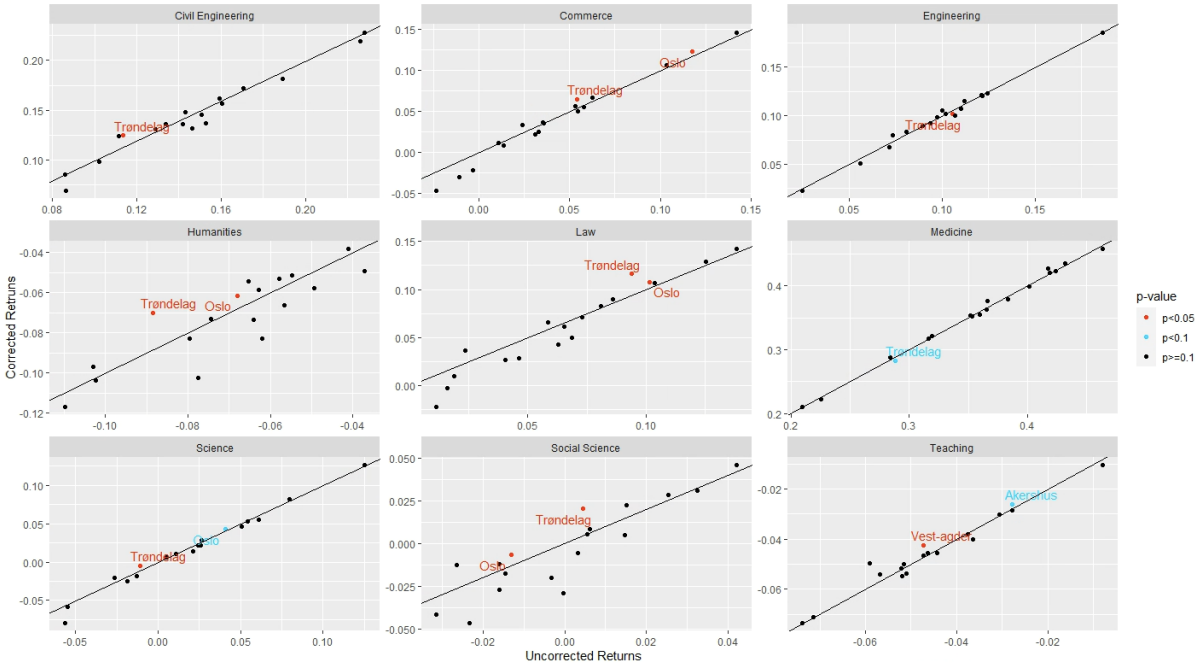

Figure 8. This plot shows the corrected returns on the y-axis and the uncorrected returns on the x-axis, broken down by field. Each point represents a county, and the black line shows the 45-degree line. The color of the points describes whether there is a statistically significant difference between the corrected and uncorrected results using the percentile method on the bootstrapped estimate of the difference as described above.

Figure 8 plots the corrected returns on the y-axis and the uncorrected returns on the x-axis, broken down by field. Each point represents a county, and the black line shows the 45-degree line. The color of the points describes whether there is a statistically significant difference between the corrected and uncorrected results using the percentile method on the bootstrapped estimate of the difference as described above.

If self-selected migration were a major driver of the between-region wage differences within fields, we would expect the points to form a less steep pattern than the 45-degree line, narrowing the gap between high and low-paying regions. This is however not the case, as the plot indicates a strong correlation between the corrected and uncorrected results, with no clear pattern of over or underestimation within specific fields. If anything, the points form a steeper pattern, indicating an increasing difference in the returns after correcting for self-selection.

That said, the most noticeable observation is the lack of significant changes in the individual parameters in most regions, with Trøndelag, Oslo, and Vest-Agder experiencing effects that are significant at the 5% level or below and Akershus experiencing effects that are significant at the 10% level of statistical significance with regarding teaching. Out of all the changes that rise to the level of statistical significance, all but two indicate that the uncorrected model underestimates the returns to certain fields compared to the reference category.

Figure 9 provides further details of to what extent correcting for self-selection changes the estimates on the coefficients of interest. It is constructed in the same way as the figure 7. Outside of the counties we already have discussed, there is little evidence that correcting for self-selection has significant effects on the coefficients of interest in most of the regions. This can be seen by the black line firmly cutting through the grey region for most coefficients.

The exceptions where we see some movement in the coefficients can be observed is Finnmark, Hedmark, and Møre og Romsdal, which all experience moves in coefficients that are around one standard deviation or more. The distributions suggest that the uncorrected model overestimates the returns to certain fields. This is the case for law in Møre og Romsdal, social science, science, and commerce in Finnmarke, and social science, law, humanities, and commerce in Hedmark. That said, none of the changes are significant at normal levels of statistical significance.

The plot also illustrates well the differing levels of precision, difference between the corrected and uncorrected models are estimated with. Take for example the difference between teaching in Oslo, where most of the mass of the estimated density lies within a 0.5 percentage point range, compared to teaching in Finnmark, which stretches over a span of around 8 percentage points. This difference in precision seems (not surprisingly) to go hand in hand with the number of observations a certain region had in 2018.

