Predictive modelling of student progression at the University of Venda using statistical and machine learning techniques

by

Phindulo Pretty Muthundinne

submitted in fulfilment of the requirements for the degree of Master of Science in Statistics

in the Faculty of Science, Engineering and Agriculture, Department of Mathematical and Computational Sciences

at the University of Venda.

Supervisor: **Dr Alphonce Bere**

Co-Supervisor: **Dr Tshilidzi B. Mulaudzi**

Submitted on: July 19, 2025

© University of Venda

University of Venda
Creating Future Leaders

Abstract

One of the challenges facing higher education is the steadily rising number of university dropouts. Over the years, survival analysis has been used in order to address the issue of student's dropout. In developed countries, machine learning methods have gained more attention on solving the problem of student's dropout. The main motivation is the lack of application of both the discrete time statistical and discrete time machine learning methods when analysing student academic outcomes. This study built both the discrete time competing risk model and discrete time machine learning models for the time from registration until graduation or dropout for students at the University of Venda. These two approaches were compared(in terms of calibration and discrimination) to check which one works best. The proposed methodology implemented the application of statistical methods (discrete time survival model for single risk and competing risk) and the machine learning models(Classification trees for competing risk) using the R Statistical Software. For the competing risk models, we considered the time intervals 3 up to 6, since the possibility of graduation starts ate the third year. This study used comparison measures like Brier Score and C-Index to evaluate the models. Results show that the discrete cause-specific model and decision tree for competing risks showed a higher discrimination ability about the students progression. However, the decision tree model seemed to be the best model than the cause-specific model since the C-index is higher. While the results showed that male students are more likely to dropout and less likely to graduate, They also showed that female students are more likely to graduate. Students with an average mark of 70+ have 48.2% higher odds

© University of Venda

University of Venda
Creating Future Leaders

of graduating compared to those with an average below 50. Students in the faculty of Human and Social Sciences are less likely to dropout as compared to those in the faculty of Science, Engineering and Agriculture. However, HSS students do not significantly differ from FSEA students in graduation odds(SE = 0.073, OR=0.904, 95% CI(0.784; 1.042) and p-value= 0.165). The Faculty of Commerce, Management, and Law (FMCL) does not significantly differ from FSEA in either dropout(p-value=0.766) or graduation(p-value=0.072). This study found that older students are more likely to dropout than younger ones. This study suggests that using a decision tree model is more efficient than standard approaches for analyzing student dropout and academic results and recommends that it should therefore be used for analysing academic outcomes. Interventions for reducing dropout rates and shortening the time from first registration to graduation should target the identified high risk groups such as male and older students.

Key words: Dropout; Graduation; Machine Learning; Survival Analysis

© University of Venda

University of Venda
Creating Future Leaders

Declaration

I, Muthundinne Phindulo Pretty [Student Number: 17000912] declare that the dissertation for the Master Science degree at the University of Venda, hereby submitted by me, has not previously been submitted for a degree at this or any other university, and that it is my own work in design and execution and that all reference material contained therein has been duly acknowledged.

Signature: Date: July 19, 2025

© University of Venda

University of Venda
Creating Future Leaders

Acknowledgements

First and foremost, I want to thank God for allowing me to complete this dissertation.

Secondly, I would like to express how grateful I am to my supervisors Dr Alphonce Bere and Dr Tshilidzi.B Mulaudzi for the support they showed me throughout my research project. Your unwavering support came in very handy.

I am grateful that the University of venda provided data used for this dissertation. Mr Tebello Mofokeng, thank you, for extracting of data from the ITS.

Lastly, I'm grateful to my friends, family and those close to me for showing emotional and financial support.

© University of Venda

University of Venda
Creating Future Leaders

Dedication

My family has been my biggest source of strength during this journey, and I would want to dedicate my research project to them for their continuous support and encouragement. To my spouse, Mr. L. Manavhela, your faith in me has been a continual source of inspiration, and I will always be appreciative of your support.

To my father, Mr. TC. Muthundinne, thank you for your unending sacrifices, love, and guidance that have molded me into who I am today.

My research has been greatly influenced by the insight and direction of my supervisors, particularly Dr. A. Bere and Dr. T.B. Mulaudzi.

Finally, I would want to thank my wonderful friends and coworkers who have supported me throughout my academic journey; your friendship and encouragement have made this experience memorable and fulfilling.

Your love, faith, and steadfast support are demonstrated in this work.

© University of Venda

University of Venda
Creating Future Leaders

Contents

Abstract ii

Declaration iv

Acknowledgement v

Dedication vi

Abbreviations xiv

1 Introduction 1

1.1 Background 1

1.2 Statement of the problem 3

1.3 Aim 4

1.4 Objectives 4

1.5 Proposed methodology 4

1.6 Overview of the study 5

2 Literature Review 6

2.1 Introduction 6

2.2 Logistic regression 6

2.3 Cox Proportional Hazard model 7

2.3.1 The partial likelihood of the Cox model 8

2.3.2 Handling of ties 8

© University of Venda

University of Venda
Creating Future Leaders

2.3.3 Variable selection 9
2.3.4 Application of the Cox model 9
2.4 Discrete time survival model 11
2.4.1 Development of the discrete time survival model 12
2.5 Discrete time competing risk models 14
2.6 Subdistribution hazard model 16
2.7 Frailty models 17
2.8 Machine Learning Methods 18
2.8.1 Machine Learning Methods in Survival Analysis 18
2.8.2 Machine Learning Methods for Competing Risks 19
2.8.3 Recursive partitioning 21
2.9 Statistical versus Machine learning methods 22
2.10 Conclusion 23

3 Research methodology **24**
3.1 Fundamentals of survival and event history analysis 24
3.2 Continuous Versus Discrete Survival 25
3.3 Continuous-time survival models 27
3.3.1 The Cox proportional hazards regression model 28
3.4 Discrete-time models 32
3.4.1 Models to estimate discrete-time data: logit, probit and complementary log-log functions 33
3.5 Evaluation and Model Choice 35
3.5.1 Relevance of Predictors: Tests 35
3.5.2 Residuals and Goodness-of-Fit 37
3.5.3 Deviance in the Case of Censoring 40
3.5.4 Martingale Residuals 41
3.5.5 Measuring Predictive Performance 42
3.6 Discrete-time competing risk model 51

© University of Venda

University of Venda
Creating Future Leaders

3.6.1 Introduction . . . 51
3.6.2 Parametric Models . . . 51
3.6.3 Maximum Likelihood Estimation . . . 55
3.6.4 Variable Selection . . . 58
3.7 Machine Learning Methods . . . 60
3.8 Recursive partitioning in discrete competing risks . . . 60
3.8.1 Classification Trees for discrete competing risks . . . 61
3.9 Performance Measures . . . 66
3.9.1 Discrimination Measures . . . 66
3.9.2 Calibration Measures . . . 68
3.10 Procedure about data processing and manipulation . . . 70
3.10.1 Introduction . . . 70
3.10.2 Enrolment dataset . . . 70
3.10.3 Other datasets . . . 71
3.11 Model fitting and R packages . . . 72

4 Data analysis and Results 74
4.1 Introduction . . . 74
4.2 Data description and exploratory data analysis . . . 75
4.2.1 Data description . . . 75
4.2.2 Exploratory data analysis . . . 75
4.3 Single risk model . . . 79
4.3.1 Baseline hazard . . . 80
4.3.2 Effect of Age . . . 80
4.3.3 Effect of Average Mark . . . 81
4.3.4 Effect of Faculty . . . 82
4.3.5 Effect of Gender . . . 82
4.3.6 Conclusion . . . 82
4.4 Discrete time cause specific for competing risks . . . 83

