University of Economics, Prague

Faculty of Informatics and Statistics

Study program: Quantitative Methods in Economics

Field of study: Quantitative Economic Analysis

The Impact of Socioeconomic Status and Social Exclusion in Higher Education.

MASTER THESIS

Author: Samuel Ryckenberg

Supervisor: Ing. Kristýna Vltavská, Ph.D

Declaration: I hereby declare that I am the sole author of the thesis entitled "The Impact of Socioeconomic Status and Social Exclusion in Higher Education". I duly marked out all quotations. The used literature and sources are stated in the attached list of references.

In Prague on ....

Signature

Samuel Ryckenberg

Acknowledgement

I hereby wish to express my appreciation and gratitude to the supervisor of my thesis, Ing. Kristýna Vltavská, Ph. D., for her continued support and guidance.

Title: The Impact of Socioeconomic Status and Social Exclusion in Higher Education

Author: Samuel Ryckenberg

Study program: Quantitative Methods in Economics

Supervisor: Ing. Kristýna Vltavská, Ph.D

Abstract: The purpose of this thesis is to gain a deeper understanding of socioeconomic status impact in higher education and how social exclusion in higher education may be reduced. It aims to answer which socioeconomic variables that impact a person's decision to pursue a higher education. The thesis also discusses how social exclusion in higher education can be reduced. The thesis practical analysis is based on data from the EU-Statistics on Income and Living Conditions. Three countries are included in the analysis; the Czech Republic, Sweden and the United Kingdom. The main statistical methods used in the thesis are the independent samples t-test, the Mann Whitney U test and binary logistic regression. The results obtained by the thesis are discussed and compared to previous research conducted within the same study field.

Keywords: binary logistic regression, EU-SILC, higher education, inequality, social exclusion, socioeconomic status, tertiary education, university education

Content

1. Introduction

Fighting poverty and social exclusion as well as improving higher education has been two areas that the European union has been working on for the last decade. Both areas are main targets included in the Europe 2020 strategy which is an agenda for growth in the European union (European Commission, 2018). When working against poverty, social exclusion, and the improvement of higher education, it is of interest to understand the relationship between the two areas. Studying the impact of socioeconomic status and social exclusion in higher education may provide a better understanding on how the situation is today and what can be done to improve it in the future.

To monitor the targets of the Europe 2020 strategy, statistics on income, social inclusion and living conditions are evaluated. The main source of these statistics is the EU-Statistics on Income and Living Conditions, also known as the EU-SILC. The EU-SILC collects microdata within the areas of income, poverty, social exclusion, education, etc. The EU-SILC is a framework providing guidelines and procedures to allow a collection of harmonized primary and secondary variables (Eurostat, 2018).

Previous research on socioeconomic status impact in higher education agree on that the impact is significant. What characterizes the different studies is the type of data used and which socioeconomic variable were found significant describing the impact. Several studies found that parental education level impacts higher education participation (Bowden & Doughney, 2012; Chevalier, Harmon, O´Sullivan & Walker, 2013; McLachlan, Gilfillan & Gordon, 2013; Triventi, 2011). Studies that included the financial area of socioeconomic status found that variables connected to family income was impacting participation in higher education (Bowden & Doughney, 2012; Chevalier et al., 2013; Chowdry, Crawford, Dearden, Goodman & Vignoles, 2010). Parental occupation was another variable found significant describing the impact (Falck & Salih, 2017). Other research included less frequently studied variables like study aspiration (Falck & Salih, 2017) and sibling correlations (Björklund & Salvanes, 2010) which were also found significant describing socioeconomic status impact in higher education.

The purpose of this thesis is to gain a deeper understanding of socioeconomic status impact in higher education and how social exclusion in higher education may be reduced. It aims to investigate which socioeconomic variables that impacts a student´s decision to pursue a higher education. Based on the results obtained, the thesis intends to discuss how social exclusion in higher education can be reduced. Previous research has agreed on that the area of socioeconomic status impact in higher education is complex and that a few variables on its own cannot explain the whole impact. Therefor more studies based on new data and different variables were proposed as further research (Björlund & Salvanes, 2010; Falck & Salih, 2017; Triventi, 2011). The main contribution of this thesis lies in the further and more detailed explanation of socioeconomic status impact in higher education and what policy makers can do to reduce social exclusion in higher education. This thesis analysis is unique in the way that no studies on the topic and countries included have previously been conducted based on the EU-SILC data. The exception is a research project at the University of Economics, Prague (2017-2018), where I was part of the research team. The project focused on a student's social and financial background and how it influenced a student's decision to pursue a university education or not. The project analyzed data from the EU-SILC on households in the Czech Republic from the years of 2009-2015. Involvement in the project inspired this thesis to further and more in depth investigate socioeconomic status impact in higher education as well as how social exclusion in higher education may be reduced. This thesis practical analysis is based on the EU-SILC data from the Czech Republic, Sweden and the United Kingdom (EU-SILC Longitudinal UDB 2016 & 2012 – version of October 2017). The statistical methods used in the thesis are the independent samples t-test, the Mann Whitney U test and binary logistic regression.

The thesis purpose and aim leads down to answering how socioeconomic status impacts the decision to pursue a higher education, and how social exclusion in higher education can be reduced.

The thesis is structured as follows. The theory chapter introduces the main definitions to set a base for what the thesis is investigating. It also presents a summary on previous research conducted on socioeconomic status impact in higher education.

The method chapter contains a description of the EU-SILC data. It presents the main statistical methods used in the thesis as well as reasons why these are appropriate for this type of analysis and data.

The results chapter is devoted to the practical analysis of the three countries included in the thesis. Throughout the analysis the statistical methods and their assumptions are checked, and the results obtained from the respective statistical methods are interpreted.

In the discussion chapter the interpreted results for the three countries are discussed and compared to each other, as well as with previous research within the study area. Recommendations to policy makers are also presented based on the thesis findings.

The final chapter of the thesis is the conclusion. It summarizes the most important aspects of the thesis and presents the major findings. The chapter leads back to the thesis purpose. It answers how socioeconomic status impacts the decision to pursue a higher education and how social exclusion in higher education can be reduced.

2. Theory

The theory chapter introduces the definitions of higher education, socioeconomic status and social exclusion. Defining these key concepts sets a base for the thesis and clarifies what the thesis is aiming to investigate and answer. The chapter proceeds with presenting previous research done within the thesis topic. The last part of the theory chapter presents this thesis idea and approach to investigate the impact of socioeconomic status and social exclusion in higher education.

2.1. Higher Education

Higher education is also known as tertiary education and can be defined in different ways depending on the country, organization or dictionary. The most common definitions of higher education agree on that it is a post-secondary education. In other words, an additional education after completion of a secondary education. The differences in the definitions come down to which exact educations that are qualified as higher education and which institutions that can provide these. Most definitions state that for an education to be recognized as higher education, it should lead to some type of academic degree or professional certification.

The International Standard Classification of Education (ISCED) is a framework for classifying education activities into internationally agreed categories. The ISCED framework was developed by the United Nations International Family of Economic and Social Classifications and is applied in statistics worldwide (UNESCO, 2011). The EU-SILC variables connected to education are classified according to the ISCED framework. According to the ISCED, higher education is defined as education that build on secondary education and provides specializations within fields of education. It includes theoretical academic education as well as practical advanced vocational or professional education (UNESCO, 2011). The educations that are considered as higher education by the ISCED are "Short-cycle tertiary education", "bachelor's or equivalent level", "master's or equivalent level" and "doctoral or equivalent level". "Short-cycle tertiary education" is often more practically based and provide students with professional knowledge, skills and competencies. This type of education is classified lower than the bachelors level, but it is still classified as higher education by the ISCED (UNESCO, 2011).

This thesis defined higher education based on the ISCED framework. The main reason for this was that the EU-SILC data variables connected to education are based on the ISCED classifications. Using this definition gave an accurate view on what the data was measuring and how this was connected to the purpose of this thesis. The EU-SILC data enabled us to separate households based on the level of education that the student in the household was currently pursuing. This enabled the thesis to divide the households included in the data into two different categories. One group that included households containing a student in higher education and another of households containing a student in lower education. Therefor throughout the thesis, higher education is defined as an education leading to a short-cycle tertiary-, bachelors-, masters- or doctor's degree.

2.2. Socioeconomic Status

Socioeconomic status is an individuals or groups social and economic position in relation to others. What determines one's socioeconomic status are mainly the following three areas; occupation, financial and education. It can also be extended to include wealth, place of residence and other variables that can be related to the three main areas. Socioeconomic status is historically a common measure to use when dividing people in society into different groups. The traditional groups are high, middle and low class. The usage of the groups is common in various topics about society and politics. The term socioeconomic status is widely used by sociologists, economists and other researchers to describe the class of a certain group of people (Crossman, 2017). The class can be used to define people's opportunities, privileges and predicting behavior in society. Socioeconomic status can have either a positive or negative impact on a person's life (Crossman, 2017). The main areas within research where socioeconomic status has been studied are poverty, health, education and career aspirations (American Psychological Association, 2018). Understanding how socioeconomic status impacts our society can be useful when making improvements in the areas and to reduce social exclusion in a society.

Data on socioeconomic status is collected by different institutions and organizations. The data is useful for politicians, governments and researchers. The data can be used to determine which class in society an individual or family is part of. This information can be used in research about your specific area of interest. The main source of collecting socioeconomic data is a countries census (Crossman, 2017), but there are also other types of sources that can include information.

This thesis studied socioeconomic status impact in education. More explicitly, the impact in higher education. The EU-SILC data contains information about a household's socioeconomic status through different variables. These socioeconomic status variables were used in this thesis to characterize the households with a student in higher education and the households with a student in lower education.

2.3. Social Exclusion

Social exclusion is a lack of participation in society (United Nations, 2016). It is driven by unequal power relationships connected to economics, politics, social life and culture. Social exclusion can be spoken about on different levels. The different levels are for example individual level, group level and household level (World Health Organization, 2018). Social exclusion prevents certain groups in the society from engaging fully in community and social life. It can mean difficulties to get a job, access education and healthcare as well as social activities. The words "engaging fully" should be highlighted. Even if for example education is officially available for everyone in a society, different reasons can lead to a certain group not accessing it in the same extent as the rest of the society, and there for becoming socially excluded. Social exclusion is linked to inequalities in opportunities, privileges and rights.

As stated in the introduction, the Europe 2020 strategy targets poverty, social exclusion and the improvement of higher education. The targets are set up to strengthen and improve Europe's economy with a sustainable approach (European Commission, 2018). The targets are connected to each other in the way that better opportunities, privileges and rights also relate to the education area. In nearly all countries socioeconomic status identity has been one of the grounds that lead to social exclusion over time (United Nations, 2016). There is a higher risk that groups with lower socioeconomic status become socially excluded from society and in education.

Identifying and measuring social exclusion is challenging and it can be hard to determine which groups are socially excluded and which who are not. The measurement is complex and involves both quantitative and qualitative data since exclusion is partly a "personal experience" (United Nations, 2016). The quantitative variables in the EU-SILC data connected to socioeconomic status may give us indications if certain groups in society are excluded in higher education.

2.4. Socioeconomic Status Impact in Higher Education

In general, a lot of research has been conducted on how a student's background impacts the performance and choices in higher education. The reasons for this is that many governments, institutions and organizations regard the topic as one of their goals for improving economy, growth and equality (European Commission, 2018). The previous research that will be presented in this section is focused on studies that are specifically related to this thesis. Some of the previous studies presented are based on data from the United Kingdom and Sweden. When it comes to the Czech Republic, not a lot of previous research has been conducted on the topic.

In an article written by Chevalier et al. (2013) the relationship between early school leaving and parental education as well as income was studied. Early school leaving was defined in the article as leaving school at the age of 16 or earlier. This means that the student did not pursue an education further than secondary school. The background to why the article wanted to investigate the topic closer was because previous research found that educated and richer parents provided a "better" environment for their children. This leading to inequities in education which should attract political attention (McLachlan, Gilfillan & Gordon, 2013). The article by Chevalier et al. (2013) focused on the relative effects of education and income. It also allowed to investigate the separate effects of paternal and maternal education levels. The article pushed the importance of the distinction between parent's education and parent's income due to the different policy approaches that depend on their relative effects. The data that Chevalier et al. (2013) used was from a labor force survey conducted on households in the United Kingdom. The article pooled the data from households between the years of 1993-2012. The main statistical method used was ordinary least squares estimation. A basic model was created focusing on the parental income and education level. This model was later extended to also include parental occupation as a control variable. The results of Chevalier et al. (2013) showed that the parental education effects were significant. A household's income was also significant in the model. This indicated that students may be financially constrained when it comes to deciding about post-secondary education. The results support the challenges policymakers have when trying to encourage higher educational participation. The article concluded that policies of increasing family's permanent income would generate positive results. A focus on increasing the participation in higher education could generate a "multiplier" effects for future generations (Chevalier et al., 2013). Simplified, since parental education effects children's decisions, it will become a positive spiral for many generations to come and lead to higher participation in higher education.

Bowden and Doughney (2012) conducted a study on how secondary student's aspiration to attend university was related to student's socioeconomic status. The study used data on students that decided to apply to the university and on students who decided not to apply. The main variable examined to explain the aspiration to attend university was the parents educational level. The study also included economic and cultural variables that could be linked to socioeconomic status. These were for example if the student had internet access at home, if they attended a public or private secondary school and if they agree with the statement; "University is only for the wealthy". Bowden and Doughney (2012) used a binary logit model to analyze the aspirations to attend university. The total sample size in the study consisted of 2189 students from the western region of Melbourne, Australia. Two binary logit models were created. One focusing on the socioeconomic variable; parent's education level, while the second one also included the economic and cultural variables. The results from the first model showed that parental education level does impact a student's aspiration to attend university. The results showed that the impact was particularly large if the student came from a high socioeconomic background. The first model however was not able to predict when a student does not aspire to attend university (Bowden & Doughney, 2012). The second model, including the economic and cultural variables, was better at predicting this group of students. The results showed that all the economic and cultural variables were significant. Bowden and Doughney (2012) concluded that the level of the parent's education drives the students view on education, which affects the student's aspiration levels. The economic variable "having access to internet at home" was also of importance. The reason for this was because it may increase the number of students from low socio-economic backgrounds enrolled in higher education (Bowden & Doughney, 2012).

