FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Master Thesis

Anna Maria Wallner

Thea Nissen

Do living conditions affect first year dropout?

An empirical investigation of dropout from higher education in Denmark during the scholastic year 2016-2017

Supervisor: Miriam Gensowski

ECTS points: 30

Date of submission: 21/12/2018

Keystrokes: 145,384

Abstract

The purpose of this thesis is to investigate the effect of living conditions on first year dropout from higher education in Denmark. For that purpose, a data set containing information on Danish students accepted into institutions of higher education in Denmark in the summer of 2016 is employed. The students are invited to participate in surveys during the the first year, where they respond to questions about their living conditions.

To investigate the effects of living conditions, we rely on three variables that account for different aspects of living conditions. The first variable, *Distance*, measures the distance in minutes from the students home to the institution of higher education he attends. The second, *Worry*, measures how worried the students is about his living conditions on a scale from 1-5. The third variable, *Move*, is a dummy for whether the student moves at the beginning of his first semester. The data from the survey is supplemented with background variables from Statistics Denmark.

With starting point in the data, a Cox proportional hazard model is employed. It allows for investigation of how the variables for living conditions affect dropout, in particular how they change the probability of dropping out. The Cox model is extended to account for the fact that we observe living conditions over time and their value change, i.e. time varying covariates. Further, it is extended to account for ties, i.e. several students are observed to dropout at the same point in time. Finally, observed and unobserved group effects are accounted for by also conducting analyses with a stratified model and a frailty model.

Based on the extended Cox model across the different model specifications, the main finding is that living conditions do affect dropout during the first year. In particular, students with higher values of *Distance* and *Worry* experience a higher probability of dropping out. Students that move at the beginning of the semester are found to have a smaller probability of dropout.

The thesis also investigates regional and sectoral effects of the variables for living conditions. This is done based on an expectation that the effects from living conditions will be strongest in especially the Capital Region and also the Central Region because students in those areas are known to have difficulties finding housing. However, we do not find any clear pattern of evidence for specific regional effects in these regions. In particular, we find that students in the Capital Region, the Central Region and the North Region all have significant effects from *Distance* to dropout, while the effects for students in Region South and Region Zealand are insignificant. For *Worry*, the pattern is the opposite with significant effects in Region South and Region Zealand, but insignificant effects in the other regions. Finally, *Move* appears only to be significantly related to dropout in Region Zealand.

Further, we expect the effects to vary across sectors as they consist of different types of students. For the sectoral effects, we find that students at universities and university colleges experience effects from the variable for *Distance*. Regarding the variable for worries about living conditions, it is generally university and business academy students who have a significant association between the variable and dropout. On the other hand, the results on moving at the beginning of the first semester are less clear.

The importance of accounting for academic and social integration when investigating dropout has been highlighted within the field of sociology and therefore, these factors are also controlled for. While both *Distance* and *Move* are robust to this, the effect from *Worry* disappears indicating that the effect most likely goes through the variables for integration. Finally, we hypothesize that older students with children who live far away from their institution of education do so voluntarily. This is investigated by including an interaction term between a dummy for having children and a dummy for being above age 30 and the variable *Distance*. The interaction terms are insignificant which confirms our hypothesis. However, we note that this conclusion is not very strong as the number of students above 30 with children is very small.

This thesis is the result of work by both of us. We have written the abstract, the introduction, the conclusion and the abstract together. The following parts of the thesis are written by:

Anna: 2.1, 2.3, 3.2, 4.1, 4.4, 5.2, 5.4, 6.2, 6.4, 7.1, 7.3, 7.5.

Thea: 2.2, 3.1, 3.3, 4.2, 4.3, 5.1, 5.3, 6.1, 6.3, 6.5, 7.2, 7.4.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis investigates the relationship between living conditions and first-year dropout from institutions of higher education in Denmark. Below, we introduce the theme and motivate our focus.

The effects from living conditions to first year dropout in Denmark are investigated in this thesis. The debate on student accommodation is typically heated during the summer where the Danish students receive admission offers to institutions of higher education. There are many stories in the media where students claim that a difficult housing situation has them on the verge of dropping out. In the two largest cities in Denmark, Copenhagen and Aarhus, it has been estimated that there was a lack of 8,400 and 4,000 student accommodations, respectively, at the start of the fall semester 2018 (The Danish Construction Association, 2018). These numbers implies that roughly a third of students in Copenhagen or Aarhus would not have a place to stay, which makes it likely that there is some truth in the stories from the students. Although the question of relation between living conditions and dropout is old, few studies have investigated this correlation and not as a main focus. Therefore, we claim to fill a research gap with this thesis as it asks:

1. *Is there an effect from living condition on the choice of dropout from institutions of higher education among Danish first-year students?*

    (a) *Are there regional differences from living conditions?*

(b) *Are there differences across sectors from living conditions?*

This thesis considers first year dropout as it is particularly important because most students drop out during the first year (The Danish Agency for Science and Higher Education, 2018, p. 10). One could imagine, that students who drop out either do not get an education or lengthen the time until they complete another education. Both ways, public expenses increase compared to a situation without dropout. Dropout during the first year in Denmark has been stable for the last ten years at a level where 1 out of 6 students drop out according to The Danish Agency for Science and Higher Education (2018, p. 10). While one can claim that it is an issue that the level has shown persistence, it has happened during a period with a large increase in the number of students admitted into tertiary education. Among the students that drop out, half start another education the year after (The Danish Agency for Science and Higher Education, 2018, p. 10). In relation to that, it is noted that some level of dropout must be expected. Nevertheless, dropout is problematic due to its societal costs. These will be discussed in the following.

It is well known that countries that educate their citizens experience economic advantages. The benefits of education are seen when individuals enter the labor market: first of all, less public expenditure is spent on social welfare programs and secondly, the state will also experience increase revenues through taxes (OECD, 2018, p. 118). These economic benefits motivate the use of public funds to finance education. In Denmark, approximately 11 percent of total government expenditure is spent on education and this is roughly 1 percentage point above the OECD average (OECD, 2018, p. 204). Focusing only on higher education, Denmark spent approximately 2.4 percent of GDP in 2011 which is almost the double of the average across the European Union (Eurostat, 2018). Based on these expenses, it is clear that Denmark is a country that spends a large amount of money on education. Of course, the expenses should be spent in a way that makes the Danish society reap the fruits of education. This leads us to consider dropout from institutions of higher education, which to a large degree must be considered inefficient.

In spite of the magnitude of dropout, there are no estimates of the actual annual costs due to first-year dropout in Denmark. In an analysis produced by FTF (2016), they argue that if the overall dropout increases by 4,000 students annually, equal to a increase of 20 percent, this will come at the cost of 300 million DKK each year. This estimate describes the direct costs and is based on dropout among all students, that is, it also includes the effect from students at master's programs. As this is the only estimate, we do not rely too strongly on it, but merely note that the costs related to dropout are high.

Adding to that, there are indirect costs related to dropout. AU Student Council (2000, p. 3) notes that dropout leads to "considerable" indirect costs but without presenting an estimate. The idea is that a student who starts studying at one program but later decides to dropout, actually takes a spot from a student that could have continued studying and eventually completed the program. A second source of indirect costs arises as transfers from one program to another accounts for approximately a third of the extra time spent studying, i.e. it leads to a large delay (AU Student Council, 2000, p. 3). This means that the point in time where the students enter the labor market and start paying taxes is delayed. Thereby, the state loses tax-income (OECD, 2018, p. 118). This is inefficient from an economic point of view. Finally, there are also individual costs for students such as the foregone earnings compared to if the student had completed an education on time.

After this introduction, Chapter 2 presents the key reasons for dropout in the literature as well as the international and Danish literature that specifically considers the relation between living conditions and dropout. This serves to set the stage for where the research on the topic is and what results can be expected. Chapter 3 presents an economic framework in which living conditions and dropout can be examined. Chapter 4 presents the data and descriptive statistics. The empirical strategy is presented in Chapter 5. We conduct a duration analysis and the chapter presents this in detail and outlines the specifications we add to make it suitable for the data at our disposal. Chapter 6 presents and discusses the results and Chapter 7 discusses to what degree we can answer our research question with the available data and the applied method. Further, the chapter also contains policy recommendations. Finally, Chapter 8 concludes.

Chapter 2

Literature review

This chapter reviews the findings on the dropout phenomena at institutions of higher education and its possible association with students' living conditions. Despite the interest in the causes of dropout in Denmark, few studies have addressed the question of whether there is an association between living conditions and dropout. On the other hand, in the international research, some papers have had a focus on living conditions which is why we begin reviewing this international literature. These are supplemented with an overview of the Danish literature as we study dropout in a Danish setting and despite the descriptive nature of the studies, they can point to local tendencies.

2.1 Key reasons for dropout

The literature on dropout is relatively extensive and various theoretical approaches have been applied, among them the economic approach (Chen, 2008, p. 209). This section presents the main reasons for dropout presented in the literature and connects them to the research question in this thesis. It is noted that living conditions are not among the key reasons. By this thesis, we merely point towards a possible explanation that has not been fully investigated.

In general, the literature suggests that personal and background variables as well as variables related to the life as a student can explain a part of dropout. In particular, there is strong evidence that pre-college preparedness, measured by high school GPA or results from admission tests, is an important explanation (DesJardins, Ahlburg and McCall, 1999; Light and Strayer, 2000). As mentioned, family background is also noted to be of importance, e.g. a family's socioeconomic status seems to be inversely related to dropout. That is, the higher level of education within the family, the lower the risk of dropping out (Tinto, 1975; Ishitani and DesJardins, 2002).

Besides these factors, the literature from sociologist Vincent Tinto (Tinto, 1975) is of great importance as also noted by e.g. Smith and Naylor (2001) and DesJardins et al. (1999). His model highlights the importance of what he refers to as academic and social integration for the dropout decision. The academic and social integration has to do with how well the student integrates into the academic and social systems. The academic system at an institution of higher education is related to formal education of students, including what happens in classrooms, lecture halls and it involves faculty and staff members who engage in the education of students. The social system is related to the daily life ”outside” the academic system, such as the recurring interactions among students, between students and faculty members (Tinto, 2012). With the above findings in mind, we now turn to what literature has found in relation to living conditions.

2.2 International literature on living conditions and dropout

This section focuses on findings in international studies in order to present current knowledge about the relation between living conditions and dropout and motivate the importance of more research on the topic. The structure in this section is as follows; first, the papers and their results will be presented along with the methods applied. Hereafter, the samples will be discussed shortly.

Surprisingly few papers consider the effect of living conditions on dropout. Among the papers that do, living conditions are measured by variables such as living at parental home or whether a student lives on or off campus. While there is evidence that living closer to the institution decreases the risk of dropout, the results are not robust between the papers. Three papers find that the effect of living at parental home or off campus increases the likelihood of first-year dropout significantly (Bozick, 2007; Smith and Naylor, 2001; Gury, 2011). Another paper finds both insignificant and significant increases in the risk of dropout related to a variable for residing in the town where the institution is located in relation to first-year dropout (Lassibille and Navarro Gómez, 2008). Finally, two papers find insignificant effects from living at parental home or off campus in well-specified models (Schudde, 2011; Arulampalam, Naylor and Smith, 2004).

As for the papers with significant results, Bozick (2007) finds that first-year students living with parents or off campus are 41 percent less likely to persists than if they had lived on campus based on logistic regression, while Smith and Naylor (2001) find similar results for first-year dropout among university students based on a binary regression analysis. They estimated that the risk of dropping out is approximately 2-2.5 percent higher for students who live at home compared to students living on campus. For students who lived off campus, but not at parental address, they estimated that these students had a risk of dropping out around 5 percent above students living on campus. Finally, Gury (2011) applies a discrete semi-parametric duration model with time-varying effects and finds that living at home significantly increases the risk of dropout during the first year at the same level as Bozick (2007).

The three papers mentioned in the previous paragraph rely on appropriate statistical methods to analyze student dropout. Nevertheless, only Gury (2011) accounts for the temporal aspect of dropout through a duration model which according to Chen (2008, p. 231) is of importance. Not doing so means that dropout is modeled as a static process, while the duration models accounts for the temporal dimension of dropout. This is discussed in detail in Section 5.1.

Let us now turn to the paper that found did not find a clear pattern of living conditions and dropout. Based on discrete-time hazard models fitted through separate logistic regressions, Lassibille and Navarro Gómez (2008) do not find well-determined effects from living conditions to dropout. The effect depends on the type of institution and model specification; for higher technical schools, living in the institutions home town significantly decreases the risk of dropout for first-year students. The same effect was found when accounting for unobserved heterogeneity. On the other hand, the effect was insignificant for university schools and only significant in a model without unobserved heterogeneity for university faculties. As the living arrangements are measured at the time of entrance at university, a changing living arrangement during the first scholastic year is not modelled which is a point of critique.

Finally, the papers that only find insignificant results in meaningful regressions are based on propensity score matching in a series of logistic regressions (Schudde, 2011) and extreme value, probit and logit models (Arulampalam et al., 2004). The method by Arulampalam et al. (2004) has similarities with the methods applied in the papers that found significant results. Schudde (2011) finds that living on campus decreases the dropout rate during the first year in simple models with no or few explanatory variables, but as more background variables related to prior school results, work and parents background are added, the significance disappears. This highlights the importance of controlling for background variables related to the students' parents and ability. Finally, Arulampalam et al. (2004) find that the effect of living at home and living on campus during the first year are insignificant in explaining dropout from a medical degree at any point in various models, including ones that account for unobserved heterogeneity. All covariates are measured at the start of the first year, so they do not account for changes in the covariates over time. However, it would seem likely that covariates can change over the 5 years when students where analyzed which may have an effect of the results. We note that neither of these papers apply duration-like models, i.e. they do not account for the temporal nature of dropout.

In the introduction of this section, we mentioned that the samples would be presented. It is clear that the papers consider different populations in different countries, but all of them have the strength of having data from quite big samples (especially compared to the Danish literature). The smallest sample consists of 3,500 students, while the largest has more than 76,000 students. Many of the cohorts considered in the samples are relatively old; some consider students who became enrolled at institutions of higher education is the 80's. This is important because the living conditions and dropout students face today may very likely have changed. The newest paper consider a representative sample American students who enrolled in the mid 00's (Schudde, 2011).

To sum up, few researchers have addressed the question of whether living conditions causes student dropout. Among the presented international papers, it has not been established clearly whether their is such an effect. What the papers do find, is that there seems to be an effect from living at the parental home or off campus, but given applied methods there are some deviations. It is noted that no papers find the opposite conclusion, namely that living with parents or off campus during the first year decreases the risk of dropout. The significant results are based on binary regression models (logit and probit) as well as duration analysis, which according to Chen (2008) is most appropriate for analysis of dropout. Although these findings are interesting, the focus in this thesis is Danish students and the Danish educational system. Therefore, we not turn to findings analyzing Danish students.

2.3 Literature on living conditions and dropout in Denmark

This section presents the Danish literature that analyzes the impact of living conditions on dropout from higher educational institutions in Denmark. The Danish educational system may differ from the setups presented in international papers. Further, in the countries considered in the papers, it might be more normal to live at the parental home for a longer period. Finally, we note that living on campus is not that relevant in a Danish setting, but we can consider the variable in terms of what it means to live close to the educational institution. Therefore, it is important also to consider the Danish literature.

The studies presented below have been included as we believe that they are the most important in a Danish context due to their focus, methodology and results. The structure in this section is as follows; first, the papers and their results will be presented along with the methods applied. Hereafter definitions of dropout and samples will be discussed shortly. As will become clear, compared to the international literature, the Danish papers are mainly descriptive and lack causal analysis.

The overall conclusion from the Danish literature, is that there is no general agreement on the impact on living conditions. Interestingly, AU Student Council (2000) find that satisfaction concerning living conditions is not significantly correlated to dropout, while the remaining papers do find an impact from living conditions on dropout (DMA Research, 2002; The Danish Agency for Science and Higher Education, 2018; Hoff and Demirtas, 2009; Holm, Laursen and Winsløw, 2008).

Some of the studies that found significant correlations between living conditions and dropout need to be interpreted with caution since the applied methods are descriptive (DMA Research, 2002; The Danish Agency for Science and Higher Education, 2018; Holm et al., 2008). Holm et al. (2008) and DMA Research (2002) used questionnaires and interviews with students who had dropped out. Holm et al. (2008) found that out of the 18 mathematics students who completed the interviews, 3 reported that one of the reasons behind their dropout decision was related to long transportation time and/or difficulties in combining studying with family life. The results from DMA Research (2002) were more general as they found that an unsettled living situation (meaning living far from the institution, the bad housing conditions or lack of time to focus on studying) is a reason for dropout in Copenhagen, but not in Aarhus.

In resemblance to both Holm et al. (2008) and DMA Research (2002), the study population in The Danish Agency for Science and Higher Education (2018) consists of the group of students who had dropped out. The students were asked to state the reason for the dropout decision and 18 percent replied that a ”too long” transportation time was a reason for dropout. High transportation costs, bad location of housing, the general condition of the housing and the inability to find permanent housing are also mentioned as reasons for dropout. Distance was also present as an explanatory factor in the study by Hoff and Demirtas (2009) and based on logistic regression, their findings are that ethnic students who live at their parental home are 2.6 times more likely to drop out. The results presented by AU Student Council (2000) and Hoff and Demirtas (2009) are interesting given that they use logistic regression which lies relatively close to the method applied later in this thesis. However, the papers can be criticized as e.g. they do not present standard errors for their estimated effects.

With the results presented, we shortly notice the difference in the reporting of dropout, applied data and methods. First of all, we note that the papers differ in whether the dropout is self-reported by the students or based on registration made by the institutions. This is important in an analysis that accounts for the timing of dropout as a student might have been inactive for some time before it is registered by the institution. Further, the sample sizes vary substantially from 23 to almost 4000 students. All papers have a clear focus on dropout in a Danish context, but they differ in the population that is analyze. One paper has a narrow focus and analyzes students enrolled at the bachelor's programme of Mathematics at University of Copenhagen (Holm et al., 2008), while another study focuses on ethnic minority students at the 5 largest Danish universities (Hoff and Demirtas, 2009).

The remaining three papers have a broader focus and consider either all students who started at Aarhus University at a certain time (AU Student Council (2000)), or students from three different programs at three different universities (DMA Research, 2002) and finally, dropouts from all institutions of higher education from 2015-2016 (The Danish Agency for Science and Higher Education, 2018). In other words, the populations considered vary in size and representativity of the entire body of students. Similar to the population in our thesis, it is noticed that The Danish Agency for Science and Higher Education (2018) as the only paper considers a population outside the universities by including University Colleges and Business Academies.