© University of Venda

University of Venda
Creating Future Leaders

4.4.1 Baseline hazard 83
4.4.2 Effect of Gender 85
4.4.3 Effect of Age 85
4.4.4 Effect of Faculty 85
4.4.5 Effect of Average Mark 86
4.4.6 Conclusion 86
4.5 Decision tree for competing risks 89
4.5.1 Nodes with Higher Risk of Dropout 92
4.5.2 Nodes with Higher Probability of Graduation 92
4.5.3 Conclusion 93
4.6 Comparison of Decision tree results with the Discrete time Competing risk survival model results 95
4.7 Single model with age and average mark as continuous variables 97
4.7.1 Effect of explanatory variables(Age, Gender, Average mark and Faculty) on dropout 97
4.8 Cause specific hazard model with age and average mark as continuous variables 98
4.8.1 Effect of Age 99
4.8.2 Effect of Average mark 99
4.8.3 Effect of Faculty 99
4.8.4 Effect of Gender 100
4.9 Model evaluation 100
4.9.1 Concordance Index 100
4.10 Discussion 102
4.11 Determinants of student progression 102
4.12 Multiple imputation 104

5 Conclusion **107**

5.1 Introduction 107

© University of Venda

University of Venda
Creating Future Leaders

5.2 Conclusion 108
5.3 Recommendation 109
5.4 Limitation of the study 109
5.5 Future Research 109

References **121**

© University of Venda

University of Venda
Creating Future Leaders

List of Figures

4.1 Violin plot of Age at first year 79
4.2 Baseline hazard for dropout 80
4.3 Baseline hazard for graduation and dropout 84
4.4 BIC plot 90
4.5 Decision tree 91

© University of Venda

University of Venda
Creating Future Leaders

List of Tables

4.1 Variables used in the study. 76
4.2 Frequency distribution of qualitative explanatory variables 77
4.3 Summary Statistics of `Age at first year` 78
4.4 Logistic Regression Results 81
4.5 Multinomial Logistic Regression Results 88
4.6 Node probabilities of dropping out(P(D)), graduating(P(G)) and still enrolled(P(SE)) 94
4.7 Comparison of Key Results Between the discrete time Cause-specific hazard model and the Decision Tree model 96
4.8 Single Risk model Results 97
4.9 Cause specific model Results 98
4.10 C indexes for the models 101
4.11 Cause specific model results from Multiple imputation 106

© University of Venda

University of Venda
Creating Future Leaders

Abbreviations

HE	Higher Education
HEI	Higher Education Institution
SVM	Support Vector Machines
PH	Proportional Hazard
ANN	Artificial Neural Networks
NB	Naive Bayes
BN	Bayesian Network
ANN	Artificial Neural Networks
PLANN	Partial Logistic Artificial Neural Network
PLANNCR	Partial Logistic Artificial Neural Networks for Competing Risks
RSF	Random Survival Forest
RSFCR	Random survival forests for competing risks

© University of Venda

University of Venda
Creating Future Leaders

Chapter 1

Introduction

1.1 Background

For administrators in higher education, the steadily rising number of university dropouts continues to be a problem (Ramokolo). Though some of the students end up graduating, the time they take to complete their studies is beyond the stipulated normal time. Students who take a long time to complete their studies waste the resources of parents and the institution. Like other nations, South Africa continues to have high dropout rates, which lead to low graduation and retention rates.

Leaving university without a degree has far-reaching consequences that extend beyond the individual student, impacting both the university and the government. For the government, a high dropout rate leads to a rise in unemployment, which in turn puts significant strain on the economy. Increased unemployment reduces consumer spending, hampers economic growth, and lowers tax revenues, making it more difficult for the government to fund essential services. This study will use survival analysis techniques to assess factors affecting the time to graduation or dropout for students at the University of Venda.

© University of Venda

University of Venda
Creating Future Leaders

Wang et al. (2019) defined survival analysis as a branch of statistics that seeks to model and analyze data with the end result being the amount of time until an event of interest takes place. Numerous issues may be resolved with survival analysis, and dropouts can be examined using this method (Wang et al., 2019). The existence of individuals whose occurences become unobservable beyond a specific time point or in rare cases who do not experience any events throughout the monitoring period is one of the primary obstacles in survival analysis. This is called censoring, and can most effectively be handled using survival analysis techniques.

Survival time can be discrete or continuous. In continuous survival time, the exact timing is known (Odeniya et al., 2013). Discrete time arises naturally and the exact timing is not known since it is measured in intervals. In discrete time, the most commonly used model is the discrete time survival model which analyses the likelihood of an event to occur when the time is measured in intervals. The most commonly used model in continuous time is the Cox proportional hazard model (Cox, 1972). This model derives the robust estimates of covariate effects using the proportional hazards assumption.

Machine Learning (ML) methods have also been applied to survival analysis. Machine Learning has proven to be similarly useful for predicting the success of online education (Gitinabard et al., 2018). Examples include survival trees, neural networks (Rosenblatt, 1958) and (Biganzoli et al., 1998) and Random Survival Forest (Ishwaran et al., 2014).

Competing risks modeling is, according to Janitza and Tutz (2015) the modeling of event times in the presence of multiple events. There are statistical and machine learning competing risk models in survival analysis. Examples includes discrete time competing risk model and partial logistic artificial neural network for competing risks. This study will focus on both the statistical and machine learning competing risks models.

© University of Venda

University of Venda
Creating Future Leaders

1.2 Statement of the problem

Student drop out is one of the biggest problems in the Higher Education Institutions, significantly affecting student success, institutional efficiency, and national human capital development (Ramokolo). Accurately predicting the likelihood of student dropout and identifying at-risk students is essential for developing effective early intervention strategies.. Some studies looked at student dropout but were restricted to logistic regression (Zewotir et al., 2011), single event and two competing events (Ndlovu, 2015).

Ishitani and DesJardins (2002) used the discrete time survival model to investigate the departure behavior of college students in the United States. Ramokolo also used the discrete time survival model for single risk and competing risk to model the occurrence and timing of undergraduate engineering students at Tshwane University of Technology (TUT). Our study draws inspiration from earlier research efforts where discrete time statistical methods and discrete time machine learning methods were used to model and analyse student progression.

However, although some international studies have compared machine learning and statistical methods in survival analysis contexts (Frydman and Matuszyk, 2022; Chen et al., 2018; Suresh et al., 2022), there is a noticeable absence of such comparative studies within the South African higher education context. This lack of comparative analysis represents a significant gap in the literature. Therefore, this study aims to fill this gap by comparing discrete time machine learning and statistical methods for competing risks using data from the University of Venda. The findings are expected to inform institutional decision-making and improve targeted interventions for student retention.

© University of Venda

University of Venda
Creating Future Leaders

1.3 Aim

The aim of this study is to construct and evaluate statistical and machine learning discrete-time competing risk models for analysing time from enrollment to either graduation or dropout at University of Venda.

1.4 Objectives

The objectives of the study are to:

(i) Build a discrete time competing risk model to identify factors which determine the time to graduation or dropout,

(ii) Develop a discrete time machine learning model(decision tree) to identify the factors which determine the time to graduation or dropout,

(iii) Compare the discrimination and calibration performance of the machine learning and statistical models for time to graduation or dropout,

(iv) Make recommendations on the strategies that universities can use to reduce the time from registration to graduation and to reduce dropout rates among their students.

1.5 Proposed methodology

We are going to compare the use of statistical methods with machine learning methods and see which models predict the time to graduation or dropout more. The statistical model to be used will be the discrete time cause specific competing risk model. The C-Index and the Brier Score will be used as a comparison criteria for these two models. In machine learning models, we will use the classification decision trees model.

© University of Venda

University of Venda
Creating Future Leaders

1.6 Overview of the study

This research project is structured into five chapters. Chapter 1 introduces the subject of the study, while Chapter 2 presents a review of the relevant literature. Chapter 3 outlines the detailed methodology employed to address the research aims and objectives. Chapter 4 focuses on data analysis and the interpretation of findings. Finally, Chapter 5 summarizes the study, draws conclusions, and offers recommendations.