Triventi (2011) authored an article which main goal was to explore stratification within higher education and its relationship with social inequality. The article was comparative and based on data of tertiary graduates from 11 different European countries. The United Kingdom and the Czech Republic were included in the data. The article investigated both vertical and horizontal stratification in higher education. It included different lengths of study programs; "short" equivalent to bachelors and "long" equivalent to masters and PhD´s. The data used in the article came from the "Research into Employment and professional FLEXibility" survey. The main statistical method used was binary logistic regression. Triventi (2011) created two models, one with parental education level including basic control variables and a second which included variables related to the secondary school attended. The results from the analysis showed that parental education was strongly associated with attaining different types of qualifications in higher education (Triventi, 2011). When it came to different types of degrees which the student attended, parental education was significant in the decision between pursuing a “short” or “long” tertiary education. However, it was not significant when it came to the decision if to pursue a PhD. degree. Triventi (2011) stated that the reason behind this may be because at the PhD. transition, students are less dependent on their family origins. The study concluded that there was evidence that social inequality in higher education was linked to the percentage of higher education graduates.

In a study at Stockholm University, Falck and Salih (2017) investigated why students from high social classes tend to pursue a higher education in a larger extent compared to students from lower classes. The study focused on family and social background of the student. The aim of the study was to explain the relationship between the parent’s class and the student’s education level pursued. The data used was Swedish and was collected from a survey on living standards conducted by Stockholm University in partnership with the government agency “Statistics Sweden”. The data included information on student’s socio-economic status and education conditions. Logistic regression was the main statistical method used in the analysis by Falck and Salih (2017). Three different models, containing different independent variables were created and analyzed. The most relevant socioeconomic variables included in the analysis were parental occupation, students study aspirations and academical socialization. The groups of parental occupations that represented the social class included farmer/own business, non-qualified worker, qualified worker, low and middle white-collar worker and upper white-collar worker. A student’s study aspiration was based on if the student thought attending higher education was important or not. The academical socialization variable related to how the parents attitude was towards education and school assignments when the student was younger. The results from the study showed that the social class of the student was significant in the decision to move on to a higher education. The study aspirations of a student were also significant in the decision, while academical socialization was not significant (Falck & Salih, 2017). The study concluded that the social class does effect if the student participates in higher education. The study aspirations of a student effects participation in higher education to a certain degree. The authors believe that there are other variables that were not included in the study that also impacts the relationship. Falck and Salih (2017) highlighted that that the only variable that turned out significant in all the three models created was a student’s study aspiration. They believed the reason for this was that a student's study aspiration was the only variable directly connected to the student. The other variables were connected to the student's family and parents. Therefor the student may have difficulties to answer those type of questions correctly (Falck & Salih, 2017).

An article by Björlund and Salvanes (2010) summarized and evaluated research that had been conducted on education and family background. There were two main focuses in the summarization and evaluation. The first one was on equality of opportunity and the second was on the child development perspective. In the relation to this thesis topic the focus on the equality of opportunity is the one that is relevant to be presented here. The questions that the article discussed were; "How important a determinant of educational attainment is family background", "what are the mechanics that make family background important" and if "specific policy reforms have been successful in reducing the impact of family background on educational achievement" (Björlund & Salvanes, 2010). The article studied how family background effected educational attainment by investigating sibling correlations. The logic behind it was that siblings come from the same family background (parental education, income, etc.). Through studying how much siblings have in common, generalizations can be made on how much family background matters when it comes to educational attainment. The data that was used for the analysis was taken from other previous articles on sibling correlation estimates. Sweden and the Great Britain were included in this data. Björlund and Salvanes (2010) stated that from an inequality point of view it is common to consider family background as a set of factors that the student has had no influence on itself. The student can therefore not be held accountable for these factors and that's why it's important to highlight those when conducting research on the topic. The results from studying sibling correlations showed significant results from an inequality point of view when it came to explain a student's education attainment based on family background. With that said, parental education and income could only account for one third of what siblings had in common on family background factors. Björlund and Salvanes (2010) stated that there is a gap in research since previous research mainly focused on parental education and income variables. Therefore, the authors stated, it would be relevant to widen the research and include more variables than these two.

Chowdry et al. (2010) conducted a study to understand the determinants of participation in higher education among students from low socioeconomic backgrounds. Socioeconomic status in the study was based on if the student was eligible to free school meals. Being eligible for free school meals is related to a family's economic situation and income. The data used in the study was on education from various sources combined, creating a sort of census for English secondary students. The data used differs from other data sets previous researched because it also contained detailed information about the students previous academical results (Chowdry et al., 2010). This advantage made it possible for the authors to investigate both how socioeconomic status impacts the participations in higher education and how school results from younger ages impacts participation in higher education. To perform the analysis, logistic regression models with random effects were used. The results from the study showed that students from lower socioeconomic backgrounds were less likely to participate in higher education compared to students from higher socioeconomic backgrounds (Chowdry et al., 2010). These differences in higher education participation did not emerge at the time when a student took the decision to pursue a higher education or not. It emerged earlier during the student's school years. The reason for this was that a student's performance in secondary school had a large impact on if they would decide to move on to the university level or not. Students from lower socioeconomic backgrounds did not perform as well as other students in secondary school which led to the skewed participation in higher education (Chowdry et al., 2010). The authors stated that a socio-economic difference in participation did remain on entry to university, but it was relatively small in comparison to prior achievements in for example secondary school. The study concluded that if policy makers want to reduce the skewed participation in higher education and increase the number of students from lower socioeconomic backgrounds, then they should intervene earlier in a student's schooling years to maximize the possible impact (Chowdry et al., 2010).

The previous research that has been presented so far in the theory chapter mainly measured socioeconomic status based on objective variables representing the parental education level, income, parental occupation or other economic aspects. As a criticism to the objective measurements of social class and socioeconomic status in higher education, Rubin et al. (2014) presented a different approach. The authors argued that it is important to include a subjective approach together with the traditional objective measurements. This would mean to include variables connected to one´s self-definition of social class. The authors discussed strengths and limitations of objective measures, and the validity of subjective measures. Rubin et al. (2014) stated that the strengths of the traditional variables, for example parental income, occupation and education level, was that they limit the influence of subjective biases. One limitation found was that objective measures of social class and socioeconomic status need to be benchmarked and interpreted relatively to population standards (Rubin et al., 2014). Another limitation was that since a student has not yet established an income etc., the variables are usually connected to the student's parent's information. This may lead to more knowledge about the student's parents than about the student itself. Depending on the age of a student its parental education and parental income could describe the student's socioeconomic status accurately or not. For a young student living at home the parent's information most likely reflects the student's situation well. For an older student that has moved away from home and started its own family, the parent's information may not reflect the student's situation accurately (Rubin et al., 2014). Turning to the subjective approach of measuring social classes, a common concern is its validity. This is mainly because people do not know what category of social class that they belong to and many people would simply consider themselves as middle class (Rubin et al., 2014). On the other hand, subjective measurements may be more reliable and predictive than objective measurements. The reason for this is that the measurements will relate more to the student and not to the parents of the student, and therefor they represent a student's socioeconomic status and social class more accurately. The subjective measures also tend to be more sensitive over time and that is the reason they are better predicters in the field of higher education (Ostrove & Long, 2007). To conclude the study, Rubin et al. (2014), highlighted that they do not propose to replace the objective measurements of socioeconomic status and social class. Instead they proposed that researchers within the field of higher education should supplement the objective measures with subjective ones.

This thesis idea and approach to the topic of socioeconomic status impact in higher education is based on objective variables, in the same line as most of the previous research. The idea is to have a more detailed approach to this complex topic by including variables representing not only one part of socioeconomic status but many aspects of it as possible. The EU-SILC data provides wide information on a household's income, housing, education, material deprivation and labor, all included in one data set. The enables the thesis to include six different independent variables that together represents socioeconomic status. By doing this the thesis answers how socioeconomic status impacts the decision of pursuing a higher education. The approach will lead to a deeper understanding within the topic and provide an insight on how social exclusion may be reduced in higher education.

3. Method

The method chapter includes a data description of the EU-SILC emphasized on how it was extracted and modified for the practical purposes of the thesis. After the data description the statistical methods used in the practical analysis are presented. This together with the reasons why they were appropriate for the type of analysis conducted. The final part of the method chapter is devoted to the research design.

3.1. The European Union Statistics on Income and Living Conditions

The EU-SILC project started in 2003 by six-member states of the European union. The aim behind the project was to collect output-harmonized data on income and living conditions. The project now involves over 30 countries in Europe including all the 28 EU-member states (Eurostat, 2018). Selecting stable and constant variables from an official European Union institution like Eurostat strengthens the reliability of the thesis results which is connected to if the results found would be confirmed if the analysis was replicated (Bryman & Bell, 2011). The EU-SILC instrument provides two dimensions of microdata on income, poverty, social exclusion and living conditions. The first type is cross-sectional and is linked to a given time. The second type is longitudinal and includes individual changes over time (Eurostat, 2018). Micro data of the EU-SILC type is only granted access for scientific purposes to organizations recognized as research entities (European Commission, 2017). This thesis was granted access to the data through representing the University of Economics, Prague.

The data used in this thesis was extracted from the longitudinal EU-SILC data files. The reason for this was because the cross-sectional data does not track the household id over time. Since a large sample size was required for the analysis, more than one year of data was combined. Using the cross-sectional data may have led to problems concerning one household representing more than one observation in the data. By using the longitudinal data, tracking the household ID (Mack, 2016), it enabled us to make sure that a specific household did not represent several observations. Note that even though the data was extracted from the longitudinal EU-SILC data, the thesis has a cross sectional research design. The data used in the thesis is from 2016 and 2012. 2016 years data was the latest available EU-SILC longitudinal data. Year 2012 was selected because the household ID in the EU-SILC is tracked during four years of time (Mack, 2016). Using 2016 and 2012 ensured that a household was not included more than once in the data set. When selecting how many years of data to include there was a tradeoff between sample size and the relevance of older data. Using the years of 2016 and 2012 gave a balance where the results would be up to date and the sample size would be large enough to perform a relevant analysis. To insure disclosure control and confidentiality the EU-SILC data educational variables were top/bottom coded (European Commission, 2016). This led to the thesis analysis not being able to investigate differences between the different study levels included in higher education (e.g. differences between bachelor's level and master´s level).

The structure of the EU-SILC data divides the data into files containing household information and personal information. Every household has a household ID and every person has a person ID connected to the household ID (Mack, 2016). This thesis based its analysis on comparing two categories of different households. Households containing a student currently participating in higher education, and households containing a student currently participating in an education lower than higher education. This was done to investigate the differences between a population that has decided to pursue a higher education and a population that has not yet had to make the decision. In practice to be able to create these two categories for the EU-SILC data, the household ID had to be matched with the person ID. A filter was created were only households containing persons that were currently in education were kept in the data set. To divide the categories based on the current level of education of the student, a new variable was created. The new variable gave information on what level the student in the household currently studied at, which divided the households into the two categories. Note that in some cases a household contained more than one person currently in education (either higher or lower). In those cases, the variable created only contained the information of the student that studies on the highest level in the household. The reason why this had to be performed was because a specific household may not be included in both categories. This would have broken the assumptions for logistic regression regarding independence of observation as well as mutually exclusive and exhaustive categories. (Ri. Burns & Ro. Burns, 2008). Another filter added, was that the person currently in education had to be 27 years old or younger. The reason for this was to focus the analysis on younger students that in general are more dependent on the household they are in, compared to older students who may already have worked for several years and may not be very connected to the household (Rubin et al. 2014).

The last step of the data extraction was to combine the information from the household data with the information of the head of the household from the personal file. The head of the household was considered by this thesis as the person in charge of responding to the household questionnaire. We believed that the head of the household was the person who represented the household in the most accurate way. A data file containing both information about the household and personal information about the head of the household was created. A filter was applied to remove the observations where the head of the household was the same person as the student representing the household. The reason for this was to avoid issues for the dependent and independents variables relationship. After completing the procedures above the data was imported to the computer software SPSS where the analysis was conducted.

Studies on this thesis topic and countries have previously not been done using the EU-SILC data, excluding the project at the University of Economics, Prague. The EU-SILC data contains large amounts of information on socioeconomic status, social exclusion and higher education. The usage of the EU-SILC data together with the methodology of this thesis therefor led to a unique analysis of the impact of socioeconomic status and social exclusion in higher education.

3.2. Variables

The dependent variable in the thesis was “study level” which referred to if the student in the household currently participated in higher education or lower education. The independent variables included in the thesis were the following; “household gross income”, “social exclusion benefits”, “head of household education”, “material deprivation”, “financial burden of housing” and “occupation”. The variables were selected to represent socioeconomic status in the most accurate way possible. This to make sure that the thesis variables measure what they intent to do, leading to trustworthy and valid results (Bryman & Bell, 2011). A detailed description of the variables and how they were modified to fit the statistical methods used can be found in the results and discussion chapter.

3.3. Statistical Methods

3.3.1. Parametric and Non-Parametric Tests

Parametric statistical tests rely on assumptions connected to the data’s shape of distribution and the parameters of that assumed distribution. The most typical parametric assumption is that the data is approximately normally distributed (Hoskins, n.d.). Nonparametric statistical tests rely on fewer assumptions and less strict ones. Since the nonparametric tests has fewer assumptions,

they are convenient for researchers to use. On the other hand, the parametric tests are often more powerful and easier to interpret (Hoskin, n.d.). The question is when it is appropriate to use parametric tests respectively nonparametric tests.
There are different advantages of parametric tests and they can provide trustworthy results even if the distribution of data is skewed and non-normal if the sample size is large enough. One advantage of parametric tests is that the two groups analyzed do not have to have the same dispersion. Another advantage is that parametric tests have greater statistical power (Frost, 2017). The nonparametric tests also have advantages. They assess properties of the median instead of the mean, which can be an advantage depending on the data set and the aim of the analysis. Another advantage is that nonparametric tests work well even if the sample size is small and the data is potentially non-normal. Frost (2017) concludes that the main aspects to consider when deciding between parametric and nonparametric tests is the sample size and if the median or mean is the most appropriate measurement. Fagerland (2012) states that nonparametric tests have become more common at the same time as the average median sample sizes of research studies have gone up. This creates a paradox since nonparametric tests in general should be used when small sample sizes occur. Fagerland (2012) conducted a simulation study to compare rejection rates (statistical power) for the nonparametric Wilcoxon-Mann-Whitney test and the parametric two-sample t-test. The conclusion from the simulation study was that nonparametric tests are useful for studies with small sample sizes while the parametric tests should be used for studies with large sample sizes. The use of parametric tests in the case of large sample sizes apply even if the data is heavily skewed. The reason for this was that researchers using nonparametric tests in studies with large sample sizes may be answering the wrong research questions (Fagerland, 2012).

