DOI: 10.20472/IAC.2016.022.037

MAJA MIHALJEVIC KOSOR
Faculty of Economics, University of Split, Croatia

STUDENT DROPOUT IN HIGHER EDUCATION: AN APPLICATION OF HAZARD FUNCTIONS

Abstract:

Hazard functions are a part of survival analysis which is a branch of statistics dealing with failure in mechanical systems and death in biological organisms e.g. lifetime or reliability of machine components, survival times of patients in clinical trials. Here, the interest is focused on a group of individuals, for which there is a defined point event, often referred to as failure, arising after a length of time, referred to as the failure time. To gain more insight into student dropout we examine the application of hazard functions in higher education. In such a model, the probability is investigated that the student will complete/leave a degree in a given year conditional on him/her having 'survived' the programme up to that point. This may allow a wider analysis as it captures both students who have and have not completed their studies and examines the impact of selected variables for the duration of student's higher education course.

Keywords:

hazard functions, student droput, duration analysis

JEL Classification: I23, I29, C40

1. Introduction to Hazard Functions

In the following sections, we examine a hazard model for assessing the probability of student dropout in higher education (HE). In such a model, the probability is investigated that the student will complete/leave a degree in a given year conditional on him/her having 'survived' the programme up to that point. This may allow a wider analysis as it captures both students who have and have not completed their studies and examines the impact of selected variables for the duration of student's HE course.

First, some of the general characteristics of duration analysis and hazard functions are briefly presented. Following this is a literature review of hazard functions and their application in the economics of higher education. At the end of the paper, given the characteristics of the duration analysis and hazard functions and their disadvantages, a rationale is presented as to whether the examination of student non-completion would benefit from introducing this type of analysis.

2. Hazard functions

Hazard functions are a part of survival analysis which is a branch of statistics dealing with failure in mechanical systems and death in biological organisms e.g. lifetime or reliability of machine components, survival times of patients in clinical trials. Here, the interest is focused on a group of individuals, for which there is a defined point event, often referred to as failure, arising after a length of time, referred to as the failure time (Cox and Oakes, 1984). Three conditions need to be satisfied in order to determine the failure time: the starting time must be unambiguously defined, a scale for measuring the passage of time must be agreed and the meaning of failure must be completely clear (Cox and Oakes, 1984).

The duration modelling literature in economics has addressed specific issues with duration data. One of the more extensively studied duration data in economics are data on the length of spells of unemployment where Lancaster (1979) and Nickell (1979) applied hazard function methods to examine unemployment duration. Some additional areas of application of duration models in economics are the time taken to adopt new technology, survival of firms, time to invention from research investment and student persistence in tertiary education.

The focus in duration modelling is not the unconditional probability of an event occurring (e.g. the probability of a student completing his/her HE in exactly 5 years) but the conditional probability (e.g. the probability that a student will drop out in the fifth year of study given that he/she persisted for four years already). Although these two types of probabilities are mathematically similar, their conceptual difference is of relevance in duration modelling. As summarised by Kiefer (1988), duration analysis has advantages in interpreting the type of data for which the representation in terms of conditional

probabilities is theoretically and intuitively supported. Here, the hazard function specification serves to highlight the conditional probabilities while the probability distribution stresses unconditional probabilities. The literature on duration modelling and the hazard function mostly relies on three probability distributions, exponential, Weibull and extreme-value distributions. Hence, the specification of a hazard function is an alternative to specification of a probability density function where economic data can be modelled as generated by a series of sequential decisions (Kiefer, 1988).

In the duration modelling literature the hazard function is also known as the hazard rate, failure rate and force of mortality, however the term hazard function is used in this paper. The hazard function can simply be defined as the ratio of the probability density function to the survival function where the survival function describes a probability that a random variable takes on a value equal to or greater than some value t (Evans et al., 2000). Next, this relationship is presented in more detail drawing on the work of Cox and Oakes (1984) and Keifer (1988) where a population of individuals is considered and each one has a 'failure time'. For research in this paper it is useful to consider this 'failure time' as the moment when a student 'dropped out'.