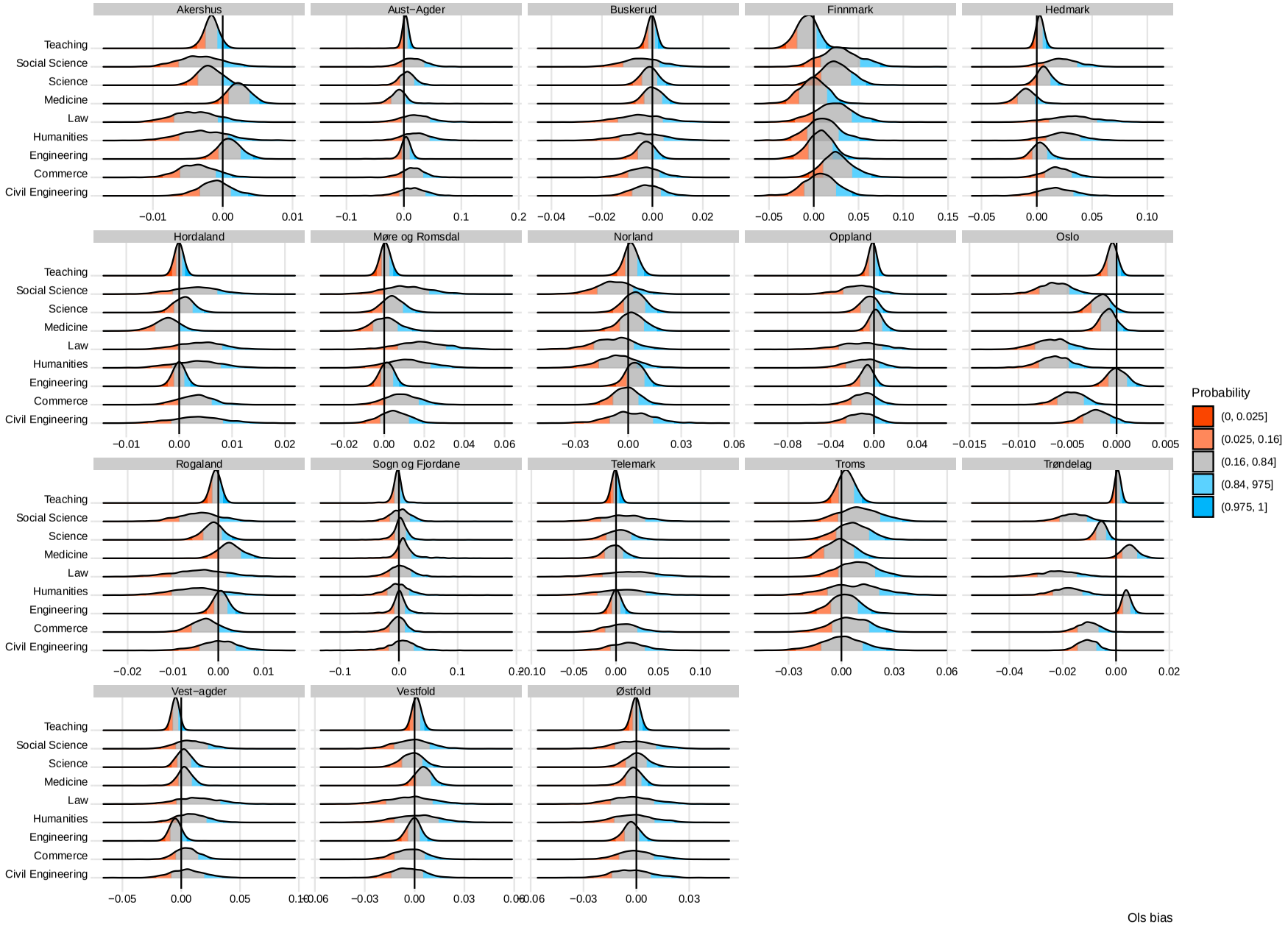

Figure 9. This plot shows the density of 1000 bootstrapped resamples where for each iteration I take the difference $\beta_{uncorrected} - \beta_{corrected}$. The colored regions show the probability density for different percentiles.

7 Discussion

If we truly believe that the model is appropriately specified and that the instrument is exogenous and strong, the result above points in the direction of self-selected migration not being an important driver of the differences in wage premia observed across counties within fields in Norway. With payoffs changing significantly within only a small subset of regions, ranging between 0.5 to 2.2 percentage points compared to the reference category.

It is possible that the reason significant effects were not found for more regions is due to a lack of power. With only 95,542 observations, 10 fields, and 18 counties, the data is spread thin. Additionally, the main migration flows are towards more densely populated areas, which means there is likely little variation in the estimated migration probabilities for smaller regions.

To clarify, even though the instruments used to predict people's propensity to stay are strong, if people are only moving out from a region, the estimated retention probability is high, and all other estimated probabilities are relatively similar and low for all individuals staying. Let's assume that no one is migrating into the region. In that case, all the people who have stayed have potentially fairly similar estimated probabilities for being seen in all 18 regions. As a result, this leaves us with too little variation to estimate the correction terms to the necessary level of precision. Though all migration routes have been observed, the same problem arises when there are only a few people migrating from a densely populated area, such as Oslo, to a less populated one, like Sogn of Fjordane.

Another reason might be the possibility of an incorrectly specified selection function. Apart from the problem of important left-out variables and the violation of the assumptions of the IIA test, functional form decisions could also be a contributing factor. For instance, it could be that gender and age should have interacted or that the application score should have been entered as a second-degree polynomial. One potential alteration that was tried was to interact the region at 16 with field, which made the model more flexible at the cost of drastically increased computation time by increasing the number of parameters estimated from around 700 to around 4000. However, this change did not significantly impact the results, except for making them more noisily estimated.

7.1 Instruments and the exclusion restriction

One of the early uses of distance instruments as cost shifters can be found in Card (1993). He uses college proximity from where an individual grew up as an instrument for whether people attend college, with the idea being that if you grow up closer to a college, the "cost" of attending will be lower, significantly increasing the probability of attendance. When considering this variation as exogenous, the estimates from his instrumental variable approach lead to returns to schooling results that are 25-60% higher compared to the conventional OLS.

One criticism of this approach is that it is not clear whether the exclusion restriction holds, even when family background characteristics like education level are taken into account. If the measured variables do not fully account for the effects of family background on education and wages, the instrument is not valid. Additionally, other unobservable factors, such as the cultural attitudes of parents towards education, could influence where you grow up near a college and future earnings.

The reason for discussing the paper by Card is that both Dahl (2002) and the current paper are applications of distance instruments, treating where one is born/grew up as exogenous with regard to earnings in a given labor market, only entering the selection equation as a cost shifter, changing the probability of observing an individual in a certain region. The insight from section 2 is that we are not observing a random sample of individuals in any given labor market but that it is an outcome of utility maximization. But if this is the case for the individual in our sample, it is reasonably also the case for their parents. One of the assumptions here is that skill and other wage-earning factors are equally distributed across all counties at birth, and this might be the case. Still, the problem is that the if characteristics of the parents affect, affecting both the type of education they obtain and their labor market outcomes. Controlling for parental characteristics mitigates some of these concerns, but unobserved characteristics like attitudes towards types of education, earning, and migration are not observed, potentially violating the excludability of the instrument.