This thesis conducted parametric tests for two of the independent variables and non-parametric tests for three of the independent variables. The aim was to use the parametric tests on as many of the variables as possible because our sample size is considered in large in this context. The reason why only two of the variables were able to be analyzed with the parametric tests came down to assumptions connected to how the variables were measured. The two-sample t-test, known as the independent samples t-test in SPSS, was the parametric test that was performed in this thesis analysis. The test variable for the independent samples t-test should be measured on a continuous scale while the grouping variable should contain two categories. Only two of our independent variables were measured on a continuous scale and therefor appropriate for the independent samples t-test. There were also other assumptions to be met. According to Laerd

Statistics (2018) the data used for the independent samples t-test should not contain any significant outlier since this could reduce the validity of the results. There also needs to be homogeneity of variances which is tested with the Levene’s test (Laerd Statistics, 2018). The independent samples t-test is used when comparing the means of two independent groups to determine if they are significantly different from one another (Field, 2009). In this thesis we were comparing the means of the two independent categorical groups; households containing a student currently participating in higher education, and households containing a student currently in an education lower than higher education. The two groups were independent since a specific household could not belong to both the categorical groups. The two-sample t-test can be performed even if the categories do not contain an equal number of observations (Field, 2009). The equation for the two-sample test statistics is presented as follows:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{estimate\ of\ the\ standard\ error} \tag{1}$$

The t-statistic for the two-sample t-test (1) is obtained by calculating the difference between the overall sample means of the two samples and comparing (dividing) them with the estimate of the standard error. The estimate of the standard error would be the differences we would expect between the means of the two populations from which the samples come from. (Field, 2009).

The non-parametric test used in this thesis was the Mann Whitney U test which is an equivalence to the parametric independent samples t-test (Frost, 2017; Field, 2009). Three of the variables were analyzed using the Mann Whitney U test. These three variables were of ordinal measurement and not measured on a continuous scale. Therefor they were not appropriate for the independent samples t-test. For the Mann Whitney U test the test variables can be measured on either ordinal or continuous level. The grouping variable should have two independent categorical groups, just as for the independent samples t-test. There are also assumptions regarding independence of observations and distribution shapes of the two groups (Laerd Statistics, 2018). The Mann Whitney U test is used to test differences between two conditions and is based on the principle of ranking the data to test if the differences are significant (Field, 2009). The Mann Whitney U test calculates the sum of ranks and mean ranks for the two categorical groups. Then a U-statistic is calculated based on the sum of ranks to determine the significance. The U-statistic is calculated using the following equation:

$$U = n_1 n_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \quad (2)$$

The U-statistic in equation (2) is calculated based on the sample sizes of the two categorical groups and the sum of ranks (Field, 2009). Equation (2) shows how it is calculated for the first group, this should also be done in the same way for the second group. In practice when using SPSS, it does not matter if the U-statistic is calculated for the first or second group since they have a direct relationship (Field, 2009).

3.3.2 Logistic Regression

The main statistical method used in the thesis practical part was binary logistic regression analysis. Logistic regression can be used when the dependent variable is categorical, and the independent variables are either continuous, ordinal or categorical (Field, 2009). When it comes to studies in socio-economics, it is common that the variables are often categorical rather than of interval scale (Tranmer & Elliot, 2008). In this thesis we had a categorical dependent variable and both continuous, ordinal and categorical independent variables. Therefore, logistic regression was an appropriate method to use in the analysis. Logistic regression can provide information on which category a person is more likely to belong to given other information. In this thesis we are investigating a dichotomous dependent variable. A dichotomous variable is a variable that can only take two possible values (Field, 2009). Since there are only two possible values these can be assigned as one and zero which leads to a binary state. In our case the two possible values for the dependent variable were; household with student currently participating in higher education (one) and household with student currently participating in education in lower education (zero). This was the reason why binary logistic regression was performed in the analysis and not for example multinomial logistic regression.

To explain logistic regression further in detail and how it differs from normal linear regression a few equations will be presented. In multiple linear regression, the dependent variable is explained by several independent variables. The dependent variable Y is predicted from a combination of each independent variable multiplied by its respective coefficient (Field, 2009). The equations for multiple linear regression and logistic regression are expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \ldots + \beta_n X_{ni} + \varepsilon_i \tag{3}$$

$$P(event\ Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots + \beta_n X_{ni})}} \tag{4}$$

In the multiple linear regression (3), we predict the value of the dependent variable Y. In logistic regression (4) we instead predict the probability of the dependent variable Y (Field, 2009). In the practical analysis predicting the probability of the dependent variable Y was extended to calculating odds ratios. The main point to highlight is that logistic regression is about probability and not a simple value. Note that both the multiple linear regression equation (3) and the logistic regression (4) include several independent variables explaining the dependent variable. The reason why this thesis uses logistic regression and not regular multiple regression comes back to the fact that we have a categorical dependent variable. One of the assumptions of linear regression is that the relationship between the dependent and independent variables needs to be linear. If we had a categorical dependent variable this assumption would not hold and using multiple linear regression would not be appropriate (Berry, 1993). The logistic regression equation instead uses a logarithmic transformation. It expresses the multiple linear regression equation (3) in logarithmic terms which results in equation (4). This is called the logit and overcomes the problem of breaking the assumption of linearity (Field, 2009). The assumption for the binary logistic regression instead states that there must be a linear relationship between the independent variables and the logit transformation of the dependent variable (Laerd, 2018). In addition to the assumption of linearity, multicollinearity should also be checked for the variables since it can affect the parameters of a regression model in a negative way (Field, 2009).

The outcome predicted in the logistic regression equation (4) is the probability of the dependent variable Y. When we performed logistic regression in our practical analysis the probabilities were used to calculate odds ratios. The odds ratios are the most important when interpreting the results in logistic regression. In the practical part of this thesis, SPSS denotes the odds ratio as Exp(B). The Exp(B) is the change in odds resulting from one-unit change in the predictor (dependent variable) (Field, 2009). The following equations explain the relationship between the logistic regression equation, the odds and the Exp(B):

$$\text{odds} = \frac{P(event\ Y)}{P(no\ event\ Y)} = \frac{\left(\frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}\right)}{1-P(event\ Y)} \tag{5}$$

$$\text{Exp(B)} = \Delta\text{odds} = \frac{odds\ if\ we\ have\ a\ unit\ change\ in\ the\ predictor}{odds} \tag{6}$$

As shown above, the probability of the variable Y is used to calculate the odds (5). The odds are in the next step used to calculate the Exp(B) in equation (6). The Exp(B) is interpreted in the terms of change in odds. A value higher than one, indicates that as the predictor (independent variable) increases, the odds of the outcome Y increases. A value lower than one indicates that as the predictor (independent variable) increases the odds of the outcome Y decreases (Field, 2009). This lead to speaking in terms of “more likely” and “less likely” when interpreting the results in logistic regression (Tranmer & Elliot, 2008).

3.4. Research Design

3.4.1. Quantitative Approach

A quantitative approach was selected for this thesis. Quantitative research has a focus on measurements and explaining relationships through logic and facts (Bryman & Bell, 2011). The selection of a quantitative approach was based on the thesis purpose which was to gain a deeper understanding of socioeconomic status impact in higher education, not how socioeconomic status of specific individuals impact higher education. To be able to fulfill the purpose a large amount of observations was needed to be analyzed. Previous research trying to answer similar questions have also mainly used quantitative research designs. In a few cases a mixed approach has been taken where the analysis was quantitative but included subjective variables. Therefor selecting a quantitative approach in this thesis also enabled an easier and more relevant comparison of the thesis results with findings of previous researchers. The EU-SILC data set was well suited for a quantitative approach since the data is survey based and contains a large amount of observations that could be imported and analyzed in a statistical software.

3.4.2. Cross Sectional Design

The thesis has a cross sectional design and was based on data from three European countries. Cross sectional studies rely on existing differences, they do not have a time dimension and groups are selected based on existing differences and not random allocation (University of Southern California, 2018). The reason why a cross sectional design was chosen was that the thesis aimed to measure current differences between two distinct groups. Cross sectional studies provide a clear picture of the outcome and the characteristics of the results (University of Southern California, 2018). This suits the purpose of the thesis and what the thesis aims to answer. As stated in the data description, the EU-SILC data was extracted from the longitudinal data files because of methodological reasons. Even though it was extracted from the longitudinal files the research design was cross sectional. Observations were not included more than once in the data set and changes over time were not analyzed in this thesis.

The countries included in the thesis were the Czech Republic, Sweden and the United Kingdom. The main reason why these countries were selected was that we wanted to include three countries that were from different parts of Europe to be able explore similarities and differences. The countries selected differ in culture, behavior and living conditions. Selecting countries that differ in these aspects enabled an interesting discussion of the results. There were also other advantages linked to selecting these three specific countries. The Czech Republic was the country analyzed by the project that inspired this thesis, therefor we already had some previous knowledge regarding the Czech EU-SILC data. Selecting Sweden gave an advantage on finding previous research done on Scandinavian studies, that were not necessarily written in English, due to my Swedish origin. Selecting the United Kingdom came with the advantage of being able to compare this thesis results with a wide range of previous research already conducted on data from the United Kingdom.

4. Results

The results chapter introduces the thesis data sample and the variables included in the thesis. Then the three countries; the Czech Republic, Sweden and the United Kingdom are analyzed using the independent samples t-test, the Mann Whitney U test and binary logistic regression. During the analysis the different statistical methods and it´s assumptions are discussed. The results chapter also includes interpretations of the results obtained from the different statistical methods. The aim of the results chapter is to present results that can answer how socioeconomic status impacts the decision to pursue a higher education, and how social exclusion in higher education can be reduced.

4.1. Sample and Variables

The sample of the EU-SILC data was divided on the three countries included in the thesis according to the frequency tables presented below. The sample sizes presented are the ones after the filters during the data extraction process were applied (see method chapter). There for they are the actual sample sizes used for the practical analysis in this chapter. Table 1 shows the different sample sizes based on each country. What could be observed was that the sample sizes for the Czech Republic and the United Kingdom were larger than for Sweden. Many Swedish observations were excluded when the filter regarding that the head of the household should not be the same person as the student currently in education was applied. The reason behind this may be that many young people in Sweden move away from their family's households early in life to live on their own. Eurostat (2014) published data on the share of young people (age 16-29) living with their parents. The results showed that approximately 35 % of young people in Sweden live with their parents. This can be compared to the United Kingdom where the same number was approximately 50 % and for the Czech Republic where the number was around 70 %. This socioeconomic phenomenon resulted in many students from Sweden becoming both the head of the household and the student of the household in our data. Even though a lot of observations were lost because of this filter, especially from Sweden, it was necessary to apply it, to correctly describe the independent variables impact on the dependent variable.

Table 1. Sample sizes of the countries

	Sample size	%
Czech Republic	1816	39,7
Sweden	950	20,8
United Kingdom	1806	39,5
Total	4572	100,0

Source: Own calculation, data EU-SILC

Table 2. Sample sizes divided between the categories of the dependent variable

	Lower Education	Higher Education	Total
Czech Republic	56.67%	43.33%	100%
Sweden	79.68%	20.32%	100%
United Kingdom	64.78%	35,22%	100%

Source: Own calculation, data EU-SILC

Table 2 shows how the sample sizes expressed in percentage were divided between the two categories representing the binary dependent variable for each country. The dependent variable of the thesis was the current study level of the student in the household. The two categories, lower education and higher education, were based on if the student in the household was currently participating in higher education or currently participating in lower education. The number of households containing a student in lower education was consequently larger than the number of households containing a student in lower education for all three countries. Figure 1 clearly illustrates this. From the figure we can especially observe the low number of Swedish households containing a student in higher education.

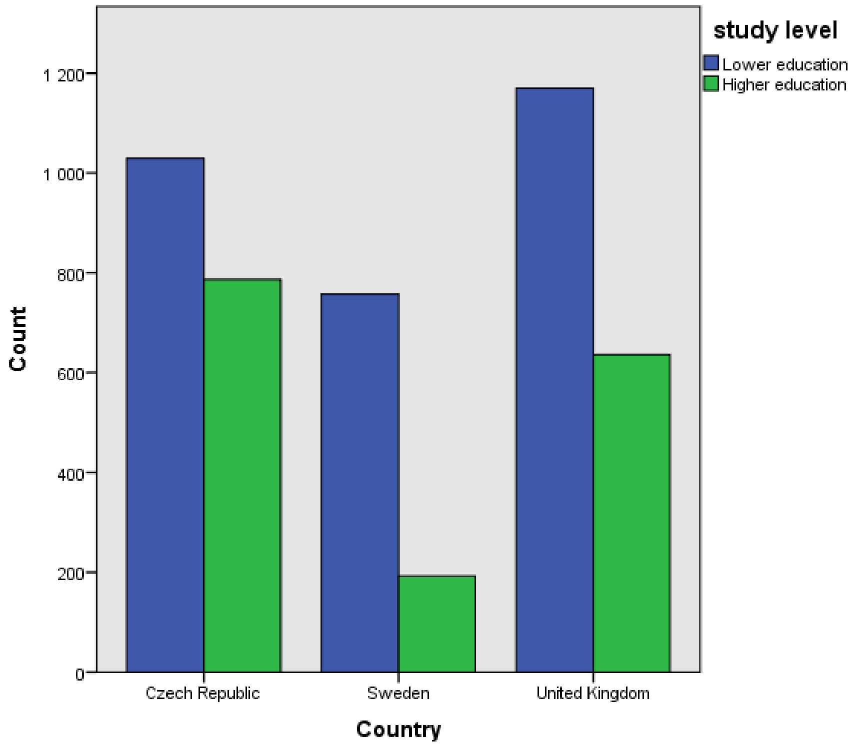


Figure 1. Sample sizes of the countries and the categories of the dependent variable

Source: Own calculation, data EU-SILC

When performing tests and the binary logistic regression analysis, assumptions connected to the data sample had to be checked. This was to ensure that the data was appropriate for the statistical methods used. The main assumptions connected to our statistical methods were mentioned in the method chapter. When the tests and binary logistic regression are conducted in this chapter the assumptions are also discussed.

