Bachelor's project

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

Nazifa Mohyuddin & Tommy Brodersen

The correlation between parents and firstborn child's level of education

Bachelor's project in Social economics
Supervisor: Per Tovmo
May 2021

NTNU
Norwegian University of Science and Technology

Nazifa Mohyuddin & Tommy Brodersen

The correlation between parents and firstborn child's level of education

Bachelor's project in Social economics
Supervisor: Per Tovmo
May 2021

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

NTNU
Norwegian University of
Science and Technology

Bachelor thesis in social economics:

The correlation between parents and firstborn child's level of education.

13.05.2021

Sammendrag

Hensikten med denne oppgaven er å undersøke korrelasjonen mellom foreldrenes og det førstefødte barns utdanningsnivå. Denne studien undersøker om høyere utdanning hos forelder fører til høyere utdanning hos førstefødte barn. Denne oppgaven tar for seg dataen som ble samlet inn i 2017, ved hjelp av runde 3 i livsstudien bestående av tre runder. For å undersøke avhandlingsspørsmålet vil vi hovedsakelig bruke OLS-regresjon, og se på de forskjellige aspektene ved oppgavespørsmålet. Vi undersøker blant annet effekten av å dele foreldrenes og barnas utdanningsnivå inn i kategorier, og kjører OLS-regresjoner blant samme utdanningsgruppe, og på tvers av de forskjellige utdanningsgruppene.

Våre estimeringsresultater viser tydelig at ulike utdanningsnivåer blant foreldre påvirker det førstefødte barnet ulikt. Derimot, antyder testene at vi ikke kan se på denne korrelasjonen uten å vurdere andre signifikante variabler.

Summary

The purpose of this thesis is to examine the correlation between the parent's and the firstborn child's level of education. This study examines whether higher levels of education in parents leads to higher education in their firstborn child. This thesis considers the data collected in 2017 using round 3 of the life study consisting of three rounds.

To investigate the thesis question, we will mainly use the OLS-regression and look at the different aspects of the thesis question. We investigate, among other things the effect of dividing the parent's and child's education level into categories and running OLS-regression among the same education group, and across the different education groups.

Our estimation results clearly show that different levels of education among parents effects the firstborn child differently. When that is said, our tests insinuate that we cannot look at this correlation without considering other significant variables.

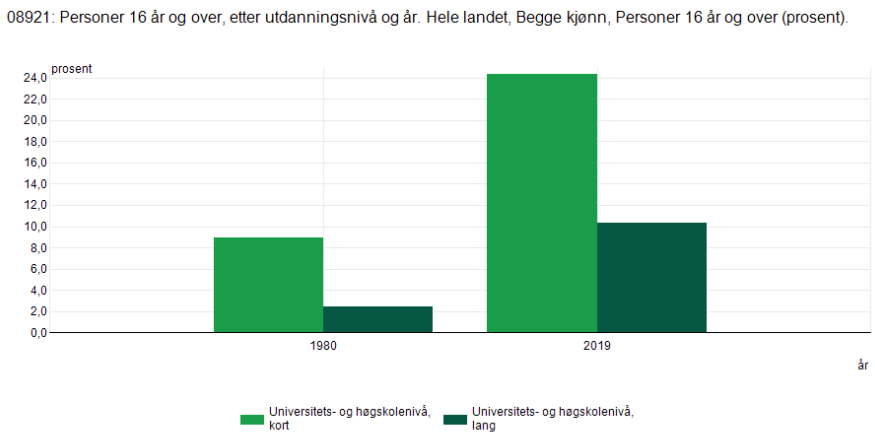
Table of contents

1 Introduction

In this thesis we are looking at the education level of parents and their firstborn child in Norway. We are looking specifically at the correlation between the parent's education level and if it has an impact on the education level on their firstborn child. In 1980 10,3 % of Norway's population had a college education and in 2019 that number had increased to 34,6%. (Statistisk sentralbyrå, 2020) More jobs today require a college degree of some sorts to get a job, while in the 1980's more people would work their way up in their company to get the experience and abilities to qualify for the job.

With the progress of technology, knowledge, and higher requirements, the demand for better qualified workers has also risen. In 1980 one could build a house on their own, but today you must follow different regulations and be certified to do different jobs which requires specialization in different fields where the work must be signed off and be up to code. In Norway, the education is free and accessible to everybody and the state offers scholarships and loans during your time at college.

With Norway being one of the richest countries in the world per capita (Worldbank, 2019) the need for streamlining the work is absolutely required to compete with the international marked who can pay less in both wages and production costs to do the same job.



Source: Statistisk sentralbyrå (SSB)

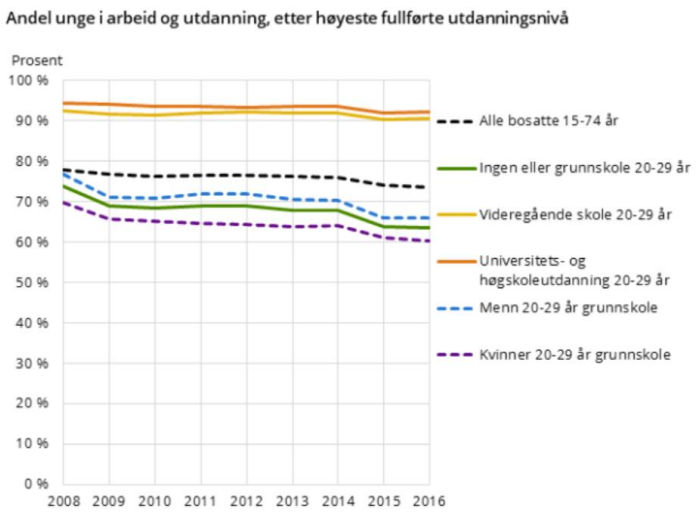
1.1 Norway

Norway is a big producer of oil and fish, and the production of oil and farmed fish represents over 60% of the nation's export (World's top exports, 2021). This has made Norway one of the richest countries in the world but also one of the most expensive to live in (Numbeo, 2021). All of Norway's oil is at sea and with that comes higher production cost compared to other oil producers who have their oil on land. Therefore, the need for innovation and streamlining the production is very important for companies in Norway who want to compete internationally against other producers with lower wages and production costs. With the need for innovation and optimization comes the need for expertise and specialized education in the respective fields.