Another reason why the exclusion restriction migth violated is if the quality of schooling differs between regions. Card and Krueger (1992) find that the return of an additional year of schooling varies based on the quality of schooling. They also show that rates of returns to education is higher when students have better educated teachers. If this is the case, where someone grew up will carry information on their wage-earning potential.

When comparing college to high school returns Dahl (2002) finds that the OLS is, on average, biased upwards about nine percent, with the change in coefficient being significant at the 5% level or lower in around half the states. My findings are much less drastic, and in most of the places, I find significant bias the OLS returns underestimate the payoff. One divergence between Dahl and myself is that he uses family circumstances as an instrument for self-selection. As discussed, he estimates migration probabilities by cell means. After having grouped people into levels of education attainment. "Married movers are then divided based on whether children 18 years or younger are present in the home. Nonmarried movers are grouped based on whether they live with extended family."

In an ordinary least squares regression, many of these variables could be considered potentially bad controls, as they are messured after the migration decision and might be both an outcome of the earnings potential and the migration decision. In a control function approach, the instrument needs to be excludable. That said it is not unlikely that the family circumstances carry additional information about both wage-earning potential and level of education. The fact that one is married or, for that case, has children most definitely can be

an outcome of wage-earning potential. One story could be that people who earn more are more attractive in the marriage market. If having a higher level of education makes someone more attractive in the marriage market at the same time as increasing the migration propensity, the exclusion restriction might no longer hold.

Wrapping up the discussion of the instrument, I would like to suggest a natural extension of the current study. We can use more robust instruments not only for self-selected migration but also for selection into the field. The current dataset is potentially suitable for a regression discontinuity approach, as demonstrated by Kirkeboen et al. (2016) for the field of study. They used the fact that having an application score marginally above the cut-off increases the probability of completing studies within a particular field. By conditioning on individuals' next-best field, they approximated people's choice margins, leading to a robust identification strategy for the payoffs for the field, compared to their next-best alternative.

The discontinuities present in the data do not only shift people between fields but also reasonably shift people between regions. When people apply, they specify field and university/university college combination. This means that people marginally above and below the cut-off will be shifted along two potential margins, creating credible instruments both for location and field. I explored this angle early in the master's process. I found that crossing the application threshold between schools located in two different regions significantly increased the probability of living in the region where one was assigned, showing at least some instrument relevance.

The reason why this thesis does not delve further into this aspect and present results using discontinuities as instruments is due to the relatively short time constraint of writing the master thesis during the fall semester. My idea was to use the Dahl approach as a baseline and then compare the results from the analysis done with more credible instruments. This, however, I see in hindsight was overly ambitious and will be left for future research.

7.2 Sample differences

Another obvious difference between Dahl's paper and mine is the populations that we study are different. He studies white men between 25 and 34, while my sample is noticeably older and does not make restrictions on race or gender. There is also an almost 30-year time difference between our cross sections. With so many moving parts, there are many avenues for investigating the differences in results.

A clear distinction with respect to the sample is that I only look at variation for people with higher education and do not compare the wage premium of college vs no college. Comparing the importance of self-selected migration on wages and the effect within education level to between education levels, it might be reasonable that correcting for self-selection leads to a smaller change in payoffs.

A more natural comparison than Dahl in this regard is Ransom (2021) who uses a similar

approach as Dahl, but only looks at people with tertiary education in an American setting, using data from 2010 - 2019. He uses teaching as his comparison group and finds that the returns to STEM and business are biased upwards by 15% at the median.

One difference is, of course, the difference between the Norwegian and American labor markets, where the wage structure in Norway is much more compressed due to the nature of how centralized wage bargaining works. This could be a reason for the difference in results, but another could come down to what is controlled for.

7.3 Does it all come down to ability?

The big advantage in my dataset compared to both Ransom (2021) and Dahl (2002) is that I have a fairly good measure of ability in application score. There are many aspects of self-selection that could lead to wages being biased in a certain region. Still, one fairly obvious one could be that ability determines wage-earning potential and correlates with the probability of self-selected migration. Not controlling for ability in this setting will lead to omitted variable bias.

Assuming there is some dependence between the propensity to migrate and ability, one can partially control for the effect of ability on wages indirectly by including the propensity to migrate. As my baseline comparison before correcting for self-selection includes a measure of ability, one could argue that the residual bias that can be corrected for by using the control function is relatively minor.

Let's assume column 2 in table 4 was the baseline. This is the table that demonstrates the importance of covariates in the case of Oslo, and in column 2, I only correct for age as dummies interacted with gender. If we assume this would be our baseline and we compare it to the corrected column for Oslo in table 8, the results start to look much more like what is observed by Ransom (2021), with the returns to, for example, medicine and civil engineering being biased upwards by around 7 percentage points compared to the reference category.

A test to see whether using where someone grew up purges the ability bias from the coefficients would be to run two regressions, both using the control function approach implemented in this paper, with the only difference being whether application score is included as a measure of ability or not.

7.4 Some further discussion on labor market partition and field aggregation

As has been discussed to some extent already, is what definition of the local labor market should be used. The first iteration of the analysis was done using an updated version of Bhuller (2009) 46 local labor markets. The pros of using such a partition is that it more closely captures the nature of the local labor market, with for example the greater Oslo metropolitan

area as its own region.

The challenge of using this definition was that I needed to aggregate the fields further due to the limited size of the sample. The dichotomy I used was to make the distinction between STEM (science, technology, engineering and mathematics) and non-stem (the other fields)[22].

Controlling for the full set of variables and comparing the effect of including the control functions showed no evidence of self-selection driving wage premiums. Of course, there could be many reasons for this, but one interpretation is that the aggregation glossed over important in-group differences, in returns, and migration preferences. Take, for example, Trøndelag where civil engineering and science are underestimated and engineering is overestimated, while when running the STEM vs non-STEM regression implemented as in this paper, showed no significant changes for Trondheim (the biggest city in Trøndelag).