As a natural extension of the variation in the number of students, the applied methods differ so that the smaller studies rely on e.g. group interviews, while the larger studies draw conclusions based on larger surveys and register data. Finally, it is worth noticing the difference in when dropout occurs. In accordance with the focus in this master's thesis, (The Danish Agency for Science and Higher Education, 2018; AU Student Council, 2000) analyzes which factors seem to have an impact on first-year dropout, while Holm et al., 2008 considers "the first couple of years", DMA Research, 2002 considers dropout before finishing the BA and Hoff and Demirtas, 2009 during the bachelor's or master's program.

To sum up on the Danish literature on living conditions and dropout, the studies are quite heterogeneous both in sample size, study population and method applied. Nevertheless, most of the papers present results that may suggest a relationship from especially transportation time to dropout. The results are primarily drawn based on university students. Regarding the methods, most applied methods are of descriptive character which limits causal interpretation. Unfortunately, the papers are generally not peer-reviewed. The lack of significant conclusions and correct statistical approaches, underlines the research gap we aim to fill with this thesis.

Chapter 3

Economic framework

While there are numerous theories concerning the demand for human capital, the economic theories on dropout are relatively limited and to our knowledge, none of these incorporate the effects of living conditions. Therefore, this section will present a simple economic intuition that the later analysis can build on. First, the ideal identification and ideal framework are presented. Hereafter, in order to put the intuition into a meaningful framework, the empirical model is shortly presented. This is followed by a thorough discussion of what effects that can be expected from the variables for living conditions and the control variables that will be used in the thesis.

3.1 Ideal identification

In the best of worlds, this thesis would rely on a randomized experiment in order to causally determine the effect from living conditions to dropout. With such a setting, students would be assigned different living conditions randomly. As examples, some students would be given housing very close to the educational institutions, others further away and some students would be offered housing from the beginning of the semester, while others would not and so on. If the students are assigned these different living conditions randomly, then the living conditions should be the only thing to vary systematically between the students. Of course, the students should not be able to reject the living conditions they are offered or to search for housing themselves, because that would lead to a situation with self-selection.

While the described experiment would allow for causal analysis, it is not possible to conduct it for obvious ethical and practical reasons. Students cannot be forced to accept the living conditions that would be imposed on them under this experiment. Instead, we have to find away around the concern that living conditions among students are most likely not random. Also, living conditions are possibly an outcome of individual characteristics, that is, how good each student is at finding a place to stay. One possible strategy to solve this is to include relevant and sufficient controls to provide a setting where we can control for the fact that living conditions are not random.

3.2 Ideal economic framework

In this section, we consider what we would want from an economic model that investigates the relation between living conditions and dropout. This should serve as a starting point for the later analysis in highlighting important features and important variables.

An optimal economic framework, i.e. ”a model that is faithful to the evidence” (Cunha and Heckman, 2007), to study the effects from living conditions on dropout should account for the following:

1. Most students drop out during the first year, so this is an interesting period to consider (The Danish Agency for Science and Higher Education, 2018). This is accounted for as we only consider first year dropout.
2. Personal and family background should be controlled for as emphasized in Chapter 2. Such variables will be incorporated into the model and the expected effects will be presented and discussed in the following subsection.
3. Time should be incorporated into the model. Dropout is not static, but a result of conditions that may vary over time. This motivates the use of time-varying covariates when possible. Further, one could argue that the effect of the variables should be

allowed to vary over time, but this is not incorporated as we have few observations over time and they are measured imprecisely.

4. Based on the different conclusions in the literature on Denmark presented in Section 2.3, there might be differences across sectors and regions. This will also be taken into account through analyses with focus on how the variables for living conditions vary across the sectors and regions.

The above features will be incorporated into the empirical model presented shortly in the following section and in detail in Chapter 5.

3.3 Discussion of variables

This section presents the explanatory variables employed in the thesis and discusses how they can be expected to affect dropout. In order to discuss the variables, we shortly present the empirical model in which they are included. Hereafter, the key variables for living conditions are considered and finally, we consider the background variables.

For the empirical estimation, we rely on estimation of equation (3.1). The empirical model behind the equation is introduced in detail in Chapter 5. For now, we note that it estimates the hazard function, which is defined as the instantaneous probability of dropping out. On the right hand side, it contains $\lambda_0(t)$, which is a baseline hazard that is common to all students and will not be estimated. Further, explanatory variables that will be presented in this chapter are included. That way, it can be investigated how the covariates affect the overall hazard of dropping out.

$$\begin{aligned}\lambda(t|\mathbf{x},\beta) = \lambda_0(t)\exp(&\beta_1\text{Female} + \beta_2\text{Parental education} + \beta_3\text{High School GPA}\\ &+\beta_4\text{Age} + \beta_5\text{Age}^2 + \beta_6\text{Distance}_t + \beta_7\text{Move} + \beta_8\text{Worry}_t + \beta_9\text{Dum\_dist}_t)\end{aligned} \tag{3.1}$$

As seen from Equation 3.1, the following variables for living conditions are included: *Distance*, which measures distance in minutes from the student's residential address to the educational institution, *Worry*, which is a scale that measures worries concerning living situation and *Move* which is a dummy for whether the student moves at the beginning of the semester. These are expected to influence the dropout decision along with personal and family background variables, in particular gender, parental education, high school GPA and age as can be seen from the equation.

One would expect a positive association between distance and the probability of dropping out during the first year as well as a positive association between how worried a student is and his probability of dropping out. For *Distance*, the intuition is the following: if one thinks of student's available time as constrained, then spending more time on transportation should, *ceteris paribus*, reduce time studying. As for the variable *Worry*, the intuition is that a student who is more worried, should, *ceteris paribus*, have less mental energy and focus to absorb the syllabus compared to a student who is not concerned, also such a student will spend time being worried instead of studying.

Move, the third variable that controls for living conditions, could be thought of the student adjusting his living conditions to student life. This variable is expected to have a negative association with dropout. The idea is that a student that moves should have better possibilities to be part of the environment at the institution because the student does not have to spend time searching for housing and further, the student has likely moved closer to the institution.

As mentioned, we believe that students' living conditions are unlikely to be random; they exhibit selection effects. By including relevant control variables, we attempt to control for these selection effects. If e.g. *Distance* was the only explanatory variable determining the risk of dropout, there could be a concern that what actually drives dropout and residential accommodation is related to e.g. family background. Thereby, the estimated hazard would not be the true effect from the variable. The included control variables are gender, parental education, high school GPA and age. We discuss what we expect of these variables in the the following paragraphs.

For parental education, the intuition is that if your parents have a high level of education, you might be raised in a different way making it easier for you to continue studying and further, it might be an expectation for you to complete your own higher education. There might also be a role model effect; you see what your parents lives are like and decide if you want to pursue something similar. As a measure of how well the students have done in high school, we expect student with higher high school GPA to be less likely to drop out. Further, we expect students with educated parents and higher high school GPA to have less problems with their living conditions. Intuitively, since earnings typically increase with educational level, well-educated parents are believed to have more resources to economically support their children. One example could be that these parents can buy apartments for their children or contribute to the rent. Secondly, more able students might also struggle less to find housing because they have a larger network and have more energy to search for a place to stay. In other words, these variables are negatively correlated with both housing problems and dropout which underlines the importance of controlling for these variables to answer our research question.

In relation to gender, we would expect that women are less likely to dropout as this has been pointed towards in the papers presented in Section 2.2. The students' age is expected to be of importance as older students typically are more mature, i.e. determined and aware of their skills, which matters for staying enrolled. On the other hand, if a student reaches a certain age and has not completed a tertiary education yet, this might be for a reason. This non-log-linear effect is accounted for by including age as a squared term.

Besides the effects that can be investigated from Equation 3.1, there are other important aspects of relation between living conditions and dropout that will be investigated. First of all, there might be heterogeneity in the group of students that lives far away from the educational institutions. In particular, we expect older students with children to be settled, while younger students might be eager to move closer to the educational institutions. This is partly controlled for by the variable age. However, that might not be a sufficient control to investigate this particular issue. Therefore, we conduct an additional analysis, where we consider if the student is above age 30 or has children. In such a case, we would expect the effect of distance to be smaller or insignificant compared to the settled students.

As motivated in Section 2.1, it is potentially important to investigate if academic and social integration drives the dropout decision, controlled for personal characteristics and family background. The idea is that the variables for living conditions merely affect how academically and socially integrated the students can become, which is then what determines dropout. We would expect academic and social integration to be negatively correlated to dropout, so if a student is more integrated, his risk of dropout decreases. As a last comment, the variable *Dum_dist* is included for technical reasons that will be presented in Chapter 4.

To sum up, both variables for living conditions and control variables are expected to affect dropout in different ways a described above. The variables are presented in further detail in Chapter 4.

Chapter 4

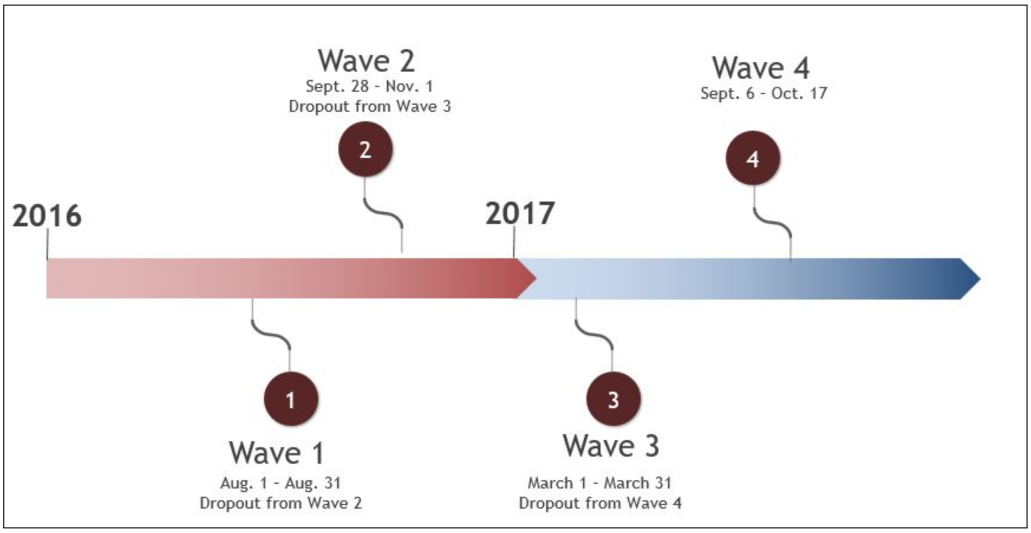
Data

In this chapter, we present the data used in the analysis and highlight the most important issues related to it. The starting point is the applied survey data which is supplemented with register data from Statistics Denmark. First, the data sources are presented, hereafter we present descriptive statistics and data management. Finally, the main challenges related to survey data are presented.

4.1 The data sources

The applied data consists of a combination of survey data and register data on individual level. The survey data was collected through 4 waves (i.e. 4 points in time) by the Danish Evaluation Institute (EVA) during the academic year 2016-2017. All first-year students who received an admission offer in July 2016 to a program of higher education were invited to participate in the survey. The survey questions were related to many aspects of the students' lives, among these living conditions and dropout. The responses concerning living conditions from the 3 first waves were used and connected to dropout in waves 2-4, cf. Figure 4.1. This means the variable for dropout was lagged one period, which is motivated later in the chapter. Students who responded in the first wave were invited to participate in all subsequent waves. Figure 4.1 shows when students participated in the surveys at each wave and that they could respond to the surveys within the given time intervals.

Figure 4.1: Timeline of survey questions



The second data source is Statistics Denmark. The retrieved register data includes information on dropout during the first year, as well as personal characteristics and family background of students.

4.2 Descriptive statistics

Even though around 60,000 students were admitted into an institution of higher education in the summer of 2016, the population we consider is smaller since we excluded international students, students who enrolled in a Master's program, student who transfer to another program at the beginning of the semester, were on a waiting list or were admitted in the summer of 2016 to start in the beginning of 2017. This allows us to analyze first-year dropout among Danish students. The final population, hereinafter population, consists of 44,496 of which 40,826 had background data such as gender, parents education and high school GPA that made them suitable for analysis.

Of these 40,826, around half replied to the survey in the first wave, which means the sample consists of 19,032 students. The analysis is restricted to students with Danish citizenship because of two reasons. First, we believe that these students have a general knowledge of the Danish education system. The second reason is related to financing of the education. Students with Danish citizenship are in general affected by the same rules regarding entitlement to educational grant and student loans.

Table 4.1 summarizes the variables of the sample and compares them to the population. Further, the table also serves to investigate the representativity of the sample over the waves, which will be discussed later in this section. Participation in the survey was voluntary and as the table shows, the actual number of respondents in each wave was lower than the potential. The table considers the participation in the 3 waves separately, which means that the number of dropouts for wave 3 is not the accumulated number, but merely the number of students that drop out in that particular wave. Due to the data structure, the dropout variable variable is lagged one period as presented in Figure 4.1 and discussed later in the chapter. This means that the dropout presented for e.g. wave 3 actually took place in wave 4.

The three variables for living conditions are *Distance*, which is the distance in minutes from the student's home to the institution of higher education, *Move*, which is a binary variable that takes the value 1 if the student reports to have moved at the beginning of the semester (i.e. during wave 1 or wave 2, see Figure 4.1) and *Worry*, a continuous variable describing the student's concerns regarding his living situation. The scale is increasing with the level of worries and takes the following values: 1) "Not at all", 2) "To a lesser degree", 3) "Do not know" 4) "To some degree", 5) "To a very large degree". Based on the table, the students are generally not worried, spend around 35 minutes on transportation each way and more than 1/3 moved at the beginning of the semester. It is noted that the sample changes over time and all of the variables for living conditions "improve" with time. That is, the average respondent is less worried, lives closer to the institution at which he studies and a larger proportion of those that respond did move at the beginning of the semester.

The table also presents the control variables *Female*, a dummy for gender, *Age*, a continuous variable for age, *High school GPA*, ranking from 2 to 13.7 and finally, *Parental education*, which is included as a dummy for the different educational levels. We see that

Table 4.1: Descriptive statistics of variables

	Population	**Wave 1**	**Wave 2**	**Wave 3**
Number of students	40,826	19,032	11,538	8,016
Active students among respondents	-	18,854	10,938	7,510
Dropouts among respondents	-	178 (0.9 %)	600 (5.2 %)	506 (6.3 %)
Living conditions				
Average worry (scale: 1-5)	-	1.9	1.8	1.6
Average distance (minutes)	-	37.3	36.3	32.7
Move	-	36.8 %	39.0 %	40.42 %
Individual characteristics				
Female	53.7 %	58.6 %*	63.8 %*	66.2 %*
Average years of age	21.7	21.9*	22.1*	22.2*
High school GPA	7.8	8.0*	8.3*	8.5*
Parental education				
Primary and secondary education	10.8 %	10.5 %	10.2 %	10.0 %*
Vocational education	34.1 %	34.1 %	32.9 %*	32.2 %*
Short-term higher education	6.5 %	6.5 %	6.5 %	6.5 %
Medium-term higher education	28.9 %	28.9 %	29.4 %	29.7 %
Long-term higher education	19.7 %	20.0 %	20.9 %	21.6 %
Higher education institution				
University	59.5 %	59.7 %	62.9 %*	64.9 %*
University college	24.6 %	24.9 %	23.8 %	23.1 %
Business academy	14.1 %	13.7 %	11.6 %*	10.4 %*
Geographic location of institution				
Capital Region	36.0 %	34.8 %*	35.5 %	35.8 %
Central Region	25.6 %	26.8 %*	27.3 %*	28.3 %*
North Region	12.7 %	12.2 %	11.7 %	11.2 %*
Region Zealand	7.7 %	7.3 %	7.2 %	6.8 %
Region of South Denmark	18.0 %	18.9 %	18.4 %	17.9 %

* The sample value is significantly different from the population value.

Note: The sectors do not sum to 100 % as maritime and artistic educations are excluded.

Source: EVA and Statistics Denmark.

mainly female students in their early 20's who have above average high school GPA and parents with vocational educations make up a large part of the sample. It can also be seen that university students by far make up the largest part of the respondents. Also, while the share of university student increases with time, the share of university college students is roughly constant and the share of business academy students decreases. As for the regions, it is noted that the variable shows where the educational institutions that the students are enrolled in are located. These shares are roughly constant over time. The largest part of the respondents attend institutions in the the Capital Region and the

Central Region, which is not a surprise as these regions are home to many institutions of higher education. As a final comment on the variables in the table, Table A.6 in the appendix compares the average values for an active student and a dropout and we note that they are surprisingly similar.

As mentioned, Table 4.1 also contains information on the representativity of the samples. The comparison is made for the averages of the variables and tested through either a t-test for continuous variables or a chi-squared test for categorical variables. The asterisk * shows for what variables the samples are significantly different from the population. This is the case for all the individual characteristics, which might be because a certain type of students, who is different from the average student in the population, responds to the surveys. Besides that, the response rate from students studying in the Central Region increases with time, making this proportion of responds significantly larger than it is in the population. These differences can be problematic when determining if the results can be extrapolated. This will be taken into account throughout the paper and discussed further in Chapter 7.

Finally, as mentioned in Chapter 3, the relation between living conditions, academic and social integration and dropout will be investigated in this thesis. We take starting point in data on integration created by EVA. The indices were created for social and academic integration based on specific question carried out in wave 2, 3 and 4. As for academic integration, the questions were ”I try to create coherence between the things I learn at the different classes at my education”, ”I try to do a bit more at my education than what is asked of me” and ”I use the feedback I get to improve”. For social integration, the questions asked were ”I feel welcome at the study”, ”The other students are generally obliging”, ”I generally feel I am on the same wavelength as the other students”. Based on these, the indexes ranging from 1-3, where a higher value means more integrated, are created.

4.2.1 Distribution of dropout during the first year

As was mentioned briefly in the introduction, the empirical strategy in this thesis is based on a duration analysis. Although the empirical strategy is described in detail in Chapter 5, we use a part of if now: non-parametric estimation. The reason is that non-parametric analysis is a descriptive tool. This section begins with a brief explanation of the method and follows by a figure that depicts the dropouts in our sample.

We use the Kaplan-Meier estimator which is an estimate of the survivor function, $S(t)$, or equivalently, the probability of dropping out after t. As will become clear from the graph below, the estimator is a decreasing step-function with a discrete jump at each wave. If students are observed at $t_1, ..., t_k$ times and k represents the number of distinct dropout times, then the Kaplan-Meier estimate at time t is given by:

$$\hat{S}(t) = \prod_{j|t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right), \tag{4.1}$$

where n_j is the number of individuals at risk at time t_j and d_j is the number of dropout at time t_j. In this analysis, n_j is the total number of survey respondents at that point in time. The Kaplan-Meier estimates are presented in Figure 4.2.