The dependent variable in the thesis was "study level" which referred to if the student in the household currently was enrolled in higher education or lower education. The IECD education levels that represented the higher education group were short-cycle tertiary, bachelors, master and doctorate. The IECD education levels representing the lower education group were post-secondary non-tertiary, upper and lower secondary, and primary. The independent variables that were used to explain the dependent variable were the following; "household gross income", "social exclusion benefits", "head of household education", "material deprivation", "financial burden of housing" and "occupation".

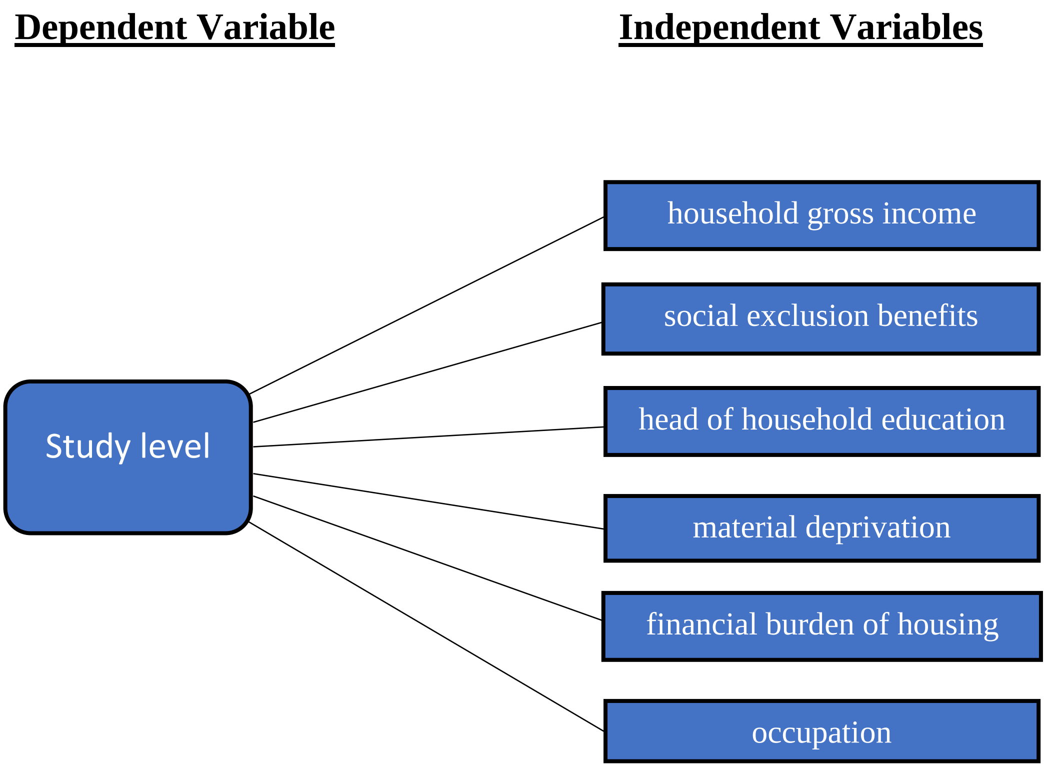


Figure 2. Variables included

Source: Own calculation, data EU-SILC

The variable "household gross income" was the total household gross income over one income reference period in national currency (Eurostat, 2017). The "social exclusion benefits" variable was the total income from support and benefits related to social exclusion that a household received during one income reference period in national currency. Social exclusion benefits are in general payed out to destitute people, migrants, refugees, drug addicts, alcoholics, victims of criminal violence (Eurostat, 2017). Because of the different currencies and income reference periods we did not focus on the actual numeric values of these two variables in this thesis, only their relationship with the dependent variable. The variable "head of the household education" was based on the five IECD levels of education and represents the highest IECD level attained by the head of the household (Eurostat, 2017). The levels included were higher education (as defined for "study level"), secondary non-tertiary, upper and lower secondary, and primary. The reason why the specific higher education levels were not included in the EU-SILC data was to insure disclosure control and confidentiality (European Commission, 2016). The variable "Material deprivation" was based on the standard nine material deprivation items specified by the European Union Social Protection Committee. It indicated how many of the 9 items a household could not afford. The variable "financial burden of housing" represented how heavy of a financial burden the total housing cost was for a household. There were three different level; heavy burden, slight burden and not a burden at all (Eurostat, 2017). The "occupation" variable was based on the International Standard Classification of Occupations (ISCO-08) coding and defines which type of occupation the head of the household currently has or had in the past (Eurostat, 2017).

4.2. Independent Samples T-test

The first statistical method performed in the practical analysis was the independent samples t-test. It is used to compare the mean of two independent groups to determine if they were significantly different from each other (Field, 2009). It is a parametric test and was performed on the variables "household gross income" and "social exclusion benefits". The reason why only these two variables were suitable for the test was because of the assumptions connected to the independent samples t-test. The test variable should be met on a continuous scale while the grouping variable should consist of two categorical groups (Field, 2009). The grouping variable in our case was "study level" which was divided into the two categories of higher education and lower education. The assumption of independence of observations (Field, 2009) was also satisfied for this thesis data, it meant that one household was not part of both groups. The first country analyzed using the independent samples t-test was the Czech Republic. Before performing the test, we checked if there were any outliers in our data since the independent sample t-tests mean may be sensitive to outliers (Field, 2009). Box plots were made in SPSS to detect outliers in the data (See Appendix B). For the variable "social exclusion benefits" no severe outliers were detected for the data of the Czech Republic. The variable "household gross income" had three severe cases of outliers that were detected for the Czech data (See Appendix B). For this thesis, regular outliers were included since they are a part of the data and are not errors. However, the few cases of severe outliers detected were most likely due to errors or very wealthy households that do not represent the general population in an accurate way. In this thesis they were considered as a disruption of the analysis. These three severe outliers for the data of the Czech Republic were omitted before conducting the independent samples t-test on the "household gross income" variable. It should also be mentioned that the data for the two variables "social exclusion benefits" and "household gross income" was not necessarily normally distributed. According to Frost (2017) the assumption of normally distributed data can however be overcome if the sample size is larger than 15 observations per group for the independent samples t-test. Due to the large sample size of our data in this context the violation of non-normality should not affect the results (Frost, 2017; Fagerland 2012). After the assumptions above were checked and corrected for, the independent sampled t-test was performed on the variables “social exclusion benefits” and “household gross income” with “study level” as the grouping variable.

Table 3. Group Statistics, “household gross income” – Czech Republic

	study level	N	Mean	Std. Dev.	Std. Error Mean
household gross income	Lower education	1026	20989,37	11215,536	350,144
	Higher education	787	25376,28	12504,836	445,749

Source: Own calculation, data EU-SILC

Table 4. Independent Samples Test, “household gross income” – Czech Republic

	Levene's Test		t-test for Equality of Means				
	F	Sig.	T	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
household gross income	7,623	,006	-7,851	1811	,000	-4386,912	558,781
			-7,739	1590,79	,000	-4386,912	566,827

Source: Own calculation, data EU-SILC

When interpreting the results from the “household gross income” variable, we could observe from the group statistics that the mean for the higher education group was higher than the mean for the lower education group. This indicated that gross income for a household containing a student in higher education was higher than for a household containing a student in lower education. Moving on to the independent samples test table, the Levene´s test showed a significant result. This meant that we could not assume equal variances. Therefor when interpreting the t-test for the equality of means we interpreted the last row in the table. The results from the independent samples t-test showed that the means significantly differed from each other. Households containing a student in higher education had a significantly higher household gross income compared to households with a student in lower education. The p-value was 0.000 which gave a significance level of 0.1 % for the results.

Table 5. Group Statistics, “social exclusion benefits” – Czech Republic

	study level	N	Mean	Std. Dev.	Std. Error Mean
social exclusion benefits	Lower education	1029	58,69	426,842	13,306
	Higher education	787	13,32	206,042	7,345

Source: Own calculation, data EU-SILC

Table 6. Independent Samples Test, “social exclusion benefits” – Czech Republic

	Levene's Test		t-test for Equality of Means				
	F	Sig.	t	Df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
social exclusion benefits	29,786	,000	2,747	1814	,006	45,372	16,516
			2,985	1560,37	,003	45,372	15,199

Source: Own calculation, data EU-SILC

The group statistics for the “social exclusion benefits” variable showed that the mean for the households containing a student in lower education was higher than the mean for households containing a student in higher education. The Levene´s test gave significant results which meant that we could not assume equal variances. The t-test for equality of means showed that the means were significantly different from each other. Households containing a student in lower education received higher amounts of social exclusion benefits compared to households containing a student in lower education. The p-value was 0.003 which gave a significance level of 1% for the results.

The independent samples t-test was performed on Sweden’s data in the same way as for the Czech Republic. The assumptions were gone through and the tests were conducted on the same test variables; “household gross income” and “social exclusion benefits” with “study level” as the grouping variable. For the Swedish data four severe outliers concerning the variable “household gross income” were detected (See Appendix B) and omitted. No severe outliers were detected for the “social exclusion benefits” variable.

Table 7. Group Statistics, “household gross income” – Sweden

	study level	N	Mean	Std. Dev.	Std. Error Mean
household gross income	Lower education	753	87849,07	44750,303	1630,792
	Higher education	193	95503,44	48288,982	3475,917

Source: Own calculation, data EU-SILC

Table 8. Independent Samples Test, “household gross income” – Sweden

	Levene's Test		t-test for Equality of Means				
	F	Sig.	T	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
household gross income	3,815	,051	-2,085	944	,037	-7654,371	3670,352
			-1,994	282,336	,047	-7654,371	3839,464

Source: Own calculation, data EU-SILC

When interpreting the group statistics and the independent samples t-test for the “household gross income” variable we could observe that the mean for the households containing a student in higher education was higher than the mean for households containing a student in lower education. The Levene’s test showed that we could assume equal variances. The results from the t-test for equality of means showed that the household gross income was significantly higher for households with a student in higher education compared to households with a student in lower education. The p-value was 0.037 which gave a significance level of 5% for the results.

Table 9. Group Statistics, “social exclusion benefits” – Sweden

	study level	N	Mean	Std. Dev.	Std. Error Mean
social exclusion benefits	Lower education	757	314,73	2045,850	74,358
	Higher education	193	108,42	805,314	57,968

Source: Own calculation, data EU-SILC

Table 10. Independent Samples Test, "social exclusion benefits" – Sweden

	Levene's Test		t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
social exclusion benefits	7,484	,006	1,374	948	,170	206,306	150,192
			2,188	796,201	,029	206,306	94,283

Source: Own calculation, data EU-SILC

The group statistics for the "social exclusion benefits" variable showed that the mean was higher for households with a student in lower education than for households with a student in higher education. According to the Levene's test we could not assume equal variances. The results from the t-test for equality of means showed that households containing a student in lower education received a significantly higher amount of social exclusion benefits compared to households containing a student in higher education. The p-value was 0.029 which gave a significance level of 5% for the results.

Following the same methodology as for the previous two countries the independent samples t-test was conducted on the data for the United Kingdom. The test variables were "household gross income" and "social exclusion benefits" and the grouping variable was "study level". When checking the assumptions on the data for the United Kingdom, an issue was found related to the "household gross income" variable. The issue was that two of the observations had negative values. This was clearly an error in the data since income may not be negative in the EU-SILC data. These two observations were omitted before performing the independent samples t-test on the "household gross income" variable. In addition to this, seven severe outliers were detected (See Appendix B) and omitted. For the "social exclusion benefits" variable no errors or severe outliers were found.

Table 11. Group Statistics, "household gross income" – United Kingdom

	study level	N	Mean	Std. Dev.	Std. Error Mean
household gross income	Lower education	1164	59867,57	46717,467	1369,313
	Higher education	633	66343,52	46367,835	1842,956

Source: Own calculation, data EU-SILC

Table 12. Independent Samples Test, “household gross income” – United Kingdom

	Levene's Test		t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
household gross income	3,760	,053	-2,814	1795	,005	-6475,954	2301,081
			-2,821	1306,09	,005	-6475,954	2295,976

Source: Own calculation, data EU-SILC

Interpreting the group statistics for the variable “household gross income” we could observe that the mean for the households with higher education students was higher than the mean for the households with lower education students. The results from the Levene´s test showed that we could assume equal variances. The results from the t-test for equality of means stated that the household gross income was significantly higher for households containing a student in higher education compared to households containing a student in lower education. The p-value was 0.005 which gave a significance level of 1% for the results.

Table 13. Group Statistics, “social exclusion benefits” – United Kingdom

	study level	N	Mean	Std. Dev.	Std. Error Mean
social exclusion benefits	Lower education	1170	901,45	2411,016	70,487
	Higher education	636	663,36	2006,466	79,562

Source: Own calculation, data EU-SIL

Table 14. Independent Samples Test, “social exclusion benefits” – United Kingdom

	Levene's Test		t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.
social exclusion benefits	13,160	,000	2,123	1804	,034	238,088	112,168
			2,240	1515,77	,025	238,088	106,294

Source: Own calculation, data EU-SILC

When performing the independent samples t-test on the “social exclusion benefits” variable, the group statistics table showed that households containing a student in lower education had a higher mean that households containing a student in higher education. The results from the

Levene's test showed that we cannot assume equal variances. According to the results of the t-test for equality of means, households with a student in lower education received significantly higher amounts of social exclusion benefits compared to households with a student in higher education. The p-value was 0.025 which gave a significance level of 5% for the results.