This begs the question of whether the definition of fields used in this thesis is to course and hide additional heterogeneity. One way of tackling the question would be to get more updated wage data, giving us four to five more years, which could up the number of observations drastically, as most of the people in the application dataset are in the younger age brackets, being excluded due to their age in 2018. This could be an avenue of future research.

8 Conclusion

The primary concern of the paper is whether self-selected migration is one of the factors driving the differences in the hourly wage premiums within fields, across regions in Norway. Building upon existing literature, particularly Dahl (2002), the study uses the multinomial logistic regression model to estimate selection probabilities and employs a control function approach to purge the self-selection bias from the coefficients of interest.

The results show a few instances of bias in the OLS returns that rise to the level of statistical significance, with estimated biases ranging from 0.4 to -2.2 percentage points. Controlling for self-selected migration does not narrow the range of wage premiums within fields across regions.

The paper also documents the differing mobility patterns between different types of tertiary education and shows how migration propensity varies based on field of study and the region they resided in when growing up. Notably, the migration propensity into different regions varies greatly depending on the type of education individuals possess.

The paper also shows how controlling for age, gender, application scores, and parental education reveals substantial variation in the returns to education across different counties, particularly pronounced in fields typical of the private sector.

Overall, this thesis sheds light on the intricate relationship between mobility patterns,

[22] In hindsight, it would have been cleaner too make only nursing or teaching the reference group.

education returns, and self-selected migration, contributing new insights to the existing literature, particularly in the context of Norway.

References

Bertrand, M., Goldin, C., and Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American economic journal: applied economics*, 2(3):228–255.

Bhuller, M., Moene, K. O., Mogstad, M., and Vestad, O. L. (2022). Facts and fantasies about wage setting and collective bargaining. *Journal of Economic Perspectives*, 36(4):29–52.

Bhuller, M. S. (2009). Inndeling av norge i arbeidsmarkedsregioner. *Statistics Norway, Memo*.

Blackburn, M. L. and Neumark, D. (1993). Omitted-ability bias and the increase in the return to schooling. *Journal of labor economics*, 11(3):521–544.

Borjas, G. J., Bronars, S. G., and Trejo, S. J. (1992). Self-selection and internal migration in the united states. *Journal of urban Economics*, 32(2):159–185.

Bourguignon, F., Fournier, M., and Gurgand, M. (2007). Selection bias corrections based on the multinomial logit model: Monte carlo comparisons. *Journal of Economic surveys*, 21(1):174–205.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling.

Card, D. and Krueger, A. B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *Journal of political Economy*, 100(1):1–40.

Croissant, Y. et al. (2012). Estimation of multinomial logit models in r: The mlogit packages. *R package version 0.2-2. URL: http://cran. r-project. org/web/packages/mlogit/vignettes/mlogit. pdf*.

Dahl, G. B. (2002). Mobility and the return to education: Testing a Roy model with multiple markets. *Econometrica*, 70(6):2367–2420.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2):294.

Dubin, J. A. and McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Erola, J., Jalonen, S., and Lehti, H. (2016). Parental education, class and income over early life course and children's achievement. *Research in Social Stratification and Mobility*, 44:33–43.

Goldin, C. (2020). Journey across a century of women. *NBER Reporter*, 3:1–7.

Goldin, C. (2021). *Career and family: Women's century-long journey toward equity*. Princeton University Press.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 52(5):1219–1240.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hellstad, V. (2010). " nordlendinger uønsket": en studie av nordnorsk identitet i møte med oslo. Master's thesis.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Kirkeboen, L. J., Leuven, E., and Mogstad, M. (2016). Field of study, earnings, and self-selection. *Quarterly Journal of Economics*, 131(3):1057–1111.

Kleven, H., Landais, C., and Søgaard, J. E. (2019). Children and gender inequality: Evidence from denmark. *American Economic Journal: Applied Economics*, 11(4):181–209.

Korpi, M. and Clark, W. A. (2015). Internal migration and human capital theory: To what extent is it selective? *Economics Letters*, 136:31–34.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica*, 51(2):507–512.

Machado, J. A. F. and Parente, P. (2005). Bootstrap estimation of covariance matrices via the percentile method. *The Econometrics Journal*, 8(1):70–78.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

Midtbøen, A. H. (2015). *Diskriminering av samer, nasjonale minoriteter og innvandrere i Norge: En kunnskapsgjennomgang*.

Murphy, K. M. and Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97.

Ransom, T. (2021). Selective migration, occupational choice, and the wage returns to college majors. *Annals of Economics and Statistics*, (142):45–110.

Ripley, B., Venables, W., and Ripley, M. B. (2016). Package 'nnet'. *R package version*, 7(3-12):700.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.

Staiger, D. O. and Stock, J. H. (1994). Instrumental variables regression with weak instruments.