FIGURE 4.2: Distribution of survival probabilities over time

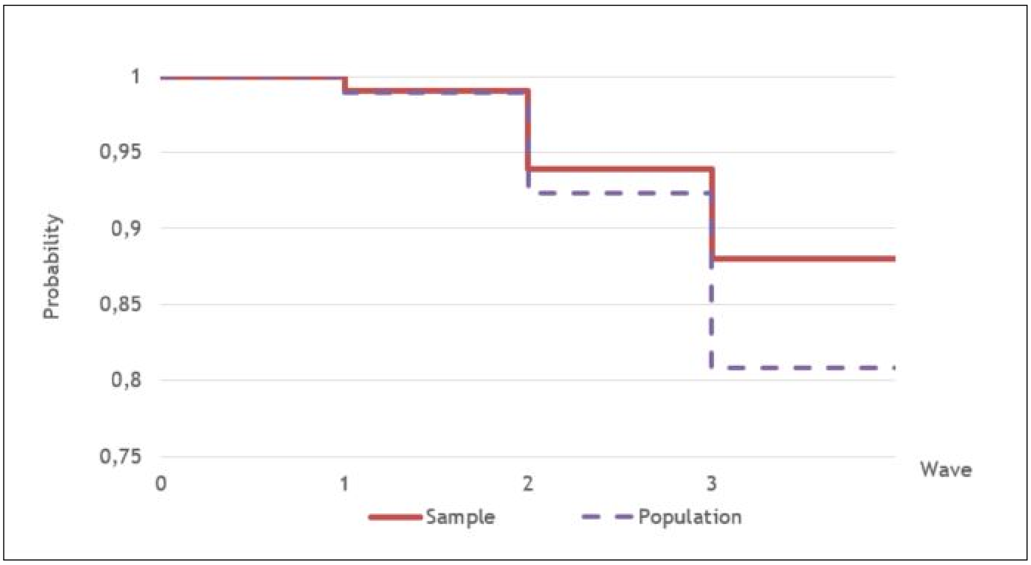


Figure 4.2 shows how the probability of remaining enrolled is distributed with our data on first-year Danish students. In the figure, two curves are depicted that give the estimates for students in our sample and the population, respectively. This shows that the level of dropout in the sample is not as high as in the population. As an example, the probability of remaining enrolled in wave 2 is approximately 94 percent for students in our sample, compared to 92 percent in the population. The difference in dropout between the two groups will be accounted for later in this thesis.

4.3 Data management of the sample

The sample presented in Table 4.1 has been modified to facilitate the subsequent analysis. The variable *Distance* was restricted to have a maximum value of 200 minutes per way. This is more than the time it takes to go by train from Copenhagen to Aarhus, which we believe is a unrealistic long transportation time. Self-reported values above this maximum were replaced by the average distance for students reporting distances below the maximum. These students were given a dummy variable, *Dum_dist*, indicating that their value of distance had been replaced by the average, which was included in the regressions.

As mentioned, the dependent variable *Dropout* has also been modified. When a student drops out, he is given a different survey and therefore, there are only active students with information on living conditions in each wave. This means that the regressions are only run for active students, which does not make sense. Therefore, *Dropout* is lagged one period. This can also be argued to be meaningful to ensure the causal order. From an economic point of view, one can argue that the situation described by students in e.g. wave 1 affects the choice of dropout in the following wave. This is discussed further in Chapter 7.

Further, the variable for dropout is a mixture of information from the surveys and from Statistics Denmark. This is because the survey data is based on self-reported dropout from the respondents. In case of students only responding in one wave and dropping out in the next without responding to the survey, we would have a missing data problem. As an example, student i reports being active in the second wave but does not respond to the survey in the following wave where he drops out. To circumvent the missing data problem, we rely on data from Statistics Denmark to determine his status. To be clear, we only rely on dropout data from Statistics Denmark, when the survey data is not available. If the student had remained active and simply had not responded to the survey, he would simply have left the sample.

4.4 Empirical challenges

This section presents the challenges related to the population and the sample. These will be taken into account throughout the thesis and discussed in detail in Chapter 7. There are two challenges related to the data set. The first is that the sample is unlikely to be random and thereby representative of the population due to the voluntary participation in the surveys and attrition. The second is that time is only observed in intervals. Finally, we shortly discuss the validity of survey data.

As for the first issue, it arises because participation in the survey is voluntary, which is likely to lead to self-selection into the sample. As can also be seen from Table 4.1, the characteristics of the sample differ more and more from the characteristics of the population with time, i.e. there are issues with attrition that are likely to make the sample in the third wave less representative than the sample in the first wave.

There are several factors that can affect self-selection into the survey as well as attrition. For one thing, it is noted that all the surveys are relatively long, which probably can explain a large degree of the non-response. Further, there are several surveys to respond to during the year and we must expect those that respond to be relatively patient. On the other hand, the surveys in question are carried out online, which gives the students more time to respond compared to e.g. a survey carried out on the phone. Further, the students are motivated to respond as they could win 1000 DKK by participating in each survey and finally, the respondents are anonymous which might increase the response rate. The non-representativity is discussed further in the next subsection that suggests weighting as a solution.

The second issue is that the students are only observed in few waves and as a result, the time of dropout is not known exactly, it is only known to lie in a given interval. This is also referred to as interval-censored data (Cameron and Trivedi, 2005, p. 588). These could be an argument for either considering a model that accounts for this interval-censoring. Another option is to consider time as being discrete. These two are taken into account during the analysis and discussed in detail in Chapter 7.

Finally, we note that as a large part of the data comes from surveys, this generates some other challenges than register data. One challenge is that between the respondents there might be large differences in their perception of the questions. Especially for a variable such as *Worry*, where the students are asked to indicate how worried they are on a scale from 1-5, the perception is important. In other words, the survey data is to some degree subjective. Further, the response given by the students must be trusted and there are a few examples related to e.g. *Distance* that are not realistic. We believe it unlikely that is a problem as students will not spend time on providing wrong answers. Finally, we note that there are also advantages related to the use of survey data. For example, it can provide information that can not be found in the registers such as information on how worried a student is.

4.4.1 Weighting

This subsection discusses weighting as a potential solution to the non-randomness in the sample noted above. In particular, it is investigated if weighting changes the significance or the estimated effects, which turns out not be the case. However, if it had been so, it would point towards the survey sample being to different from the population to extrapolate the obtained results. In the following, the applied weighting procedure is presented and discussed.

The idea of implementing inverse probability weighting is to make a potentially non-representative sample representative of the population. In this thesis, the cause of non-representativity of the initial sample is believed to be non-response to the survey as discussed above. Two solutions could be argued for in this case of missing data: multiple imputation and weighting strategies (Chambers, 2003, p. 278). However, we use weighting since multiple imputation should not be implemented if more than 50 pct. of information is missing which is the case with our data (Höfler, Pfister, Lieb and Wittchen, 2005).

The method works by weighting students by the inverse of the probability of responding to the survey in the first wave. This implies that students who are over-represented in the sample relative to the population, are given a lower weight in the sample (Wang and Aban, 2015). In practice, weights can be obtained from a model where the response indicator (i.e. an indicator for whether they respond) is the outcome variable and the regressors are available information on the entire population (Höfler et al., 2005, p. 293). We apply logistic regression to estimate the weights as it typically yields the smoothest fit to data (McCullagh and Nelder, 1989). For further details of the method, we refer to Appendix A.3.

The advantage of the approach is that it may correct for non-response bias if the weights are correct in the sense that they give larger weights to the respondents who are underrepresented in the sample relative to the population. On the other hand, a drawback is that weighting may yield estimators with high variance. This is due to the fact that respondents with very low estimated response probability, receive large non-response weights and may be to influential in estimates of means and totals (Little and Rubin, 2002, p. 49). A final drawback is the risk of model misspecification. The strengths and weaknesses captures the issues of bias and efficiency: in certain senses the use of survey weights may reduce bias but also reduce efficiency (Chambers, 2003, p. 83).

As was mentioned in the beginning, weighting was applied to the sample and the effects were estimated for the sample, cf. Appendix A.3. Weighting did not change the estimated effects nor their significance. Therefore, we rely on the non-weighted sample in the analysis. This will be discussed further in Chapter 7.

Chapter 5

Empirical strategy

This chapter presents the main econometric model used in this master thesis, namely the Cox proportional hazard (PH) model. Before it is presented, we will motivate the use of this duration model and present the theory behind semi-parametric duration analysis. Hereafter, we present the extended Cox model, which accounts for data specific features and add further extensions to account for group effects. Finally, we discuss assumptions for the model. The empirical framework presented here is the foundation for the results presented in Chapter 6.

5.1 Motivation for duration analysis

In order to analyze the association between living conditions and first-year dropout, we use a duration model. This model is chosen because it allows us to follow students over time and their transition from being active students to dropouts (DesJardins, 2003). The main advantage of using a duration model compared to running separate logistic regressions for each wave, is how the former method incorporates time, leading to more accurate results. In what follows, we will explain why duration analysis is preferred in our set-up.

Let us assume that students are observed several times and can potentially drop out at each of these points in time, as in the case for the applied data. By using a logistic regression model, one could perform an alternative duration analysis where the students are followed until they drop out. As an example, if the interest lies in estimating the risk of dropout at the second observation time, one would condition on students who were observed at that point as being active or having dropped out at that point. That way, separate analyses could be conducted for each point in time. In this alternative duration analysis, however, one would ignore that some students drop out before and some drop out later. Clearly, that approach is not the most optimal since it does not use all available information on students. This is where the duration model is preferred relative to separate logistic regressions as it uses all information.

A second alternative method could be to use a pooled logistic regression using information from all dropout times, i.e. a static approach. However, this approach would not account for the temporal dimension, i.e. it would not account for the fact that students drop out over time. This is important because it decreases the number of students that are at risk of dropping out and this information is valuable. Again, ignoring this by using a pooled logistic leads to a situation where all available information is not employed.

With the above arguments in mind, we rely on the the duration model in this thesis. Thereby, we can estimate the students' risk of dropping out and incorporate time which returns more accurate results. This will be explained in further detail in the subsequent sections.

5.2 The conditional hazard function

A key concept in duration analysis is the conditional hazard function, $\lambda(t|\mathbf{x}(t))$. It defines the instantaneous probability of dropping out conditional both on having been an active student to time t and the covariates of the students, $\mathbf{x}(t)$ (Cameron and Trivedi, 2005, p. 599). The conditional hazard function for a student observed at time t is given by:

$$\lambda(t|\mathbf{x}(t)) = \underset{\Delta t \to 0}{lim} \frac{Pr[t \leq T < t + \Delta t)|\mathbf{x}(t), T \geq t]}{\Delta t}. \tag{5.1}$$

The hazard function can vary from zero, meaning no risk, to infinity where there is certainty of failure at every instant. If t is the length of time in school, measured in waves, then $\lambda(2)$ is (approximately) the probability of dropping out between time 2 and 3, i.e. the risk of dropping out between these periods conditional on having stayed enrolled until then (Wooldridge, 2010, p. 985). The advantage of talking of hazard functions rather than the traditional density and cumulative density functions is that hazard functions give a more natural way to interpret the process that generates dropout and regression models for duration data are more easily grasped by observing how covariates affect the hazard (Cleves, Gould, Gutierrez and Marchenko, 2010, p. 13).

5.3 Semi-parametric duration modelling

We apply a semi-parametric model and this choice is based on Figure 4.2, which described the distribution of dropouts in our sample. The figure demonstrated that the risk of dropout decreases over time. However, with only 3 dropout points in time during one year, we argue to have too little information to determine the baseline hazard function, which is an important component of the model. It is modelled in a parametric duration model, but left unspecified in the semi-parametric setting. The semi-parametric model does not require a specific distribution of dropout time and thereby, we avoid possible misspecification. The specific model is the Cox proportional hazard model which is introduced below.

5.3.1 The Cox proportional hazard model

In order to investigate how living conditions affect dropout, the Cox PH model is used. This model has proven to be very useful within duration analysis, such that is has become the standard method for survival data (Cameron and Trivedi, 2005, p. 593). For the applied data, however, the model must be extended to account for ties and time-varying covariates. These challenges will be discussed after the presentation of the model.

The model is given by equation (5.2), from which it is clear that the conditional hazard rate is the product of two functions. The first function is the baseline hazard, $\lambda_0(t)$, which only depends on time and the second is a person-specific and non-negative function that accounts for the effect of the individual characteristics, $\phi(\mathbf{x}(t), \beta)$ on the conditional hazard. The non-negativity is important as dropouts cannot unhappen and with ϕ parameterized in the exponential form, this is ensured. As mentioned, the baseline hazard function is not estimated in the Cox model.

$$\lambda(t|\mathbf{x}(t), \beta) = \lambda_0(t) \exp(\mathbf{x}(t), \beta) \tag{5.2}$$

The baseline hazard function is the risk of dropping out for (hypothetical) individuals with $\mathbf{x}(t) = 0$, that is, when only time is controlled for. These students serve as a reference group and the expression $\exp(\mathbf{x}(t), \beta)$ is an adjustment that depends on the characteristics $\mathbf{x}$. The name, "proportional hazard", is related to the assumption underlying the model, namely that the risk is proportional between individuals. What this assumption essentially states, is that the effect from covariates is constant over time. Let us assume that we are comparing two students who only differ in one dimension, e.g. their high school GPA. At time t, their difference in dropout probability will then be caught by β_{grade}.

As mentioned, the standard Cox model is extended so that it can model ties and time-varying covariates. Ties are defined as two or more students dropping out at the same time. We observe ties in our data because students are only observed at three times during the first year, instead of e.g. on a daily or monthly basis. A second feature of our data, is that some covariates vary over time. Treating these variables as time-invariant would be a clear misspecification and would mean excluding data from the model.

Time-varying covariates are included in equation (5.2) through the vector $\mathbf{x}(t)$ and as a restriction, only the current value of the covariates matters, rather than the entire history (Cameron and Trivedi, 2005, p. 599). Further, it is assumed that the time-varying covariates are strictly exogenous meaning that the variables are not allowed to "exhibit feedback" on the outcome variable (Cameron and Trivedi, 2005, p. 598), an issue discussed in Chapter 7. Finally, in the model, the underlying time is assumed to be continuous although students are observed at discrete points in time. Discrete data is also referred to as interval-censoring within the literature on duration modelling and will be discussed later.

5.3.2 Partial likelihood estimation

The model presented in the previous section is fitted using partial likelihood estimation. The partial likelihood estimation estimates the β's without requiring simultaneous estimation of $\lambda_0(t)$, i.e. it is a limited information likelihood. It has been shown mathematically, that the parameter estimates obtained from partial likelihood estimation have the same distributional properties as full maximum likelihood estimators (Hosmer, Lemeshow and May, 2011, p. 74).

For simplicity, we first present the log-likelihood for data without ties and hereafter, we present the approximation used to account for ties. It is assumed in the estimation that the sample consists of individuals that are independent of each other (Hosmer et al., 2011, p. 72). To derive the log-likelihood function, the risk set, $R(t_j)$, is defined as the set of students who are at risk of dropping out before the jth ordered failure. By ordered failures we note that registration of dropouts is ordered with the first wave, second and third wave. With that in place, the probability that a dropout occurs at time t_j, or the likelihood contribution, is given by:

$$\begin{aligned}
Pr[T_j = t_j | R(t_j)] &= \frac{Pr[T_j = t_j | T_j \geq t_j]}{\sum_{l \in R(t_j)} Pr[T_l = t_l | T_l \geq t_l]} \\
&= \frac{\lambda_j(t_j | \mathbf{x}_j(t_j), \beta)}{\sum_{l \in R(t_j)} \lambda_l(t_j | \mathbf{x}_l(t_j), \beta)} \\
&= \frac{\lambda_0(t) \exp(\mathbf{x}_j(t_j), \beta)}{\sum_{l \in R(t_j)} \lambda_0(t) \exp(\mathbf{x}_l(t_j), \beta)} \\
&= \frac{\exp(\mathbf{x}_j(t_j), \beta)}{\sum_{l \in R(t_j)} \exp(\mathbf{x}_l(t_j), \beta)} \qquad (5.3)
\end{aligned}$$

The vector $\mathbf{x}_j(t_j)$ denotes the value of the covariates for the subject with the ordered dropout time t_j. From equation (5.3), we note that the baseline hazard rate has dropped out, as a consequence of the proportional hazard assumption which holds at time t. This holds generally for the time-constant variables and conditional on time for the time-varying covariates (Therneau and Grambsch, 2000, p. 127; Hosmer et al., 2011, p. 216). Equation (5.3) gives the probability that spell j is the actual spell that ends, i.e. the conditional probability that student j drops out divided by the conditional probability that any student in $R(t_j)$ drops out, i.e. the sum of the conditional probability of failure for each student in $R(t_j)$.

Given the assumption of k distinct ordered dropout times and the assumption of independent observations, (5.3) can be rewritten as the joint product of each contribution (Cameron and Trivedi, 2005, p. 595):

$$L_p(\beta) = \prod_{j=1}^{k} \left(\frac{\exp(\mathbf{x}_j(t_j), \beta)}{\sum_{l \in R_j} \exp(\mathbf{x}_l(t_j), \beta)} \right) \tag{5.4}$$

As mentioned, the expression can be treated like a likelihood function so that maximization of it leads to coefficients that are asymptotically normal with mean β_x (Cleves et al., 2010, p. 146). The likelihood function is often expressed in terms of log-likelihood:

$$\ln L_p(\beta) = \sum_{j=1}^{k} \left((\mathbf{x}_j(t_j), \beta) - \ln \left(\sum_{l \in R(t_j)} \exp(\mathbf{x}_l(t_j), \beta) \right) \right) \tag{5.5}$$

Note, that equation (5.5) does not account for ties which was argued before is necessary. Therefore, we turn to the Efron approximation for handling ties as this method is appropriate given the small proportion of ties relative to the risk set in our data (Therneau and Grambsch, 2000, p. 48). Based on data from Figure 4.1, the proportion was calculated as $\frac{\text{dropouts}}{\text{number of students}}$ and we note that the largest proportion is in wave 3 of approximately 6.3 percent, which is clearly not a relatively large share. The approximation adjusts the risk set that follow after the first tied dropout to account for the risk set becoming smaller.