3. Advantages in using hazard functions

From the discussion above some of the advantages of using duration analysis over more conventional regression analysis may be emphasised. The principal advantage is related to the characteristics of duration data, e.g. if the interest lies in student persistence in HE the duration of which is a dependent variable and its determinants are measured by a set of exogenous variables across a number of individuals, then the problem that arises in conventional regression analysis is how to measure the explanatory variables whose values change during the time that the student is still enrolled. Given that the focus is on conditional probabilities in some time frame (typically a year), both completers and non-completers in the year, as well as those whose study is still in progress, are included in the analysis. As Ehrenberg and Mavros (1995) indicate, because each individual/year observation is considered as a separate observation, using hazard functions permits time varying covariates and there is also a possibility to include additional information (e.g. on student's financial support or labour market conditions) that a student deals with each year. Hence, the advantage of a duration model when compared to the regression framework is that time-variant covariates can be introduced without conceptual difficulties and that right-censored observations can be simply handled (Häkkinen and Uusitalo, 2003). However, as becomes apparent in the following section, in practice authors rarely use time-varying covariates in duration modelling and furthermore, the set of variables employed in this modelling is much smaller that the one emphasised by the theory of student persistence and non-completion presented above.

4. Literature Review of the Application of Hazard Functions in HE

In this section empirical work on higher education which employs hazard functions as an estimation technique is reviewed. A general overview of the topics and models considered by authors is provided, the choice of explanatory variables is discussed and potential limitations of the research are addressed. A general conclusion on the papers considered and the assessment as to whether to employ hazard functions is presented at the end of this section.

Modelling the time-to-complete using duration or hazard models was first applied in the UK data to doctoral students (in Booth and Satchell, 1995). It was later applied to other levels of tertiary education however the application of duration analysis to PhD level education still dominates in the literature. All models examined in this section estimate the probability that a student will complete a degree in a given year or that he/she will drop out in a given year conditional on him/her having 'survived' until that point in time . Hence, both students who have or have not completed are included in the data set as well as the students whose study is still in progress, making it possible to follow the impact of chosen variables on a student during his/her study. In such a situation a competing risks model is specified (e.g. in Cox and Oakes, 1984) with right censoring for all students who did not complete or did not withdraw . The research undertaken of PhD study is similar in the choice of variables and techniques to the work on HE (undergraduate programmes), hence some of these studies are examined in more detail.

Booth and Satchell (1995) used data on about 480 entrants to UK PhD programmes in 1980. They examined PhD completion and dropout rates and used their results to suggest that there are problems in using PhD completion rates as performance indicators for HEI's academic departments. Their dataset was nationally based, individual institutions were not identified and all fields of study were specified so the authors could not estimate whether there were different effects on dropout across fields. Furthermore, their dataset was rather limited as it only covered one entry cohort, thus preventing an examination of how, for example changing labour market conditions may influence the length of study and completion rates.

Building on the work of Booth and Satchell (1995) is the work of Ehrenberg and Mavros (1995) that uses individual level data from one US HEI awarding doctorates in four fields (Economics, English, Mathematics and Physics) over a 25-year period (from 1962-1986). They use data on student ability (proxied by the graduate record examination score - GRE), gender, citizenship and nationality, whether the student completed a master's degree prior to enrolment, the type of financial support received by students during the first six years they were enrolled in the programme and data on labour market opportunities by fields and years. From the discussion above it may be argued that Ehrenberg and Mavros (1995) also used a rather limited set of explanatory variables.

Primarily the authors did not fully capture student ability since it was proxied only by the GRE score which is a limited measure of student ability to complete PhD studies. This problem is made more important given the lack of a variable on prior schooling and attainment (other than a dummy variable indicating whether the student had an MA degree prior to enrolling on a PhD). Additional information on the type of the MA programme attended and the GPA at the Masters level or during the PhD studies would improve the model along with any other grades or scores prior to entering the PhD programme. Additionally, no peer effects were estimated although their importance is emphasised in the literature on student dropout in higher education.

Additional research on student persistence on PhD programmes is presented by van Ours and Ridder (2003) for Dutch PhD students in Economics. They use a hazard function estimation based on the models of Ehrenberg and Mavros (1995), i.e. the competing risks estimation where the outcome variables are the duration until completion of the dissertation and the duration until dropout. Van Ours and Ridder (2003) use data on 250 PhD students who enrolled by January 1993 and finished their studies by January 1998. The minimum duration of PhD studies is 4 years but most of the students finish within five to seven years from the start. However, it is unclear from the article how the authors are monitoring student persistence after the fifth year since the dataset ends in 1998. Given this, as the authors themselves warn, student dropout is adequately registered in their dataset only within the first four years, after which it is not possible to distinguish non-completers from completers in their dataset. This is a serious concern when using duration models since the data on the outcome variable is incomplete. Furthermore, from the starting sample of 250 students the authors removed incomplete observations leaving 200 students in the final sample. The final concern is related to the explanatory variables used, i.e. the authors include gender, duration of the undergraduate study, a dummy if the undergraduate degree is from the same university that employs the student's current PhD supervisor, the field of undergraduate degree and a dummy variable indicating whether the supervisor is a research fellow (i.e. if the supervisor is a successful researcher which in turn may deter dropout).