1.2 Education level

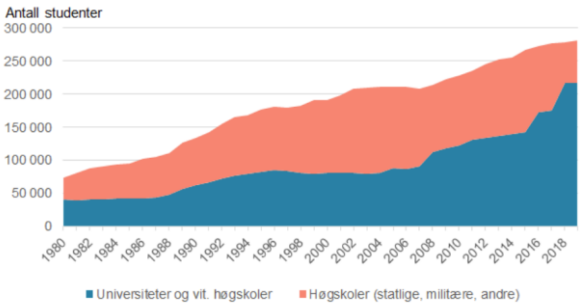
In Norway higher education is divided into three categories, university, scientific university college and university college. There are 10 universities, 9 scientific university college's and 13 university colleges that have accreditation to educate students, and a handful of private schools that have some accredited studies but are not accredited as a school. (Studentum, 2021)

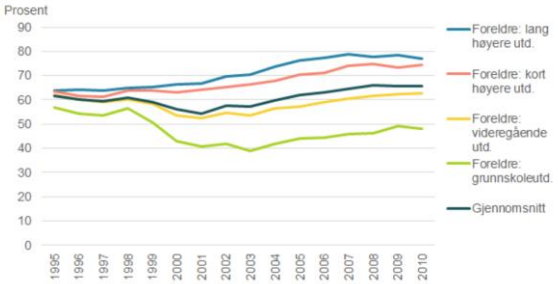
In 2018 OECD published a report called "Investing in youth Norway" (OECD, 2018) which shows that lack of education is the biggest risk factor to ending up outside the work marked. 90% of all youth aged 20-29 with a higher education have a job, where those who do not have a degree is it only 74% who have a job, a number that has decreased by 10% from 2008-2018 (Statistisk sentralbyrå, 2021)



Source: SSB

Since the 1980's the number of students who take higher education has tripled from 73 000 in 1980 to 273 000 in 2016. The share of students with a parent with higher education has risen the last 20 years from 40% to over 50% in 2017. Statistics also show that the education level of a parent has an impact on a student's completion rate within the first 8 years of starting a degree. For higher education, a student finishes their degree within 8 years 75% of the time and for a parent with primary school, just under 50% finishes in that time. (Forskningsrådet, 2021)





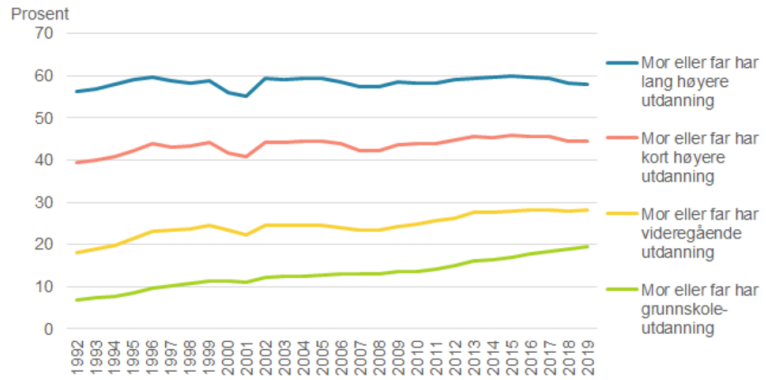
Source: SSB

1.3 Parents

In 1980 78,3% of men aged 15-74 years old were working and only 53,8% of all women aged 15-74 years. In 2019 those numbers had changed to 70,5% and 65,1% for men and women. and the population has increased from 4 078 900 in 1980 to 5 328 212 in 2019, with the workforce going from 1 902 000 to 2 724 000.

With the big growth in students in higher education the importance of a parent's education level is still relevant, but it was 8 times higher chance of becoming a student in 1992 has been reduced to 3 times higher chance in 2019. If a parent has long higher education between 55-60% of their children becomes students and for short higher education 40-45%. This number has been stable from 1992-2019. For a parent with primary school these numbers were 7% in 1992 and 19% in 2019. (Forskningsrådet, 2021)

In the table below we interpret the expressions "lang høyere utdanning" as a parent having a master/PhD degree, and "kort høyere utdanning" as a parent with bachelor's degree.



Source: SSB

1.4 Reasoning

The reason for choosing our thesis question was to see if there is a correlation between a parent's and their first-born child's level of education. We also want to study the relationship between the different levels of education between parent and child.

Previous research such as "indikatorrapporten", and SSB with their data on the field shows that there is a correlation between the parent's and child's education level. Studies like "Birth order Matters" (Booth, 2009) shows that a parent has the biggest impact on their first-born child.

With more children, the effects cannot only be based on the parent's level of education. With firstborns and Norway as a starting point, we will keep the unobserved variables under control. These unobserved variables are, for example, tuition, fees, scholarships and student loans. In Norway, this is something that is available to everyone, and therefore more people will have the opportunity to go to school, and this will give a better result to build on for future studies.

2 Data material

The data we collected for our thesis is based on a sample from a life study from NSD (Norsk senter for forskningsdata) carried out over three rounds of the survey (2002, 2007 and 2017), going under the name NorLAG. NorLAG gathered the information from men and women over time from their 40's and onwards. All participants are born between 1922 and 1966. It gathers information on work, family, health, and life quality. In the survey they have also given information on their education level, gender, wages, personality and more. NorLAG

gathered their information over three rounds, but we are looking at the reported results specifically from round 3 to get the newest results we have access to.

NorLAG lets you build your own dataset by choosing the variables you need, and every response is identified with a specific ID-number to identify who have given the different information across the variables.

In this life survey the parents are the participants and not the children. The information on the children is given by the parents. We must therefore assume that the answers given are not only correct but also completed degrees. We do not have the age of the children so any results we get would only be stronger if one had the chance to take out children under the age of 22 which is a student doing a bachelor's degree on normed time would be age wise. (Norsk senter for forskningsdata, 2021)

3 Theoretical frameworks

Previous literature in the field is "Indikatorrapporten" (Forskningsrådet, 2021) which is a yearly report on the Norwegian research and innovation system. Here they show the statistics of how many aged 19-24 are in higher education and what their parents education level is. It also shows several students who come from the same household. This differs from what our thesis is focusing on, which is the correlation between firstborn child and parent's education level, assuming finished degrees.

In 2016 Alissa Jo Combs-Draughn of Walden University wrote a doctor dissertation "The impact of psychological birth order on academic achievement and motivation" (Combs-Draughn, 2016) which takes into consideration the studies of the effect of being a firstborn and the effect of being the psychological first born in a family, and what impact this has on academic achievements and motivation. This gives great insight into both the impact of being the firstborn child and feeling like one is, and the implications this brings. While we are looking at the actual first born it is important to consider other factors that could influence the tests we are doing. An example of this is if a divorce happens in a family, and the new family composition leads to that the child who has previously the firstborn, no longer is in the new family. In situations like this, Combs-Draughn studies the consequences this could have.

Since we are looking at the correlation between the parents' education level and what impact this has on their first-born child's education, we must therefore look at previous literature of both the parents' impact and the effect of being a first-born.

The indicator report shows that the chances of a student becoming a student is, and how well they perform is impacted by the education level of the parent. Even with a 300% increase in students from 1980-2019 the share of students having a parent with higher education is over 50% of all the students in Norway.

In Alissa Jo Combs-Draughn's research she investigates previous studies on the effect of the birth order, and she finds that the psychological order where a child feels like the firstborn effects their motivation, but she cannot find conclusive evidence with regards to academic achievements. Previous research does however find proof that the birth order does in fact have an impact on a child's academic achievements and motivation. (Booth, 2009)

4 Econometric model

In this thesis we are going to look at how one and more independent variables ($x_1 + x_2 + \cdots + x_k$) are related to our dependent variable y, which is an individual's education level. We are therefore going to use the *Multiple Linear Regression Model.* The goals of the multiple regression model (MLR) are to model the linear relationship between the explanatory variables ($x_1 + x_2 + \cdots + x_k$) and the explained variable (y).