4.3. Mann Whitney U test

The next statistical method performed was the Mann Whitney U test. It is the non-parametric alternative to the independent samples t-test. The test was used to compare two samples and determine if they were significantly different from each other. The variables analyzed using this test were "head of household education", "material deprivation" and "financial burden of housing". The reason why these variables were not appropriate for the independent samples t-test was because they were ordinal scale of measurement and not continuous (Field, 2009). The assumptions of the Mann Whitney U test states that the test variable should be measured either on the continuous or ordinal level and the grouping variable should consist of two categorical groups. In addition to this there should also be independence of observations, the same assumption as for the independent samples t-test (Laerd Statistics, 2018). In our case "study level" was the categorical grouping variable and "head of household education", "material deprivation" and "financial burden of housing" were the test variables that were measured at the ordinal level. The assumption that indicates how the results for the Mann Whitney U test should be interpreted was that the distributions of both groups in the categorical variable should have the same distribution shape. The distributions of the two groups for all the three test variables in this thesis were of slightly different shapes for the data of the Czech Republic as well as for the data of Sweden and the United Kingdom. Therefor the assumption of same distribution shapes was broken which led us to comparing mean ranks and not medians when interpreting the Mann Whitney U test (Laerd Statistics, 2018). Once the assumptions were checked, the Mann Whitney U test was conducted on the variables "head of household education", "material deprivation" and "financial burden of housing" with "study level" as the grouping variable on the data of the Czech Republic.

When interpreting the output of the Mann Whitney U test for the variable "head of household education" on the data of Czech Republic we started by investigating the ranks table (See Appendix A). It indicated which group had the higher mean rank for the test variable. The ranks table showed that the households with a student in higher education had a higher mean rank than the households with a student in lower education (See Appendix A).

Table 15. Test Statistics, "head of household education" – Czech Republic

	head of household education
Mann-Whitney U	327207,500
Wilcoxon W	857142,500
Z	-9,212
Asymp. Sig. (2-tailed)	,000

Source: Own calculation, data EU-SILC

To determine if the difference between mean ranks was significant or not we moved on to the test statistics table. The results showed that the head of the household's education level was significantly higher for households containing a student in higher education compared to households containing a student in lower education. The p-value was 0.000 which gave a significance level of 0.1% for the results.

Next was the Mann Whitney U test for the "material deprivation" variable. The ranks table for the "material deprivation" variable showed that the mean rank for the households with a student in higher education was lower than the mean rank for the households with a student in lower education (See Appendix A).

Table 16. Test Statistics, "material deprivation" – Czech Republic

	material deprivation
Mann-Whitney U	313561,500
Wilcoxon W	623639,500
Z	-8,885
Asymp. Sig. (2-tailed)	,000

Source: Own calculation, data EU-SILC

The results from the test statistics showed that households containing a student in lower education had a significantly higher degree of material deprivation compared to households containing a student in higher education. The p-value was 0.000 which gave a significance level of 0.1% for the results.

The third Mann Whitney U test for the Czech Republic was for the variable “financial burden of housing”. The ranks table showed that households with a student in higher education had a lower mean rank that households with a student in lower education (See Appendix A).

Table 17. Test Statistics, “financial burden of housing” – Czech Republic

	financial burden of housing
Mann-Whitney U	348037,000
Wilcoxon W	658115,000
Z	-6,224
Asymp. Sig. (2-tailed)	,000

Source: Own calculation, data EU-SILC

According to the results from the test statistic there was a significantly higher financial burden of housing costs for households with a student in lower education compared to households with a student in higher education. The p-value was 0.000 which gave a significance level of 0.1% for the results.

The second country that the Mann Whitney U test was conducted for was Sweden. The variables tested for the Mann Whitney U test were “head of household education”, “material deprivation” and “financial burden of housing with “study level” as the grouping variable. The first variable tested was “head of household education” and the Mann Whitney U test ranks table showed that households that included a student in higher education had a higher mean rank than households that included a student in lower education (See Appendix A).

Table 18. Test Statistics, "head of household education" – Sweden

	head of household education
Mann-Whitney U	66699,000
Wilcoxon W	352089,000
Z	-1,844
Asymp. Sig. (2-tailed)	,065

Source: Own calculation, data EU-SILC

However, when checking the test statistics table, the results showed that this difference in mean ranks was not significant. The p-value was 0.065 and the results were therefore not significant on the 5% level.

Similar results were found when conducting the Mann Whitney U test on the "material deprivation" variable. The mean rank for households containing a student in higher education was lower than the mean rank for households containing a student in lower education (See Appendix A).

Table 19. Test Statistics, "material deprivation" – Sweden

	material deprivation
Mann-Whitney U	69552,000
Wilcoxon W	88273,000
Z	-1,396
Asymp. Sig. (2-tailed)	,163

Source: Own calculation, data EU-SILC

When interpreting the test statistics table, we could observe that the difference in mean ranks was not significant. The p-value was 0.163 and therefor the results were not significant on the 5% level.

The last Mann Whitney U test performed for the Swedish data was on the variable "financial burden of housing". The ranks table showed that households with a student in higher education had a lower mean rank than the households with a student in lower education (See Appendix A).

Table 20. Test Statistics, “financial burden of housing” – Sweden

	financial burden of housing
Mann-Whitney U	65206,500
Wilcoxon W	83734,500
Z	-2,092
Asymp. Sig. (2-tailed)	,036

Source: Own calculation, data EU-SILC

For “financial burden of housing” the difference of the mean ranks was significant according to the test statistics. The results showed that households that included a student in lower education had a higher financial burden of housing compared to households that included a student in higher education.

The United Kingdom was the last country for the non-parametric Mann Whitney U test to be conducted on. Again, the test was performed on the test variables “head of household education”, “material deprivation” and “financial burden of housing with “study level” as the grouping variable. The first variable tested was “head of household education”. The mean rank was higher for the households containing a student in higher education than for the households containing a student in lower education (See Appendix A).

Table 21. Test Statistics, “head of household education” – United Kingdom

	head of household education
Mann-Whitney U	294761,000
Wilcoxon W	915816,000
Z	-3,858
Asymp. Sig. (2-tailed)	,000

Source: Own calculation, data EU-SILC

The difference in mean ranks turned out to be significant on the 0.1 % level, the p-value was 0.000. These results showed that the head of the household’s education level was higher for households with a student in higher education compared to households with a student in lower education.

The Mann Whitney U test was then conducted on the “material deprivation” variable. The mean rank for the households with a student in higher education was lower than the mean rank for the households with a student in lower education (See Appendix A).

Table 22. Test Statistics, “material deprivation” – United Kingdom

	material deprivation
Mann-Whitney U	342196,000
Wilcoxon W	544762,000
Z	-2,982
Asymp. Sig. (2-tailed)	,003

Source: Own calculation, data EU-SILC

The results from the test statistics showed that households containing a student in lower education had a significantly higher degree of material deprivation compared to households containing a student in higher education. The p-value was 0.003 which gave a significance level of 1% for the results.

The third and last Mann Whitney U test for the United Kingdom was performed on the “financial burden of housing” variable. According to the rank table the mean rank for the households with a student in higher education was lower than the mean rank for households with a student in lower education (See Appendix A).

Table 23. Test Statistics, “financial burden of housing” – United Kingdom

	financial burden of housing
Mann-Whitney U	334429,500
Wilcoxon W	536995,500
Z	-3,818
Asymp. Sig. (2-tailed)	,000

Source: Own calculation, data EU-SILC

The test statistics showed a significant result at the 0.1% level with the p-value 0.000. Therefor households containing a student in lower education had a significantly higher financial housing burden compared to households containing a student in higher education.

What could be observed after conducting the independent samples t-tests and the Mann Whitney U tests was that all the variables tested for the Czech Republic and the United Kingdom were significant. For Sweden, two of the variables, “head of household education” and “material deprivation”, were non-significant when conducting the Mann Whitney U test. In overall the significance levels for the tests were higher for the Czech Republic and the United Kingdom compared to Sweden. Investigating the means and mean ranks, the largest relative differences between the groups of households containing a student in higher education and households containing a student in lower education, were observed for the Czech Republic followed by the United Kingdom. Sweden had the lowest relative differences in the means and mean ranks.

4.4. Binary Logistic Regression

Binary logistic regression was the last but main statistical method to be performed. For binary logistic regression the dependent variable should be measured on a dichotomous scale and the independent variables can be either continuous or categorical (Field, 2009). This meant that all the six independent variables included in the thesis could be included in the binary logistic regression models. Before creating the binary logistic regression models another two assumptions that were mentioned in the method chapter had to be checked in SPSS. The first was that there should be no multicollinearity between the predictors (Field, 2009). This was checked by investigating the collinearity statistics and interpreting the variance inflation factor (VIF) values. According to Myers (1990) a VIF value above 10 should cause concerns for multicollinearity. Based on the low VIF values in our output there were no signs of multicollinearity among the predictors included in this thesis (See Appendix C). The assumption of linearity between any continuous independent variables and the logit transformation (Laerd, 2018) was checked by performing the Tidwell box test. This was checked for the continuous independent variables “household gross income” and “social exclusion benefits” by including interaction terms between each variable and the logarithm of itself. The results showed that the assumption of linearity was met since the interaction terms included in the model were not significant (See Appendix D). In addition to the assumptions, the correlation between the independent variables were also analyzed. The output showed that we could observe some degree of correlation between the variables (See Appendix E). Even though it did not lead to multicollinearity, it may still be of relevance when discussing the results of the binary logistic regression models. For improved interpretation and sample sizes, the values of the two categorical variables “occupation” and “head of household education” were re-grouped. The “occupation” variable initially included many specific occupations from the ISCO-08 coding and was sorted into four larger groups. The first group consisted of highly skilled professionals and managers, the second consisted of office occupations, the third consisted of practical occupations and the fourth consisted of elementary occupations.

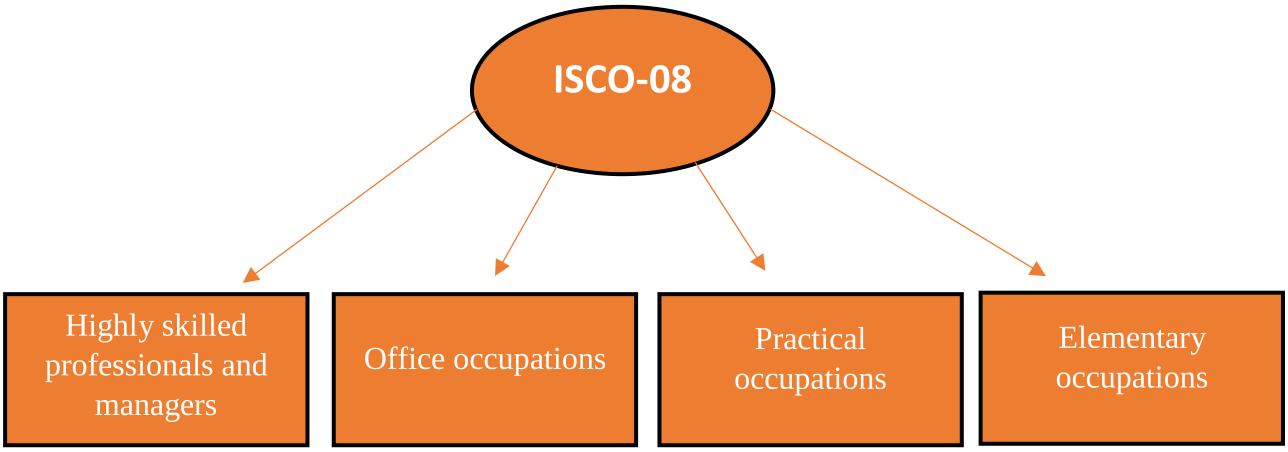


Figure 3. Grouping of occupations

Source: Own calculation, data EU-SILC

The variable “head of household education” was regrouped into two categories. The two categories were higher education and lower education. The two variables “household gross income” and “social exclusion benefits” measured on a continuous scale also had to be modified. The reason for this was that the Exp(B) represents the change of odds resulting from one-unit change in the independent variable (Field, 2009). A change of one currency unit for a household in our scale variables would naturally not lead to any changes in odds because the change would in proportion be irrelevant. Therefor the variable “household gross income” was divided by 10 000. This led to interpreting the EXP(B) as the change of odds if a household earned 10 000 currency units more in gross income during a reference period. The variable “social exclusion benefits” was divided by 10 which led to the interpretation of the EXP(B) as the change of odds if a household received 10 more currency units in social exclusion benefits during a reference period. These modifications led to more relevant and easier interpretation of the Exp(B) values.

When the binary logistic regression models were created in SPSS the dependent variable and the categorical independent variables became coded. The reference categories for these variables were also selected. The coding and reference categories for these variables were the same for all the three countries binary logistic regression models. The dependent variable ”study level” had lower education as its reference category. The categorical independent variable “occupation” had four groups and the reference category was highly skilled professionals and managers. The categorical independent variable “head of household education” only had two groups and the reference category was lower education.

When creating the binary logistic models there was a discussion about either creating full models containing all the predictors or smaller models that did not include non-significant predictors. The question was which of these models would be most appropriate explaining socioeconomic status impact in higher education. According to Harrell (2015) the full models may not be as parsimonious as smaller models where the non-significant predictors have been removed. Removing non-significant predictors can however lead to a variety of problems and properties lost and there for the full models should be considered the “Gold Standard”. This especially applies when it comes to making formal inferences (Harrell, 2015), which was the aim of this thesis. Based on the modeling strategy recommended by Harrell (2015), full binary logistic regression models containing all the independent variables were created for the three countries included in the thesis. The first binary logistic model created was the model for the Czech Republic.

4.4.1. The Czech Republic

Table 24. Omnibus Tests of Model Coefficients – Czech Republic

	Chi-square	df	Sig.
Step	176,873	8	,000
Block	176,873	8	,000
Model	176,873	8	,000

Source: Own calculation, data EU-SILC

The omnibus tests of model coefficients is an overall test of the model. It tests if the model that includes the independent variables is an improvement from a model only including an intercept. The omnibus test gives a first indication if the independent socioeconomic variables impact the dependent variable "study level". The results showed that including the socioeconomic independent variables in the model for the Czech Republic significantly improved the model. This indicated that the independent variables impact "study level".