© University of Venda

University of Venda
Creating Future Leaders

Chapter 2

Literature Review

2.1 Introduction

What follows in this section is the literature review of previous studies that used statistical and machine learning methods in survival analysis.

2.2 Logistic regression

When multiple explanatory variables are involved, logistic regression is a suitable method for estimating the odds ratios associated with each predictor, (Sperandei, 2014). While it shares similarities with multiple linear regression in terms of incorporating several independent variables, logistic regression differs in that the response variable is binary. This allows researchers to assess how each variable influences the likelihood of the event of interest occurring.

This method has been effectively applied in educational research. For instance, Bokana and Tewari (2014) employed both Ordinary Least Squares (OLS) and logistic regression to examine the factors influencing student success in management studies at a South African university. Their analysis revealed that speaking English as a first language at home, higher total matriculation points, and strong performance in

© University of Venda

University of Venda
Creating Future Leaders

matriculation-level mathematics and English were significant predictors of academic achievement. These findings illustrate how logistic regression can be used to uncover meaningful relationships between background characteristics and educational outcomes.

Building on this, Zewotir et al. (2011) used a multi-level logistic regression to identify the factors that lead to the successful progression of students to higher level modules at the University of Kwazulu-Natal. According to the findings, students' success in various faculties is influenced by their first attempt at the course, while accommodations (on or off campus) also played a part. Additionally, high school academic performance (both in English and mathematics) was found to be helpful in fostering success in certain faculties. This multi-level approach reinforced the role of both individual and contextual factors in shaping academic trajectories.

Similarly, Rohr (2012) conducted a study at Liberal Arts College in Indiana using data of 803 students. A logistic regression analysis was used to examine the effect that college preparatory GPA and aggregate SAT score have on the retention of science, technology, engineering, mathematics, and business students. It was discovered that these two factors predicted the retention of students studying business, science, technology, engineering, and mathematics. The likelihood that the student would be kept more than doubled for each point rise in GPA. For every SAT point that was gained, retention increased by 0.3%. These results highlight the predictive value of academic preparedness indicators in supporting persistence in higher education.

2.3 Cox Proportional Hazard model

Survival time can either be discrete or continuous. According to Odeniya et al. (2013), continuous-time survival models are frequently used when the exact timing of an event is known. The Cox Proportional Hazards (PH) survival model, also known as the Cox regression model, is one of the most popular models for doing survival analysis

© University of Venda

University of Venda
Creating Future Leaders

(time to event analysis). This model is semi-parametric, meaning it does not require precise knowledge of the underlying distribution of survival times, but it does make assumptions about it. In medical research, it is frequently used to analyze survival data (such time-to-death) and find risk variables that could be connected to the event of interest.

Cox (1972) introduced the Cox Proportional Hazards Model in his seminal paper, "Regression Models and Life-Tables". The model was developed to analyze survival data and determine the relationship between covariates and the hazard function. Cox (1972) further explained how the Cox Proportional Hazard model can be used to analyze the time-to-event data when there are censored variables.

2.3.1 The partial likelihood of the Cox model

A new approach of analysing the Cox model was developed by Cox (1975), this approach is known as Partial Likelihood. The partial likelihood method was a breakthrough in survival analysis, as it allowed for efficient estimation of regression coefficients without specifying the baseline hazard function. This technique involves constructing a likelihood function based only on the order of event times, leading to more computational efficiency. Cox (1972) demonstrated how to estimate the regression parameters by maximising the partial likelihood, under the assumption that no tied event times exist.

2.3.2 Handling of ties

Some observations usually have the same survival time in survival analysis. This is commonly known as "ties". When two occurrences occur simultaneously in the data, the data is said to be tied. Partial likelihood algorithms must be modified to deal with ties since a significant number of tied events results in estimations of regression coefficients that are skewed towards zero. Several methods for managing ties have

© University of Venda

University of Venda
Creating Future Leaders

been proposed within the Cox model framework. Tie-handling approximations of the partial likelihood have been created by Breslow (1974), where the approach assumes a simple approximation of the likelihood function and Efron (1977) where a more accurate correction is provided by considering different possible risk set contributions. When there are few ties, all approaches will typically produce results that are similar, but when there is a large number of ties, their performance differs significantly. This may be the situation, for instance, in clinical trials where a sizable study population and a common outcome are present, or in sizable cohort studies where time is expressed in months.

2.3.3 Variable selection

Tibshirani (1997) suggested a novel technique for shrinkage and variable selection in Cox's proportional hazards model that minimizes the log partial probability while requiring that the total of the parameters' absolute values be constrained by a constant. Coefficients are reduced to zero by the nature of this restriction. Consequently, it lowers the estimation variance and produces a final model that can be understood. The approach is a variant of Tibshirani (1996)'s 'lasso' concept, which was created within the setting of linear regression. The motivation for LASSO arose from the need to handle high-dimensional data, where traditional regression methods, such as ordinary least squares (OLS), fail due to multi-collinearity and over-fitting.

2.3.4 Application of the Cox model

The research conducted by Mphekgwana et al. (2020) identified factors contributing to the survival rates of post-graduate students registered from 2011 to 2017 at a Historically Black University located in the Limpopo Province in South Africa. They identified the factors influencing postgraduate students' survival using the Cox proportional hazards model. According to the results of their study, being female, receiving financial aid, getting enrolled beyond the age of thirty, and living on campus are all as-

© University of Venda

University of Venda
Creating Future Leaders

sociated with higher dropout rates in postgraduate programs. Students who exhibited these traits were less likely to discontinue their education. Additionally, it was shown that pupils' academic achievement was impacted by a lack of funding.

Ndlovu (2015)'s study was conducted in order to model the time to graduation of students at the Durban University of Technology. The observation period was fixed at 5 years, from 2004 to 2008. To find out whether graduation and other explanatory factors like faculty, gender, and race are related, they employed the Cox model.

Ameri et al. (2016) developed a survival analysis framework for early prediction of student dropout using Cox proportional hazards model in the united states. The time-dependent Cox model, which takes into account variables that change over time and may be used to forecast student dropout rates more precisely, was used. Their research focused on many sets of characteristics, including demographics, family history, finances, high school information, college enrollment, and semester-by-semester credits, utilizing actual student data gathered at Wayne State University. The difficulty of forecasting dropout students and the semester in which they will occur can both be solved by the suggested approach.

Zewotir et al. (2015) conducted a study so as to estimate the time it takes for students to successfully complete or drop out from a master's programme at the University of Kwazulu-Natal. The Cox model was used to identify the factors which affect the response variable. According to their findings, receiving financial help of any kind reduced the time it took to ultimately leave a master's program. Students from the faculties of management or the humanities left their programs earlier than those from the faculties of science and agriculture. The time required to successfully finish a master's program was greatly reduced for students who were of international origin, had some kind of financial aid, and/or were enrolled in the faculties of management, law, health, or humanities. Compared to students from the Faculty of Science and Agriculture, students from the Faculty of Medicine took longer to finish their programs. Using race

© University of Venda

University of Venda
Creating Future Leaders

as a baseline, Black Africans graduated more quickly compared to other races. Older students took longer to complete their programmes.

In Chimka et al. (2007), the Cox Proportional Hazards (PH) model was used to analyze graduation trends among engineering college students. The time until graduation was regarded as continuous, and the models were estimated using the Breslow (1974) approach, which handles tied survival times.

Restaino (2008) analysed the interval between the first enrolment at university and the first occurrence of non-enrolment. It was assumed in his analysis that there were no relationships between the dropout dates, that the variables were time-invariant, and that the time to dropout was continuous.