Linear regression is an approach of modeling the relationship between a dependent variable and one or more independent variables.
We hypothesize that the relationship is:

$$yi = \hat{\beta}_0 + \hat{\beta}_j x_i + u_i \quad (1)$$

This is a procedure of fitting a line through the data points. The next step is to find an objective rule that will deliver estimates for β_0 and β_1, this is to determine the intercept (β_0) and slope (β_1)of that line we are trying to fit.

yi = dependent variable

x_i= explanatory variables

$\hat{\beta}_0$= the intercept. Measures the expected y when the explanatory variables are equal to zero, ceteris paribus.

$\hat{\beta}_j$= measures the expected change in the dependent for a unit change in an explanatory variable, ceteris paribus.

u_i= the models error term. It represents the difference between actual values and the estimated values of a regression.

The OLS estimator is a rule based on minimizing the sum of squared residuals which is:

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (2)$$

Were,

$\hat{u}_i$= are the residuals squared.

(1) Shows the residual for each observation i and the difference between the observed data point for the explanatory variable, and what we would predict based on the following model:

$$yi = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i \quad (3)$$

The sum of squared residuals is the unexplained variation in the model. This is anything that you can confidentially say is not due to the regression model itself. It can be factors or variables that we have not included in our regression model.

By Gauss-Markov theorem, we know that if the assumptions of the OLS holds, the OLS is the best linear unbiased estimator (BLUE). We choose therefore to use the OLS estimator to estimate parameters of simple and multiple linear regressions. When that is said, we can use the OLS estimator regardless of whether the assumptions hold. The impact the assumptions of the OLS estimator has on our regression model is if our model delivers good and realistic answers or not. For a regression model to deliver good and realistic answers, the assumptions of the OLS-estimator must be fulfilled, when the assumptions are fulfilled, we can then say that our OLS-estimator is unbiased.

4.1 Assumptions of the OLS estimator

The *Multiple Linear Regression* model (MLR) assumes the following:

MLR.1: Linearity.

We could assume that the model is linear. This means that we assume there exists a linear relationship between the dependent variable and the independent variables.

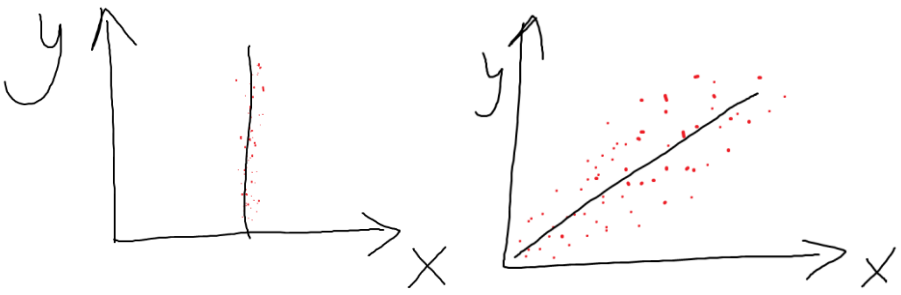
MLR.2: Random sampling.

This means that each observation has the same probability of being selected. In other words, this means that the sample is independent and identically distributed. In our dataset the number of individuals is randomly chosen within some restrictions. From our first dataset the restrictions on choosing the individuals were that these individuals have children. The reason for this is that we want to study the relationship between the parents and the child's level of education.

MLR.3: Enough variation and no perfect collinearity.

We need to have enough variation in our independent variables. This means that the variance of $(x_1 + x_2 + \cdots + x_k)$ must be different from zero. $v(x_i) \neq 0$

Graphically we cannot draw the variation across our sample like the left-handed figure. Since the observations need to be more spread out to have a variance different from zero.



No perfect collinearity means that we cannot use the other independent variables to describe one independent variable. But that said, we do not require uncorrelation between regressors. High correlation also works, but this assumption rules out perfect correlation. This also means that the R-squared value cannot equal 1! This assumption is usually satisfied. In our model we have included dummy variables, if we do not enter the dummy variable trap, our model will have no perfect collinearity.

We have five categories of education level. One of these categories for example «master's degree» must be in the reference group, since if none of them are in the reference group, then we will have nothing to compare the other parent's education levels with. A parent must be in one of the five categories. If we would have decided to include all the five categories, this would have been seen as the «no perfect linear combination» condition, and it is called a dummy variable trap.

MLR.4: Zero conditional mean.

This assumption says that the expected value of the error term, given any of the independent variables x_i must equal to zero. $E[u_i|x_i] = 0$. If this is no longer held, this means that the OLS estimator is biased. This means that our $\widetilde{\beta_j}$ given any of the independent variables does not equal to the population parameter β. The reason for this is sampling error. My sample will not necessarily represent what happens in the population. Another way of writing $E[u_i|x_i] = 0$, is $\text{cov}(u_i, x_i) = 0$. This means that there is no relationship between u_i and x_i. If this assumption were violated, we would not have a normal distribution of the error terms. The distribution would have been normal up to a point, but beyond that point it is going to be different, maybe go in one direction. In our model we are looking isolated on how the parent's education level affects the child's education level.

Under assumptions MLR.1-4, we can say that the OLS estimator is unbiased.

$$E(\hat{\beta}_j) = E(\beta_j) \tag{4}$$

This means that on average the expected value of the sample parameter $\widetilde{\hat{\beta}_j}$ is equal to the population parameter β_j . We can then conclude that the OLS estimator BLUE.

MLR.5: Homoscedasticity.

The error term has the same variance given any values of the explanatory variables. In other words,

$$\text{var}(u|x_1 \dots x_k) = \sigma^2 \tag{5}$$

This assumption means that the variance of the error term, conditional on the explanatory variables, is the same for all combinations of outcomes of the explanatory variables. If this

assumption does not hold, this means that the model exhibits heteroscedasticity. In our model heteroscedasticity would mean that the error term depends on the parent's level of education.

MLR.6: Normality.

This assumption says that the population error term is independent of the explanatory variables ($x_1 + x_2 + \cdots + x_k$) , and that u is normally distributed with zero mean and variance σ2. Which are the criteria for a term to be normally distributed. This assumption implies a stronger efficiency property of OLS; the OLS estimator have the smallest variance among all unbiased estimators.

5 Theory about hypothesis testing

Hypothesis testing in statistics is a way for you to test the results of your sample data, to see if you have meaningful results. So, by using this method we are testing whether our results are valid by figuring out the odds that our results have happened by chance. If our results have happened by chance, we can then say that our regression models have little use.

Hypothesis tests in the general case are about examining whether it is sufficient statistical evidence than an original hypothesis, often called the null hypothesis is correct, or whether this null hypothesis must be rejected in favor of an alternative and contradictory hypothesis. In this thesis we are going to use "Single Hypothesis Test" and "Joint Hypothesis Test".