Some of the general conclusions and limitations that can be drawn from all the previously discussed studies on PhD duration are presented at the end of this section along with main issues and limitations in using duration analysis for analysing student persistence. As previously stated there are a small number of applications of duration analysis for HE and two papers will be presented dealing with hazard functions in the HE framework: Arulampalam et al. (2004) for UK medical schools, and Häkkinen and Uusitalo (2003) for Finnish HE.

Arulampalam et al. (2004) use duration analysis to model the probability of dropout in UK medical schools. The authors use variables reflecting the student's personal characteristics (age, gender, marital status, nationality), prior attainment (A or H-level

scores, A-level subjects), previous schooling (type of secondary school attended), socio-economic background (parental social class) and institutional characteristics. Their sample consists of two cohorts of medical students who enrolled in 1985 and 1986 and had completed/not-completed by 1993 when the last data was available. Hence, the students in the sample had a minimum of eight years for the 1985 cohort, and seven years for 1986 cohort, to complete their education or to 'drop out', although the standard length of the medical programme is five years. Dropout is defined as leaving the medical programme for whatever reason and completion is defined as obtaining a medical degree by the end of five to seven years (eight years for the 1985 cohort). Their results suggest that that the decision to withdraw from medical school is strongly influenced by prior attainment and the student's personal characteristics i.e. students with a higher A-level score and science subjects were markedly less likely to leave their HE studies prior to completion. Also, gender is statistically significant with male students being more likely to leave. On the other hand, the type of secondary school attended and socio-economic status were found not to be significant in influencing the probability of dropout.

Given the above, some of the limitations of empirical model by Arulampalam et al. (2004) may be noted. This is primarily related to the method used in analysing dropout where it is unclear how the authors are capturing students who completed, withdrew or who are still studying given that the model is estimated using logit with a year dummy equal to one if a student withdrew during the first four years. The authors stack their data and all students taking more than 5 years are lumped together. The analysis used employs no time-varying covariates, thus, not capturing how the student environment changes and what effect it may have on the probability of (non)completion. Furthermore, there is a lack of variables (e.g. peer effects and student effort) that have previously been found to be relevant in modelling student attainment as well as student dropout in HE.

The next paper is by Häkkinen and Uusitalo (2003) who examine how the reform of a student aid programme has influenced the duration of studies of 9,350 Finnish students since 1992 when a loan-based system was replaced by a system that relies on student grants. Also, the maximum duration of student aid was reduced. The duration of studies is measured as the number of months starting from September of the first enrolment year (from 1987-1997) until the last observation date – August, 2000. The official duration of most of the university programs in Finland is 5 years. The authors include the following explanatory variables: gender, whether the student has children, intersection term of the presence of children and gender, marital status, language group, age at entry, student-teacher ratio, local unemployment rate, unemployment rate in the specific field, 20 dummy variables indicating the field of study, year dummies, and an indicator if the student changed the field of study during the first four years (N=9350). The authors also estimate an additional model for a smaller sample of students for which they obtained information on parents' income and students' mean score in the matriculation exam (N=4600). In the empirical work Häkkinen and Uusitalo (2003) investigate only the

completion hazard, i.e. the probability of graduating after t months of study given that the student is still enrolled.

The authors find that older, married and female students have higher completion hazards. The coefficient on the year dummies indicates that the cohorts that entered after the student aid reform of 1992 have significantly higher completion hazards. Also a higher local unemployment rate increases the completion hazard. The main problem related to this work is in distinguishing how the non-completers were defined in the dataset, since at one point the authors state that their dataset cannot define non-completers (p. 8) and on p. 10 for the purpose of descriptive analysis of the dataset non-completers were defined as all students who have not yet completed their studies which would present an obvious overstatement of dropout in their dataset. For this research it may also be argued that the variables used were inadequate in modelling dropout and that the length of the dataset prevented the analysis of the characteristics of the later cohorts, i.e. students that enrolled in 1997 had only three years during which they were monitored, after that time the dataset does not provide information whether they completed or discontinued their studies.