Other studies like Chen et al. (2018), Ndlovu (2015) and many more used the applied the Cox model in the student academic outcomes area.

2.4 Discrete time survival model

According to Odeniya et al. (2013), discrete-time survival models are commonly used when merely the interval in which an event occurs or the event itself occurs at discrete intervals are known, typically, length is expressed in months, years, or days. As a result, although the change between states occurs at a certain moment, the precise moment is typically unknown. The information that is accessible provides a summary of the events that occurred within a certain time period. Studies with an intrinsically discrete time scale, such as clinical trials involving the time to pregnancy (where the observation time is frequently determined by the number of cycles of menstruation) and the analysis of university outcomes in academic research (where time is, for instance, measured in semesters), are examples of studies that encounter discrete observation times ((Fehring et al., 2013) and (Vallejos and Steel, 2017)). Compared to a continuous survival model, which approximates the observed data, discrete time-to-event models are more accurate

© University of Venda

University of Venda
Creating Future Leaders

and, hence, more suitable. Models with discrete event times do not lead to issues with ties, unlike survival models for continuous time. The discrete time survival model is the most often used technique for discrete time data.

When the time variable is monitored discretely, the probability that an event will occur is analyzed using the discrete time survival model. The longitudinal aspect of institutional data may be more effectively utilized with discrete-time survival analysis approaches, which enable the study to account for time-dependent variables, time-varying factors, and time-invariant factors (Ramokolo).

Discrete-time survival models have gained prominence as flexible tools for analyzing time-to-event data when event times are recorded in fixed intervals (Singer and Willett, 1993). Unlike continuous-time models such as the Cox proportional hazards model (Cox, 1972), discrete-time models rely on logistic or complementary log-log regression, making them well-suited for applications where the exact event time is unknown. These models have been widely applied in social sciences, medicine, and finance (Jenkins, 1995).

2.4.1 Development of the discrete time survival model

The concept of discrete-time survival analysis was initially proposed by Andersen (1970) where he presented a method for estimating the probability of an event occurring within a given time interval, conditional on the event not having occurred in any prior intervals. This methodology is now commonly referred to as the "discrete-time hazard model" (Andersen, 1970).

Thompson Jr (1977) was among the first to examine discrete survival data in biometrics, focusing on the logistic model since he believed that, "grouping rather than discrete time is the right way to handle tie". Meanwhile, Prentice and Gloeckler (1978) introduced the grouped Cox model and developed a score test for hypothesis testing.

© University of Venda

University of Venda
Creating Future Leaders

To deal with complicated data structures including competing risks, time-varying variables, and censored (missing) data, researchers began creating extensions of the discrete-time survival model. Allison (1984) expanded the discrete-time survival model. In contrast to modeling the time until the event occurred, he proposed modeling the occurrence of an event as a binary outcome at a specific time interval.

Variable selection

As with other regression-based models, selecting relevant variables is critical in discrete survival analysis. Standard approaches, such as stepwise selection or regularization methods like the LASSO (Least Absolute Shrinkage and Selection Operator), have been used to improve model interpretability and predictive performance (Tibshirani, 1996). However, these methods often fail to account for unobserved heterogeneity, which can introduce bias and inefficiency in estimating hazard probabilities (Vaupel et al., 1979).

Unobserved heterogeneity, often referred to as frailty in survival analysis, is a critical issue when modeling event times. The common method for accounting for unobserved variability is to employ frailty models or random effects models. The term frailty was first introduced by Vaupel et al. (1979). Several approaches have been proposed to integrate heterogeneity in survival models, including random effects and mixture models (Hougaard, 1995). However, these methods do not always perform well in high-dimensional settings, necessitating more sophisticated variable selection techniques.

To overcome this problem, Groll and Tutz (2017) proposed a penalized likelihood approach that integrates variable selection with the explicit modeling of heterogeneity. This method combines ridge and LASSO-type penalties, allowing for efficient selection of relevant covariates while mitigating the impact of unobserved heterogeneity. This approach is particularly useful in contexts where large numbers of potential predictors exist.

The incidence and timing of undergraduate engineering student dropouts at TUT between 2010 and 2014 were modeled by Ramokolo. The study's primary objectives were

© University of Venda

University of Venda
Creating Future Leaders

to compare the single risk discrete-time model against a competing risk discrete-time model, test for the influence of unobserved heterogeneity, and examine the incidence and drivers of dropout.

Ishitani and DesJardins (2002) used the discrete time survival model to investigate the departure behavior of college students in the United States and their findings showed that students who receive financial aid had lower dropout rates than those who do not receive financial aid. They also found that a student's initial level of commitment is thought to influence how integrated they become into the institution's social and academic understanding, and that level of integration is thought to influence a student's decision to stay in university or leave.

Several studies such as Letseka and Cosser (2010) , have applied discrete-time survival models in educational research. The discrete time survival models has also been used in other areas like medicine, see Scheike and Jensen (1997) and Odeniya et al. (2013).

2.5 Discrete time competing risk models

Simple survival analysis applications make the assumption that individuals are only at risk of encountering one event. However, there is a chance that individuals will experience two or more related occurrences in many applications. Single event models fail to take into consideration potential dependencies between competing outcomes. The results that are competing with the event of interest are considered censored, and the censorship is presumed to be random and non-informative. DesJardins (2003) asserts that because dropout and graduation may be negatively associated, estimating dropout as a single risk may lead to model mis-specification. In survival analysis, failure to identify competing events may result in wrong conclusions (Allison, 2014). Competing events are mutually exclusive outcomes that are statistically dependent,

© University of Venda

University of Venda
Creating Future Leaders

as the occurrence of one event can influence the likelihood or timing of the others. This effectively indicates that having experienced one event eliminates the possibility of having experienced another event during the observation period (Allison, 1984).

Competing risks modeling is, according to Janitza and Tutz (2015), the modeling of event times in the presence of multiple events. Janitza and Tutz (2015) further explained that even though the literature on competing risks mostly deals with cases where time is measured as a continuous variable, time may occasionally be expressed on a discrete scale, such as weeks, months, or years. The discrete time variable is not a coarsened version of an underlying continuous time variable when the time until the occurrence of a particular event is organically discrete. This is evident when calculating the time needed to graduate or leave college, which is expressed in semesters. Fitting a multinomial logit model, which connects an individual's variables to the likelihood of seeing a particular event, is the conventional method for modeling discrete-time competing hazards.

Narendranathan and Stewart (1993) examined the use of discrete competing risks models to analyze unemployment durations while incorporating flexible baseline hazards. Fahrmeir and Wagenpfeil (1996) introduced an approach to estimate smooth hazard functions and time-varying effects using posterior modes. Steele et al. (2004) developed a comprehensive multilevel multistate competing risks model to account for repeated episodes.

A summary of statistical techniques for analyzing discrete failure times with competing events is provided by Schmid and Berger (2021a). They discussed discrete versions of the cause-specific risks model and the sub-distribution hazard model, two of the most popular modeling techniques for discrete-time data. They offered strategies for nonparametric estimates and model validation in addition to going over the features of these techniques.

© University of Venda

University of Venda
Creating Future Leaders

Singer and Willett (1993) showed how educational statisticians can examine event occurrence with the best framework thanks to the techniques of discrete-time survival analysis. They developed maximum likelihood estimators for the parameters of a discrete-time hazard model using longitudinal data on the career trajectories of 3,941 special educators as a starting point. They also demonstrated how typical logistic regression software may be used to fit the model.

Some of the studies that used the discrete time survival model in the context of education are; Ramokolo, Ishitani and DesJardins (2002) and Letseka and Cosser (2010). The discrete time survival models has also been used in other areas like medicine, see Scheike and Jensen (1997) and Odeniya et al. (2013).