5.1 Single hypothesis test

Single hypothesis tests are also known as Student's t-test or t-test under the OLS. This type of test uses the t-distribution. The t-distribution is essentially the normal distribution, but for small samples. This means that as the sample size grows, the t-distribution converges towards the normal distribution. The distinction between these distributions is very important for the inference we want to draw about the true population, using our sample. The reason for this is because we infer real life relationships between variables, then we run regressions on the sample drawn from the population, and thereafter test if the relationship between two or more variables are statistically significant. An important notice is that when we talk about the distribution of our sample, we are usually referring to the error term u. So, when we are doing a single hypothesis test, we are assuming that the error term is normally distributed. This is also the assumption of MLR.6, normality of errors. This assumption is very strict, but we can also use the Central Limit Theorem. The Central Limit Theorem says that since u, the error

term is the sum of many different unobservable that affect y, it will have an approximately normal distribution.

Under the *Multiple Linear Regression* model assumptions MLR.1-MLR.6, the t-distribution of a standardized estimator is equal to:

$$t_{n-k-1} = t_{df} \quad (6)$$

where df stands for the degrees of freedom. This is equal to the number of observations(n) subtracted by the number of slope parameters(k), and the intercept (1). The t-test can be stated as the following:

$$\text{TS}= \frac{\tilde{\beta}_1 - \beta_1}{se(\tilde{\beta}_1)} \sim t_{n-k-1} = t_{df} \quad (7)$$

Where $\hat{\beta}_j$ is computed through OLS using the sample regression model. β_j is the true population parameter. The standard error of our $\hat{\beta}_j$ is the estimate we get from our sample regression model. So, the nature of this test is to test if our estimate is statistically different from the true population parameter. The value of this is chosen by what we want to test, and this choice represents our null hypothesis(H0). Our alternative hypothesis (HA) represents a specified deviation from the null hypothesis.

5.2 Joint hypothesis test

F-tests are used to test hypotheses containing multiple variables. The computation of a F-test is different from a t-test since we now estimate two models instead of one. These two models are called the restricted and unrestricted models. The restricted model is just like the unrestricted model, except that we have removed the variables we want to test from the restricted model. We will then end up with two models, where they both have their own SSR, R^2, and their own degrees of freedom.

The test statistic can either be constructed around the SSR, or the R^2. The test will measure how much the power of our regression is reduced, when we exclude the variables, we are testing in the restricted model.

The F-statistic for the R-squared formula will look like the following:

$$\text{F-stat}= \frac{R_u^2 - R_r^2}{1 - R_u^2} * \frac{n-k-1}{q} \quad (8)$$

Were:

R_u^2= R^2 from the unrestricted model

$R_r^2 = R^2$ from the restricted model

n= number of observations

k= number of parameters

1= intercept (B0)

q= number of restrictions

The SSR-formula will look like the following:

F-stat=$\frac{SSR_R - SSR_U}{SSR_U} * \frac{n-k-1}{q}$ (9)

SSR_R= SSR from the restricted model

SSR_U= SSR from the unrestricted model

The distribution of the test statistic is:

F-stat $-F_q: n-k-1$ (10)

The rejection region is:

F-stat> critical value c

We know that the F-statistic always will be positive, since variances always are positive, both the numerator and denominator for F must therefore always be positive.

6 Descriptive statistics on the levels of education of parents and children

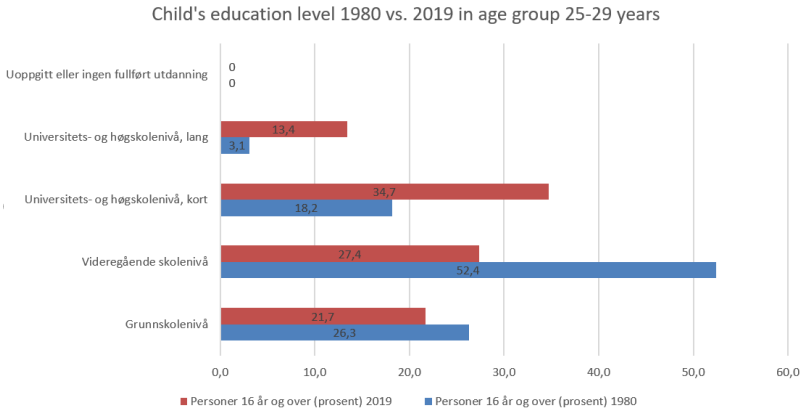
In this section we are going to look at descriptive statistics over the development in the education levels for both parents and children. We are using SSB's statistics over how the education levels have developed from 1980 to 2019. (Statistisk sentralbyrå, 2021)

6.1 Child's education level development 1980 compared to 2019 in percentage in Norway.

We can see that taking higher education among younger people has become an increasing trend compared to 1980. There are many explanatory reasons for that. The main reason is that higher positions in the labor market are adapted to individuals with higher education. There is an OECD report called "Investing in youth Norway" which shows that the lack of education is

the biggest risk factor to ending outside the labor market. (OECD, 2018) This report strengthens the incentive for younger people to take higher education.

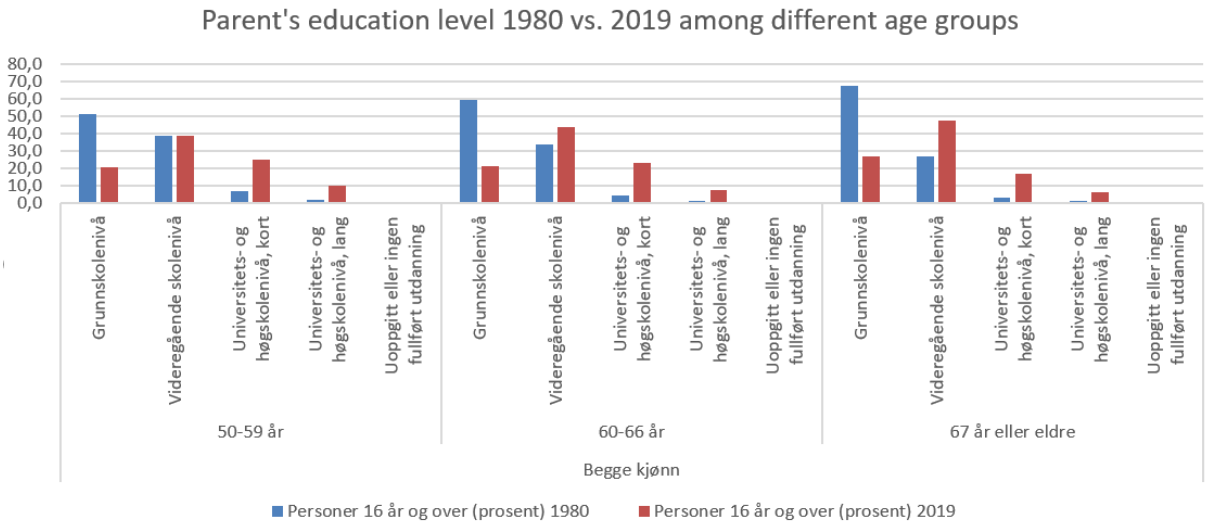
Another factor is that since we are looking at the statistics in Norway, we do not have to control for tuition fees, since education in Norway is free. When that is said, the demand after higher education has changed a lot since 1980. The technological development did not accelerate until the 1990s, so the demand for higher education is greater today than it was in 1980. So, this is also a big reason for higher education.