To understand how the log-likelihood function is modified to take ties into account, two concepts are introduced. $D(t_j)$ is the set of subjects that drop out at time t_j and d_j denotes the number that drop out at time t_j. Including the two definitions, the partial log-likelihood, $\ln L_E(\beta)$, is changed to the following (Hosmer et al., 2011, p. 86):

$$\ln L_E(\beta) = \sum_{j=1}^{k} \sum_{i \in D_{t_j}} \left[(\mathbf{x}_j(t_j), \beta) - \sum_{k=1}^{d_i} \ln \left\{ \sum_{j \in R_{t(j)}} \exp(\mathbf{x}_l(t_j), \beta) - \frac{k-1}{d_i} \sum_{j \in D_{t(j)}} \exp(\mathbf{x}_l(t_j), \beta) \right\} \right] \tag{5.6}$$

The log-likelihood in equation (5.6) has changed compared to (5.5); we both sum over the the ordered failure times and the the set D_j. For the first dropout at each tied dropout time, the last term in (5.6) will be equal to zero, but from the next dropout, the risk pool is corrected for becoming smaller. It is noted that the Efron method is an approximation that acknowledges that the risk pool decreases within a period with more failures by including all possible risk pools and their probability. This makes it a more accurate approximation of the exact marginal likelihood compared to default method.

5.3.3 Interpretation

This section considers the interpretation of the estimated β's described in the previous section. For ease of interpretation, all estimated effects in Chapter 6 are reported as hazard ratios. The hazard ratio basically compares the hazard functions for two students, i and j, with different values of a specific covariate at time t_j. The hazard ratio is given by (Hosmer et al., 2011, p. 39):

$$HR \equiv \frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp(\mathbf{x_i}(t_j), \beta)}{\lambda_0(t)\exp(\mathbf{x_j}(t_j), \beta)} \tag{5.7}$$

For example, comparing the effect of an increase of 1 year of age in the covariate *Age* on the hazard ratio, we re-write (5.7) such that the hazard function in the nominator is increased:

$$\begin{aligned} HR &= \frac{\lambda_0(t)\exp(\mathbf{x_i}(t_j), \beta)}{\lambda_0(t)\exp(\mathbf{x_j}(t_j), \beta)} \\ &= \frac{\lambda_0(t)\exp(\beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{age}(x_{i,age} + 1) + ... + \beta_k x_{ik})}{\lambda_0(t)\exp(\beta_1 x_{j,1} + \beta_2 x_{j,2} + \beta_{age} x_{j,age} + ... + \beta_k x_{jk})} \\ &= \exp(\beta_{age}) \end{aligned} \tag{5.8}$$

From (5.8), an increase in students age by one year affects the hazard ratio by $\exp(\beta_{age})$. If β_{age} is estimated to -0.07, then the hazard ratio is found by exponentiating the coefficient, i.e. $\exp(-0.07)$, which gives as hazard ratio of 0.932. The effect of age increasing by one year implies that the risk of dropping out decreases by $1 - 0.932 = 0.068$ or 6.8 percent. Note, that a hazard ratio close to 1 corresponds to a coefficient close to 0 (as $\exp(0) = 1$), meaning a very limited effect.

As stated in the previous section, the underlying assumption in the Cox model is that the estimated effect is constant over time. In other words, comparing student i and j at each wave, we should find that the effect from age has the same effect on their probability of dropping out. In a model with time-varying covariates, the effect of the covariates are no longer proportional to a function of time only through the baseline hazard, $\lambda_0(t)$, but also through the time-varying covariates (Hosmer et al., 2011, p. 215). In this case, where *Distance* and *Worry* vary, it means that for exemplified change in *Age*, it is not only proportional to the baseline hazard, but also to the time varying covariates. However, this only matters for interpretation in that we need to condition on t, i.e. the effect would be $\exp(\beta_{age})$ conditional on t.

5.3.4 Statistical inference

For valid statistical inference of the estimates, we rely on a robust estimate of variance that also adjusts for clustering. This is done as we believe that there may be dependence among students enrolled in the same program at a give institution. By dependence, we refer to dependence in the observed dropout time. Intuitively, students studying at the same institution and program, are likely to be affected in roughly the same manner by e.g. the timing of exams, courses, etc., which may affect their choice of continuing studying or dropping out.

It is clear from the above motivation that we cluster on student groups rather than at an individual level. The reason for this is that the dependence is related to the time spent as an active student: a student can only drop out once in our data. Since each individual therefore only consists of one observation of how long he was active, it is not meaningful to talk about dependence. On the other hand, if one were to observe students over time that could drop out from one program, start on another and possibly drop out from that program as well, it would make sense to cluster on individual level.

Further, the results are presented as hazard ratios for easier interpretation, where the hazard rations are equivalent to the exponentiated coefficients. The standard errors for the hazard ratios are calculated by applying the delta method to the original standard-error estimate for the coefficients. The method calculates the standard error of the exponentiated coefficients by calculation of the variance of the corresponding first-order Taylor expansion. For $\exp(\beta_1)$, this corresponds to multiplying the standard error obtained for β_1 with $\exp(\beta_1)$ (Cleves et al., 2010, p. 133). As will be seen from the presented results in Chapter 6, this often leads to quite small standard errors.

5.4 Group effects in duration models

In the model presented so far, we have not accounted for the fact that students are enrolled at different institutions and programs. We think it likely that there might be group effects on this level, i.e. an effect of a program within an institution. In terms of duration models, two ways are highlighted to account for group effects, namely stratification and frailty models. They are introduced in the following subsections and the results the models generate are presented in Chapter 6.

5.4.1 Strata: modelling group-specific observable heterogeneity

As mentioned in Section 5.3.1, the key assumption in the model is proportional hazards across all students. The data in this study contains students from different programs and institutions which differ in location and type of institution. We argue that this is good level to control for group effect on because we can account for the fact that students could be different between educations and also between institutions. By forcing students at different institutions and programs to have the same baseline hazard, this key assumption may be violated. The assumption with respect to stratified models is that within each stratum, the assumption should not be violated. This is tested formally in Section 6.5. Through the use of stratification, the potential violation can be solved with a model that might give a more adequate fit to data (Hosmer et al., 2011, p. 207).

The idea with stratification is to model observed group effects assuming that the baseline hazard which in the previous model was common to everybody, is now specific to the strata-variable. The variable that we stratify on differs for each program at each institution. The introduction of stratification changes the conditional hazard function from (5.2) to (5.9) presented below, which underlines the idea of letting the baseline hazard differ for each stratum, s,

$$\lambda(t|\mathbf{x}(t), \beta) = \lambda_{0s}(t)\phi(\mathbf{x}(t), \beta). \tag{5.9}$$

For each stratum, the underlying risk profile is allowed to have a different shape. The stratified analysis is equivalent to fitting separate Cox PH models for each stratum, i.e. the partial likelihood is estimated for each stratum and the full stratified log partial likelihood sums over the contributions from each strata. This is done under the constraint that the estimated coefficients are equal across strata and the baseline hazard functions are not. That is, the effect from e.g. age is assumed to be the same for all students, but the underlying risk profile of dropping out is specific for each strata (Hosmer et al., 2011, p. 208). The interpretation of the hazard ratio is the same as describe in section 5.3.3 since we are still interested in the effects from the covariates, and not on the effect from the strata-variable. The advantage of stratifying is that it gives greater flexibility in modelling the shape of the hazard function for the groups, thereby in theory capturing a better fit of the model.

5.4.2 Shared frailty: modelling group-specific unobserved heterogeneity

Another way to account for group effects is through the frailty model which is known as the ”random effects” for duration models (Cleves et al., 2010, p. 156). The idea is that there may be a potential correlation among students which can be assumed to be induced by a latent effect on program and institution level. This latent effect is also referred to as unobserved heterogeneity. The model is given by equation (5.10) which incorporates the unobserved heterogeneity through a multiplicative effect, α_s from each group, s, on the baseline hazard.

$$\lambda(t|\mathbf{x}(t), \beta) = \lambda_0(t)\alpha_s\phi(\mathbf{x}(t), \beta) \tag{5.10}$$

As mentioned, the baseline hazard function must be positive, which motivates the use of the Gamma distribution, which has mean 1 and variance θ to model α_s. One may believe that individuals have different unobserved characteristics and that those who are most frail will drop out earlier than others and this is caught by the term α_s. An important assumption is that the unobserved heterogeneity, α_s is independent of any censoring that may take place and is independent of the covariates (Hosmer et al., 2011, p. 297; Allison, 2009, p. 76). This is a strong parametric assumption.

Chapter 6

Results

In this section, we present the results from the extended Cox model. As throughout the thesis, the focus is on the effects from living conditions to dropout. The results are divided into subsections with different focal points. First, the overall results are presented and hereafter region- and sector-specific effects for the variables for living conditions are presented. Besides these results, we also examine whether the living conditions are affected when controlling for academic and social integration. Further, an analysis accounting for potential heterogeneity the the type of student that lives far away from the educational institutions is presented. Finally, specification tests of the models are presented.

6.1 Overall results

The first step towards answering the main research question is to examine the effects on students across all of Denmark. The results suggest that living conditions are significantly associated with the dropout decision among first-year students. The estimated equation is presented again below and the hazard function consists of the covariates that were defined in section 4.2. It is noted that the presented model is the baseline model, i.e. without frailty or stratification.

$$\begin{aligned}\lambda(t|\mathbf{x},\beta) = \lambda_0(t)\exp(&\beta_1\text{Female} + \beta_2\text{Parental education} + \beta_3\text{High School GPA} \\ &+\beta_4\text{Age} + \beta_5\text{Age}^2 + \beta_6\text{Distance}_t + \beta_7\text{Move} + \beta_8\text{Worry}_t + \beta_9\text{Dum\_dist}_t)\end{aligned} \tag{6.1}$$

With this in mind, let us now turn to interpretation of the results. The overall results are presented in Table 6.1. For this first model, the results are described and discussed in large detail to set the scene. As shown in the table, there are three columns where each columns represents a model specification. Column 1 is the standard extended Cox referred to as baseline model. Column 2 describes the results from a stratified model and the column 3 shows the results from a shared frailty model. The standard errors are clustered and robust standard across the models, cf. Section 5.3.4. As mentioned, there may be a correlation between the time of dropout for students within the same education at the same institution, which is the argument for using clustered standard errors. We note that the estimated significance levels are robust to clustering. Table 6.1 is interesting in several ways. First, we see that the estimated effects of the variables related to living conditions, *Distance*, *Move* and *Worry*, are significant and roughly the same across the models. If *Distance* increases by 10 minutes in a given wave, the risk of drop out during the first year increases by 5 percent. To exemplify, if two students with the same background characteristics only differ in

Table 6.1: Dropout risk across all students

	(1)	(2)	(3)
VARIABLES	Baseline	Strata	Frailty
Female	0.956	1.047	0.981
	(0.060)	(0.082)	(0.060)
Vocational[†]	0.943	0.960	0.943
	(0.093)	(0.102)	(0.091)
Short-term higher education[†]	0.749**	0.734**	0.757*
	(0.107)	(0.109)	(0.109)
Medium-term higher education[†]	0.756***	0.802**	0.766***
	(0.078)	(0.088)	(0.078)
Long-term higher education[†]	0.813*	0.854	0.832*
	(0.091)	(0.103)	(0.091)
High School GPA	0.956***	0.978	0.966***
	(0.013)	(0.017)	(0.012)
Age	0.933***	0.997	0.950*
	(0.023)	(0.028)	(0.025)
Age^2	1.001***	1.000	1.001*
	(0.000)	(0.000)	(0.000)
Distance	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)
Move	0.860**	0.864**	0.852***
	(0.052)	(0.054)	(0.053)
Worry	1.071**	1.085***	1.078***
	(0.030)	(0.031)	(0.027)
Wald (chi^2)	94.25	63.94	79.38
DF	12	12	12

*** p<0.01, ** p<0.05, * p<0.1. [†] The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Dum_dist is omitted from the results.
Source: EVA and Statistics Denmark.

that person A lives 60 minutes away from the institution, while person B lives right next to it (0 minutes away), then person A will have a 30 percent higher probability of dropout than person B at a given point in time. The result is not surprising; if you spend more time on transportation, all else equal, you have less time to study. The effect from *Distance* is rather large when considering the distribution of the variable. In the first wave, the average distance from home to educational institution is 37 minutes across all students. Secondly, as much as 25 percent of the students spend more than 45 minutes commuting one way.

Returning the results: for a student who moves at the beginning of the first semester, the risk of dropping out is 13.6-14.8 percent lower compared to a student who does not move. Increasing the level of worries about living condition by one unit on the scale, e.g. going from the category ”A lesser degree” to ”To a great extent”, increases the probability of dropout by 7.1-8.5 percent. However, we note that *Worry* is presented on scale from 1-5 and this is not necessarily the most natural way to consider how worried a student is. Overall, these results are highly significant and they clearly show that living conditions matter for first year dropout.

Importantly, our results regarding living conditions are in line with most of the papers presented in Section 2.3 which focus on dropout in Denmark. However, caution should be paid in comparing the results since the applied methods in the Danish papers limit causal interpretation. As mentioned in the literature review, the international papers did not control for distance as such but focused on whether students live on or off campus. We argue that this campus effect could potentially be measured by *Move* as students that move at the beginning of the first semester might leave their parental home and move closer to the educational institution. Based on the results in Table 6.1, one can argue that our results confirm the findings from the international papers. Nevertheless, our finding suggests that there is a much smaller effect from *Move* compared to Bozick (2007) who found that college and university students who live at their parents house face a 41 percent higher risk of dropping out than students who live on campus. Gury (2011) found the effect to be a 59 percent higher risk of dropping out.

It is possible that the definition of *Move* compared to living on or off campus, may explain the rather small effect in our results relative to the results presented by Bozick (2007) and Gury (2011). We acknowledge that moving at the beginning of the semester and living with your parents may not measure the exact same behavior among students. The difference in magnitude could possibly also be affected by different comparison group or the fact that the papers consider the US and France, where the educational system, the housing and the students are likely to differ from the Danish system. Finally, no papers consider what worries about living conditions mean for dropout and therefore, the obtained effect cannot be compared.

Table 6.1 also reveals that in addition to the housing variables, there are significant associations between high school grade, a student's age and whether one of the student's parents has an education above vocational level, respectively and the risk of dropping out. However, this is only the case in the baseline model and in the frailty model, i.e. it is not the case with the stratified model in column 2 of the table. Therefore, we are somewhat careful to interpret these results. Nevertheless, the significant findings are consistent with the key reasons for dropout described in Section 2.1. As an example of interpretation, consider column 3. An increase in the high school grade by one unit, reduces the probability of dropout by 3.4 percent according to the model. This suggests that a student who is more prepared from high school, has smaller risk of dropping out, possibly because the student has better conditions for acquiring new knowledge.

Family background was also found to have an impact, if the educational level was above vocational. The effects from parents with short and medium-term higher educations are robust across all 3 model specifications. Based on column 3, a student who has a parent with a short-term tertiary education has a 24.3 percent lower risk of dropping out, compared to a student who has a parent with primary or secondary school (the baseline level of education). It is interesting that students whose parents have a long-term higher education are not better off regarding dropping out during the first year. That is, the risk of dropping out for this group is 16.8 percent and it is not significant across all the models. The difference in magnitude may be a result of higher motivation and support from parents with lower educational level. These parents may have experienced that education is a positive thing and do therefore support their children more. It could also be that students with highly educated parents do not spend much time with their parents since their parents have demanding job, requiring many hours. Overall, the results indicate that students are positively affected if their parents have an educational level above vocational.

Age is modelled non-log-linearly, thereby taking into account a diminishing effect with increasing age. The variable decreases the probability of dropping out during the first year in the baseline and frailty model specifications. The squared term means that we have a minimum, that is, the risk of dropping out is decreasing up to a given age and thereafter, the risk is increased. In the baseline model, the largest effect from age was found to be approximately 35 years, while in the frailty model is was around 25 years of age. The calculation of the values are given by the standard formula presented in Appendix A.4. The variable is insignificant in the stratified model which is a general feature of most of the background variable. This may seem strange at first sight, but we argue later in this section that there is a good explanation.

The difference of around 10 years implies that caution should be paid in claiming when the largest effect is. Therefore, we restrict our interpretation in noting, that age does have an effect on first-year dropout and the effect is most likely not linear. Intuitively, young students may be undecided regarding their choice of education but as they grown older, they become more determined of what they want to study. This could explain why we see a decreasing risk of dropping out in "early" years. On the other hand, older students may have different outside options or responsibilities such as having small children. This may therefore explain why the effect from age on risk of dropping out is declining. Remarkably, gender is insignificant across the three models in Table 6.1 which is in contrast to Gury (2011), Arulampalam et al. (2004) and Lassibille and Navarro Gómez (2008). Gury (2011) mentions that men and women do not generally exhibit the same dropout behavior yet our results suggest that gender do not affect the dropout probability. He claims that women can assess their academic performance faster and therefore, they self-select out of the education early on. If men generally have a higher dropout rate, this could lead to similar dropout rates during the first year and this could be the reason for the insignificant effects.

The available data made it possible to control for effects on a rather detailed level: for specific programme at institutional levels, a "group", and this information is exploited in the strata model. As mentioned in Chapter 5, the stratified model allows for different baseline hazard functions across groups. This gives a more flexible functional form, which is the main advantage of applying the model. On the other hand, we control for unobserved heterogeneity on a group level, which also is of importance. That way the the two models complements each other.

As mentioned before, it can appear as a puzzle that many of the standard control variables are insignificant in the stratified model. Nevertheless, let us briefly mention some features of the stratified model, which could most likely explain these findings. First, we remember from Section 5.4.1 that the full log partial likelihood for a stratified model sums over the log partial likelihood for each stratum. As the stratas are educations within educational institutions, it seems likely that the students within each stratum are quite homogeneous. While this is a good argument for why they should have the same baseline hazard, this means that there is potentially not much variation among the covariates within the stratas. As examples, certain educations are quite gender specific, students within the same education will tend to have similar high school grade average and ages and potentially also parents with similar background. Nevertheless, living conditions are more likely to vary for students within the same stratas. Further, both *Distance* and *Worry* are time-varying, allowing for more variation compared to the time-invariant background controls.

Finally, the test statistics from the Wald test indicate the likelihood of the estimated model compared to a model without covariates. Due to the specification of clustered standard errors, the test is a Wald test instead of the usual likelihood ratio test. The statistic has an approximate chi-square distribution under the null-hypothesis. We test 12 exclusion restrictions, corresponding to the 12 included covariates. For all the estimated models, they are significantly preferred over models with no covariates.

To sum up, the results from the overall regression suggest that there is a higher risk of dropping out during the first year among students that live further away from their educational institution and are worried regarding their living condition. On the other hand, a student that moves at the beginning of the first semester has a smaller probability of dropout. This means that living conditions do matter for dropout during the first year at an institution of higher education.