A Appendix

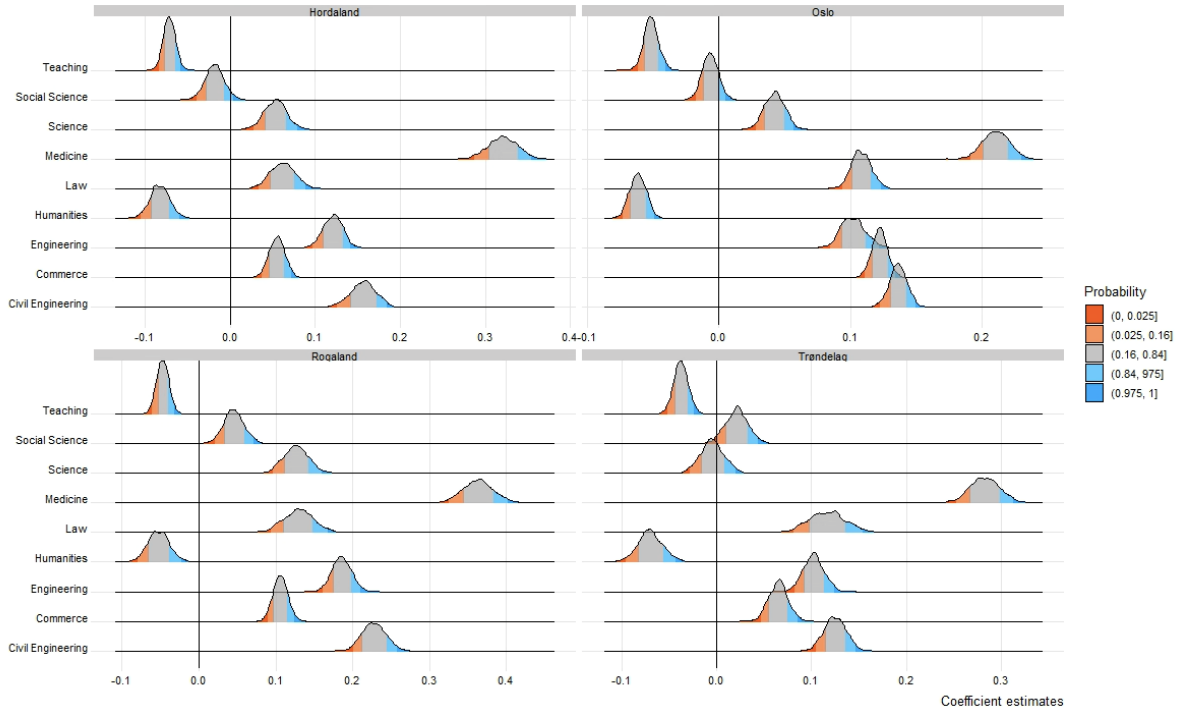

Figure 10. This figure shows the densities plot of the estimated coefficients for the corrected returns in four regions. The densities are based on 1000 bootstrapped and show the uncertainty with which the coefficients are estimated. The black line represents health, the reference category. Most coefficients have fairly normal-looking distributions.

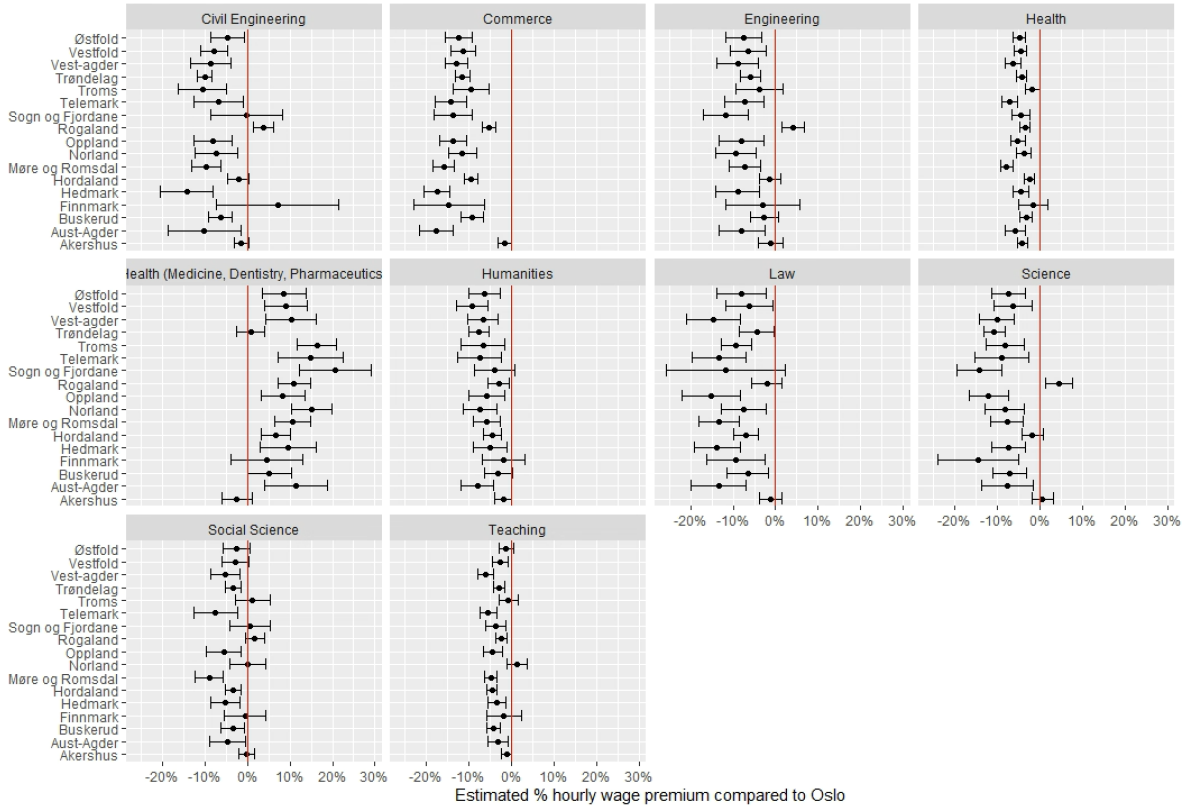

Figure 11. This plot shows the estimated percentage hourly wage premium compared to Oslo. The estimation is done field by field, controlling for age as dummies interacted with gender, application score, and parental higher education. The points are spot estimates and the errorbars are 95% confidence intervals

Table 7. The table shows a subset of the estimated coefficients and standard errors from the multinomial logistic regression. Oslo is the reference outcome. The columns are outcomes, and the rows are coefficients for where someone lived at age 16.