Source: SSB.no

6.2 Parent's education level development from 1980-2019 in percentage in Norway.

The development of the parent's education level varies among the age groups. When that is said, we can see that the common factor is that there is a large proportion of parents compared to 1980 who take higher education. We know that there are many social factors that affect parent's ability to pursue higher education, but the common denominator here is that higher education today has become more accessible and has a lower alternative cost. By alternative cost, we mean that there is less loss in the form of income, which makes the choice to take higher education more desirable. In addition to this, we have many social schemes and benefits for individuals who wants to take higher education today, and this is something that the labor market aspires to. (Statistisk sentralbyrå, 2021)



Source: SSB.no

7 Estimation results

In our OLS regression we have regressed a simple linear regression (SLR) on child's education level in terms of the parent's education level. These variables are not continuous, but categorical. We have five categories. These are the following:

- Elementary school/ no education (pschool)
- High School (hschool)
- Supplementary studies qualifying for higher education (hpschool)
- Bachelor's degree (universityl)
- Master/PhD (universityh)

The relationship between child's and parent's education level looks like the following:

$$yi = \hat{\beta}_0 + \hat{\beta}_1 ed\_ioedu + u_i \qquad (11)$$

A very strong assumption for this regression is that since ch281 and ed_ioedu are categorical variables, we are assuming a constant change in the variables. This means that we are saying that if a parent goes from having high school completed to having a bachelor's degree has the same effect on the child's education as a parent who goes from having no education to having a high school degree.

	Model 1	Model 2
ch281		
ed_ioedu	0.288 (0.003)**	
hschool		0.359 (0.013)**
hpschool		0.498 (0.013)**
universityl		0.939 (0.013)**
universityh		1.120 (0.018)**
_cons	2.622 (0.010)**	2.894 (0.009)**
R^2	0.0973	0.1002
N	66,323	66,323

**Standard errors in parantheses.

Table 7.1: Correlation between parent's and child's education level.

In table 6.1 our simple linear regression model is model 1. Model 2 is our multiple linear regression where we divided the explanatory variable ed_ioedu into the different education levels.

7.1 Interpretation of the SLR model

If the parent's education increases my one category, the child's education will increase by 0,28 categories, ceteris paribus. That is what our slope parameter $\hat{\beta}_1$ represents. The constant $\hat{\beta}_0$ is 2,62, which shows the value of ch281 when ed_ioedu is zero. $\hat{\beta}_0$ is also represented as the intercept parameter.

The number of observations is 66,323, this is our sample size which is randomly drawn from the population.

The *standard error* indicates how accurate the mean of any given sample from that population is likely to be compared to the true population mean. This means that when the standard error increases, the means are more spread out. This again means that it becomes more likely that any given mean is an inaccurate representation of the true population mean. In this simple linear regression model, the standard error is 0,003. Which is low, this means that we have a pretty accurate mean.

The p-value $P>|t|$ for a t-statistic will correspond to the result of a t-test in the following way: The p-value is the probability of observing a t-statistic as extreme as we did if the null hypothesis was true. If we get a p-value of, say 0.1, then the probability of observing a t-statistic we did is 10 percent, assuming that the null hypothesis is true.

The integrated null and alternative hypothesis in Stata is:

$H0: \beta_1 = 0$

$HA: \beta_1 != 0$

The test statistic is:

TS= $\frac{0,288-0}{0,003} = 96$

The t-statistic tells us that if H0 was true, the probability of observing a t-stat= 96 on $\tilde{\beta}_1$. We are testing if the parent's education level has an impact on the child's education level at a 5 % significance level. We get a t statistic of 96 with a probability of 0,00% from the regression table in Stata. This means that the probability of observing a t-statistic as extreme as 96 is 0 % if H0 was true. This means that we have enough evidence to reject H0 and say that we are 95% certain that parent's education level has an impact on the child's education level.

The table also provides the R^2which is 0,0973 which equals 9,73%. This shows us how close the data are to the fitted regression line. It tells us the percentage of variation in ch281, explained by the regression on ed_ioedu. It only tells us the strength of the correlation it does not tell us if the correlation between ch281 and ed_ioedu is positive or negative. So, since we have a relatively small R-squared, we can say that 90,27% of the variation is due the factors outside our regression model.

The *SSR* is also called the sum of squared residuals. This is the unexplained variation in the model. This is anything that you can confidentially say is not due to the regression model itself. It can be variables or factors that are not included in the regression model or anything that is not from the model itself. In this model the value of the SSR is 87260,3.

7.2 Interpretation of the MLR model

We have now introduced dummy variables into our OLS- regression. Dummy variables are not quantitative, they are qualitative. This is the reason for comparing all the variables to the category "elementary school/no education because this is our reference group. We have done this to prevent the dummy variable trap.

The coefficient on *hschool* is 0,359. This means that a parent with a high school degree effects the child's education level more than parents with elementary school/no education by 0,359 categories, ceteris paribus.

The coefficient on *hpschool* represents that a parent with supplementary studies qualifying for higher education effects the child's education level more than parents with elementary school/no education by 0,498 categories, ceteris paribus.

The interpretation of the coefficient on *universityl* shows that if a parent had a bachelor's degree compared to having finished elementary school/no education increases the child's education level more by 0,939 categories, ceteris paribus.

The coefficient on *universityh* represents that if a child's parent a master/PHD degree, compared to elementary school/no education increases the child's education level more by 1,128 categories, ceteris paribus.

We observe that the R^2 has increased by 0,29% compared to our SLR model. This implies that in our MLR model it is a larger percentage of variation in the child's education level that is explained by the regression on the different categories of parent's education level. The reason for this can be that we now have separated ed_ioedu in categories, we prevent the different ranges of variation to cancel each other out.

7.3 Hypothesis testing

In this section we want to test if the parent's education level is jointly statistically significant or not to determine if there is a constant change from one category to another. We perform therefore the following test:

$HO: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

HA: not HO

Stata has an integrated command called *"test"*. We use this command with the categorical dummies we want to test. When we use the command *"test hschool hpschool universityl university"* Stata report the F-statistic F (4, 66318) = 1847,27, with Prob>F= 0

This means that the chance of seeing a F-statistic as extreme as 1847,27 is 0% if H0 was true. We can therefore conclude that we will reject H0 at a 95% level. This means that there is 5% chance of being wrong. Further on this means that the expected change from one category to another is not constant.

In conclusion this means these categorical dummies are jointly statistically significant. Since the individual dummy variables have high t statistics, with p-values equal to zero, this result is as expected. This may be because these variables are collinear with each other, this means that they have relatively high levels of correlations with each other. We can therefore say that these variables have statistically insignificant t scores but are jointly statistically significant.