Discrete-time survival models have evolved from simple logistic regression approaches to sophisticated frameworks incorporating time-varying covariates, competing risks, machine learning, and multilevel modeling. These developments have expanded their applications across various fields, including medicine, education, and finance, making them indispensable tools for time-to-event analysis.Some benefits of discrete event timings were demonstrated by Tutz et al. (2016). The ability to represent hazards using conditional probabilities is one advantage of discrete time to event models.

2.6 Subdistribution hazard model

The proportional sub-distribution hazards model by Fine and Gray (1999), which is based on the proportional hazards specification by Cox (1972), is a widely used modeling technique for competing risks analysis in longitudinal research. It does not apply when event timings are given on a discrete time scale since it is thought that survival times are based on random factors assessed on a continuous scale. However, the lack of accurate (continuous) event timings in several clinical and epidemiological investigations remains a chronic difficulty. Rather, the information that is now accessible could simply show

© University of Venda

University of Venda
Creating Future Leaders

that certain events took place inside particular time frames between successive periods in time. In order to get around this, Berger et al. (2019) suggested an approach based on a weighted machine learning estimation methodology for binary regression for modeling the sub-distribution risks in discrete time. In the absence of competing events, the discrete subdistribution hazard model simplifies to the fundamental discrete hazard model outlined in Tutz et al. (2016).

2.7 Frailty models

Incorporating a frailty term helps capture unobserved heterogeneity within the population, making the model more reflective of the true data-generating process. However, including a frailty component can lead to model misspecification if its assumed distribution does not reflect the true underlying structure. They were among the first to look at how heterogeneity affected survival (Vaupel et al. (1979)). It was also examined from an econometric standpoint by Heckman and Singer (1984b) and he analytically demonstrated that neglecting unobserved heterogeneity leads to biased hazard estimates, specifically causing an artificial appearance of negative duration dependence. Negative duration dependence refers to a declining hazard function over time, meaning that the probability of an event occurring decreases as the duration increases.

Lancaster (1990) demonstrated that ignoring heterogeneity in survival models can lead to inconsistent parameter estimates. Specifically, in a Weibull distribution model without censoring, he showed that the estimated parameters tend to be asymptotically biased toward zero. This suggests that failing to account for unobserved heterogeneity may underestimate the true effects of covariates on the hazard function.

Heckman and Singer (1984b) examined a Weibull duration model and found that estimation results and inference were highly sensitive to the choice of the mixing distribution. This suggests that using an incorrect functional form for the mixing distribution

© University of Venda

University of Venda
Creating Future Leaders

can significantly impact parameter estimates. However, since estimation issues can also arise from misspecified hazard functions, some of their findings may have been influenced by the lack-of-fit of the Weibull model. To address this issue, Heckman and Singer (1984a) introduced a consistent nonparametric maximum likelihood estimator for the mixing distribution, providing a more flexible and robust approach to modeling heterogeneity.

2.8 Machine Learning Methods

Machine learning models are computational approaches that allow systems to learn patterns and make forecasts from data without explicit programming. Similarly, machine learning has shown promise in forecasting the success of online learning (Gitinabard et al., 2018).

2.8.1 Machine Learning Methods in Survival Analysis

Survival analysis has traditionally been dominated by statistical methods such as the Cox (1972) and the Subdistribution hazard model (Fine and Gray, 1999). However, machine learning techniques have increasingly been employed to handle complex, high-dimensional survival data. One of the earliest applications of machine learning in survival analysis was the use of decision trees (Ishwaran et al., 2008), which led to the development of Random Survival Forests (RSF) by Ishwaran et al. (2014). Building on the decision tree framework, RSF constructs an ensemble of survival trees by drawing a predetermined number of bootstrap samples (B) from the data. For each sample, a tree is grown using randomly selected subsets of variables at each node to determine the optimal split, thereby introducing randomness that improves generalization and reduces overfitting.

© University of Venda

University of Venda
Creating Future Leaders

2.8.2 Machine Learning Methods for Competing Risks

Competing risks models account for situations where multiple failure types may occur, requiring methods that can simultaneously handle different types of events. Traditional methods, such as the Fine and Gray subdistribution model (Fine and Gray, 1999), provided a foundation for competing risks analysis. Machine learning methods have extended these approaches, offering flexibility in modeling complex dependencies.

Survival random forests were adapted to competing risks by incorporating event-specific risk estimations (Ishwaran and Kogalur, 2010). According to Ishwaran et al. (2014) , a non-parametric tree-based estimate and prediction technique called Random Survival Forest for Competing Risks (RSFCR) was introduced in 2014 and through simulations, the method was demonstrated to accurately forecast the cumulative probability function of each risk.

Rosenblatt (1958) published the first study on artificial neural networks (ANN), introducing the perceptron, a computational model inspired by the human brain's neural architecture. His work formalized how machines could learn from data by adjusting the weights of connections between artificial neurons. The perceptron was capable of binary classification and could be trained using a supervised learning algorithm. Rosenblatt's work laid the foundational architecture for modern neural networks. By linking the simple artificial nodes known as "neurons" based on a weighted link, this technique(ANN) builds a network that resembles a neural network. Biganzoli et al. (1998) improved the prediction of ANN model, they proposed the partial logistic artificial neural network (PLANN) method in order to examine the link between the features and the survival times. Biganzoli et al. (2006) proposed an extension of the partial logistic artificial neural network for the joint modelling of discrete cause-specifc hazards. They named the model 'partial logistic artificial neural network for the competing risks (PLANNCR)'. Later in 2009, Lisboa et al. (2009) implemented the partial logistic artificial neural network extension (PLANNCR-ARD) under a bayesian regularisation

© University of Venda

University of Venda
Creating Future Leaders

framework with the standard approximation of the evidence to implement automatic relevance determination(PLANNCR-ARD).

Gitinabard et al. (2018) used data from two different offerings of the Big Data in Education Massive Open Online Course (BDE MOOC) from 2013 and 2015 to conduct a survival analysis with the aim to identify students that are likely to dropout. they then evaluated two kinds of variables, social and behavioral, and employed a mix of modeling and feature-selection approaches to determine the most significant aspects to forecast both dropout and certification. They assessed the usefulness of three distinct model types and checked the influence of different definitions of dropout on the predictors.Lastly, by determining whether or not models from week 1 could forecast dropout in week 2, and so on, they evaluated the models' dependability across time and found that both the behavioral and social aspects had significant p-values and a high hazard ratio.

Costa et al. (2018) used data from 1202 first-year students at a Brazilian Federal University between 2004 and 2009 to examine the factors that influence the dropout rate and retention of students in undergraduate management programs. These pupils were observed until 2013. They gathered information on the student's sociodemographics, program characteristics, graduation or dropout status, and participation in the program. A statistical method for assessing the variables influencing program completion time and the change in dropout risk was the survival analysis. Course completion duration and dropout risk were impacted by the number of semesters, student grades, gender, and the number of failures or dropouts per course. Age, marital status, race, and high school experience (public or private) had no effect on graduation time or dropout rate.

The default prediction performance of RSFCR and the Cox proportional hazard model for competing risks (CR Cox) were compared by Frydman and Matuszyk (2022).

© University of Venda

University of Venda
Creating Future Leaders

According to the results, RSFCR outperforms CR Cox in terms of prediction accuracy as determined by the integrated Brier score, suggesting that RSFCR might be a beneficial substitute for the current competing risk models. RSFCR's nonparametric nature and capacity to handle high-dimensional data make it an attractive alternative to the current competing risk approaches (Frydman and Matuszyk, 2022).

2.8.3 Recursive partitioning

There are various proven ways for simulating cause-specific risks in discrete time. Schmid et al. (2016) provides a detailed treatment of statistical technique. A significant portion of this methodology involves parametric regression models. However, parametric models can be overly restrictive in many applications, particularly when complex interactions among explanatory variables exist. Additionally, specifying a parametric model, such as the multi-nomial logit model, often results in an excessive number of parameters relative to the available sample size. To overcome these limitations, recursive partitioning or tree-based modeling can be employed. Tree-based methods offer a flexible alternative, especially for discrete time-to-event data with a single type of event (Berger et al., 2019). Such approaches, commonly referred to as survival trees, have been explored in previous studies, including Schmid et al. (2016).