Table 25. Binary Logistic Regression Model – Czech Republic

	B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
household gross income	,105	,048	4,679	1	,031	1,110	1,010	1,221
social exclusion benefits	-,001	,002	,145	1	,703	,999	,995	1,003
head of household educ., higher	,517	,142	13,312	1	,000	1,676	1,270	2,212
material deprivation	-,178	,051	12,231	1	,000	,837	,757	,925
financial burden of housing	-,151	,108	1,942	1	,163	,860	,696	1,063
occupation			33,242	3	,000			
office occupations	-,028	,166	,028	1	,867	,973	,703	1,346
practical occupations	-,599	,125	23,110	1	,000	,549	,430	,701
elementary occupations	-,902	,253	12,684	1	,000	,406	,247	,667
Constant	,206	,290	,507	1	,477	1,229		

Source: Own calculation, data EU-SILC

The results from the binary logistic regression model for the Czech Republic showed that four out of the six independent variables were significant. The four significant variables were "household gross income", "head of household education", "material deprivation" and "occupation". However, the significance of a variable may not be the only interesting aspect to focus on (Gupta, 2012). The confidence intervals for the Exp(B) could give indications that non-significant variables may be of importance explaining the impact of socioeconomic status in higher education. The interval could also indicate that even if a variable is significant, it might not be of a major importance. The reasons for this is that the size of the effects a variable has can sometimes be more relevant than the actual significance (Gupta, 2012). The Exp (B) of the significant variable "household gross income" showed that if a household earns an additional 10 000 currency units, it was between 1.010 and 1.221 times significantly more likely to have a student in higher education. The mean income for a Czech household was 22 893.67 currency units according based on the thesis data. An additional 10 000 currency units in gross income is there for a relatively high change in gross income and still it did not result in a large Exp(B) value. The “Head of household education” variable was also significant. The Exp(B) interval showed that if the head of the household has a higher education, then the household was between 1.270 and 2.212 times significantly more likely to have a student in higher education compared to if the head of the household had a lower education. When it came to “material deprivation”, households were significantly less likely to have student in higher education for every additional item that they could not afford. The interval for the Exp(B) was between 0,757 and 0.925. The variable “financial burden of housing” was not significant in the model but was worth interpreting due to the Exp(B) confidence intervals. The intervals were between 0.696 and 1.063 which resulted in non-significant results since the value 1 was included in the interval. The interval still indicated that households with a higher financial burden of housing were less likely to have a student in higher education compared to households with a less financial burden, even though this was not proved significantly. The last variable in the model, “occupation” had an overall significance. The results showed that if the head of the household had a practical occupation the household was significantly less likely to have a student in higher education compared to if the head of the family had a highly skilled professionals and manager occupation. The Exp(B) interval was between 0.430 and 0.701. The results also showed that if the head of the household had an elementary occupation, the household was significantly less likely to have a student in higher education compared to if the head of the family had a highly skilled professionals and manager occupation. These results showed an even larger impact with an Exp(B) interval between 0.247 and 0.667.

4.4.2. Sweden

The binary logistic regression model for Sweden was created including the same variables as the one for the Czech Republic. As shown in the model output tables below, the dependent variables in the Swedish binary regression model were not significant.

Table 26. Omnibus Tests of Model Coefficients – Sweden

	Chi-square	df	Sig.
Step	9,386	8	,311
Block	9,386	8	,311
Model	9,386	8	,311

Source: Own calculation, data EU-SILC

The omnibus tests of model coefficients showed that the model including the independent variables was not a significant improvement from the model only including the intercept. This indicated that the socioeconomic variables in the model do not impact the dependent variable "study level".

Table 27. Binary Logistic Regression Model – Sweden

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
household gross income	,013	,019	,478	1	,489	1,014	,976	1,053
social exclusion benefits	,000	,001	,182	1	,670	1,000	,998	1,001
head of household educ., higher	,004	,197	,000	1	,985	1,004	,683	1,476
material deprivation	-,012	,120	,010	1	,922	,988	,781	1,251
financial burden of housing	-,239	,149	2,581	1	,108	,787	,588	1,054
occupation			2,202	3	,532			
office occupations	-,003	,350	,000	1	,994	,997	,503	1,979
practical occupations	-,283	,216	1,712	1	,191	,753	,493	1,151
elementary occupations	-,400	,474	,713	1	,398	,670	,265	1,697
Constant	1,005	,338	8,821	1	,003	,366		

Source: Own calculation, data EU-SILC

Even though none of the variables were significant in the Swedish binary logistic model, two of the variables may still be of importance based on the Exp(B) confidence intervals. These two variables were "financial burden of housing" and "occupation". We could observe that the Exp(B) confidence interval for "financial burden of housing" was between 0.588 and 1.054.

This indicated that households with a higher financial housing burden were less likely to have a student in higher education in comparison with households with a lower financial burden, even though it could not be proven significantly. For the variable “occupation” there were indications that households where the head of the family had a practical occupation were less likely to have a student in higher education compared to households where the head of the family had a highly skilled professionals and manager occupation. The interval for the Exp(B) was between 0.493 and 1.151. These indications were not significant but may still be of importance. There could be several reasons why none of the independent variables were significant in the Swedish binary logistic regression model and why the omnibus tests showed non-significant results. One reason could be that socioeconomic status simply does not impact the decision to pursue a higher education in Sweden and therefor the socioeconomic variables in this thesis were non-significant. Conducting the independent samples t-tests and the Mann Whitney U tests on the variables we could observe that the means and mean ranks for Sweden did not differ relatively as much between the two dependent variable groups as they did for the United Kingdom and the Czech Republic. The variables “head of household education” and “material deprivation” did not even show significance when conducting the Mann Whitney U test. This may have been an indication that socioeconomic status might not impact the decision to pursue a higher education to the same extent in Sweden as it does in the United Kingdom and the Czech Republic. That none of the independent variables were significant in the binary logistic regression model for Sweden would however go against previous research conducted on the topic in general as well as research specifically conducted on Swedish data (Falk & Salih, 2017; Björlund & Salvanes, 2010). Another explanation might be that the sample size for Sweden was not large enough in this thesis to generate significant results. A combination of the smaller relative differences between the two dependent variable groups together with the small sample size could have led to the non-significant predictors in the Swedish binary logistic regression model. The independent variables could also be competing to explain the variance of the dependent variable due to being slightly correlated to each other (See Appendix E), even though significant multicollinearity was not detected. This would lead to higher p-values and non-significant results. It could explain why some of the variables were significant when conducting the independent samples t-tests and the Mann Whitney U tests but not in the binary logistic regression model.

4.4.3. The United Kingdom

The last binary logistic regression model created was for the United Kingdom. The results from the Omnibus tests for the United Kingdom showed that the model including the socioeconomic independent variables was a significant improvement compared the model only including an intercept. This indicated that the independent variables have an impact on the dependent variable “study level”.

Table 28. Omnibus Tests of Model Coefficients – United Kingdom

	Chi-square	Df	Sig.
Step	37,410	8	,000
Block	37,410	8	,000
Model	37,410	8	,000

Source: Own calculation, data EU-SILC

Table 29. Binary Logistic Regression model – United Kingdom

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
household gross income	,004	,012	,113	1	,737	1,004	,980	1,029
social exclusion benefits	,000	,000	,006	1	,941	1,000	,999	1,000
head of household educ., higher	,276	,121	5,224	1	,022	1,317	1,040	1,669
material deprivation	,017	,051	,109	1	,742	1,017	,921	1,123
financial burden of housing	-,187	,079	5,595	1	,018	,830	,711	,969
occupation			8,381	3	,039			
office occupations	-,273	,203	1,814	1	,178	,761	,511	1,132
practical occupations	-,383	,138	7,763	1	,005	,682	,521	,893
elementary occupations	-,127	,195	,426	1	,514	,881	,601	1,290
Constant	-,227	,219	1,071	1	,301	,797		

Source: Own calculation, data EU-SILC

The results showed that three out of the six independent variables were significant in the United Kingdom's binary logistic regression model. The three significant variables were "head of household education", "financial burden of housing" and "occupation". For the variable "head of household education" the Exp(B) interval showed that if the head of the household has a higher education, then the household was between 1.040 and 1.669 times significantly more likely to have a student in higher education compared to if the head of the household had a lower education. When it came to "financial burden of housing", households with a higher financial housing burden were significantly less likely to have a student in higher education in comparison with households with a lower financial burden. The Exp(B) interval was between 0.711 and 0.969. The "occupation" variable was significant overall but when interpreting the groups separate, the only occupation group that was significantly different compared to the reference category was the practical occupations. The results showed that if the head of the household had a practical occupation the household was significantly less likely to have a student in higher education compared to if the head of the family had a highly skilled professionals and manager occupation. The Exp(B) interval was between 0.521 and 0.893. There were also indications that if the head of the household had an office occupation the household was significantly less likely to have a student in higher education compared to if the head of the family had a highly skilled professionals and manager occupation. These results could be of importance but were however not significant since the Exp(B) interval was between 0.511 and 1.132.

5. Discussion

Interpreting the results from the independent samples t-tests, the Mann Whitney U tests and the binary logistic regression models for the Czech Republic, Sweden and the United Kingdom, we could observe that the results differed between the countries. The Czech Republic and the United Kingdom presented several significant findings while the results for Sweden were weaker. The table below presents which variables that were significant when conducting the independent samples t-tests, the Mann Whitney U tests and creating the binary logistic regression models for the countries.

Table 30. Significant Variables based on Statistical Methods

Statistical method	Czech Republic	Sweden	United Kingdom
Independent Samples t-test	household gross income, social exclusion benefits	household gross income, social exclusion benefits	household gross income, social exclusion benefits
Mann Whiney U test	head of household education, material deprivation, financial burden of housing	financial burden of housing	head of household education, material deprivation, financial burden of housing
Binary Logistic Regression	household gross income, head of household education, material deprivation, occupation		head of household education, financial burden of housing, occupation

Source: Own calculation, data EU-SILC

For Sweden significant results were found when conducting the independent samples t-test and when conducting the Mann Whitney U test but not in the binary logistic model. Only slight indications of importance for “financial burden of housing” and “occupation” were found. What we could say from the results for Sweden was that there were significant differences between the households containing a student in higher education and a student in lower education concerning the variables “household gross income”, “social exclusion benefits” and “financial burden of housing”. However, these socioeconomic differences between the groups did not have a significant impact on the decision of pursuing a higher education. The results go against the findings of Björlund and Salvanes (2010) who´s study included data from Sweden and the United Kingdom. The authors found significant results that family background affected a student's educational attainment from an in-equality point of view (Björlund & Salvanes 2010). The reason the results differ may be because Björlund and Salvanes (2010) study was based on different data and variables compared to this thesis. The results that they were able to find based on their data on sibling correlation, we were not able to confirm by using the independent variables in the EU-SILC data. The results from this thesis should also be compared to Falck and Salih (2017) who had a focus on parental occupation, a variable that was closely related to the "occupation" variable included in this thesis. Their study was conducted on Swedish data and concluded that parental occupation effects participation in higher education. The reason why Falck and Salih (2017) results differed from the findings in this thesis may be because of two reasons. One was that their sample size was of a much larger size compared to this thesis and the second was that their grouping of occupations differed from this thesis. This thesis results gave indications that households where the head of the family had a practical occupation were less likely to have a student in higher education compared to households where the head of the family had a highly skilled professionals and manager occupation. With a larger sample size these indications might have become significant. What should also be mentioned was that these indications in our results were found between the two occupation groups that had the most observations in them. This goes in line with our reasoning about the sample size issue. Despite the uncertain results found for Sweden there was still an interesting aspect that could be observed in overall. All the three variables that were found significant when conducting the independent samples t-tests and Mann Whitney U tests were variables connected to the financial aspects of socioeconomic status. Even though the results for Sweden did not point towards socioeconomic status having a substantial impact in higher education, it could still be relevant for policy makers to focus their efforts on households with financial constraints. In the case for Sweden, reducing the socioeconomic differences within the financial area between households could lower social exclusion in higher education even further.

The thesis obtained several significant results for the United Kingdom, both from the independent samples t-tests and the Mann Whitney U test, as well as for the binary logistic regression model created. Previous research on socioeconomic status impact in higher education included several studies for the United Kingdom. This thesis results showed that there were clear differences in socioeconomic status between households with a student in higher education and households with a student in lower education. Moreover, the socioeconomic status variables “head of household education”, “financial burden of housing”, “occupation” were significantly impacting the decision of pursuing a higher education.

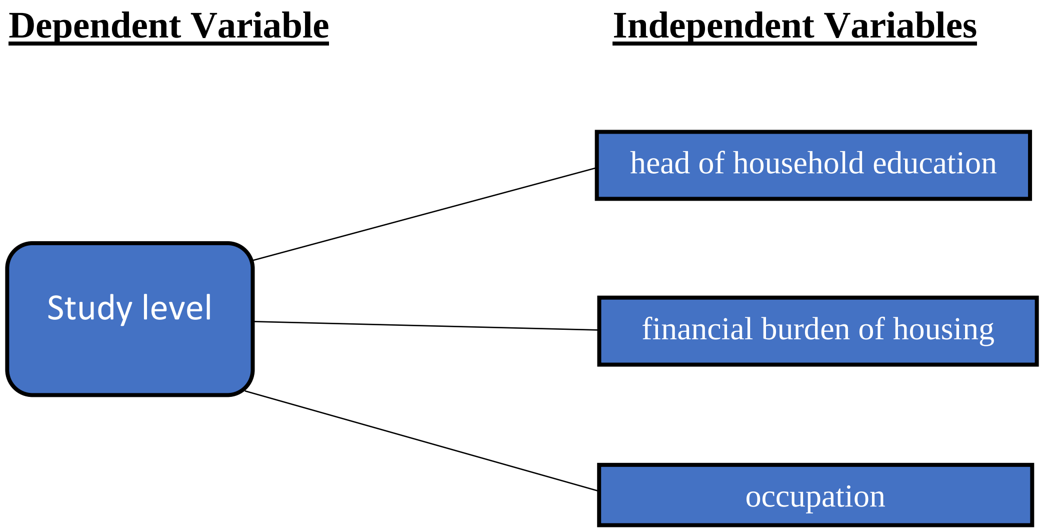


Figure 4. Variables with Significant Impact – United Kingdom

Source: Own calculation, data EU-SILC

The significance of “head of household education” went in line with previous research found on parental educations impact in higher education on data from the United Kingdom (Chevalier et al., 2013; Triventi, 2011). The “financial burden of housing”, a variable connected to the financial aspects of socioeconomic status could be compared to the findings of Chevalier et al. (2013) and Chowdry et al. (2010) who both had data on the United Kingdom. Their results showed that financial socioeconomic variables impact higher education in a significant way. Even though previous research may not have investigated the exact same variables as this thesis, we can still draw conclusions from the similarities in the findings. The fact that different variables were used could even give a deeper knowledge and strengthen the results that were found by previous research. This thesis also adds the significant results from the “occupation” variable to the discussion. Labor “variables” were not studied by the previous research using data from the United Kingdom presented in the theory chapter. What this thesis found for the United Kingdom was that if the head of the household had a practical occupation the household was significantly less likely to have a student in higher education compared to if the head of the family had a highly skilled professionals and manager occupation. In other words, students from households with higher socioeconomic status are more likely to participate in higher education. This is an aspect that could lead to social exclusion in higher education. Summing up the results for the United Kingdom it was seen that several areas of socioeconomic status together impact the decision to pursue a higher education. This complexity for policymakers was mentioned by previous studies and the authors did not necessarily agree on how policymakers should act. While Chevalier et al. (2013) argues for policies on permanent income, Chowdry et al. (2010) instead highlights the importance of intervening earlier in a student's school years for maximum effect. What could be said based on this thesis results was that policymakers should not have a narrow focus when trying to reduce social exclusion in higher education for the United Kingdom. They must have in mind that several different socioeconomic factors together impact higher education.