Number of obs = 95,542 LR chi2(765) = 208185.09 Prob > chi2 = 0.0000 Log likelihood = -128134.57 Pseudo R2 = 0.4482																	
	County 2018																
	Akershus	Aust Agder	Buskerud	Finnmark	Hedmark	Hordaland	Møre og Romsdal	Norland	Oppland	Rogaland	Sogn og Fjordane	Telemark	Troms	Trøndelag	Vest agder	Vestfold	Østfold
County age 16																	
Oslo	-1.949	-1.296	-1.591	-0.614	-1.479	-0.897	-0.967	-0.932	-1.172	-0.881	-1.083	-0.918	-0.498	-1.051	-1.096	-0.979	-1.298
	(.038)	(.321)	(.113)	(.282)	(.192)	(.119)	(.218)	(.237)	(.169)	(.157)	(.310)	(.251)	(.207)	(.114)	(.206)	(.142)	(.126)
Rogaland	-1.867	1.134	-0.34	1.311	-0.36	2.647	1.344	1.182	0.041	5.629	1.357	1.173	1.237	1.165	2.086	0.447	-0.368
	(.0797)	(.272)	(.146)	(.284)	(.258)	(.089)	(.199)	(.228)	(.218)	(.104)	(.268)	(.241)	(.223)	(.105)	(.151)	(.166)	(.181)
Sogn og Fjordane	-1.792	1.429	-0.173	0.79	0.699	3.731	2.671	0.666	0.739	2.088	6.03	0.404	1.085	1.467	0.746	0.208	-0.371
	(.117)	(.326)	(.201)	(.489)	(.241)	(.097)	(.185)	(.386)	(.236)	(.154)	(.193)	(.440)	(.327)	(.132)	(.290)	(.257)	(.265)
Telemark	-1.652	1.619	0.225	0.578	-0.467	0.93	0.226	0.594	-0.489	1.202	0.508	5.159	0.58	0.62	1.693	1.369	-0.408
	(.080)	(.248)	(.125)	(.385)	(.287)	(.132)	(.301)	(.292)	(.287)	(.157)	(.367)	(.164)	(.291)	(.132)	(.169)	(.132)	(.194)
Troms	-1.431	0.371	0.007	2.862	0.438	1.158	1.203	3.02	0.269	1.679	-0.188	0.553	5.813	2.199	0.312	-0.215	0.107
	(.113)	(.534)	(.202)	(.262)	(.291)	(.175)	(.297)	(.197)	(.310)	(.184)	(.733)	(.441)	(.159)	(.117)	(.378)	(.333)	(.232)
Trøndelag	-1.622	0.682	-0.191	1.645	0.671	1.268	2.199	1.984	0.678	1.37	0.913	0.786	1.675	4.484	0.435	0.109	-0.138
	(.064)	(.279)	(.122)	(.248)	(.158)	(.106)	(.159)	(.179)	(.156)	(.133)	(.276)	(.243)	(.188)	(.074)	(.208)	(.164)	(.144)
Vest-agder	-1.956	2.802	-0.757	0.898	-0.384	1.184	0.04	0.656	-0.262	2.526	0.308	1.16	0.917	0.254	5.047	-0.233	-0.551
	(.096)	(.209)	(.201)	(.358)	(.297)	(.128)	(.346)	(.299)	(.279)	(.124)	(.422)	(.264)	(.269)	(.159)	(.128)	(.243)	(.223)
Vestfold	-1.821	0.895	0.443	0.406	-0.219	0.618	0.5	0.385	-0.73	0.669	0.303	1.689	0.596	0.364	0.626	3.911	-0.338
	(.070)	(.272)	(.101)	(.346)	(.222)	(.126)	(.236)	(.269)	(.270)	(.1628)	(.343)	(.202)	(.249)	(.123)	(.201)	(.094)	(.163)
Østfold	-1.614	0.382	-0.539	0.177	-0.031	0.304	-0.24	-0.114	-0.187	0.215	0.142	-0.197	0.527	0.065	0.315	0.18	3.504
	(.065)	(.308)	(.140)	(.370)	(.200)	(.138)	(.300)	(.315)	(.209)	(.185)	(.354)	(.341)	(.249)	(.135)	(.218)	(.163)	(.079)
Aust-Agder	-1.838	5.488	-0.175	0.812	-0.65	1.03	0.911	0.518	-0.411	1.825	-0.613	2.056	0.949	0.129	3.162	0.469	-0.723
	(.108)	(.187)	(.182)	(.425)	(.393)	(.153)	(.281)	(.367)	(.350)	(.155)	(.732)	(.228)	(.308)	(.194)	(.146)	(.210)	(.280)
Buskerud	-1.519	0.345	3.134	0.312	-0.037	0.83	0.319	0.326	0.455	0.713	0.366	0.978	0.793	0.378	0.605	1.157	-0.611
	(.064)	(.322)	(.067)	(.357)	(.207)	(.118)	(.250)	(.274)	(.171)	(.161)	(.333)	(.236)	(.233)	(.122)	(.202)	(.126)	(.181)
Finnmark	-1.456	1.062	-0.083	6.216	-0.094	1.248	1.56	2.369	0.716	1.593	0.798	1.151	4.287	2.059	0.651	0.671	-0.039
	(.185)	(.612)	(.335)	(.221)	(.594)	(.264)	(.390)	(.320)	(.402)	(.279)	(.737)	(.533)	(.203)	(.179)	(.523)	(.358)	(.397)
Hedmark	-1.48	-0.081	-0.48	-0.087	4.047	0.168	0.355	0.462	1.497	0.444	0.051	0.273	0.813	1.26	-0.076	-0.294	-0.634
	(.068)	(.397)	(.150)	(.451)	(.108)	(.157)	(.263)	(.279)	(.136)	(.184)	(.399)	(.312)	(.246)	(.102)	(.273)	(.211)	(.193)
Hordaland	-1.617	1.202	-0.119	1.307	0.075	5.104	1.587	1.183	0.149	3.019	2.767	0.957	1.407	1.167	1.279	0.435	-0.421
	(.072)	(.264)	(.133)	(.291)	(.216)	(.081)	(.186)	(.228)	(.207)	(.112)	(.209)	(.253)	(.214)	(.104)	(.178)	(.1643)	(.182)
Møre og Romsdal	-1.715	0.739	-0.523	0.875	0.297	2.078	5.255	0.955	0.075	1.612	1.943	0.969	0.785	2.297	0.614	0.204	-0.458
	(.071)	(.292)	(.149)	(.322)	(.190)	(.095)	(.140)	(.239)	(.205)	(.132)	(.231)	(.246)	(.253)	(.084)	(.208)	(.171)	(.176)
Norland	-1.462	0.744	0.185	2.224	0.401	1.221	1.661	5.672	-0.04	2.054	0.375	1.12	3.868	2.961	0.757	0.178	-0.141
	(.090)	(.366)	(.150)	(.265)	(.236)	(.140)	(.212)	(.155)	(.280)	(.143)	(.450)	(.287)	(.163)	(.090)	(.253)	(.223)	(.204)
Oppland	-1.51	0.051	-0.058	0.458	1.402	0.464	0.827	0.486	4.055	0.535	0.972	0.628	0.731	1.092	0.585	-0.106	-0.308
	(.068)	(.379)	(.127)	(.370)	(.141)	(.142)	(.225)	(.279)	(.106)	(.179)	(.291)	(.275)	(.256)	(.107)	(.214)	(.195)	(.170)
Intercept (Akershus)	1.122	-2.678	-0.616	-3.311	-1.691	-2.382	-2.108	-2.346	-1.433	-2.404	-3.13	-2.549	-2.468	-2.017	-2.541	-1.277	-0.67
	(.100)	(.316)	(.167)	(.396)	(.227)	(.145)	(.222)	(.254)	(.220)	(.169)	(.307)	(.274)	(.258)	(.138)	(.229)	(.195)	(.193)