8 Robustness

In this section we are going to discuss how robust our regression model is. We are therefore going to check if our results are robust to the possibility that one of our assumptions might not be true.

8.1 Is there certain education level groups that have more mobility than others?

Here we are checking correlations between the probability that parents and children have the same level of education. Such correlations may be stronger for certain levels of education than others. Therefore, we have now divided our dependent variable child's education level into 5 categories and ran regressions among the same education level between parent and child.

	childpschool	childhschool	childhpschool	childuniversityl	childuniversityh
pschool	0.115 (38.05)**				
hschool		0.051 (30.31)**			
hpschool			0.046 (10.59)**		
universityl				0.072 (16.74)**	
universityh					0.210 (39.79)**
_cons	0.090 (66.09)**	0.024 (28.66)**	0.322 (154.71)**	0.341 (159.40)**	0.143 (97.80)**
R^2	0.02	0.01	0.00	0.00	0.02
N	66,323	66,323	66,323	66,323	66,323

* $p<0.05$; ** $p<0.01$

Table 8.1: Correlation between the probability that parents and children have the same level of education.

We can see from the t statistics that correlations between the probability that parents and children have the same level of education vary. These correlations are stronger for certain levels of education than others. The highest t-statistic is between parents who have a high university degree and children who have the same level of education compared to the other categories. From table 8.1 we can see that the correlation between a high university degree for both parent and child is 39,79, with probability of 0% if H0 was true. We will therefore reject the null hypothesis at a 5 precent significance level and say that we are 95% confident that children with parents who have a master or PhD degree, have higher probability of taking a master/PhD degree.

Another observation from table 8.1 is that the t-statistic between parents with elementary school/no education and children with the same level of education is relatively high at 38,05 with probability of 0% if H0 was true. Therefore, we can conclude that we reject $H0$ with 95% certainty and say that children with parents with elementary school/no education, have a higher probability of taking no education/ elementary school.

We can also see that the R^2 value on the education level groups master/PhD and elementary school/no education level is higher compared to the other regression models in table 7.1. This means that there is higher variation in the dependent variable child's education which is explained by the regression on parent's education compared to the other regression models in

table 8.1. Further this implies that master/PhD and elementary school/ no education level groups have more mobility than the other education level groups.

8.2 Cross sectional regressions, to see if there are certain education level groups that have more mobility than others.

In this section we are checking the correlation between the probability that parents and children have different levels of education. We are therefore running cross sectional regressions. We have chosen to look at children with a master/ PhD degree, since we saw more extreme values in this specific education group. So, our dependent variable in the model below is children with master/PhD, and we are looking at the differences among the different education levels for the parents.

	childuniversityh			
pschool	-0.130 (37.31)**			
hschool		-0.065 (19.93)**		
hpschool			-0.061 (18.13)**	
universityl				0.158 (48.79)**
_cons	0.185 (117.70)**	0.175 (107.29)**	0.173 (107.14)**	0.120 (74.87)**
R^2	0.02	0.01	0.00	0.03
N	66,323	66,323	66,323	66,323

* $p<0.05$; ** $p<0.01$

Table 8.2: Correlation between the probability that children with master/PhD and parents have different levels of education.

We observe that the highest t-statistic is between parents with a bachelor's degree and children with a master/PhD. We get a t-statistic of 48,79, with a probability of 0% if H0 was true. We will reject the null hypothesis at a 5 precent significance level and say that we are 95% confident that a child has a higher probability of taking a master/PhD if the parent has a bachelor's degree compared to having elementary school/ no education, high school, and supplementary studies.

Since we looked at the correlation between children with master/PhD and parent with bachelor's degree, we want to look at the correlation both ways. In the regression between

children with bachelor's degree and parents with master/PhD we get the same results as we did between parents with bachelor's degree and children with master/PhD. We can therefore conclude that for each jump in education level, the correlation between parent and child's education increases.

8.3 Testing for homoskedasticity

My analysis assumes that the variance of the error term u_i is constant, and unrelated to the independent variables ($x_1 + x_2 + \cdots + x_k$). This means that we assume that the error term is homoskedasticity. Heteroskedasticity is when the variance of the error term u_i is related to one of the independent variables ($x_1 + x_2 + \cdots + x_k$) in your regression. If my error term is heteroscedasticity, then my results may have incorrect standard errors. We suspect that the error term may be heteroscedasticity in my analysis because we have only included one independent variable which is the parent's education level divided into categories. Intuitively we know that there are many other factors and variables that may affect a child's education level. Therefore, we may say that there are factors in the error term that may be related to the parent's education level, like for example income.

We want therefore to test for homoscedasticity. We are going to use δ as an expression to check if our error term is homoskedastic.
H0: δ =0 (homoskedasticity)
H0 says that the variance of the error term u_i is constant, and unrelated to the independent variables ($x_1 + x_2 + \cdots + x_k$).
H1:δ!=0 (heteroskedasticity)
H1 says that the variance of the error term u_i is related to one of the independent variables ($x_1 + x_2 + \cdots + x_k$).
We use the *Multiple Linear Regression Model*

$$yi = \hat{\beta}_0 + \hat{\beta}_1 hschool + \hat{\beta}_2 hpschool + \hat{\beta}_3 universityl + \hat{\beta}_4 universityh + u_i \qquad (12)$$

The reason for using this model is because from the hypothesis test, we know that there is not a constant change from one category to another in parent's education level. This means that our first model, which was a SLR model is not good enough to use for testing for homoscedasticity, since it assumes constant change from one category to another. We choose therefore to run the homoscedasticity test on our MLR model, which is shown in (12).

To check for homoscedasticity in Stata, we firstly run our MLR model, and then use the command *"hettest, rhs iid"*. This is a Breusch-Pagan/ Cook-Weisberg test for heteroskedasticity. This test is distributed as a Chi-square distribution. Which is the distribution that both F and t- distribution is derived from. From Stata vi get the Chi-square statistic 194,85, with a probability of 0% if H0 was true. This means that we will reject H0. Since the probability of observing such an extreme chi- square statistic is 0%. This means that MLR.5 does not hold in this model and we can expect heteroscedasticity.

The reason for heteroskedasticity is that the variance of the error term is increasing along with our independent variables, which are categories for the parent's education level. An example of factors that can contribute to increasing variance of the error term is if some parents motivate their children not to take higher education, because they need their children to work at their family company or that the parents need their children to contribute financially to meet the family's needs. Another example is that some parents encourage their children to take higher education, since they do not have it themselves, and are therefore experiences the disadvantages of it. These are all reason for increasing variance of the error term along with the education-level categories.

8.4 Is the OLS estimator unbiased?

In our model we are looking isolated on how the parent's education level affects their child's education level. In practice we know that there are many other factors that may affect a parent's level of education. For example, the family's economy, personal experiences etc. It is therefore possible that the model suffers from endogeneity. This means that our explanatory variable is correlated with the error term. This violates MLR.4-Zero conditional mean, which says that there is not a relationship between the error term u and our explanatory variables xi. This is called omitted variable bias and occurs when a statistical model leaves out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to those that were included.