Berger et al. (2019)introduced a novel extension of the method developed by Schmid et al. (2016) to handle discrete time-to-event data with competing events. Their approach modeled cause-specific hazards using a classification tree with a multi-categorical outcome, providing a flexible framework for analyzing competing risks in survival analysis. The foundation of recursive partitioning originates from automatic interaction detection, with one of the most widely recognized versions being the Classification and Regression Trees (CART) method developed by Breiman (2017). This approach, which serves as the basis for the proposed method, follows a straightforward principle: the predictor space, defined by the explanatory variables and encompassing all observations, is systematically divided into distinct, non-overlapping subgroups through a

© University of Venda

University of Venda
Creating Future Leaders

series of binary splits. Doove et al. (2014) provided a comparison of a number of recent advancements in recursive partitioning techniques.

2.9 Statistical versus Machine learning methods

The comparison between machine learning and traditional statistical methods in competing risks analysis revolves around interpretability, flexibility, and predictive performance. Traditional statistical models, such as Cox regression (Cox, 1972) and the Fine-Gray model (Fine and Gray, 1999), offer interpretability and well-established theoretical foundations. However, they often rely on strong assumptions, such as proportional hazards or specific functional forms of covariate effects, which may not always hold in real-world data (Kalbfleisch and Prentice, 2002). Machine learning models, on the other hand, provide greater flexibility in capturing complex, nonlinear relationships without requiring strict parametric assumptions. Studies comparing these approaches have shown that machine learning models, such as survival forests and deep learning models, often outperform traditional methods in predictive accuracy while being more adaptable to high-dimensional data (Ishwaran et al., 2014). However, machine learning models can suffer from reduced interpretability, making their application less transparent in clinical and epidemiological studies. Hybrid approaches that integrate statistical and machine learning techniques are emerging to balance interpretability and performance in competing risks analysis.

Suresh et al. (2022) compared the predictive performance of statistical(Cox PH and the discrete time model) and machine learning(Support Vector Machine, Neural Network and Random Survival Forest) methods using different datasets in USA. The study was conducted in the context of medicine.

Chen et al. (2018) compared the performance of statistical survival analysis approaches to machine learning approaches like decision trees and boosting and it was seen that machine learning methods perform best, producing better results.

© University of Venda

University of Venda
Creating Future Leaders

Frydman and Matuszyk (2022) contrasted the Cox proportional hazard model for competing risks (CR Cox) with the default prediction performance of CR RSF. The data set, which included 7,874 auto leases with 34 time-fixed variables for each lease (CL data) provided by a Polish financial institution in New York, was subjected to both of these approaches. According to the integrated Brier score, the results indicated that CR RSf outperforms CR Cox in terms of prediction accuracy.

Monterrubio-Gómez et al. (2024) provided a summary of a range of competing risks modelling techniques, covering the statistical and machine learning literature. They facilitated the adoption of competing risks survival methods by presenting a concise overview with consistent notation and interpretation across different approaches. They also highlighted available software and, where possible, demonstrates their application through reproducible R vignettes. Additionally, the paper addressed two key challenges that can impact benchmark studies in this field: the selection of performance metrics and issues related to reproducibility.

2.10 Conclusion

Most of the studies used logistic regression (e.g. (Zewotir et al., 2011)). Some authors looked at only one outcome event. Some authors looked at both outcome events but they only used competing risks models without including machine learning methods (e.g. (Ndlovu, 2015) and (Zewotir et al., 2015)). While some studies across the world compared the machine learning methods with the statistical methods, there is no literature on studies that compared machine learning with statistical methods in the context of education in South Africa ((Frydman and Matuszyk, 2022), (Chen et al., 2018) and (Suresh et al., 2022)). Hence this study is going to compare the use of machine learning methods and statistical methods for competing risks in the context of education in South Africa.

© University of Venda

University of Venda
Creating Future Leaders

Chapter 3

Research methodology

3.1 Fundamentals of survival and event history analysis

Survival and event history analysis is the group of statistical methods that answer questions that are related to time and the duration until an event occurs. A practical example is the modelling of time between the start of a specific treatment and the date of discharge or death. The models assess the conditional probability that an event happens at a particular time interval (t), known as the hazard rate. This hazard rate is the response variable. The time it takes for an event under study to occur is being examined, as well as the relationship between the survival time and the explanatory variables (also related to as covariates). The explanatory variables assess the impact of certain characteristics on the response variable. These explanatory variables may either be fixed (variables that do not change over time, like gender) or time varying (variables that change over time, like age) (Singer and Willett, 2003).

Survival time is how long it takes for an event to happen, also referred to as failure time.It can either be discrete or continuous. Survival time is measured on a continuous scale in cases where it is known precisely. If the time units are not known within larger

© University of Venda

University of Venda
Creating Future Leaders

amounts of years, then discrete time methods will be used. These methods are used when we only know that events happened within a certain time interval but not the exact time.

3.2 Continuous Versus Discrete Survival

Even though we think of time as a continuum, time is always measured discretely in reality. Duration is typically measured, for instance, in days, years, or months, especially in the social sciences, economics, and medical sciences (Singer and Willett, 2003). As a result, although the change between states occurs at a certain moment, the precise moment is typically unknown. The information that is accessible summarizes the events that took place within a certain time period. Time can be represented by the positive numbers 1, 2, 3,.... More formally, continuous time is divided into intervals $[0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty)$

Let us denote discrete event times by T, where $T = t$ means that the event has occurred in the interval $[a_{t-1}, a_t)$. Interval censoring or aggregated survival statistics are also discussed. Some observations usually have the same survival time in aggregated survival data. The term "ties" is typically used for this. For continuous time, some models and estimating techniques even use the assumption that the data is unrelated. However, ties do occur in real-world applications with continuous event periods, which might be seen as a clue for underlying grouping (Allison, 1982; Singer and Willett, 2003). Discrete data are relatively normal in some fields, such as demography. For instance, years are typically used as a metric for life span in life tables.

The true transition process is sometimes referred to as fundamentally discrete. Think about the time to pregnancy, for instance. Menstrual cycles, a really distinct response,

© University of Venda

University of Venda
Creating Future Leaders

are a natural indicator of the time that a couple takes to conceive. Surveys conducted on a monthly or annual basis can also yield really discrete measurements.

In conclusion, grouped data reflect events in underlying time intervals, and the response relates to an interval; discrete time-to-event data are fundamentally discrete measurements, where the measurements denote natural numbers.

In grouped data, especially when the intervals are identical in length (for instance, whether they indicate months or years), the fundamental modeling techniques are the same for both forms of data. As long as the discrete event timings are uniformly spaced, it is theoretically possible to use conventional modeling techniques for continuous time-to-event data to discrete time-to-event data as well. Statistical techniques specifically tailored for discrete event timings, however, offer several benefits in this situation:

One benefit of taking into account discrete time-to-event models is that risks may be known as conditional probabilities (Mills, 2010). Because of this, they are far easier to read than continuous hazard functions. In reality, a lot of event periods are either viewed on a discrete scale or are inherently discrete. As a result, employing discrete time-to-event models is more accurate (and hence more suitable) than using a continuous survival model to approximate the observed data. Models with discrete event times do not lead to issues with ties, unlike survival models for continuous time. The generalized linear model (GLM) framework may incorporate discrete time-to-event models. For this reason, estimating GLMs is simple and may be done using conventional tools. When subject-specific characteristics are incorporated into so-called frailty models, for instance, the approach may be used to more complex models due to its integration into the GLM framework. It is due to the above justifications that this study will consider a case where time is assumed to be discrete (Schmid and Berger, 2021b).