Table 8. This table shows the correction terms for corrected regressions in table 5. Boo strapped standard errors in parenthesis based on 1000 resamples.

	Oslo	Rogaland	Akershus	Trøndelag
Akershus	0.11824	0.05749	0.05357	0.14097
	(0.06451)	(0.17328)	(0.11199)	(0.14717)
Aust-Agder	0.05674	0.01523	0.06392	0.19470
	(0.06935)	(0.16356)	(0.12532)	(0.16481)
Buskerud	0.11789	0.00890	0.09960	0.09755
	(0.06484)	(0.17286)	(0.11326)	(0.15062)
Hedmark	0.10298		0.03920	0.13003
	(0.06627)		(0.11617)	(0.14416)
Hordaland	0.09782	−0.03569	0.09746	0.09883
	(0.06447)	(0.15450)	(0.11498)	(0.14422)
Møre og Romsdal	0.08333	−0.05091	0.04364	0.09555
	(0.06612)	(0.16243)	(0.11608)	(0.14323)
Norland	0.09426	−0.04187	0.14708	0.11444
	(0.06903)	(0.16005)	(0.11770)	(0.14583)
Oppland	0.08866	−0.08133	0.08731	0.08579
	(0.06675)	(0.20327)	(0.11543)	(0.14780)
Oslo	0.06389	−0.07089	0.05400	−0.00324
	(0.06404)	(0.16181)	(0.11282)	(0.14818)
Rogaland	0.08286	−0.03741	0.06066	0.10020
	(0.06454)	(0.15413)	(0.11619)	(0.14584)
Sogn og Fjordane	0.09090	−0.04812		0.02481
	(0.07360)	(0.16920)		(0.16147)
Telemark	0.07232	−0.07278	0.14857	0.08232
	(0.06639)	(0.16477)	(0.11902)	(0.15603)
Troms	0.10847	0.13268	0.20275	0.10127
	(0.06638)	(0.17904)	(0.12284)	(0.14503)
Trøndelag	0.10126	0.00010	0.08497	0.10772
	(0.06522)	(0.15771)	(0.11378)	(0.14238)
Vest-agder	0.10282	−0.03293	0.01438	
	(0.06632)	(0.15898)	(0.11838)	
Vestfold	0.09368	−0.08206	0.10039	0.07605
	(0.06574)	(0.16814)	(0.11728)	(0.15724)
Østfold	0.12632	−0.02470	0.10519	0.03371
	(0.06537)	(0.17564)	(0.11425)	(0.15005)
Finnmark		−0.08594	0.06732	0.06959
		(0.19938)	(0.16482)	(0.16647)
Num.Obs.	27109	8963	9980	9081
RMSE	0.26	0.26	0.25	0.24
F test, p-value on correction terms	0.0066	0.8712	0.0033	0.0416
F test degrees of freedom	(17,27092)	(17,8946)	(17,9963)	(17,9064)