The Czech Republic was the country where the largest differences between the groups were found within socioeconomic status in higher education. This was based on the significance levels from the independent samples t-tests and the Mann Whitney U tests. The binary logistic regression model also showed that several socioeconomic variables impact the decision to pursue a higher education. The significant variables were "household gross income", "head of household education", "material deprivation" and "occupation".

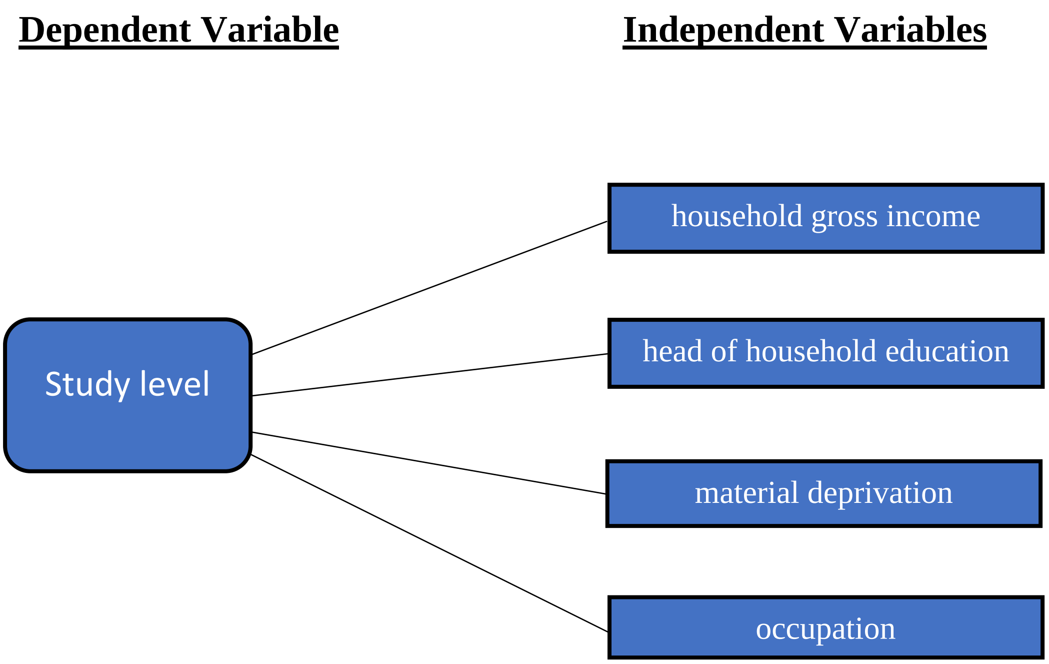


Figure 5. Variables with Significant Impact – Czech Republic

Source: Own calculation, data EU-SILC

What was challenging when it came to the Czech Republic's results was that not a lot of previous research had been conducted on the country within this study area. The exception was Triventi (2011) who included the Czech Republic in its study together with many other countries. Therefor when discussing and comparing the results from the Czech Republic we had to have a more general approach. Previous research had been conducted on similar variables as this thesis significant variables but based on other countries data. The results of this thesis went in line with previous research for both the significant variables "household gross income" (Chevalier et al., 2009; Chowdry et al., 2010) and "head of household education" (Bowden & Doughney, 2012; Chevalier et al., 2009; Triventi, 2011). What was mentioned when interpreting the model though was that the change of "household gross income" had to be very large to result in an impact on the decision to pursue a higher education. Therefor it was debatable how much impact this variable had on pursuing a higher education. When it came to this thesis significant variable "material deprivation" none of the previous studies presented in the theory chapter had investigated any similar variables. The closest to compare the results with was Bowden & Doughney (2012) who investigated how access to internet at home effected student's aspiration to attend university. Even though "material deprivation" was more standardized and general they both represented the financial aspects of socioeconomic status. The head of the household's occupation also impacted the decision to pursue a higher education. What was interesting to observe for the Czech Republic was that this impact was only significant when comparing the reference category, highly skilled professionals and managers, with the two lowest groups of occupations. It seems that the impact only applied to the lower classes of socioeconomic status. Once the head of the family had a certain level of occupation the impact was not significant any more. The Czech Republic's overall results showed that several areas of socioeconomic status impact higher education. Especially for the lowest classes characterized by a high degree of material deprivation and where the head of the family had an occupation in the lower occupation categories. Policymakers may use these findings to target their efforts towards the lower social socioeconomic classes to reduce social exclusion in higher education in the Czech Republic.

When discussing differences and similarities between the three countries included in the thesis there were some interesting aspects to be observed. Even though the significant variables in the binary logistic regression models for the Czech Republic and the United Kingdom differed slightly, the end results may be similar. The difference was that for the Czech Republic "households gross income" and "material deprivation" were significant while for the United

Kingdom “financial burden of housing” was significant. When it comes down to it, all these three variables represent the financial area of socioeconomic status and therefor one can argue that the results were in fact very similar. What could be discussed regarding differences found when conducting the independent samples t-tests and Mann Whitney U tests was that Sweden clearly had the smallest difference between households with a student in higher education and households with a student in lower education. Together with the non-significant results found in the binary logistic model it indicated that socioeconomic status does not impact higher education as much in Sweden as it does in the Czech Republic and the United Kingdom. With that said it was important to have the issue regarding the sample size in mind. The small sample size for Sweden and for the thesis in general was a weakness when conducting the analysis and obtaining the results. Another aspect that could be discussed was that this thesis only used subjective variables. This ensured that the results did not become influenced by objective biases. Subjective variables also have limitations though (Rubin et al., 2014). One of the weaknesses is that subjective variables must be benchmarked and interpreted relatively to population standards. Since the variables were taken from the EU-SILC this did not become an issue in this thesis. The EU-SILC based its questionnaire on international standards and classifications when it came to for example occupations and education. Therefor deciding on correct standards and benchmarks in the thesis newer became an issue, the EU-SILC methodology already had clear recommendations regarding that matter. Another limitation that Rubin et al. (2014) mentioned was that subjective variables are often not directly connected to the student which can lead to misinterpretations of results. This limitation was more relatable to this thesis. The fact that the data used for this thesis was based on household questionnaires and not questionnaires directly answered by the student could have led to the results giving us more knowledge about the households than the actual students. It would be interesting for future researchers to include control variables that are directly connected to the student when conducting research within the topic of socioeconomic status impact in higher education. This together with a larger sample size could generate interesting findings to further develop the area of research.

6. Conclusion

This thesis is devoted to investigating the impact of socioeconomic status and social exclusion in higher education. The theoretical chapter presented the definitions and concepts of higher education, social exclusion and socioeconomic status. Especially defining higher education was important for this thesis to make sure the practical analysis was constructed in a correct way following the thesis purpose. Regarding socioeconomic status it was stated that it is not only focused on one, but several different areas. This thesis succeeded in including and measuring the impact of several of these by using variables representing the different socioeconomic areas. The three main areas of socioeconomic status are occupation, financial and education. When speaking about social exclusion it is important to remember that it is a subjective and complex phenomenon. This thesis may only give indications on if certain groups in society are socially excluded. Therefore, we cannot say if the social exclusion is significant or not, what we can say though is that improvements can always be made. When recommendations about reducing social exclusion are given to policy makers in this thesis, these recommendations are stating how policy makers can reduce social exclusion, not how severe the situation is. The theory chapter also presented the findings of previous research. What could be concluded from previous research findings was that socioeconomic status clearly had an impact in higher education. The question was which socioeconomic variables that were of importance. Previous research was proposing more studies based on new data and different socioeconomic variables (Björlund & Salvanes, 2010; Falck & Salih, 2017; Triventi, 2011). This was exactly what this thesis was able to accomplish and contribute with. No previous studies, excluding the project at the University of Economics, Prague, had been done on socioeconomic status impact in higher education based on the EU-SILC data. In addition to this, the thesis managed to include several different socioeconomic variables in a single analysis. Accomplishing this led to a deeper understanding of socioeconomic status impact in higher education, which was the purpose of the thesis.

The method chapter presented the EU-SILC data that was analyzed in the thesis. The EU-SILC data is a large and complex data set in terms containing many variables that are connected to both household ID and personal ID. During the extraction process this complexity led to a relatively small sample size and the data from the years 2016 and 2012 being combined. This was to ensure that a specific household were not represented more than once but at the same time to increase the sample size in comparison with only selecting 2016 years data. It should be mentioned that researchers with expertise on the EU-SILC data might have been able to solve these issues during the data extraction process in a better way and through that eliminate these limitations. It would have been an improvement for this thesis methodology to have been able to increase the sample size and at the same time only use the latest available data. Another limitation concerning the data and method was that the variables used were connected to the student's household and not directly to the student itself. This is not unique for this thesis but is in general a concern when conducting studies within socioeconomic status impact in higher education using objective variables (Rubin et al., 2014). It can lead to more knowledge about the households than the actual students. Therefore, it would be interesting for future researchers to include control variables that are directly connected to the student when conducting research within the topic of socioeconomic status impact in higher education.

This thesis purpose and aim led down to answering how socioeconomic status impacts the decision to pursue a higher education, and how social exclusion in higher education can be reduced. The results chapter together with the discussion chapter had the aim to answer this question. Based on the results obtained we can say that socioeconomic status does impact the decision to pursue a higher education in the case of the Czech Republic and the United Kingdom. For the Czech Republic the impact is the most obvious. The socioeconomic variables that impacts higher education in the Czech Republic significantly represents all the three main areas of socioeconomic status; occupation, financial and education. What can specifically be observed is that the impact is the largest for students from the lowest socioeconomic status classes. To reduce social exclusion, policy makers should therefore target their efforts on students from the lowest socioeconomic status to maximize the effect.

The United Kingdom's results also clearly show that socioeconomic status impacts the decision to pursue a higher education. The significant variables represent all three main areas of socioeconomic status, but the impact is not as large as for the Czech Republic. There are no signs that one area would impact higher education in a significantly larger extent than the others. Based on these results we recommend the policy makers in the United Kingdom to have wider focus when trying to reduce social exclusion. This since several different socioeconomic factors together impact higher education.

If socioeconomic status impacts the decision to pursue a higher education in the case of Sweden is more uncertain based on the results obtained in this thesis. No significant results are found for Sweden in the binary logistic regression model. This goes against previous research conducted on socioeconomic status impact in higher education using other data from Sweden (Björklund & Salvanes 2010; Falck & Salih 2017). The reasons why no significant results were found could be because of the small sample size combined with correlated independent variables competing to explain the dependent variable. This could have led to insignificant results. Why the results differ from previous research findings could also be because they have conducted their studies on different types of data and therefor obtained different results. It would be especially interesting for future researchers to conduct studies on Swedish data containing a larger sample size to either confirm or reject this thesis results on Sweden. Even though the results for Sweden are not significant, there are still indications of impact that can be observed. The indications show that the variables representing the financial area of socioeconomic status are the ones impacting the decision to pursue a higher education the most. Based on these indications policy makers in Sweden should try to decrease the socioeconomic differences within the financial area to reduce social exclusion in higher education.

List of Tables

List of Figures

References

American Psychological Association. (2018). *Education and socioeconomic status.* Retrieved from http://www.apa.org/pi/ses/resources/publications/education.aspx

Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.

Burns, Ri., & Burns, Ro. (2008). *Business Research Methods and Statistics using SPSS.* London: Sage.

Björklund, A., & Salvanes, K. G. (2010). *Education and family background: Mechanisms and policies.* Discussion paper series, Forschungsinstitut zur Zukunft der Arbeit, No. 5002. Bonn: ZBW.

Bowden, M. P., & Doughney, J. (2012). The importance of cultural and economic influences behind the decision to attend higher education. *The Journal of Socio-Economics 41*, 95-103.

Bryman, A., & Bell, E. (2011). *Företagsekonomiska forskningsmetoder*. Stockholm: Liber.

Chevalier, A., Harmon, C., O´Sullivan, V., & Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics, 2,* 8.

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2010). Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 176(2),* 431-457.

Crossman, A. (2017, December 3). *An Introduction to Socioeconomic Status.* Retrieved from https://www.thoughtco.com/socioeconomic-status-3026599

European Commission. (2016). *DIFFERENCES BETWEEN ORIGINAL DATABASE (as described in the guidelines) AND THE ANONYMISED USER DATABASE*. Retrieved from https://www.scribd.com/document/376800569/c16-Differences-Collection-vs-Udb

European Commission. (2017). *HOW TO APPLY FOR MICRODATA?*. Retrieved from http://ec.europa.eu/eurostat/documents/203647/771732/How_to_apply_for_microdata_access.pdf

European Commission. (2018). *Europe 2020 – Overview*. Retrieved from http://ec.europa.eu/eurostat/web/europe-2020-indicators/europe-2020-strategy

Eurostat. (2014). *Young people – Social inclusion*. Retrieved from http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Young_people_-_social_inclusion&oldid=326182#Living_with_parents

Eurostat. (2017). *List of variables*. Retrieved from http://ec.europa.eu/eurostat/web/income-and-living-conditions/methodology/list-variables

Eurostat. (2018). *European union statistics on income and living conditions (EU-SILC).* Retrieved from http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

EU-SILC Longitudinal UDB 2016 and 2012 – version of October 2017.