6.2 Regional analysis

This section investigates if living conditions affect dropout differently across regions, where region refers to the location of the institution of higher education. First, we motivate the regional analyses and in the following subsections, the three variables for living conditions are interacted separately with regional dummies. Finally, the regional effects are summed up. Our hypothesis is that the effect from the housing variables on dropout is larger in the Capital Region and Central Region and this hypothesis is motivated below.

Based on the literature review, no direct evidence on regional differences were presented, yet some studies mentioned different effects across cities. While AU Student Council (2000) did not find an effect from living conditions to dropout in Aarhus, Holm et al. (2008) note that students in Copenhagen mention living conditions as a reason for dropout. Living conditions in Copenhagen and its association with dropout is supported by DMA Research (2002), who mention that there is in particular a problem with housing in Copenhagen, not in Aarhus. Besides the three mentioned analyses, one could also imagine that housing market is under more pressure in the Capital region. Even though the articles do not indicate it, we also hypothesize a stronger effect in the Central Region as Aarhus has a relatively large number of citizens, hence a housing market under pressure.

Aalborg and Odense and partly Aarhus have a so-called housing guarantee, *boliggaranti*, ensuring the students a place to live either at the beginning of or during the first semester. This is only relevant for the specific cities and not the entire regions. Therefore, this might not show in the presented results. As will become clear, we do not find any evidence to support that living conditions have a stronger effect on dropout in the Capital Region.

6.2.1 Distance

Table 6.2 shows the estimated hazard ratios for the regions interacted with *Distance* in the three model specifications. We focus our interpretation on the interaction terms since the other control variables enter in the same fashion as in the overall analysis. The same thought is applied when presenting the interaction effects on *Worry* and *Move* in the following two sections.

We found a significant association between living conditions and dropout not only in the Capital Region and the Central Region, but also in North Region. These findings are consistent across all three model specifications. Regarding the size of the effects, the effects are slightly larger in comparison with those presented in Table 6.1. This is especially the case for the North Region, where living 10 minutes further away can result in as much as a 9 percent higher probability of dropout. This is surprising given the housing guarantee in Aalborg. For the South Region, the effects are insignificant across all models. The same is found in the case for Region Zealand which is borderline significant effect in the baseline model. Taken together, these results suggest that living further away from the institution has a negative effect on the risk of dropping during the first year in the Capital, Central and North Region, but not in South Region and Region Zealand.

The average distances are roughly the same across regions. The only region that stands out is Region Zealand, where the average distance is more than 10 minutes above the average in the other regions. The distribution of distance across regions using a kernel is showed in the appendix in Figure A.1. It clearly shows that students report higher values in Region Zealand which could possibly explain the insignificant association from *Distance*. The intuition is that these students may be used to spending more time on transportation, and are therefore not affected by distance the same way as students are in other regions. Also, Odense which has a housing guarantee for students is located in this region and this might also be part of the explanation of the insignificant results. At the same time, it is a bit counter-intuitive, since one could imagine that the effect would be larger in the region due to the larger distances. Given that students in South Region report a distance which does not deviate as much as Region Zealand, the explanation behind the insignificant results is not entirely clear.

Table 6.2: Regional effects of *Distance*

VARIABLES	(1) Baseline	(2) Strata	(3) Frailty
Female	0.958	1.050	0.985
	(0.060)	(0.083)	(0.060)
Vocational[†]	0.935	0.958	0.936
	(0.093)	(0.101)	(0.090)
Short-term higher education[†]	0.750**	0.736**	0.757*
	(0.107)	(0.109)	(0.109)
Medium-term higher education[†]	0.762***	0.802**	0.770**
	(0.079)	(0.088)	(0.078)
Long-term higher education[†]	0.852	0.862	0.858
	(0.096)	(0.104)	(0.094)
High School GPA	0.963***	0.978	0.970**
	(0.014)	(0.017)	(0.013)
Age	0.934***	0.996	0.951*
	(0.023)	(0.028)	(0.025)
Age^2	1.001***	1.000	1.001*
	(0.000)	(0.000)	(0.000)
Distance*Capital Region	1.006***	1.007***	1.006***
	(0.002)	(0.002)	(0.002)
Distance*Central Region	1.006***	1.007***	1.006***
	(0.002)	(0.002)	(0.002)
Distance*North Region	1.008***	1.009***	1.008***
	(0.002)	(0.002)	(0.002)
Distance*South Region	1.002	1.000	1.002
	(0.003)	(0.003)	(0.003)
Distance*Region Zealand	1.003*	1.001	1.002
	(0.002)	(0.002)	(0.002)
Worry	1.081***	1.085***	1.083***
	(0.030)	(0.031)	(0.028)
Move	0.850***	0.860**	0.845***
	(0.052)	(0.054)	(0.053)
Wald (chi^2)	114.5	78.94	95.49
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. [†] The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Regional dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

6.2.2 Worry

This subsection considers the regional effects of the variable *Worry* presented in Table 6.3 and discussed below. Contrary to expected, we did not find that a larger effect for students in Capital Region nor in the Central Region with respect to *Worry*. *Worry* was in fact found to be insignificant in all the specifications for the Capital Region and also mainly insignificant for the Central

Table 6.3: Regional effects of *Worry*

VARIABLES	(1) Baseline	(2) Strata	(3) Frailty
Female	0.956	1.047	0.982
	(0.059)	(0.082)	(0.060)
Vocational†	0.933	0.959	0.934
	(0.093)	(0.102)	(0.090)
Short-term higher education†	0.746**	0.733**	0.753**
	(0.107)	(0.109)	(0.108)
Medium-term higher education†	0.757***	0.802**	0.766***
	(0.078)	(0.088)	(0.078)
Long-term higher education†	0.840	0.854	0.847
	(0.095)	(0.103)	(0.093)
High School GPA	0.963***	0.977	0.970**
	(0.014)	(0.017)	(0.013)
Age	0.934***	0.997	0.951*
	(0.023)	(0.028)	(0.025)
Age^2	1.001***	1.000	1.001*
	(0.000)	(0.000)	(0.000)
Worry*Capital Region	1.060	1.077	1.064
	(0.047)	(0.050)	(0.043)
Worry*Central Region	1.070	1.082*	1.073
	(0.050)	(0.051)	(0.053)
Worry*North Region	1.010	1.001	1.003
	(0.095)	(0.096)	(0.077)
Worry*South Region	1.202**	1.178**	1.199**
	(0.094)	(0.088)	(0.094)
Worry*Region Zealand	1.129**	1.106	1.132**
	(0.065)	(0.068)	(0.060)
Distance	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)
Move	0.851***	0.863**	0.847***
	(0.052)	(0.054)	(0.053)
Wald (chi^2)	112.9	68.62	93.21
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Sector dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

Region. This is despite the fact that on average, students in Capital Region report a higher level of worries in each wave, as shown in Appendix A.3. Further, this insignificance is also found for the North Region.

On the other hand, in the South Region, *Worry* is significant on a 5 percent level across all three model specifications. Compared to the overall effects presented in Table 6.1, the effect from

Worry is much larger in this region. An increase of one unit in the level of *Worry* leads to increases in the dropout probability in the range of 17.8 - 20.2 percent. Finally, Region Zealand has significant effects from *Worry* on a 5 percent level in the baseline model and the frailty model. It is unfortunate that the stratified model does not return significant results as this model does not break with the proportional hazard assumption, cf. Section 6.5. This means that the significant results for Region Zealand should be interpreted with care. Further, the very large results found for the South Region are surprising compared to Table 6.1. However, again we argue that as *Worry* is a scale from 1-5, it can be discussed what it means to experience an increased level of worries. Therefore, we primarily focus on the fact that the effect indicates that a higher level of *Worry* means a higher risk of dropout.

Overall, the significant effects from *Worry* are in the South Region and to some degree Region Zealand, while the other regions generally report insignificant results. This is interesting as it is the opposite picture of that shown in Table 6.2 for *Distance*. Finally, *Distance* and *Move* have more or less the same impact on dropout as in Table 6.1. The other control variables do also seem constant compared to both the main results and regional effects controlling for distance.

Given the above mentioned findings, one could imagine that students in different regions perceive worries regarding living conditions differently. Think of individuals who apply for a program in Region Capital. In this region and especially in and around Copenhagen, the housing market is known to be difficult for students and probably, the students are well-aware of this. Therefore, their concerns of finding a place to stay are relatively high compared to students in other regions without it being linked to dropout. From the point of view of the other regions, we expect that students in e.g. South Region should not struggle too much finding a place to live. Therefore, if they report a high level of worries, it must be because the situation is very serious and because it matters for dropout. To our knowledge, concerns regarding living conditions and dropout have not been analyzed before and therefore, which limits comparison with other findings.

6.2.3 Move

In this subsection, we examine how moving at the beginning of the first semester is associated with dropout. Table 6.4 presents the results for the three model specifications. From the table, it can be seen that *Move* is insignificant across all model specifications for the Capital Region, the Central Region and the North Region. While there is one borderline significant result for the South Region in the baseline model, we will not pay attention to it. For Region Zealand, however, the effect is

Table 6.4: Regional effects of *Move*

	(1)	(2)	(3)
VARIABLES	Baseline	Strata	Frailty
Female	0.957	1.047	0.982
	(0.059)	(0.082)	(0.060)
Vocational†	0.935	0.959	0.936
	(0.093)	(0.102)	(0.090)
Short-term higher education†	0.749**	0.734**	0.756*
	(0.107)	(0.109)	(0.108)
Medium-term higher education†	0.761***	0.801**	0.768***
	(0.078)	(0.088)	(0.078)
Long-term higher education†	0.843	0.854	0.850
	(0.095)	(0.103)	(0.093)
High School GPA	0.962***	0.978	0.970**
	(0.014)	(0.017)	(0.013)
Age	0.934***	0.999	0.951*
	(0.023)	(0.028)	(0.025)
Age^2	1.001***	1.000	1.001*
	(0.000)	(0.000)	(0.000)
Move*Capital Region	0.918	0.953	0.923
	(0.095)	(0.102)	(0.099)
Move*Central Region	0.890	0.863	0.877
	(0.097)	(0.094)	(0.098)
Move*North Region	0.855	0.857	0.844
	(0.135)	(0.140)	(0.139)
Move*South Region	0.747*	0.800	0.743
	(0.122)	(0.128)	(0.161)
Move*Region Zealand	0.755**	0.772*	0.752**
	(0.100)	(0.109)	(0.098)
Distance	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)
Worry	1.082***	1.085***	1.084***
	(0.030)	(0.031)	(0.028)
Wald (chi^2)	116.5	65.29	90.63
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Regional dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

significant across the three models. The effect from *Move* is quite large: it indicates that students that move at the beginning of the first semester have a 22.8-24.8 percent smaller probability of dropping out compared to students that did not move.

The overall conclusion for the regional analysis is not in line with the hypothesized; there is no particularly strong effects from neither the Capital Region nor Central Region. We find that *Distance* significantly increases the probability of dropout across the regions except in the South Region and Region Zealand. The largest effects across models are found in Region North Denmark. *Worry* is significantly related to dropout only in South Region and to some degree in Region Zealand. The regional analysis for *Move* suggested that there are significant decreases for students that move at the beginning of the semester only in Region Zealand, while the effects in all other regions are insignificant. The overall picture based on these findings indicates that *Distance* is important for dropout in the Capital Region, the Central Region and the North Region. *Worry* and *Move* are not of importance in these regions, but in Region South and Region Zealand. While the analysis shows there are differences in what living conditions matter across regions, there are no obvious explanations as to why this is the case.

6.3 Sector analysis

In Section 3.2, we hypothesized that students across educational sectors may have different effects from living conditions to dropout. Sectoral differences are therefore analyzed in this section. Given a limited number of students at maritime education and artistic higher education in the sample, our analysis only considers students at universities, business academies and university colleges. First, we present results from interacting with *Distance*, followed by results from the same procedure for *Worry* and *Move*.

6.3.1 Distance

Table 6.5 presents results from the interactions between *Distance* and students in the three sectors. The findings suggest that university and university college students have a negative association between *Distance* and dropout. For students at university colleges, the effect in the stratified model, however, was insignificant. The significant effects are similar to those found in Table 6.1. For students at business academies, there was only found a significant effect in the stratified model. Even with different model specifications, we note that the size of the significant effects are roughly equal across sectors. In other words, we do not seem to find evidence in favor of our hypothesis.

The findings from Table 6.5 share some similarities with the findings presented by Smith and Naylor (2001). They do not examine the effect from distance but whether university students live on or off campus. One could think of this as living close or further away from the institution. Their findings suggests that living off campus, or what we think of as having a longer distance, was negatively correlated with dropout. In this light, our findings could be thought of as being in line with

Table 6.5: Educational sector effects of *Distance*

VARIABLES	(1) Baseline	(2) Strata	(3) Frailty
Female	1.011	1.049	1.022
	(0.064)	(0.082)	(0.063)
Vocational †	0.941	0.959	0.939
	(0.094)	(0.102)	(0.090)
Short-term higher education †	0.743**	0.734**	0.750**
	(0.107)	(0.109)	(0.108)
Medium-term higher education †	0.747***	0.803**	0.756***
	(0.077)	(0.088)	(0.077)
Long-term higher education †	0.789**	0.857	0.811*
	(0.089)	(0.103)	(0.089)
High School GPA	0.943***	0.978	0.956***
	(0.014)	(0.017)	(0.013)
Age	0.949**	0.997	0.962
	(0.023)	(0.028)	(0.026)
Age^2	1.001**	1.000	1.001
	(0.000)	(0.000)	(0.000)
Distance*University	1.005***	1.006***	1.005***
	(0.001)	(0.001)	(0.001)
Distance*Business Academy	1.004	1.005*	1.004
	(0.002)	(0.003)	(0.002)
Distance*University College	1.004**	1.003	1.004**
	(0.002)	(0.002)	(0.002)
Worry	1.070**	1.083***	1.075***
	(0.029)	(0.031)	(0.027)
Move	0.853***	0.864**	0.848***
	(0.051)	(0.054)	(0.053)
Wald (chi^2)	132.5	71.26	108.7
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Sector dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

their findings, yet interpreted with caution. Further, they do not consider students in other sectors.

Another interesting comparison is with the findings for first-year students at universities, business academies and university colleges in Denmark presented in The Danish Agency for Science and Higher Education (2018). Among the survey respondents, long and expensive transportation time as well as complaints about the location of housing reported were more often mentioned as contributing reasons to dropout. Living conditions were not found to be the key reasons for dropout but are indicated to have an impact for some students. Nevertheless, their survey responses did not suggest sectoral differences regarding the effect of living conditions, cf. Appendix A.2.

6.3.2 Worry

Table 6.6 shows the effect of *Worry* on living conditions for the educational sectors. The table is interesting because it shows that there is a significant and large effect for university and business academy students. The effects are robust across all three model specifications for students at business academies and across the stratified and frailty model for students at universities, which we also consider as relatively robust.

TABLE 6.6: Educational sector effects of *Worry*

	(1)	(2)	(3)
VARIABLES	Baseline	Strata	Frailty
Female	1.012	1.048	1.023
	(0.064)	(0.082)	(0.063)
Vocational †	0.941	0.961	0.939
	(0.094)	(0.103)	(0.090)
Short-term higher education †	0.740**	0.734**	0.747**
	(0.106)	(0.109)	(0.107)
Medium-term higher education †	0.746***	0.804**	0.756***
	(0.077)	(0.089)	(0.077)
Long-term higher education †	0.788**	0.855	0.810*
	(0.088)	(0.103)	(0.088)
High School Grade	0.942***	0.978	0.955***
	(0.014)	(0.018)	(0.013)
Age	0.949**	0.998	0.962
	(0.023)	(0.028)	(0.026)
Age^2	1.001**	1.000	1.001
	(0.000)	(0.000)	(0.000)
Worry*University	1.054	1.074**	1.059*
	(0.035)	(0.037)	(0.032)
Worry*Business Academy	1.182**	1.199**	1.186***
	(0.085)	(0.096)	(0.075)
Worry*University College	1.031	1.030	1.040
	(0.063)	(0.065)	(0.060)
Distance	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)
Move	0.852***	0.864**	0.848***
	(0.051)	(0.054)	(0.053)
Wald (chi^2)	133.3	76.08	110
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Sector dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

Remarkably, the effect from *Worry* is far greater for students at business academies than for university students. The effects for university students indicate that an increase of one unit in the level of *Worry* leads to an increase of 5.9-7.4 percent in the probability of dropout, compared to 18.2-19.9 percent for students at business academies. The reason for this rather large difference is not completely clear, since students at universities and business academies report an average level with is roughly at the same level, cf. Appendix A.4. On average, students at business academies report a value of 1.75 compared to 1.84 for university students. This might indicate that students at universities can handle a larger degree of worries without letting it affect the probability of dropout compared to students at business academies.

Nevertheless, these finding thus need to be interpreted with caution because of two reasons. First, as mentioned, to our knowledge, no studies have analyzed the association between worry and dropout which limits our possibilities of comparing the size of the estimated effects. Secondly, *Worry* has been modelled as a continuous variable taking the values 1 to 5. It is plausible that another functional form of the variable could have resulted in a larger or smaller size, yet with the same correlation with dropout. Therefore, we again limit our interpretation to stating that the finding may suggest that increasing the degree of worry, is likely to increase the risk of dropping out.

6.3.3 Move

Table 6.7 shows the results from analyzing sector difference and moving at the beginning of the first semester. With a few exceptions, the findings did not show a sectoral difference with respect to *Move*. University students have a significant effect in the baseline and frailty model. Business academy students also have a significant association in the baseline model while students at university colleges show a significant effect in the frailty model. The results indicate a very limited effect from *Move* to dropout, which is surprising taking the relatively large and highly significant overall effect from Table 6.1 into account.

Comparing this to the literature, The Danish Agency for Science and Higher Education (2018, pp. 51-53) asked survey respondents whether difficulties in finding a permanent resident was a contributing reason for dropout. The findings point to this factor being relatively unimportant for dropout among students across sectors and further no clear sectoral differences. A direct comparison cannot be made as they consider finding permanent housing, while this thesis considers moving at the beginning of the first semester. Students moving during the first year are presumably first of all interested in finding a place to stay, rather than finding a permanent place.