5.Conclusion

From the above presented literature review it may be concluded that duration analysis and hazard functions have rarely been applied to analyse HE, though slightly more research exists at the PhD level. The few studies undertaken suffer from serious drawbacks and two main areas of limitations are evident. The primary limitations are related to the modelling technique and the variables used. The main problem here appears in distinguishing among three groups of interest/outcomes: students who are completing, not-completing and students who are continuing their studies with the latter group often excluded from investigation thus reducing the size of the sample. This is not a problem with the original use of hazard functions that were just concerned with survival times, but investigating dropout in education there is the third outcome (students who are continuing) and this makes the estimation problematic.

Furthermore, given the limitations of the datasets in terms of available variables and the time-frame under consideration, time-varying covariates are rarely applied in the literature, thus providing less insight into the determinants affecting student persistence. Indeed, the majority of the work presented above only uses explanatory variables observed at the time of student’s enrolment. Additionally, some of the basic variables affecting dropout such as student ability, prior attainment and peer effects are lacking or are not adequately represented. Moreover, a variable capturing student effort may have an important influence on the probability of dropout as suggested in the economics of education literature. However, this variable is rarely used in modelling dropout, mostly due to data limitations. The same is the case in the work of authors presented above.

Given the above mentioned limitations of the empirical work in this area the results need to be interpreted with caution and so do policy proposals stemming from this type of research.

The second type of limitation is related to the information obtained from the application of hazard functions and duration analysis. Here, the probability the student will complete/leave a degree in a given year conditional on him/her having 'survived' the programme up to that point is calculated. Most of the student dropout in tertiary education occurs during the first year of studies (e.g. in Mihaljevic Kosor, 2010). This indicates serious problems in dropout during the first year of studies and warrants the more detailed analysis of this issue as attempted in this paper. Therefore, it may be argued that regression analysis of student dropout in the first year of studies may reveal more information than is possible through duration analysis and a simple calculation of the conditional probability that a student will complete/not complete. Additionally, the policy proposals stemming from that type of research may be more specific, and target the determinants found to have a major impact on dropout, than if duration analysis is used especially if no time-varying covariates are used in duration modelling.

References

Arulampalam, W.; Naylor, R. and Smith, J. (2004): A Hazard Model of the Probability of Medical School Dropout in the United Kingdom, *Journal of the Royal Statistical Society* (Series A), Vol. 167, Issue 1, pp. 157-178.

Booth, A. and Satchell, S. (1995): The Hazards of Doing a PhD: An Analysis of Completion and Withdrawal Rates of British PhDs in the 1980's, *Journal of the Royal Statistical Society* (Series A), Vol. 158, Issue 2, pp. 297-318.

Cox, D. and Oakes, D. (1984): *Analysis of Survival Data* (Monographs on Statistics and Applied Probability), Chapman and Hall, London.

Ehrenberg, R. and Mavros, P. (1995): Do Doctoral Students' Financial Support Patterns Affect Their Times-to-Degree and Completion Probabilities, *Journal of Human Resources*, Vol. 30, Issue 3, pp. 581-609.

Evans, M.; Hastings, N. and Peacock, B. (2000): *Statistical Distributions*, 3rd edition, Wiley, New York.

Häkkinen, I. and Uusitalo, R. (2003): The Effect of a Student Aid Reform on Graduation: A Duration Analysis, *Working paper No. 2003:8,* Uppsala University, Department of Economics.

Kiefer, N. (1988): Economic Duration Data and Hazard Functions, *Journal of Economic Literature*, Vol. 26, Issue 2, pp. 646-679.

Lancaster, T. (1979): Econometric Methods for the Duration of Unemployment, *Econometrica*, Vol. 47, Issue 4, pp. 939-956.

Mihaljevic Kosor, M. (2010): Leaving Early: The Determinants of Student Non-Completion in Croatian Higher Education, *Revija za socijalnu politiku*, Issue 2, pp. 197-213.

Nickell, S. (1979): Estimating the Probability of Leaving Unemployment, *Econometrica*, Vol. 47, Issue 5, pp. 1249-1266.

Van Ours, J. and Ridder, G. (2003): Fast Track or Failure: A Study of the Completion Rates of Graduate Students in Economics, *Economics of Education Review*, Vol. 22, Issue 2, pp. 157-166.sectio