© University of Venda

University of Venda
Creating Future Leaders

3.3 Continuous-time survival models

The survival times are represented by the positive continuous random variable T. The value of T, represented by the symbol t, represents the actual survival time. There is a cumulative density function, $F(t)$, and a probability density function, $f(t)$, for the values of T. The following represents the random variable T's distribution function:

$$F(t) = \int_0^t f(u)d(u) = \Pr(T \leq t) \tag{3.1}$$

where $\Pr(T \leq t)$ is the probability that a survival time T is less than or equal to some value t. For all points at $F(t)$, the probability density function $f(t)$ is defined as:

$$f(t) = \frac{d(t)}{d(t)} = F'(t) \tag{3.2}$$

This implies that:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} \tag{3.3}$$

The density function $f(t)$ expresses the unconditional instantaneous probability that an event occurs in the time interval $(t, \Delta t)$ and is formally specified as:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} \tag{3.4}$$

Both of these equations demonstrate that the density function is an unconditional failure rate, meaning it characterizes the event's unconditional instantaneous probability.
Another fundamental idea in survival and event history models is the survivor function, which is defined as follows:

$$\hat{S}(t) = 1 - F(t) = \Pr(T \geq t) \tag{3.5}$$

which expresses the probability that a survival time T is equal to or greater than some time t. $\hat{S(t)}$ denotes the proportion of subjects surviving beyond t. At the origin time $t = 0$, $S(0) = 1$, which simply means that all the subjects in the study will survive at $t = 0$. $\hat{S(t)}$ is a function that strictly decreases over time as the surviving subjects

© University of Venda

University of Venda
Creating Future Leaders

fail over time.

The occurrence of an event(e.g. failure) and survival are related to each other, which is instantaneous transition or hazard function:

$$h(t) = \frac{f(t)}{\hat{S}(t)}. \tag{3.6}$$

The hazard rate indicates the rate at which subjects fail by t given that the subject has survived until t.

3.3.1 The Cox proportional hazards regression model

The Cox proportional hazards regression model with fixed covariates

In regression models we usually include both fixed and time varying covariates. Fixed covariates are are those whose values remain constatnt over time, e.g sex and place of birth. The basic Cox model with fixed covariates is expressed in scalar form as:

$$h_i(t) = h_0(t)\exp(\beta_1 X_{i1} + ... + \beta_k X_{ik}) \tag{3.7}$$

where the hazard for individual i at time t is the product of two factors, with x representing the covariates, $h_0(t)$ (sometimes referred to as $\lambda_0(t)$) is the unspecified **baseline hazard function** and can be interpreted as the hazard function for a subject whose covariates all have the same value of zero. Essentially, there are two components to the hazard: (a) the non-parametric baseline hazard $h_0(t)$, which is a function of time that is left undefined; (b) the parametric covariate effect $\exp(\beta_0 \mathbf{x})$, which, in its most basic form, excludes term t.

As a result, the equation has two parts: parametric and non-parametric, thus the term semi-parametric. Cox models do not contain an intercept term since the baseline hazard is not provided.

© University of Venda

University of Venda
Creating Future Leaders

A linear function of an exponentiated set of k fixed variables is displayed in the second section of the equation. Assuming that two individuals have the following fixed covariates at time t: $\mathbf{x}$ and $\mathbf{x}^*$, respectively, the ratio of their risks is as follows:

$$\frac{h(t,\mathbf{x})}{h(t,\mathbf{x}^*)} = \frac{h_0(t)\exp(\beta\mathbf{x})}{h_0(t)\exp(\beta\mathbf{x}^*)} = \exp[\beta(\mathbf{x}-\mathbf{x}^*)] \tag{3.8}$$

As the variables are fixed, $\exp[\beta(\mathbf{x}-\mathbf{x}^*)]$. is likewise fixed, meaning that the ratio of the two hazards remains constant throughout time t. The hazard functions are proportional to $\exp[\beta(\mathbf{x}-\mathbf{x}^*)]$. serves as the proportionality constant, thus the name "Cox's proportional hazard model" or Cox's PH model.

An individual with $\mathbf{x}=0$ has a hazard rate of $h_0(t)$, meaning that all of the variables that comprise the vector $\mathbf{x}$ are set to zero. It is common to refer to the people with $\mathbf{x}=0$ as the "reference group." Generally speaking, the i^th individual's hazard rate is determined by;

$$h_i(t) = h_0(t)\exp(\beta\mathbf{x}_i) \tag{3.9}$$

Thus, the relative risk of the i^th individual compared to the individual with $\mathbf{x}=0$ is given by;

$$\frac{h_i(t)}{h_0(t)} = \frac{h_0(t)\exp(\beta\mathbf{x_i})}{h_0(t)} = \exp(\beta\mathbf{x_i}) \tag{3.10}$$

To express the model in terms of the log of the hazard ratios, it can be rewritten as:

$$\log\frac{h_i(t)}{h_0(t)} = \beta_1 X_{i1} + ... + \beta_k X_{ik} \tag{3.11}$$

Thus the logarithm of the relative risk is linear in $\beta's$. It is therefore easier to see that the Cox model contains no constant term of β_0, which is due to the fact that this term is absorbed into the baseline hazard function. The key attraction of the Cox model is that it can take any form, this is why it is called **Semi-parametric**.

© University of Venda

University of Venda
Creating Future Leaders

The Cox proportional hazards regression model with time-varying covariates

Recall that time-varying covariates are those where the values differ overtime, e.g age or employment status. The may be continuous variables, (e.g age) or discrete variables (e.g employment status) which may change from time period to time period with no set pattern. In order to estimate a Cox model with a time-varying covariate, we need to adapt formula 3.7.

$$h_i(t) = h_0(t)\exp(\beta_1 X_{i1} + \beta_2 x_{i2}(t) ... + \beta_k X_{ik}). \tag{3.12}$$

The only difference is the addition of the (t) in $\beta_2 x_{i2}(t)$, which is included for the time-varying variables. In the equation above, the first covariate is fixed and the second is time-varying. The model states that the hazard at time t depends on the value of x_1, and on the value of x_2 at time t. The time-varying nature of $x_2(t)$ may also include lagged or cumulative values.

Why is the likelihood in the Cox model called the partial likelihood

As we know, likelihood estimation is required to fit a statistical model to data and provide estimates for the models's parameters, with the most common approach being maximum likelihood estimation. The combined chance of acquiring the data that was actually observed on the research participants as a function of the unknown parameters in the model under consideration is described by the likelihood function. Instead of using a full likelihood function, the Cox model uses a "partial" one. Without getting too technical, it implies that the likelihood formula only takes into account the probabilities of individuals who really experience the event result, not those who are censored. When evaluating the partial probability of the hazard function, Cox models only take into account the sequence of the failure times—not the information about the intervals between failure times. The partial probability can be expressed as the product of many failure times or likelihoods.

© University of Venda

University of Venda
Creating Future Leaders

Another common name for the Cox model is a **proportional hazards model**. This is due to the fact that the risk for each given person is a constant percentage of the risk for any other person. Take the ratio of the hazards (also termed the **hazards ratio**) for individuals i and j to see why the Cox model is referred to as a proportional hazards model

$$\frac{h_i(t)}{h_j(t)} = \exp \beta_1(x_{i1} - x_{j1}) + ... + \beta_k(x_{i1} - x_{j1}). \tag{3.13}$$

The ratio of hazards remains constant throughout time since $h_0(t)$ cancels out in the equation's numerator and denominator. Accordingly, the log risks of subjects i and j would be proportionate (strictly parallel) to one another if they were graphed.