To sum up on the overall findings from this sectoral analysis, in general, university students seem

Table 6.7: Educational sector effects of *Move*

VARIABLES	(1) Baseline	(2) Strata	(3) Frailty
Female	1.010	1.045	1.020
	(0.064)	(0.082)	(0.063)
Vocational	0.942	0.960	0.940
	(0.094)	(0.102)	(0.090)
Short-term higher education	0.742**	0.734**	0.750**
	(0.107)	(0.109)	(0.108)
Medium-term higher education	0.747***	0.801**	0.756***
	(0.077)	(0.088)	(0.077)
Long-term higher education	0.788**	0.855	0.810*
	(0.088)	(0.103)	(0.089)
High School Grade	0.942***	0.978	0.955***
	(0.014)	(0.017)	(0.013)
Age	0.948**	0.997	0.961
	(0.023)	(0.028)	(0.026)
Age^2	1.001**	1.000	1.001
	(0.000)	(0.000)	(0.000)
Move*University	0.878*	0.893	0.875*
	(0.060)	(0.063)	(0.064)
Move*Business Academy	0.762*	0.760	0.757
	(0.121)	(0.130)	(0.129)
Move*University College	0.787	0.785	0.780*
	(0.121)	(0.128)	(0.113)
Distance	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)
Worry	1.072**	1.085***	1.077***
	(0.030)	(0.031)	(0.027)
Wald (chi^2)	124.6	68.82	105.3
DF	20	16	20

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 38,586 observations, 19,032 individuals and 1,284 dropouts.
Sector dummies and *Dum_dist* are omitted from the results.
Source: EVA and Statistics Denmark.

to be affected by *Distance* and partly by *Worry* and *Move*. On the other hand, students at business academies are affected by *Worry* and partly affected by *Distance*, while *Move* is very weakly related to dropout. Finally, the last group of students in university colleges were partly affected by *Distance*, very weakly affected by *Move* and unaffected by *Worry*. Taken together, these findings point to a complex situation of which factors that influence students across sectors.

6.4 Additional analyses

This section presents the additional analysis that were motivated in Section 3.3. First, we investigate what happens to the variables for living conditions in a model that accounts for academic and social integration. Hereafter, the results from an analysis that considers a potential heterogeneity in the group of students that has a long transportation time are presented.

6.4.1 Academic and social integration

Table 6.8 presents results where the effects from living conditions on dropout are controlled for academic and social integration. These integration variables have been found to be important for dropout, cf. Section 2.1, and it may be possible that bad living conditions could affect the integration. Intuitively, an unsettled housing situation or long transportation time, could decrease the time spent at the institution and the energy used to integrate, which could increase the risk of dropping out. Our findings suggest that while both *Distance* and *Move* are still significant, controlling for integration results in an insignificant effect from *Worry*.

The results can be seen in Table 6.8. The table shows the results from two model specifications with a frailty model left out as it could not be estimated due to gaps in the integration variables. Irrespective of model specification, the association from *Worry* to dropout is insignificant. This is interesting as it could point to a mechanism, namely, that the effect from *Worry* goes through integration. The effect from the other two living conditions are in line with findings from Table 6.1.

Another finding from the table is that the effect from the integration variables are substantially larger compared to the other control variables. However, we should not dwell too much by the size of the estimated hazards as they both are scales from 1-3 and the interpretation is not completely clear. The primary concern is how the affect the variables for living conditions and whether they are have a significant effect which they do. This may suggest that the integration variable as well as living conditions are associated with dropout.

We briefly note that the variables for social and academic integration are only available for the second and third wave which results in a reduced sample. As can be seen from the table, this means we have much fewer individuals to run the analysis on and as a results, comparison with previous estimates is limited. Nevertheless, this is the best option for investigating the effect integration in relation to living conditions and dropout.

Table 6.8: Controlling for academic and social integration

	(1)	(2)
VARIABLES	Baseline	Strata
Female	0.946	1.063
	(0.066)	(0.096)
Vocational†	1.007	1.015
	(0.112)	(0.125)
Short-term higher education†	0.790	0.728*
	(0.134)	(0.134)
Medium-term higher education†	0.844	0.885
	(0.103)	(0.123)
Long-term higher education†	0.960	1.017
	(0.120)	(0.143)
High School GPA	0.962**	0.979
	(0.015)	(0.019)
Age	0.907***	0.976
	(0.024)	(0.029)
Age^2	1.001***	1.000
	(0.000)	(0.000)
Distance	1.004***	1.005***
	(0.001)	(0.001)
Move	0.845**	0.848**
	(0.058)	(0.063)
Worry	1.021	1.032
	(0.033)	(0.036)
Academic integration	0.459***	0.461***
	(0.033)	(0.035)
Social integration	0.398***	0.393***
	(0.027)	(0.030)
Wald (chi^2)	535.5	460.7
DF	14	14

*** p<0.01, ** p<0.05, * p<0.1. † The educational levels refer to parents education. The comparison group is primary and secondary school
Note: 17,498 observations, 10,865 individuals and 1,010 dropouts.
Dum_dist is omitted from the results.
Source: EVA and Statistics Denmark.

Summing up, the above presented findings point to the fact that living conditions as well as integration are associated with dropout. The findings may suggest that *Worry* affects the degree of integration or that the integration variables also control for some of the same things that *Worry* does. The other variables for living conditions affect dropout independently of the degree of integration. Overall, the findings strengthens our belief in effect from living conditions to dropout.

6.4.2 Heterogeneity in distance

In continuation, we now analyze whether students may voluntarily live far away. It is possible that students living far away from their educational institution might roughly consist of two subgroups. The first group are students who want to move closer to the institution and the second group comprise students who are happy and settled with a long transportation time. We believe the last group are older students, possibly with children. Further, we hypothesize that *Distance* is significantly related to dropout for the students in the first group, while the opposite is expected for the second group. The analysis was conducted by including two interaction terms. The first captured the effect of being a student above the age of 30 and *Distance*, while the second interaction term captured the effect from being a student with a child and *Distance*.

The results are primarily discussed here and we refer to the output in Appendix A.2. The findings in the table do indicate that our hypothesis might be true as neither of the interaction terms are significant. The interaction between students with children and *Distance* is borderline significant, but only in one model. The insignificant effects from the two interaction terms may suggest that transportation time has a rather restricted effect on the choice of dropping out for especially older students with children. On the other hand, there is a very limited number of these students in the sample which could explain the insignificant results. To be precise, we observe observe 14 dropouts in the first wave for this group, followed by 42 and 15 in wave 2 and 3. This means that while the insignificant effects of the interaction terms support the presented hypothesis, there are large insecurities regarding the results as a very little group of students is considered.

6.5 Sensitivity analysis

This section presents the tests that will be undertaken to ensure that the model specification is as correct as possible, so that the results presented in this chapter can be trusted. It is important that our inferential conclusions based on the fitted model are the best and most valid possible (Hosmer et al., 2011, p. 169). This section presents the results of tests for the proportional hazards assumption and discuss the functional form of the model.

6.5.1 Proportional hazards assumption

This subsection tests the proportional hazards assumption based on the so-called Schoenfeld residuals. In practice, a smooth function of time is fitted to the residuals and hereafter, the test is of whether there is a relationship between the residuals and time (Cleves et al., 2010, p. 206). Under the null hypothesis, there is a zero slope of the residuals, which is equivalent to a constant hazard ratio is over time (Cleves et al., 2010, p. 207).

As mentioned above, the interest is in testing whether the modelled specification violates the assumption. The test is performed for the overall models presented in Table 6.1, as well as the model specifications with regional and sector interactions for *Distance*, *Move* and *Worry*, respectively. The tests revealed that the assumption was generally not fulfilled in the baseline and frailty models. In the stratified models, the tests found proportional hazard within each strata at a 5 percent significance level. Based on the test results, the stratified model specification seems to be the most appropriate to fit the observed data since the assumption is fulfilled. Cleves et al. (2010, p. 203) states that the proportional hazard test can also be seen as a test of goodness-of-fit measure. That is, a test of whether the included variables and the functional form are appropriate. However, as the tests of the stratified model show that there is proportional hazard, this indirectly indicates that the model is well specified.

Although the performance of the frailty models based on the test was not ideal, we still believe that the model specification adds some valuable information. First, it is noted that the results for living conditions presented in the chapter are rather robust across the different model specifications. This is both regarding the size of the estimated effects and also with respect to significance of the variables. Secondly, frailty models have the advantage of controlling for unobserved heterogeneity which stratified models cannot. This motivates the use of frailty models. Both of these argument makes us confident that the frailty and stratified models supplement each other.

To sum up, there are challenges with the proportional hazard assumption in the baseline and frailty models. However, as the results for living conditions are robust across the models as well as the seemingly well-specified stratified model, we argue that the models can fit data adequately.

6.5.2 Functional form

It was stated in the previous section, that the test of proportional hazard can be viewed as examination of goodness-of-fit. In this section, we perform specific tests for functional form for relevant variables which is important. Our considerations regarding functional form is based on the structure presented by Wooldridge (2009, p. 659) and Cameron and Trivedi (2005, pp. 277-278).

Let us first look at *Distance* and *Worry*. In this thesis, *Distance* has been included so that it has a log-linear effect on the probability of dropping out, and this functional form has been challenged by including a squared term of the variable. This non-log-linear function of the variable could capture diminishing returns from the variable. Intuitively, moving further away may matter more in a given interval and at some point, moving e.g. five minutes further away may not have the same effect. Think of a student who on average spend 15 minutes on transportation. Moving ten minutes further away, may have a larger affect compared to a student who already lives 60 minutes away. We found that the inclusion of the squared term did not affect the other variables or the overall conclusions. Despite the fact that the squared term was highly significant, the economic significance was minor. With an estimate of the squared term which was very close to 1, this would indicate that increasing distance would have more or less the same effect as with a log-linear function. Based on the rather restricted economic significance, *Distance* was modelled without this squared term.

Worry has been tested in two ways: by adding a squared term and by including it as a dummy. When a squared term was included, the overall effect from the variable was insignificant, suggesting that this functional form is not appropriate. The second choice was to model *Worry* as a dummy variable which takes the value 1 if the student indicates to be worried and 0 otherwise. The results suggest that the effect was significant and the effect on dropout was much larger: around 25 percent larger hazard of dropout if the dummy is turned on. This might place too much importance on the variable when analyzing dropout. Neither the non-log-linear modelling nor the dummy variable are believed to give a more accurate fit of data. Therefore, the chosen functional form of *worry* is in our opinion appropriate.

Move could also have been modelled differently. In the survey, there is data on whether students intend to move soon in each wave. However, if all this information is employed, it is not obvious what effect should be expected. On one hand, moving could lead to a lower probability of dropping out if it means that the student has found a permanent place to stay. On the other hand, if the students moves often due to an unstable housing situation, this must be expected to increase the risk of dropout. This was mirrored in the results that were generally insignificant when moving was included that way. Therefore, we argue that *Move* as a dummy for whether the student moved at the beginning of the first semester is the most intuitive way to include the variable.

The control variables *High school GPA* and *Age* could potentially be modelled different. *High school GPA* was controlled for by an additional squared term, but this made the variable insignificant. This suggests that including the variable in a log-linear fashion is reasonable. Compared to *High school GPA*, *Age* was modelled non-log-linearly, see equation 3.1. This variable and its maximum are meaningful economically and the variables for age are significant in most of the model specifications.

As for the control variables parents education, they are included as a dummy for each additional level as described in Chapter 4. That way more information is included than if the variable had been included as a continuous variable as that would mean imposing that the effects of a higher education level would be the same in the bottom as in the top of the distribution, which is unlikely.

Based on the above tests for proportional hazard and form of the covariates, it can be concluded that the baseline and frailty models do meet some challenges, but the results from living conditions to dropout are roughly robust across the models and different functional forms of the variables. Also, the different model specifications account for different challenges with the data and therefore, it is comforting that the results are robust.

Chapter 7

Discussion

In this chapter, we discuss several issues related to the applied data, methods and assumptions made in order to assess to what degree the results presented in Chapter 6 can be trusted. Further, we provide policy recommendations.

7.1 Data challenges

In this section, we discuss the challenges related to the applied data and how the might have affected the obtained results. The data used in this thesis allows for investigation about issues such as worries about living conditions that are not available in registers. Further, the data has not been analyzed previously with focus on students' living conditions and dropout. With that in mind, the data allows us to fill a gap in the literature. These are some of the positive features of the data. On the other hand, it is possible that data challenges have influenced the obtained results. First, the sample is likely to be non-random due to voluntary participation in the survey. This is also indicated by the representativity analysis conducted in Figure 4.1. An additional potential source of error is due to the fact that there is attrition in the panel over time and finally, there are issues related to missing data. How these challenges may have affected the interpretation of our results and how well the research questions are answered, will be discussed below.

As mentioned, the first data challenge is that the sample is non-random as a result of voluntary participation in the surveys, which leads to self-selection. This is also supported by a comparison of the population and the respondents in different waves, cf. Table 4.1. We account for this self-selection to some degree by including the background controls such as high school GPA and parents education. Further as described in Chapter 4 and Appendix A.3, the sample is weighted relative to the population as a robustness check. The idea is to see if that changes the results and in such a case, it should lead to caution in term of extrapolation to the population.

The weighting method was implemented in accordance with Fitzmaurice (2011) who states that weighting should be performed on the sample in the first wave. Although this is the most common method, it is likely that weighting on information from e.g. the second wave might have given other results. Intuitively, if the largest effect from self-selection occurs in a later wave, the weights may not be representative. A further caveat with weighting is that we can only weight on observable characteristics available on both population and survey sample. This means that if unobserved factors determine whether you participate in the survey, the weights lack information. The weighting did not change the obtained results, cf. Appendix A.3. Therefore it seems likely that the sample is in general representative even though the findings in Chapter 4 indicate significant differences in some variables between the population and the sample.

The second possible data limitation is due to attrition, which can be a challenge if students with particular characteristics choose not to answer the survey in a nonrandom manner. Attrition basically reduces the sample and may make the sample unrepresentative over time. Based on Table 4.1, there are indications that the sample becomes less representative of the population over time. This implies that one can have biased results if observations on the dependent variable are lost in a nonrandom manner (Cameron and Trivedi, 2005, p. 801). It is likely, that students who are lost due to attrition may be students facing problems with living conditions and therefore students with higher probability of dropping out. This may imply that our results are biased but if so, we believe that the results are biased towards a hazard ratio of 1. In other words, the effects given by the hazard ratios are biased towards no effect. Therefore, the obtained results can be seen as conservative estimates of the effects of living conditions on dropout. It may be assumed that an even stronger effect could have been found using a sample with no attrition and therefore, there is reason to take the students problems with living conditions seriously.

Due to missing observations on key background variables, the method of listwise deletion has been performed on the initial population. It is a default option in statistical software, even though it means throwing away potentially useful information and might lead to bias (Cameron and Trivedi, 2005, p. 925). In particular, as mentioned in Chapter 4, the initial population of 44,496 students falls to 40,826 due to listwise deletion, i.e. 8.25 percent of the initial population is deleted. Assuming that this information is missing completely at random, the remaining population would still be random, but it might not be the case and that would mean that our final population is biased and therefore, that the group of students we weight relative to is not representative of the actual population (Cameron and Trivedi, 2005, p. 928). As the variables used to weigh are the ones that are missing, it is not possible to weigh relative to the entire population. An option would be to turn to imputation. If the data gaps can be filled by a statistically meaningful procedure, imputation can lead to a larger and possibly more representative sample, but this possible gain comes at the cost of potentially making wrong assumptions and the risk of ending up with a sample that is less representative than the non-imputed sample. In general there is not consensus on whether such a method improves or worsens the situation. Although the loss of data is not preferable, we nevertheless believe that it has a little effect on the results because it is a small share of the sample that is lost.

7.2 Duration analysis

In this section, we raise the most important issues related to our choice of empirical framework and discuss how they contribute to disentangling the effects from living conditions to dropout.

7.2.1 Time-varying covariates in duration analysis

As mentioned in Section 5.1, the duration model allows for inclusion of time-varying covariates. There are two assumptions related to the time-varying covariates, namely that they are strictly exogenous (Cameron and Trivedi, 2005, p. 598) and that the effect from the covariates are constant over time. The assumptions are discussed below.

The assumption of strict exogeneity of the time-varying covariates means that the outcome variable, *Dropout*, must not affect the included covariates. Such effects are known as feedback effects and they are related to the causality in the model. In this thesis, we have more or less implicitly assumed that the causal link goes from the variables for living conditions to dropout and not the other way around. This assumption could be violated if one thinks of students who live at dorms or other types of student housing. These students are in some cases obliged to be active students to live there which might have an impact on the choice of dropping out or continuing studying. In other words, a student may choose to remain enrolled in period t primarily to be able to live at the dorm or student housing in period $t+1$. In such a case, *Distance* in period $t+1$ is affected by *Dropout* in period t, i.e. there are feedback effects and the time-varying covariates do not fulfill the assumption of strict exogeneity. However, we believe this issue to be of smaller magnitude since only a limited amount of students live in such housing.

The second assumption states that the effect from time-varying covariates are constant over time, i.e. the estimated βs do not vary. An alternative method, which is applied by Gury (2011), is to allow the effect of the covariates to vary over time. By this method, is would be possible to determine at which point in time the effect from the covariates have the largest impact on the risk of dropping out. Further, Gury (2011) argues that allowing the effect to vary over time is favorable since this approach takes into account that the study population changes over time, and thereby the risks that they are exposed to as well.

In the paper by Gury (2011), allowing for time-varying effects may be the most appropriate method since he analyzes dropout over more than four years. We believe that our approach using time-varying covariates and time-invariant effects is meaningful in this thesis for two reasons. First, time is measured quite imprecisely as seen in Figure 4.1, which makes it hard to investigate the exact meaning of time. Secondly, we only consider living conditions and dropout during the first year, which is a quite limited period of time. Therefore, the assumption of a constant effect from living conditions during the first year seems credible.

7.2.2 Recording of time: Discrete or continuous?

In section 5.3.1, the assumption of continuous time was introduced. Even though we observe data as discrete, we assume that the underlying time is continuous and that a continuous duration model can be applied. This assumption may be argued to be violated given that students are observed at three times during the first academic year. Nevertheless, in this subsection, we argue that the assumption holds.

If one assumes discrete time, this means that the data is interval-censored, i.e. dropout occurs between two known time points. Interval-censoring may result in difficulties in the exact ordering of dropout times, which is crucial for the estimation explained in Chapter 5 (Cleves et al., 2010, p. 32). With the applied data that has 3 observations during 1 year, the exact ordering of dropout times is not obvious because the students are observed in intervals with several dropouts in each. In other words, there are ties in the data, but as mentioned in Chapter 5, the ordering is approximated by Efrons method and thereby, this should not be a problem.