As we saw above, we know that our model suffers from heteroscedasticity. Heteroskedasticity is a problem because our OLS regression assumes that all residuals drawn from the population has a constant variance. Although our OLS estimator remains unbiased under heteroskedasticity, our estimated standard errors are wrong. A consequence of this is that our hypotheses test cannot be relied on, and the OLS estimator is no longer BLUE. When that is

said, heteroskedasticity does not cause bias in our coefficient estimates, but it does make them less precise. Lower precision in our coefficients increases the likelihood that the coefficient estimates are further away from the population values.

We know that if the OLS assumption MLR.1-MLR.4 holds, we can say that the OLS estimator is unbiased and BLUE. But since MLR.4 is violated, we can conclude that the OLS estimator is biased and say that our OLS regression models does not deliver a good representation of the population.

8.5 Assumptions about our data

Our assumptions regarding the data are that the education levels given are completed. If we had not assumed that these were completed degrees, it would have overturned our entire model as it is based on the relationships between a parent and its child regarding the level of education.

Another reason for having this assumption is that we do not have the age of the child and assume that the degrees are finished, therefore our results will not show the degree of completion of educations or those who were almost finished with a bachelor's or master's degree at the time of the survey. When that is said, we do not have the data on the age of the children, so it will be appropriate to proceed with the assumptions we have made.

9 Limitations

There are many limitations that can be discussed about our model, but we choose to look more closely at the limitations that are more relevant to our thesis.

The first limitation is that there will be a variation in the parent's age as they did not have children at the same time, or at the same age. This is a variable we have in our dataset and are now going to look more closely at the effect of.
The model below consists of two models. Model 1 is our MLR model from earlier, where we have divided parent's education level into five categories. Model 2 includes the same four categories, but also controls for the parent's age which is the variable io_ioalder.

Overall, we can see that by including the control variable parents age, it has little, to almost no effect on our regression. The coefficient has either had very small increases or very small decreases. The R-squared is the same, this implies that there is not more of the variation in the model that gets explained by the regression on the independent variables by including the control variable io_ioalder.

	Model 1	Model 2
hschool	0.359	0.347
	(0.013)**	(0.013)**
hpschool	0.498	0.507
	(0.013)**	(0.013)**
universityl	0.939	0.939
	(0.013)**	(0.013)**
universityh	1.120	1.118
	(0.018)**	(0.018)**
io_ioalder		0.004
		(0.000)**
_cons	2.894	2.628
	(0.009)**	(0.031)**
R^2	0.10	0.10
N	66,323	66,323

**Standard errors in parantheses.

Table 9.1: Including parents age as a control variable.

A limitation is that we do not have the child's age. So, for example if the child is only 15 years old, it makes sense that he or she has not yet attended university. It would have been better to have an age limit, for example age>25 for ch281. If we had an age limit for the children, we could have controlled for that children who are younger than 25 did not attend our dataset, because this gives us inaccurate results. However, we do not have the data on the child's age, so we cannot control for this.

Another limitation is that we do not have the data on the parent's income. This could have been a relevant control variable because it tells us whether the child is expected to contribute financially at home or not. Further on this could affect whether the child could take higher education. Because there are not many students who manage to do a full-time study while contributing financially at home.

The last limitation we are going to discuss is that our model does not consider how an individual's place of residence effects their education level. It is often the case that if you come from a big city, the road to university/ higher education seems a lot shorter than if you came from the countryside. In the countryside it is more normal to take vocational subjects, or other subjects that have local jobs. Such as fishing in Lofoten. So, a limitation in our model is that we do not have data on the individual's place of residence, and we can therefore not control for the effect place of residence has on the parent and child's education level.

10 Discussion

We see that there are several similarities between our findings and previous findings. The common denominator is that there is clearly a correlation between parent's and children's level of education. In the Indicator report they look at parent's and children's level of education in categories, but the difference is that the children are still in education. (Forskningsrådet, 2021) While we have made a strong assumption that the education has been completed, as we discuss under "Assumptions about our data". The indicator report emphasizes that there is a positive correlation between the level of education of children and parents. While in our estimation results, we see that when we divide the child's level of education into categories and have children with master/PhD as the dependent variable, there is a negative correlation between the child's and parent's level of education, until you reach university and college degree.

This indicates that parents with an education level lower than university and college will have a negative impact on the child's education level when we divide the child's education level into categories. We also see that this applies in the Indicator report in a way that if a parent has a primary school level, then only 19% of their children will take higher education. While a parent with at least a bachelor's degree will have between 40-60% of their children taking higher education.

On the other hand, when we do not divide the child's education level into categories, we get a positive correlation between the parents and the child's level of education. The reason for this is because when we divide the child's level of education into categories where children with a master/PhD are the dependent variable, the reference group is parents with a master/PhD, as we see in Table 8.2. When we do not divide the child's education into categories like in model 2 in table 7.1, the reference group is parents with elementary school/no education. The reason

for different strength of correlations between these two regression models is because we have two different reference groups, and the estimation results are interpreted relative to the reference group.

Another interesting observation from the indicator report was that there are three times as many students taking higher education today compared to 1980, but only 50% of student's parents have higher education. This suggest that there are many other factors that influence and have a decisive effect on a student's education. This has similarity to our regression models since our models suffer from omitted variable bias. As earlier said, the reason for this is that there may be potentially relevant variables that have a decisive effect on our independent variable ed_ioedu. This is the reason for the violation of MLR.4. When that is said, we include control variables such as the parent's age and we observe that this has little, if any effect on our regression model. We also know that there exist other relevant control variables that we could have included to deal with omitted variable bias. Such as income, and place of residence. Although this is something we have variables for in our dataset, not all parents have provided data on these control variables. The consequence of including these control variables would have been that the numbers of observations had dropped drastically. We have rather prioritized a bigger sample size, to get a more precise and realistic representation of the population.

From the OECD report "Investing in youth in Norway" there is a lot of focus on how the lack of higher education excludes individuals from the labor market. (OECD, 2018) While from our estimation results, we have not considered how the level of education of the child affects her/his job opportunity in the labor market. The reason for this is because, in isolation, we wanted to look at the correlation between the parents and child's level of education without considering the consequences this leads to further in the labor market.

11 Conclusion

We have looked at previous data on the topic, done our own analyzes and done several hypothesis tests to test our results and to conclude, the results of our study show that there is a correlation between parents' level of education and their firstborn child's education.

It would be advised that further research should be done with additional several variables to find the real effect as our data led to both MLR.5 (standard error) being imprecise and that MLR.4 (ZCM) was broken which is due to the unmeasurable variables having an influence on our explanatory variable that makes it less accurate. So even though we can say with high certainty that a parent's level of education has an impact, one must include additional variables to give a more precise result.

Given the results of this study, one can look at why it is the case that parents' education has an impact on a child's choice in the future, is it due to better help from home, the environment growing up or something completely different? Is there anything schools can do to even out the difference parents' education makes?