Fagerland, M. W. (2012). T-tests, non-parametric tests, and large studies—a paradox of statistical practice?. *BMC Medical Research Methodology, 12,* 78.

Falck, K., & Salih, J. (2017). *Klass, socialisation och utbildning: En kvantitativ studie om hur uppväxtförhållanden påverkar utbildningsnivå* (Dissertation). Retrieved from Diva Portal.

Field, A. P. (2009). *Discovering Statistics using SPSS (and sex and drugs and rock 'n' roll)*. Los Angeles: Sage.

Frost, J. (2017, April 11). *Nonparametric Tests vs. Parametric Tests*. Retrieved from http://statisticsbyjim.com/hypothesis-testing/nonparametric-parametric-tests/

Gupta. S. K. (2012). The relevance of confidence interval and *P*-value in inferential statistics. Indian journal of pharmacology. 2012;44(1):143. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3271529/

Harrell. E. (2015). *Regression Modeling Strategies*. New York: Springer.

Hoskin, T. (n.d.). *Parametric and Nonparametric: Demystifying the Terms*. Mayo Clinic Department of Health Sciences Research. Retrieved from http://www.mayo.edu/research/documents/parametric-and-nonparametric-demystifying-the-terms/doc-20408960

Laerd Statistics. (2018). *The ultimate IBM SPSS Statistics Guides*. Retrieved from https://statistics.laerd.com/

Mack, A. (2016). Data Handling in EU-SILC. *GESIS – Leibniz-Institut für Sozialwissenschaften 2016(10).*

McLachlan, R., Gilfillan, G., & Gordon, J. (2013). Deep and Persistent Disadvantage in Australia. *Australian Government Productivity Commission Staff Working Paper*.

Myers, R. (1990). *Classical and modern regression with applications*. Boston: Duxbury.

Ostrove, J. M., Long, & S. M. (2007). Social class and belonging: Implications for college adjustment. *Review of Higher Education, 30*, 363-389.

Rubin, M., Denson, N., Kilpatrick, S., Matthews, K. E., Stehlik, T., & Zyngier, D. (2014). "I am working class": Subjective self-definition as a missing measure of social class and socioeconomic status in higher education research. *Educational Researcher, 43*, 196-200.

Tranmer, M., & Elliot, M. (2008). *Binary Logistic Regression*. The Cathie Marsh Centre for Census and Survey Research. Retrieved from http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-20-binary-logistic-regression.pdf

Triventi, M. (2011). Stratification in Higher Education and Its Relationship with Social Inequality: A Comparative Study of 11 European Countries. *European Sociological Review 29(3)*, 489–502.

UNESCO. (2011). *International Standard Classification of Education ISCED 2011*. Retrieved from http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

United Nations. (2016). *The Report on the World Social Situation 2016*. Retrieved from http://www.un.org/esa/socdev/rwss/2016/full-report.pdf

University of Southern California. (2018). *Research Guides: Organizing Your Social Sciences Research Paper: Types of Research Designs*. Retrieved from http://libguides.usc.edu/writingguide/researchdesigns

World Health Organization. (2018). *Social exclusion*. Retrieved from http://www.who.int/social_determinants/themes/socialexclusion/en

Appendices

Appendix A: Mann Whitney Ranks Tables

Ranks Table, “head of household education” – Czech Republic

	study level	N	Mean Rank	Sum of Ranks
head of household education	Lower education	1029	832,99	857142,50
	Higher education	787	1007,23	792693,50
	Total	1816		

Source: Own calculation, data EU-SILC

Ranks Table, “material deprivation” – Czech Republic

	study level	N	Mean Rank	Sum of Ranks
material deprivation	Lower education	1029	997,28	1026196,50
	Higher education	787	792,43	623639,50
	Total	1816		

Source: Own calculation, data EU-SILC

Ranks Table, “financial burden of housing” – Czech Republic

	study level	N	Mean Rank	Sum of Ranks
financial burden of housing	Lower education	1028	962,94	989905,00
	Higher education	787	836,23	658115,00
	Total	1815		

Source: Own calculation, data EU-SILC

Ranks Table, “head of household education” – Sweden

	study level	N	Mean Rank	Sum of Ranks
head of household education	Lower education	755	466,34	352089,00
	Higher education	192	504,11	96789,00
	Total	947		

Source: Own calculation, data EU-SILC

Ranks Table, “material deprivation” – Sweden

	study level	N	Mean Rank	Sum of Ranks
material deprivation	Lower education	757	480,12	363452,00
	Higher education	193	457,37	88273,00
	Total	950		

Source: Own calculation, data EU-SILC

Ranks Table, “financial burden of housing” – Sweden

	study level	N	Mean Rank	Sum of Ranks
financial burden of housing	Lower education	743	476,24	353845,50
	Higher education	192	436,12	83734,50
	Total	935		

Source: Own calculation, data EU-SILC

Ranks Table, “head of household education” – United Kingdom

	study level	N	Mean Rank	Sum of Ranks
head of household education	Lower education	1114	822,10	915816,00
	Higher education	591	911,25	538549,00
	Total	1705		

Source: Own calculation, data EU-SILC

Ranks Table, “material deprivation” – United Kingdom

	study level	N	Mean Rank	Sum of Ranks
material deprivation	Lower education	1170	929,02	1086959,00
	Higher education	636	856,54	544762,00
	Total	1806		

Source: Own calculation, data EU-SILC

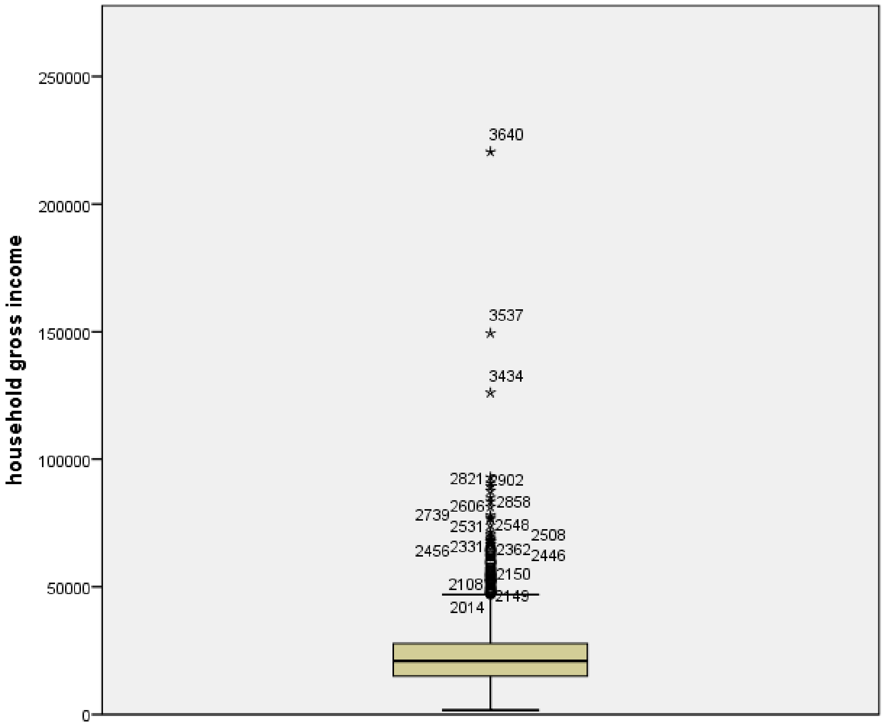
Ranks Table, “financial burden of housing” – United Kingdom

	study level	N	Mean Rank	Sum of Ranks
financial burden of housing	Lower education	1170	935,66	1094725,50
	Higher education	636	844,33	536995,50
	Total	1806		

Source: Own calculation, data EU-SILC

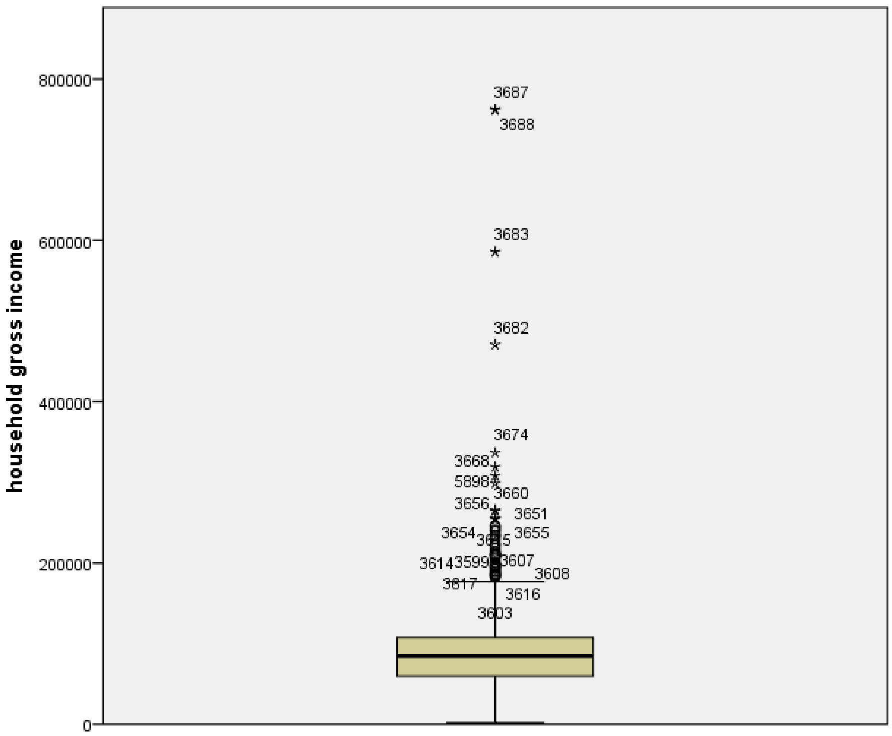
Appendix B: Outliers

Boxplot to detect Outliers – Czech Republic



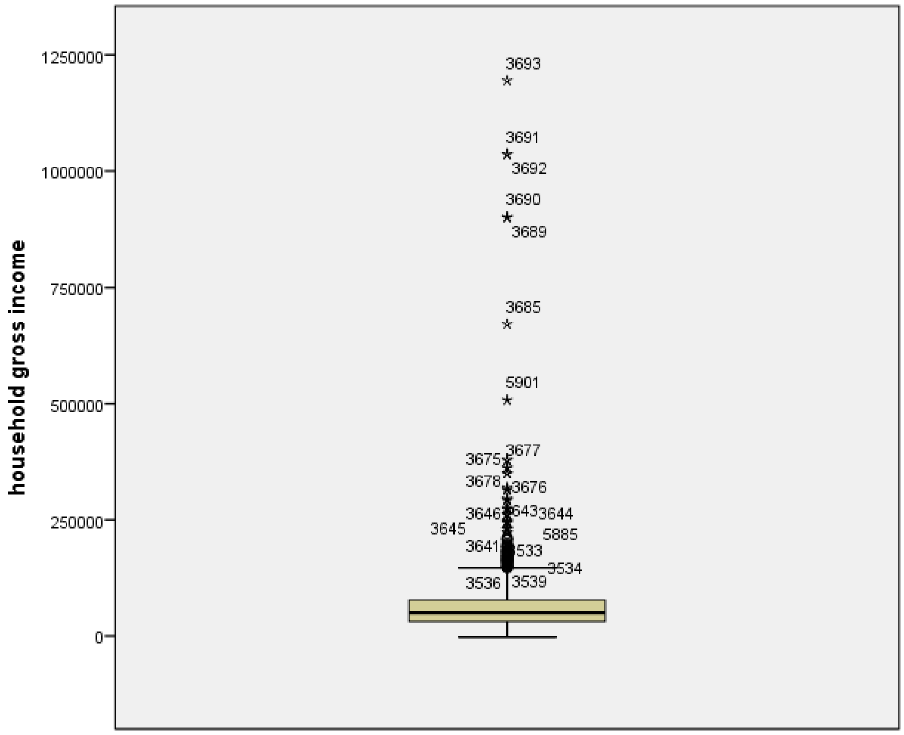
Source: Own calculation, data EU-SILC

Boxplot to detect Outliers – Sweden



Source: Own calculation, data EU-SILC

Boxplot to detect Outliers – United Kingdom



Source: Own calculation, data EU-SILC

Appendix C: Multicollinearity

Collinearity Statistics

Model	Collinearity Statistics	
	Tolerance	VIF
household gross income	,841	1,190
social exclusion benefits	,917	1,090
head of household education	,768	1,302
material deprivation	,665	1,503
financial burden of housing	,738	1,356
occupation	,720	1,388

Source: Own calculation, data EU-SILC

Appendix D: Linearity

Binary Logistic Regression Model including Interaction Terms

	B	S.E.	Wald	df	Sig.	Exp(B)
household gross income	,346	,399	,752	1	,386	1,413
social exclusion benefits	,010	,005	3,690	1	,055	1,010
head of household education	-,197	,338	,340	1	,560	,821
material deprivation	,055	,103	,283	1	,595	1,056
financial burden of housing	,061	,212	,082	1	,775	1,062
occupation	,042	,126	,113	1	,737	1,043
Log household gross income by household gross income	-,064	,161	,159	1	,690	,938
Log social exclusion benefits by social exclusion benefits	-,001	,001	3,651	1	,056	,999
Constant	-2,661	1,039	6,559	1	,010	,070

Source: Own calculation, data EU-SILC

Appendix E: Correlations

Correlations – Spearman Rho

	household gross income	social exclusion benefits	Head of household education	material deprivation	financial burden of housing	Occupation
household gross income	1,000	-,107	,313	-,432	-,420	-,306
social exclusion benefits	-,107	1,000	-,093	,332	,214	,167
head of household education	,313	-,093	1,000	-,210	-,152	-,468
material deprivation	-,432	,332	-,210	1,000	,501	,337
financial burden of housing	-,420	,214	-,152	,501	1,000	,199
occupation	-,306	,167	-,468	,337	,199	1,000

Source: Own calculation, data EU-SILC