Interval-censoring is also an issue if dropout times overlap or are uncertain, but this is not the case in this thesis since all students are observed to drop out in the same intervals. Therefore, we believe that in our setup, interval-censoring is a minor challenge. This is in line with Cameron and Trivedi (2005, p. 588), who state that the degree of interval-censoring is often assumed to be so small that it can be ignored. To sum up, we believe that the assumption of continuous time is reasonable in our setup.

7.3 Accounting for group effects

In Chapter 5, the potential need to account for group effects is noted at first. This is incorporated in the analysis as stratified and frailty models. The focus in this subsection is to discuss to what degree group-specific characteristics have been properly controlled for to avoid misleading conclusions. In particular, the following subsections discuss the chosen specifications. Hereafter, we discuss controlling for heterogeneity on an individual level instead of on a group level.

7.3.1 Stratification and frailty models

The subsection discusses whether to rely on the stratified model or the frailty model when there are differences between the estimated effects on dropout between the models. In general and in opposition to the baseline model and the frailty model, the stratified model returns significant effects for the controls. In the presentation of Table 6.1, we argue that this last issue is due to the stratas being quite homogeneous groups.

On one hand, the tests conducted in Section 6.5 speak in favor of using the stratified model as it is the only model that does not break with the assumption of proportional hazards. This is a strong argument for using the stratified model. Further, the model allows for a flexible functional form in that all stratas are allowed to have different baseline hazards. On the other hand, there is no direct estimate of the importance of the strata effect, in other words, no p-values are calculated (Therneau and Grambsch, 2000, p. 45). This means that it is not possible to do inference about the stratification variable. Therefore, one must lean on theoretical and intuitive arguments as to why the chosen variable for stratification is meaningful. We have argued for that in Section 5.4 and believe that this is reasonable. However, a point of critique when stratifying is that within some of the stratas, there might be too little variation data which results in the insignificant control variables. Nevertheless, as we are mainly interested in the effect from living conditions to dropout, it is of less importance.

Stratification only allows to control for observed heterogeneity at group level but controlling for unobserved factors is important when student dropout is analyzed. Frailty models are the standard duration models for random effects and thereby, they allow for modelling of unobserved heterogeneity. Estimating a frailty model could increase our knowledge on student dropout behavior if there is unobserved heterogeneity between groups in data. An attractive feature of frailty models is that the model output provides an estimate of the within-group correlation. The estimate was found to be significant in all the frailty models estimated in Chapter 6. This implies that there is significant unobserved heterogeneity which confirms that for the applied data, frailty modelling adds information and should be applied.

A drawback of the model, as found in Section 6.5, is that the frailty model breaks with the proportional hazard assumption. This speaks for relying on the strata model. To sum up, there are arguments for and against both models, but fortunately, they support each other on whether living conditions matter for dropout. This is the case as the majority of the results for these variables are robust across the two model specifications. On one hand, this can be surprising to the due differences between the models highlighted above and in Chapter 5. However, as this is not the case in most models, we argue that this is due to robustness of the results. Nevertheless, attention must be paid to interpreting the results when the significance differs across the models.

7.3.2 Accounting for individual-specific effects

The above mentioned controls for group-level characteristics may fail to capture some important characteristics which determine the behavior of students. In particular, one can argue that controlling for group effects does not suffice and that one should control for individual effects. Therefore, as an alternative to the applied methods, this section discusses controlling for individual effects.

The first of the two alternative approaches is to stratify on individual level where the stratifying variable is the specific id number for each student, see e.g. Allison (2009, p. 74) who refers to this as the "Cox fixed effects" model. However, that does not return a meaningful model: most of the variables are omitted. This is most likely a result of too little information within each stratum as the stratified model optimizes within each strata. In some strata, i.e. for some students, there is only one observation which implies that variation is very restricted. We therefore turn to the second alternative. Hosmer et al. (2011, p. 307) recommends to use a frailty model if the group size is less than 5. For the cases, where we have more than one observation per student, this would still be a shared frailty model but on an individual level. Nevertheless, due to computational issues, this model could not be estimated with `Stata` because the software cannot account for more than

19,000 individual frailties. Therefore, based on the relatively robust results, we argue that it is sufficient to account for unobserved heterogeneity on a group level.

To sum up, despite different controls for group-level characteristics, there was an overall agreement in the estimated results across the alternative approaches. Further, we argue that because the applied data contains many individuals and relatively few observations for each individual, accounting for individual-specific effects is not feasible practically. In our opinion, it would have been interesting to account for individual-specific effects. However, it is not possible and it seems that the analysis that accounts for group effects does come a long way.

7.4 Region and sector specific results

This subsection considers the effects of the regional and sectoral results and discusses why the overall picture based on the analyses is blurred. We do not find any clear pattern in the effects of living conditions on dropout across regions. In particular, we expected to find stronger results from living conditions to dropout in the Capital Region and the Central Region which turned out not to be the case. For the Central Region this might be because the regions is a too large and heterogeneous area to consider. Nevertheless, the same argument can not be made for the Capital Region close to all educational institutions in the region are located in Copenhagen. On the other hand, we could argue that we do not see the large effects of e.g. *Distance* we expect in the Capital Region, because the infrastructure in the city is quite good, which means that the students are not as easily affected by this than they would otherwise have been.

7.5 Policy recommendations

Before the actual policy recommendations are presented, we briefly remind the reader that our focus is on the educational sector. In other words, the recommendations given below consider the educational sector without discussing how the recommendations could potentially affect other sectors, such as e.g. the construction sector.

Regarding the results, the findings indicate that living conditions do matter for dropout among Danish students. At a general level, they point to 1) the longer transportation time a student has between his home and the institution at which he studies, the larger is his probability of dropout, 2) the more worried a student is, the larger is his probability of dropout and 3) if a student moves at the beginning of the first semester, it lowers his probability of dropout.

With these overall findings, the first recommendation is related to student accommodations. In particular, the number of available student accommodations, when they are available for the students and as well where they are located. Based on the results from *Distance* and *Move*, the student accommodation should be placed relative close to the educational institutions and further, they should be available for the students from the beginning of the first semester in order to lower first year dropout. The concern of location is closely related to the recommendation of construction of new student accommodation. The evidence presented in this thesis implies that the location is of importance: it is not enough to have a place to stay. That is, the optimal living conditions should aim at having a supply that matches demand and e.g. that the distance to institutions should not be too long. This does not necessarily mean that the student housing should be built right next to the institutions. They can be built at another location, however, in order for the students to have a relatively low transportation time, the infrastructure should be good.

While it is easy to relate *Distance* and *Move* directly to policy recommendations on student housing, *Worry* is also important. Improving students living conditions e.g. by building more student accommodations could potentially reduce dropout if that would make students worry less about their housing situation. An analysis by The Danish Construction Association (2018) estimates that there was a shortage of approximately 22,000 student accommodations in Denmark by the beginning of September 2018. This means that the demand for student housing is much higher than the supply.

The very clear recommendations that can be made based on the overall results, become more nuanced the regional and sectoral analysis are taken into account. Based these, the overall picture of how living conditions affect dropout is more complex which should be taken into account in the recommendations. As an example findings suggested that *Distance* had a significant effect on students in Capital Region, Central Region and North Region, but not in the South Region and Region Zealand. As there is no obvious explanation for this, we are careful to provide policy advice based on the regional and sector analysis. However, as for the regional analysis, an interesting finding, in contrary to the expected, was that students in Capital Region and Central Region were not at a higher risk of dropping out due to living conditions. Besides that, the analysis on regions and sectors did not find any clear pattern in evidence that students where affected differently as also discussed above. For that reason, we base our policy recommendations on the findings on all students, across regions and sectors.

The previous paragraphs highlighted that at the moment, there is a larger demand than supply for student accommodation. In the near future, the demand is likely to fall due to demographics. On the other hand, demand might increase as a result of changed rules for loans regarding parental purchase of apartments. This point is elaborated in the following. If the number of individuals at the normal age for starting a higher education is reduced in the future, this could have an impact on the number of students applying for a higher education, *ceteris paribus.* This may decrease the demand for student housing. As mentioned, another factor that can influence the demand is changed regulation for loans. This could potentially reduce the number of apartments bought, which could increase the demand for student accommodations. It also seems plausible that although there may be a decreased number of students in the future, the large cities will still experience a housing shortage as the students will prefer the large institutions in the large cities. Therefore, we still argue that there may be a shortage in the future so the policy advice on constructing more housing remains.

Chapter 8

Conclusion

The purpose of this thesis is to investigate the effect of living conditions on first year dropout from higher education in Denmark. In order to do so, we employ the following variables for living conditions; distance measured in minutes between the students' home and educational institution, how worried the student is about his housing situation and whether the student moved at the beginning of the semester. These variables are incorporated in a Cox proportional hazard model with additional control variables to account for selection. The model is extended to incorporate time-varying covariates and ties as both are present in the applied data. Further, in order to account for observed and unobserved group effects, the relationship between living conditions and dropout is also investigated based on both a stratified and a frailty Cox model, respectively.

The overall conclusion is that living conditions do matter for drop out and this conclusion is relatively robust across the model specifications. In particular, we find that a student that lives an additional 10 minutes away from his institution has a 5 percent higher probability of dropping out. Also, we find that if a student increases his levels of worries about living conditions on a scale from 1-5 by one unit, it will lead to an increase in the probability of dropout of around 7-8 percent. Finally, we find that students who move at the beginning of the first semester are around 14 percent less likely to drop out than their peers who do not move at the beginning of the semester.

Regional and sectoral effects are also investigated but findings suggest no clear pattern that can be explained. This indicates a more complex association between living conditions and dropout. In other words, contrary to our expectations, we did not find evidence for a stronger relationship between living conditions and dropout in the Capital Region or the Central Region. In particular, the results showed that students in the Capital Region, the Central Region and the North Region all have significant effects from *Distance* to dropout, while the effects are insignificant in Region South and Region Zealand. For *Worry*, the pattern was the opposite with significant effects in especially Region South and also Region Zealand, but insignificant effects in the remaining regions. Finally, *Move* appeared only to be significantly related to dropout in Region Zealand. For the sectoral effects, we found that while students at universities and university colleges experience effects from the variable for distance, this is not the case for business academy students. However, students at business academies were found to have a significant association between *Worry* and dropout, which was not the case for university college students. The effect for university students was less clear. Further, the results suggest that primarily students at universities seem to have an effect from moving at the beginning of the first semester.

It is also investigated if academic or social integration removes the effect from living conditions to dropout. We find this to only be the case for the level of worries, i.e. *Distance* and *Move* are still significant in a model where the variables for integration enter. Finally, we investigated heterogeneity in the group of students that live far away. It is hypothesized that the group consists of older students with children that have settled down and younger students that are eager to move closer to the institutions where they study. As expected, interaction terms between distance and a dummy for having children and being above age 30 return insignificant results, which is in line with the hypothesis. However, we note that this conclusion is not very strong as the number of students above 30 with children is very small.

Overall, the thesis found a strong effect from living conditions to dropout on a national level and it is not affected by inclusion of controls for academic and social integration. Relying on the these results and acknowledging our contribution to the research gap in the area, it therefore seems reasonable to state that living conditions do have an impact on first year dropout from institutions of higher education in Denmark.

Future work could address the effects from living conditions to dropout in Aarhus and Copenhagen more explicitly. While the regional effects were not as expected, effects on a city-level might be more plausible. One could imagine, that educational institutions located in these cities could show a larger effect.

Appendix A

Appendix

A.1 Data management on housing variables

Values of the covariate *Distance* was replaced for values above 200 minutes by mean substitution. Mean substitution meant that the average was calculate on students reporting a distance of 200 minutes or below. This average then replaces values that where reported to be above 200 minutes, and these students were also given a dummy variable taking the value 1, if the former value has been replaced with the average. A benefit of mean substitution is that it does not change the sample mean for *Distance*. On the other hand, the method attenuates any correlation from the variable that was substituted, but we believe to be able to account for this by using the created dummy.

An empirical challenges regarding the time-varying variables, *Distance*, *Move* and *Worry* is that there are gaps in the data set. The gaps are results of nonresponse from students or that the answer ”I do not know” is treated as a missing value. We have filled some gaps with a the following assumption: when a student reports the same value of distance in the first and third wave, the missing value in the second wave is assumed to be the same. Intuitively, this seem to us to be a realistic assumption. This procedure has been implemented for variables $Distance$, $Move$ and $Worry$. The advantages of using this procedure to fill the gaps is that this leads to a larger sample since Stata otherwise would perform listwise deletion on each student with gaps. Thereby, we could improve the precision of inference (Cameron and Trivedi, 2005, p. 923). We argue that it is meaningful to obtain more observations in cases where it should be quite obvious what value is missing.

A.2 Data management for parental education

For parental education, we have treated the category ”unknown” of values of parents education as the lowest possible educational level. Our idea is that there may be parents who have not studies in Denmark and therefore their education has not been registered in KOT. we believe that if they had an education below the lowest level, we would know this. we are talking about approximately 1800 missing values for parental education.

A.3 Weighting

For the weighting, we follow the method outlined by Little and Rubin (2002, pp. 48-49), which is supported by Kessler, Little and Groves (1995) and Fitzmaurice (2011)). The method is referred to as propensity weighting. It basically considers the probability of responding conditional on characteristics known for both respondents and non-respondents. The procedure is the following:

1. Estimate $p(X)$ as $\tilde{p}(X)$ by logistic regression, using the indicator of missingness, M and letting known controls for both population and sample enter

$$Pr(M_i = 1) = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(x'\beta)} \tag{A.1}$$

2. Predict the probability of missingness based on (A.1).
3. Form inverse probability scores by taking the inverse of the predicted probabilities, $[\tilde{p}(X_i)]^{-1}$ (Little and Rubin, 2002, p. 49).

Thereafter, each individual responding in the first wave is weighted. It is most common to rely on the first wave when weighing longitudinal data (Fitzmaurice, 2011). A technical detail is the `Stata` software of including weights is not compatible with the preferred method controlling for ties, Efron’s method. One resolution is to implement weights by using Breslow’s method for ties. The results using the weighted sample is presented below.

The results in the the table are close to those presented in Table 6.1. Only the baseline model and the stratified model are presented. This is because it was not possible to estimate the frailty model in `Stata` with weights. However, given that the results for the two other model specifications are so similar to the un-weighted results, we trust that this would also be the case for a weighted frailty model.

Table A.1: Weighted overall results

VARIABLES	(1) Baseline	(2) Strata
Female	0.953	1.041
	(0.058)	(0.079)
Vocational †	0.945	0.960
	(0.091)	(0.099)
Short-term higher education †	0.761*	0.744**
	(0.106)	(0.107)
Medium-term higher education †	0.754***	0.804**
	(0.076)	(0.086)
Long-term higher education †	0.813*	0.848
	(0.089)	(0.099)
High School GPA	0.952***	0.974
	(0.013)	(0.017)
Age	0.923***	0.989
	(0.022)	(0.027)
Age^2	1.001***	1.000
	(0.000)	(0.000)
Distance	1.005***	1.005***
	(0.001)	(0.001)
Move	0.867**	0.865**
	(0.051)	(0.052)
Worry	1.071**	1.085***
	(0.029)	(0.030)
Wald (chi^2)	97.52	63.90
DF	12	12

*** p<0.01, ** p<0.05, * p<0.1.
† The educational levels refer to parents education.
Note: 38,586 observations, 35,491 individuals and 2,365 dropouts.
The number of individuals and dropouts appear higher
due to the weighting.
Dum_dist is omitted from the results.
Source: EVA and Statistics Denmark.

A.4 Quadratic function

We find the extremum for age using the formula presented by equation (A.3) (Wooldridge, 2010, pp. 704-705; Cleves et al., 2010, pp. 180-181).

$$y = \beta_1 age + \beta_2 age^2 \tag{A.2}$$

$$age_{\text{extremum}} = \frac{\beta_1}{(-2\beta_2)} \tag{A.3}$$

A.5 Analysis on students above age 30 with children

Table A.2: Controlling for students at 30 or above with children

VARIABLES	(1) Baseline	(2) Baseline	(3) Strata	(4) Strata
Female	0.949	0.947	1.034	1.027
	(0.060)	(0.060)	(0.082)	(0.081)
Vocational[†]	0.941	0.940	0.963	0.961
	(0.093)	(0.092)	(0.103)	(0.102)
Short-term higher education[†]	0.743**	0.742**	0.734**	0.730**
	(0.106)	(0.106)	(0.109)	(0.109)
Medium-term higher education[†]	0.757***	0.759***	0.805**	0.805**
	(0.078)	(0.078)	(0.089)	(0.089)
Long-term higher education[†]	0.816*	0.817*	0.861	0.861
	(0.091)	(0.091)	(0.104)	(0.104)
High School GPA	0.955***	0.952***	0.979	0.976
	(0.013)	(0.013)	(0.018)	(0.017)
Age	0.923***	0.895***	0.981	0.953
	(0.023)	(0.024)	(0.028)	(0.029)
Age^2	1.001***	1.001***	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)
Distance	1.005***	1.005***	1.005***	1.005***
	(0.001)	(0.001)	(0.001)	(0.001)
Move	0.847***	0.843***	0.852**	0.848***
	(0.051)	(0.051)	(0.053)	(0.053)
Worry	1.068**	1.070**	1.083***	1.085***
	(0.030)	(0.030)	(0.031)	(0.031)
Distance*Child*Age30		0.996		0.995
		(0.004)		(0.004)
Child	1.381	1.252	1.777**	1.413**
	(0.306)	(0.221)	(0.430)	(0.249)
Age30		1.773***		1.769***
		(0.339)		(0.368)
Distance*Child	0.997		0.994*	
	(0.003)		(0.004)	
Wald (chi^2)	96.53	106.9	70.99	78.42
DF	14	15	14	15

*** p<0.01, ** p<0.05, * p<0.1. [†] The educational levels refer to parents education. The comparison group is primary and secondary school

Note: 38,205 observations, 19,032 individuals and 1,284 dropouts.

Dum_dist is omitted from the results.

Source: EVA and Statistics Denmark.