Further studies are also recommended to look at a larger geographical area to find out if other external influences such as tuition fees, access to education and what this would entail. Although this study did not find a precise answer to the correlation, we can still say with great certainty that it has a positive effect the higher education a parent has on the child's educational level.

12 Reference list

Combs-Draughn, A. J. (2016). *Scholarworks Walden University.* Hentet fra Scholar works: https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=3632&context=dissertations

Forskningsrådet. (2021, March 10). *Indikatorrapporten.* Hentet fra Indikatorrapporten: https://www.forskningsradet.no/indikatorrapporten/indikatorrapporten-dokument/menneskelige-ressurser/utdanning/?fbclid=IwAR0DaPNE0vHItpKfhL2VOlOjMR7-Pl3G3_KzQn1wFId9-cxMO75Gl8POOGY

Huntington-Klein, N. (2021, May 10). *nickchk*. Hentet fra https://www.nickchk.com/robustness.html

Norsk senter for forskningsdata. (2021, April 12). *NorLAG.* Hentet fra NorLAG: https://norlag.nsd.no/about

Numbeo. (2021, March 15). *Numbeo.* Hentet fra Cost of living index by country 2021: https://www.numbeo.com/cost-of-living/rankings_by_country.jsp

OECD. (2018, April 5). *Organisation for Economic Co-operation and Development.* Hentet fra OECD: https://www.oecd.org/publications/investing-in-youth-norway-9789264283671-en.htm

Statistisk sentralbyrå. (2020, August 12). Tabell 08921. Oslo, Oslo, Norge.

Statistisk sentralbyrå. (2021, April 3). *SSB.* Hentet fra SSB: https://www.ssb.no/arbeid-og-lonn/artikler-og-publikasjoner/tyngre-vei-inn-pa-arbeidsmarkedet-for-unge-med-lav-

utdanning?fbclid=IwAR282FEohF63C7f3cNXOZFqdajv6TAM2y4DfiVxOFU8X65xGNU5HDL3a5o8

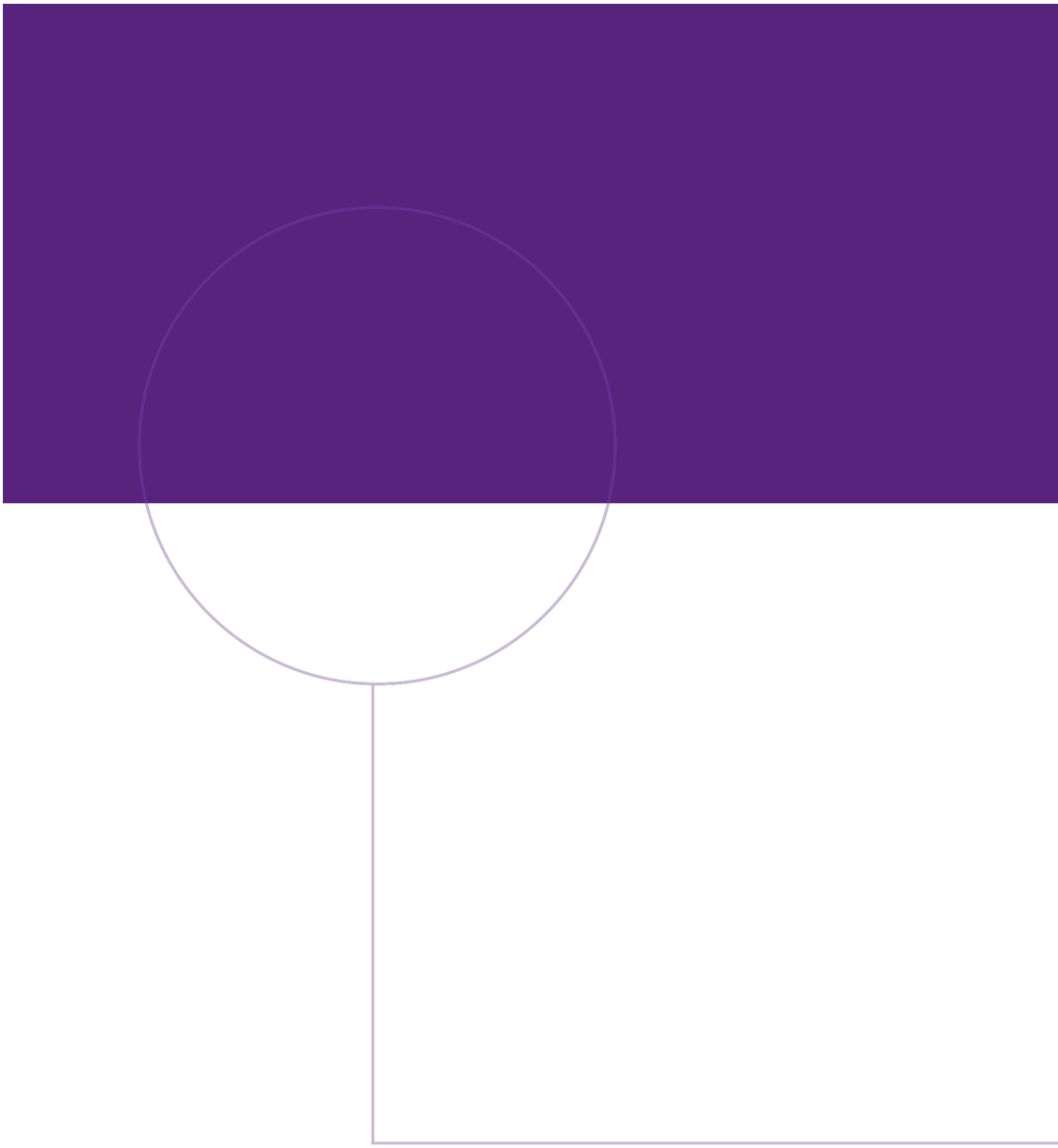
Studentum. (2021, March 20). *Studentum.* Hentet fra Universiteter og høgskoler i Norge: https://www.studentum.no/universitet-og-hoyskoler-i-norge-8136

Valenta, R. (2021, April 07). *Tutorial 5- Joint hypothesis Testing.* Hentet fra https://learn-eu-central-1-prod-fleet01-xythos.learn.cloudflare.blackboardcdn.com/5def77a38a2f7/9471909?X-Blackboard-Expiration=1620928800000&X-Blackboard-Signature=PCVBUMEsCtYHwmIK27Z%2BRAXDZR8SFy582qge%2Fzci4Ak%3D&X-Blackboard-Client-Id=303508&response-

Wooldridge, J. M. (2013). *Introductory Econometrics A modern Approch 5th Edition.* Mason : South-Western Cengage Learning.

Worldbank. (2019). *World bank.* Hentet fra Data for GDP per capita: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?most_recent_value_desc=true

World's top exports. (2021, March 15). *World's top exports.* Hentet fra Norway's top 10 exports: https://www.worldstopexports.com/norways-top-10-exports/



NTNU

Norwegian University of
Science and Technology