B Estimation code

Beneath, I attach the code for estimating the corrected models. If there is a need for looking at other parts of the code please send an email to snorre.b.skagseth@gmail.com.

```
setwd("N:\\durable\\projects\\p23snorresk")
library(data.table)
library(tidyverse)
library(broom)
library(fixest)
library(furrr)
library(nnet)

df_boot <- fread("df_boot.csv")
df_boot_fylke$field <-  factor(df_boot_fylke$field,
    levels =c("Health", "Humanities", "Social␣Science",
    "Engineering", "Commerce", "Teaching",
    "Civil␣Engineering", "Science", "Law",
    "Health␣(Medicine,␣Dentistry,␣Pharmaceutics)"))


df_boot_fylke$fylke_16 <- as.factor(df_boot_fylke$fylke_16)
df_boot_fylke$fylke_2018 <- as.factor(df_boot_fylke$fylke_2018)
df_boot_fylke$age <- as.factor(df_boot_fylke$age)

#####Fitting the corrected model
#First stage
fml_fylke_basic <- fylke_2018 ~ female + age + application_score + field +
fylke_16 + far_higher_edu + mor_higher_edu

reg_vars_inst_basic <- c("hourly_wage", "female", "age",
"application_score", "field", "far_higher_edu", "mor_higher_edu")

#This is the multinomial logistic regression
m_logit <-  multinom(fml_fylke_basic, data = df_boot_fylke, maxit = 1000,
                     MaxNWts =10000000)

#Extracting selection probabilities
selection_probs <- m_logit$fitted.values %>% as_tibble()
```

```
#choisng the covariates for the second stage
df_fit <- df_boot_fylke %>%
  select("fylke_2018","log_hourly_wage", "female", "age",
"application_score", "field", "far_higher_edu", "mor_higher_edu")

#Adding in selection probabilities.
df_fit <- cbind(df_fit, selection_probs)

#Getting a vector of all the counties
county <- df_fit$fylke_2018 %>% unique() %>% as.character()

#Looping over all the counties
model_corrected_soseco<- map(county ,~{df_fit %>%
    filter(fylke_2018 == .x) %>%
    select(-fylke_2018) %>%
    lm_robust(log_hourly_wage~. + age:female,data =.)})


names(model_corrected_soseco) <- county


###Bootstrapping standard errors
#A function that eestimates a linear model given all it gets as an input
estimate_model <- function(data, LLM, name) {
    #Linear model with log(hourly_wage) as lefthand
    #side and all other data on the right
  lm_robust(log(hourly_wage) ~. + age:female, data = data) %>%
    broom::tidy() %>% #Ectracts coefficients
    select(term, estimate) %>% #Selecting the term and the estimate
    mutate(LLM_2018 = LLM,
           correction_type = name)
}

#A function that draws a draws n observations with resampling
#and fits the corrected and uncorrected models
boot_logistic <- function(.x, multinom_fml, reg_variables, maxitter = 1){
  #Draws a random sample from df_boot
  temp_df  <- sample_n(df_boot, size = nrow(df_boot), replace = T)

  #Computes the multinomial logit
  m_logit <-  suppressMessages(multinom(multinom_fml, data = temp_df,
```

```
                                        MaxNWts =10000000, maxit = maxitter))

  #Extracting the fitted values from the multinomial logit
  selection_probs <- m_logit$fitted.values %>% as_tibble()

  #Makes a vector of all the LLMs in the sample
  LLMs <- temp_df$LLM_2018 %>% unique()

  #A loop the runs over all the LLMs and
  #estimates both the OLS for the correction functions

  map_dfr(LLMs, ~ { #.x are the elements that are looped over one by one.
        #Subsets the data so one only look at
        #people who live in a given LLM in 2018
    sub_temp_df<- temp_df[temp_df$LLM_2018 == .x,] %>%
        #Selecting all variables in the vector reg_vars
      select(all_of(reg_variables)) %>%
      mutate(index = row_number()) #ads index numberss as ids.

    #extracts a vector of all probabilities for
    #each individual living in a given LLM in 2018.
    selection_probs_temp<- selection_probs[temp_df$LLM_2018 == .x,]

    #Computing the model, the polynomial expantion
    #combines the selection probabilities and the data
    raw_p_model <- cbind(sub_temp_df, selection_probs_temp) %>%
      select(-index) %>% #Takes out index
        #Estimates a linear model.
      estimate_model(data = ., LLM = .x, name = "logit_probs")

    #Computing the model with the MDF structure
    #Log of the probability for the utility-maximizing LLM
    p_1<- log(selection_probs_temp[,.x]) %>%
      rename(p_1 = 1) %>% #Renaming the probability
      #Making an index for the join.
      mutate(index = row_number())

# Combining p_1 and the other selection probabilities
    selection_probs_temp_dmf <- cbind(p_1, selection_probs_temp) %>%
    #removing the current LLM from the probabilities, but keeping p_1
      select(-all_of(.x)) %>%
```

```
    #Make the dataframe long to make the next calculation easier
    pivot_longer(-c(index, p_1), values_to = "p_j", names_to = "name") %>%
#p_j are probs for all LLMs not chosen, p_1 is for the chosen LLM
    mutate(p = (p_j*log(p_j))/(1 - p_j) + p_1) %>%
    select(index, name, p) %>%
    #Making data frame wide again.
    pivot_wider(values_from = p, names_from = name)

 #Estimating the second step
 mdf_model <- left_join(sub_temp_df, selection_probs_temp_dmf,
 by = "index") %>% #Joining the data and the correction terms
    select(-index) %>% #Taking out index
    #Estimating the second stage.
    estimate_model(data = ., LLM = .x, name = "DMF")

 #computing dmf model, Edwins example
 p_1 <- -log(selection_probs_temp[,.x]) %>%

 #Same precedure as above. Taking the
 #log of the probability of the chosen LLM
    rename(p_1 = 1)

 selection_probs_temp_edw <- selection_probs_temp %>%
    select(-all_of(.x)) %>%  #Takes out the chosen LLM
    mutate(index = row_number()) %>%  #Makes index
    #Makes a long dataframe
    pivot_longer(-index, values_to = "p_j", names_to = "name") %>%

    #p_j is the probabilities for all non-chosen LLMs
    mutate(p = (p_j*log(p_j))/(1 - p_j)) %>%
    select(index, name, p) %>%
    #Makes a wide data fram again
    pivot_wider(values_from = p, names_from = name) %>%
    #adds in the -log of the probability of
    #the LLM individual i has choosen.
    cbind(p_1)


 mdf_model_edw <- left_join(sub_temp_df, selection_probs_temp_edw,
 by = "index") %>% #Combining the data and the corrections
```

```
      select(-index) %>% #Taking out index
      #Estimating the second step.
      estimate_model(data = ., LLM = .x, name = "DMF_edw")

    #Estimating ols on the same sample,
    #to be able to do inference on changes
    #in parameters.
    ols <- sub_temp_df %>%
      select(-index) %>%
      estimate_model(data = ., LLM = .x, name = "ols")

    #Combining all the estimates into one long data frame
    rbind(ols, raw_p_model,mdf_model, mdf_model_edw) %>%
      return()

  }) %>%
    mutate(itter = .x) %>% #Adding the iteration of the bootstrap
    return()

}

#Number of bootstraps.
n_bootstrap_resamples <- 1000
plan(multisession, workers = 4)
#Runs the bootstraps in parralell
#Still, 1000 resamples take 25 hours.
model_logistic_fylke_basic <-
future_map_dfr(1:n_bootstrap_resamples,
    ~boot_logistic_fylke(.x, multinom_fml =
    fml_fylke_basic,
    reg_variables = reg_vars_inst_basic,
    maxitter = 350),
    .options = furrr_options(seed = TRUE))

#Computes the statistics from the boostrap.
boot_stats <- model_logistic_fylke_boot %>%
  na.omit() %>%
  group_by(term, LLM_2018, correction_type) %>%
  summarise(std.error = sd(estimate),
            conf.low = quantile(estimate, probs = 0.025),
            conf.high = quantile(estimate, probs = 0.975),
```

```
    estimate = median(estimate))
```