A.6 Distribution of distance

FIGURE A.1: Distribution of distance across regions

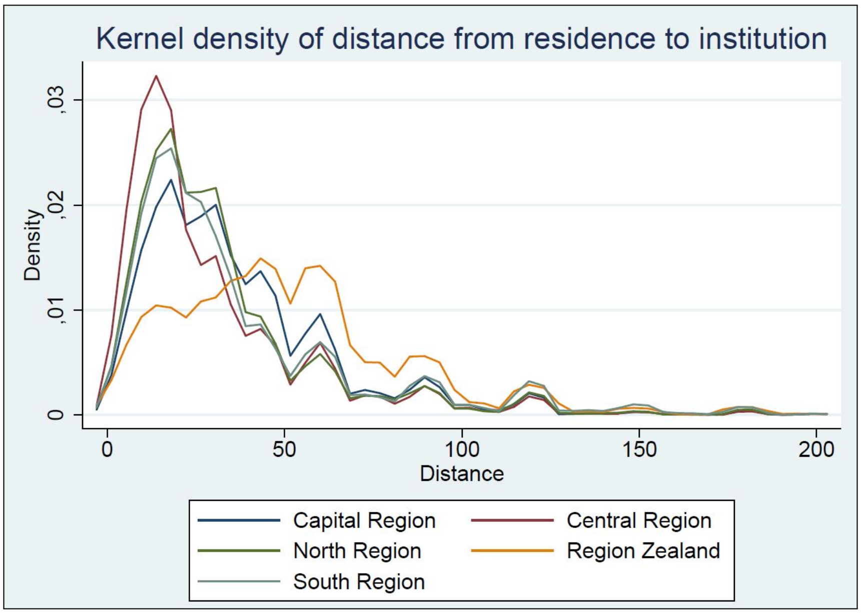


FIGURE A.2: Distribution of distance across sectors

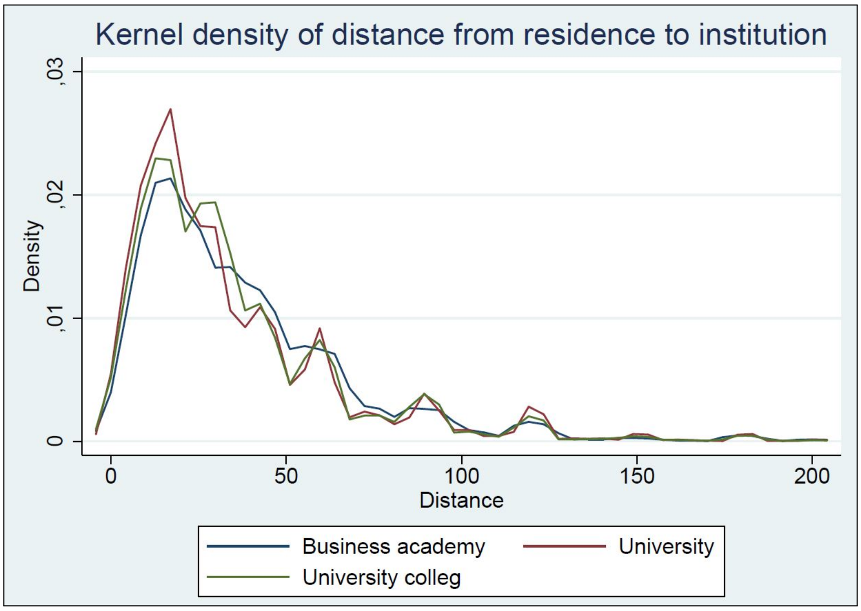


Table A.3: Respondents reported distance across regions and waves

	Wave 1	**Wave 2**	**Wave 3**
Capital Region	38.7 (6.622)	37.7 (4,099)	34.3 (2,873)
Central Region	31.6 (5,101)	30.3 (3,105)	27.2 (2,267)
North Region	34.5 (2,324)	33.3 (1,345)	30.3 (899)
Region Zealand	52.5 (1,384)	50.9 (826)	47.1 (545)
South Region	39.0 (3,601)	38.7 (2,118)	34.3 (1,432)

Note: Respondents in parentheses
Source: EVA and Statistics Denmark

A.7 Distribution of *Worry*

Figure A.3: Distribution of worry across regions

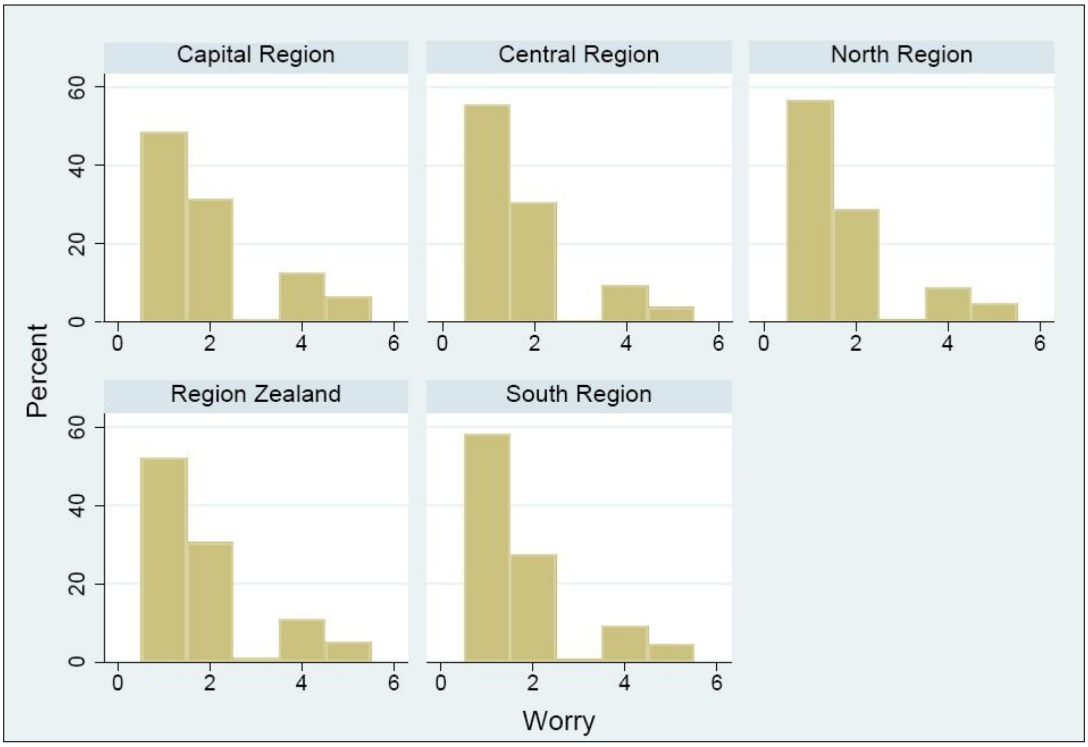
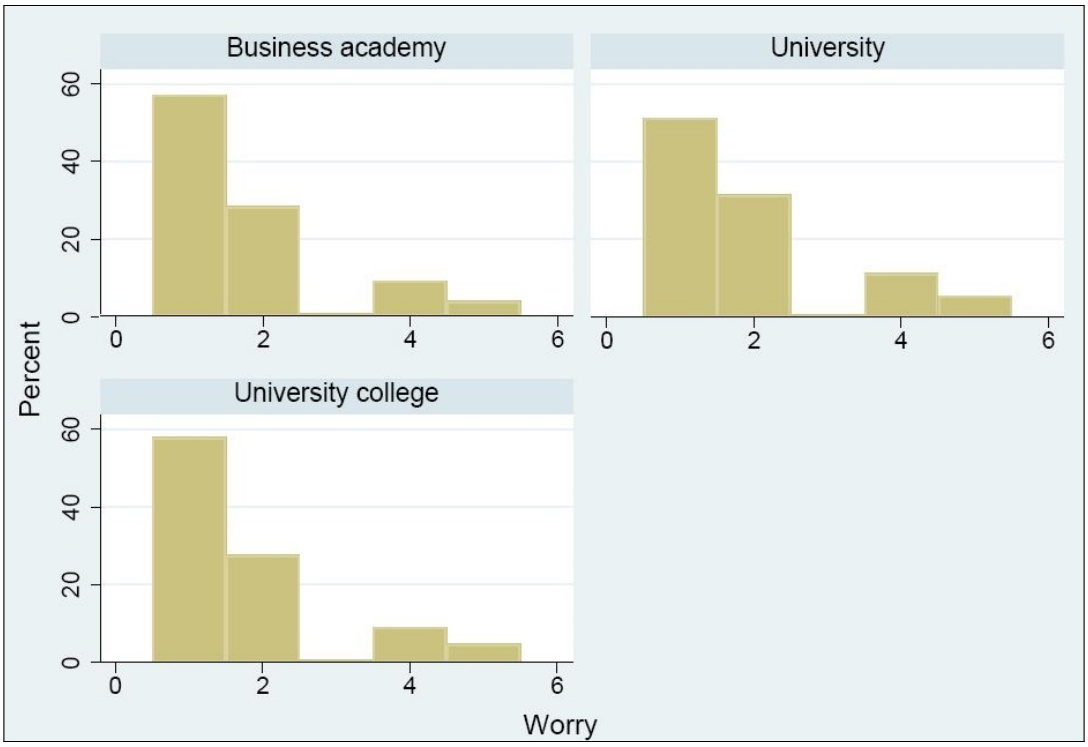


Figure A.4: Distribution of worry across sectors



A.8 Dropout rates

This sections breaks the dropout rates presented in Table 4.1 down into sectoral and region dropout rates.

Table A.4: Sectoral dropout rates

	Wave 1	**Wave 2**	**Wave 3**
University	0.9 %	5.1 %	7.2 %
University College	0.9 %	4.5 %	4.4 %
Business Academies	1.3 %	7.1 %	6 %

Source: EVA and Statistics Denmark.

Table A.5: Regional dropout rates

	Wave 1	**Wave 2**	**Wave 3**
Capital Region	0.8 %	4.8 %	5.1 %
Central Region	1 %	4.9 %	6.8 %
North Region	0.9 %	5.3 %	8.1 %
Region Zealand	1.6 %	6.2 %	5 %
Region of South Denmark	1 %	6.1 %	7.3 %

Source: EVA and Statistics Denmark.

Table A.6: Characteristics

	Dropouts	**Actives**
Distance	40.82	37.3
Worry	1.82	1.83
Move	37 %	38 %
Female	63 %	61 %
Age	22.24	22.04
High School GPA	8.1	8.23

Source: EVA and Statistics Denmark.

A.9 Distribution of *Move*

Figure A.5: Mean of move across regions

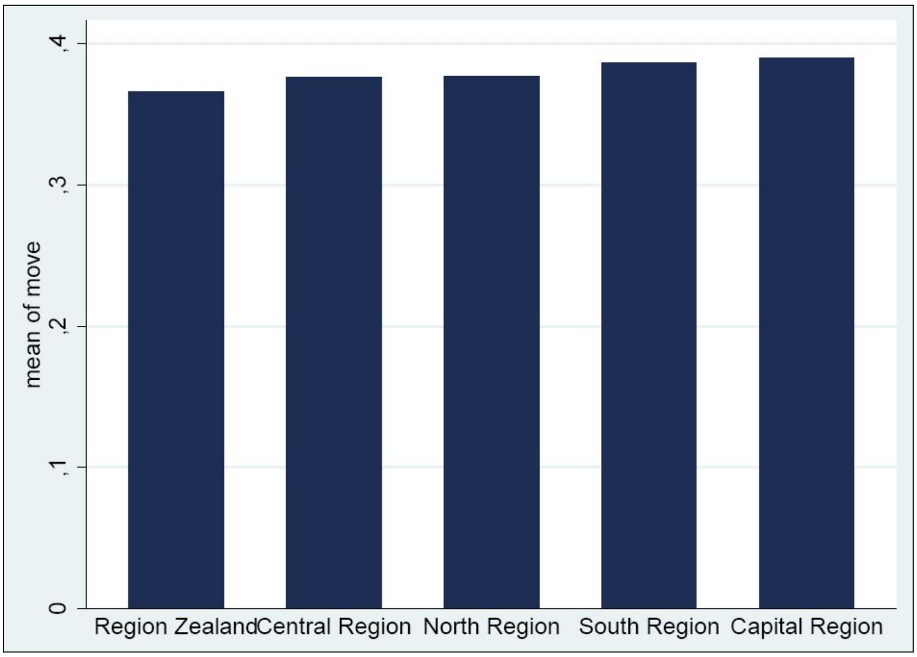
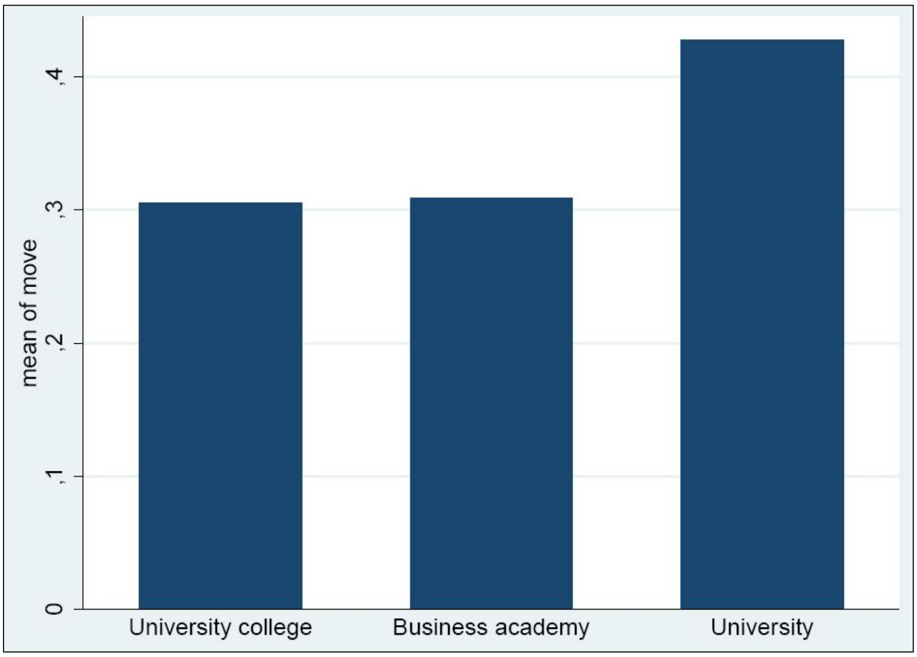


Figure A.6: Mean of move across sectors



Bibliography

Allison, P. D. (2009). *Fixed effects regression models.* SAGE Publications.

Arulampalam, W., Naylor, R. A. & Smith, J. P. (2004). A hazard model of the probability of medical school drop-out in the uk. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *167*(1), 157–178.

AU Student Council. (2000). Frafald og studiemiljø. Retrieved from http://sr.au.dk/PDF/frafald/FFrapport.pdf

Bozick, R. (2007). Making it through the first year of college: The role of students' economic resources, employment, and living arrangements. *Sociology of education*, *80*(3), 261–285.

Cameron, A. C. & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge University Press.

Chambers, R. L. (2003). *Analysis of survey data.* Wiley series in survey methodology. Chichester, England Hoboken, NJ: John Wiley.

Chen, R. (2008). Financial aid and student dropout in higher education: A heterogeneous research approach. In *Higher education* (pp. 209–239). Springer.

Cleves, M., Gould, W., Gutierrez, R. B. & Marchenko, Y. V. (2010). *An introduction to survival analysis using stata.* Stata Press, 3rd edition.

Cunha, F. & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47. Retrieved from http://www.aeaweb.org/articles?id=10.1257/aer.97.2.31

DesJardins, S. L. (2003). Event history methods: Conceptual issues and an application to student departure from college. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 421–471). Springer.

DesJardins, S. L., Ahlburg, D. A. & McCall, B. P. (1999). An event history model of student departure. *Economics of education review*, *18*(3), 375–390.

DMA Research. (2002). Frafald på lange, videregående uddannelser. frafaldsårsager.

Eurostat. (2018). *Data on expenditure on education - table educ_figdp*. Retrieved from http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=educ_figdp&lang=en

Fitzmaurice, G. M. (2011). *Applied longitudinal analysis* (2nd ed..). Wiley.

FTF. (2016). *Su-nedskæringer kan skabe øget frafald for 300 mio. kr. årligt*. Retrieved from https://www.ftf.dk/aktuelt/ftf-analyse/artikel/su-nedskaeringer-kan-skabe-oeget-frafald-for-300-mio-kr-aarligt

Gury, N. (2011). Dropping out of higher education in france: A micro-economic approach using survival analysis. *Education Economics*, *19*(1), 51–64.

Hoff, J. V. & Demirtas, M. (2009). *Frafald blandt etniske minoritetsstuderende på universitetsuddannelserne i danmark*. Forlaget Politiske Studier.

Höfler, M., Pfister, H., Lieb, R. & Wittchen, H.-U. (2005). The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology*, *40*(4), 291–299.

Holm, C., Laursen, K. B. & Winsløw, C. (2008). Hvorfor gik de ud? en analyse af frafald på årgang 2006 af matematikstudiet. Retrieved from https://www.ind.ku.dk/udvikling/projekter/IMF-frafald/

Hosmer, D. W., Lemeshow, S. & May, S. (2011). *Applied survival analysis: Regression modeling of time-to-event data*. Wiley-Interscience.

Ishitani, T. T. & DesJardins, S. L. (2002). A longitudinal investigation of dropout from college in the united states. *Journal of college student retention: research, theory & Practice*, *4*(2), 173–201.

Kessler, R. C., Little, R. J. & Groves, R. M. (1995). Advances in strategies for minimizing and adjusting for survey nonresponse. *Epidemiologic Reviews*, *17*, 192–204.

Lassibille, G. & Navarro Gómez, L. (2008). Why do higher education students drop out? evidence from spain. *Education Economics*, *16*(1), 89–105.

Light, A. & Strayer, W. (2000). Determinants of college completion: School quality or student ability? *Journal of Human Resources*, 299–332.

Little, R. & Rubin, D. (2002). *Statistical analysis with missing data, second edition*. Jon Wiley and Sons.

McCullagh, P. & Nelder, J. (1989). *Generalized linear models* (2. ed.). Chapman and Hall.

OECD. (2018). *Education at a glance 2018*. Retrieved from https://www.oecd-ilibrary.org/content/publication/eag-2018-en

Schudde, L. T. (2011). The causal effect of campus residency on college student retention. *The Review of Higher Education*, *34*(4), 581–610.

Smith, J. P. & Naylor, R. A. (2001). Dropping out of university: A statistical analysis of the probability of withdrawal for uk university students. *Royal Statistical Society*.

The Danish Agency for Science and Higher Education. (2018). Frafald og studieskift. Retrieved from https://ufm.dk/publikationer/2018/frafald-og-studieskift/

The Danish Construction Association. (2018). Stadig alt for få boliger til de mange nye studerende. Retrieved from https://www.danskbyggeri.dk/presse-politik/nyheder/2018/stadig-alt-for-faa-boliger-til-de-mange-nye-studerende/

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling survival data, extending the cox model*. Statistics for biology and health. New York: Springer.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45*(1), 89–125.

Tinto, V. (2012). *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago Press.

Wang, G. & Aban, I. (2015). Application of inverse probability weights in survival analysis. *Journal of Nuclear Cardiology*, *22*(4).

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach*. Nelson Education.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.