Assumptions of the Cox PH model

The Cox proportional hazards model is a semi-parametric model that does not assume a specific shape for the baseline hazard function. It assumes the following;

- the hazard functions for different strata are proportional over time, this means that the relative hazard is constant,
- the log hazard has a linear relationship with each covariate, which can be checked with residual plots,
- all individuals have a common baseline hazard rate at any time and that the survival times are independent.

Violations of these assumption can be formally assessed using Schoenfeld residuals, as proposed by Grambsch and Therneau (1994), which provide both graphical diagnostics and a global test for time-dependence. A significant test result or a non-random pattern in residual plots suggests a violation of proportionality. The violations of the assumption of independent survival times can arise in clustered data (e.g., patients within hospitals

© University of Venda

University of Venda
Creating Future Leaders

or families). In such cases, incorporating frailty terms or using clustered standard errors can help account for within-group correlation (Hougaard, 1995).

3.4 Discrete-time models

We have previously concentrated on continuous-time models and the dependence of the hazard function $h(t)$ on variables. We estimate the probability that the event will occur at time t in both continuous and discrete-time models. The odds (if using typical logit/probit models) are the dependent variable in a discrete-time model, whereas the hazard rate is the dependent variable in a continuous-time model. The primary distinction with discrete-time models is that the **discrete-time hazard function** is the probability of an event occurring during interval t, conditional on the fact that the event did not occur before t, written as:

$$h(t) = Pr(T = t|T \geq t) \tag{3.14}$$

where T is the event time. Conversely, the **survival function** is represented by:

$$\hat{S}(t) = Pr(T > t|T \geq t) = 1 - h(t) \tag{3.15}$$

which is the probability that an event did not occur before time t. The **cumulative probability density function**, which measures the probability that an event occurs before time t (i.e. the probability of failure), is written as:

$$F(t) = Pr(T < t) = 1 - \hat{S}(t) \tag{3.16}$$

The discrete survival model will be used to analyse the hazard of dropout for year 1 and year2 of the students enrollment.

© University of Venda

University of Venda
Creating Future Leaders

3.4.1 Models to estimate discrete-time data: logit, probit and complementary log-log functions

Numerous modeling options are available due to the fact that discrete-time data is represented as a categorical variable, which is often binary. It is relatively easy to conclude covariates in the hazard function 9.1 by treating the probability of failue as conditioal not only on survival, but also on the covariates:

$$h(t) = Pr(T = t|T \geq t, x) \tag{3.17}$$

In the event where the response variable is binary, popular models like the logistic or probit models can be estimated. You must choose an appropriate distribution function for your data when the dependent variable is binary. This might be either the ordinary normal distribution (also known as the probit model) or a logistic distribution (also known as the logit model).

The response variable in discrete-time models represents the likelihood that an event will occur contingent on survival and covariates with time t. The likelihood of an event happening is expressed as $Pr(y_{it} = 1) = h_i$ for ease of presentation, while the chance of it not happening is expressed as $Pr(y_{it} = 0) = 1 - h_i$. Additionally, it is assumed that this probability depends on the variables, x.

Logit specification

One of the most commonly used functions of this model is the logit function or logit link:

$$logit[h_i(t)] = \log \frac{h_i(t)}{1 - h_i(t)} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} \tag{3.18}$$

where h_i is specified as the log-odds ratio of the probability of an event occurring to the probability of non-occurrence. The logit coefficients β_k are interpreted in terms of

© University of Venda

University of Venda
Creating Future Leaders

their relationship to the log-odds of event of occurrence. When $\beta_{>}0$, the log of the odds ratio increases as the covariate increases, and when $\beta_{<}0$ it decreases. Since the log-odds ratios are not as intuitive to interpret, we often use the predicted probability of an event of occurrence, which is written as:

$$\hat{h_i} = \frac{\exp \beta\prime \mathbf{x}}{1 + \exp \beta\prime \mathbf{x}} \tag{3.19}$$

where $\exp(\beta\prime \mathbf{x})$ represents the exponentiated logit parameters for a given covariate.

This study is going to use the logit link function when fitting the discrete time survival model.

Probit specification

A binary model can alternatively be specified as a probit function, which has the following form:

$$\phi^{-\mathbf{1}}[\hat{h_i}] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} \tag{3.20}$$

and with the probability estimated as:

$$\hat{h_i} = \phi(\beta\prime \mathbf{x}) \tag{3.21}$$

ϕ represents the ordinary normal cumulative distribution function.

The complementary log-log specification

The complementary log-log model (cloglog) is used less frequently than the logit or probit, and the event history form of

$$\log[-\log(1 - h_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} \tag{3.22}$$

© University of Venda

University of Venda
Creating Future Leaders

The probability can be computed by:

$$h_i = 1 - \exp[-\exp(\beta \prime \mathbf{x})] \tag{3.23}$$

The cloglog model can occasionally provide significantly different estimates than the logit and probit specifications. The complementary log-log model's response curves deviate slowly from $h_i = 0$ and rapidly approach $\pi = 1$. It is also related to the knowledge that the cloglog function is asymmetric, which implies that in data with few failures, the results may vary the greatest. The model chosen (also known as the link function) has little influence on your outcomes and is typically based on ease of interpretation and familiarity with specific models, which are frequently entrenched in discipline preferences. The odds ratio makes the logit model easier to read, hence it is frequently used for this reason. For practical reasons, one reason to prefer the probit model over the logit model is if you are confident that the underlying distribution of the procedure you are modeling is normal, or if the event under consideration is a percentage rather than a binary result.

3.5 Evaluation and Model Choice

Basic hazard model estimating approaches for discrete survival were introduced. We shall now discuss diagnostic techniques for these models. We cover test statistics for evaluating predictor significance, including goodness-of-fit tests, residuals, predictive performance metrics, and flexible connections.

3.5.1 Relevance of Predictors: Tests

The form of the discrete survival model discussed in the preceding section is

$$\lambda(t|\mathbf{x}_i) = h(\gamma_{0t} + \mathbf{x}_i^T \boldsymbol{\gamma}_r) \tag{3.24}$$

© University of Venda

University of Venda
Creating Future Leaders

where $\boldsymbol{\gamma}_r$ indicates the weights of the explanatory variables. When modeling survival, explanatory factors play an important role. In treatment trials, it is important to determine if the treatment has an impact on survival rates. It's important to assess the relevance of other factors in the model that are responsible for population heterogeneity. To assess the significance of variable j in the model, one approach is to utilize null hypothesis testing.

$H_0 : \gamma_j = 0 against H_1 : \gamma_j \neq 0$

However, this hypothesis only applies when a variable is represented by a single parameter. When using a factorial explanatory variable or quadratic terms in a continuous variable, all associated parameters must be tested concurrently to ensure they are zero. A more broader pair of hypotheses that covers these instances are the linear hypotheses of the form:

$H_0 : \mathbf{C}\beta = \varepsilon$ against $H_1 : \mathbf{C}\beta \neq \varepsilon$,

where $\mathbf{C}$ is a fixed matrix of full rank $s \leq p$ and ε is a fixed vector. The vector β in the linear hypothesis $\mathbf{C}\beta = \varepsilon$ collects all the parameters of the model, that is, $\beta^T = (\gamma_{01}, ..., \gamma_{0q}, \gamma^T)$. For instance, suppose the model has three groups(represented by the coefficients $\gamma^T = (\gamma_1, \gamma_2)$ and a single factorial explanatory variable and if the aim is to test this covariate against zero, $\mathbf{C}$ becomes

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

with $\beta^T = (\gamma_{01}, ..., \gamma_{0q}, \gamma_1, \gamma_2)$ and $\varepsilon = (0, 0)^T$.

Tests may be created using the binomial representation of log-likelihood. This method allows you to use techniques from binary regression. The likelihood ratio test is a commonly used test statistic for linear hypotheses. This analysis compares two models: one without constraints and one with linear constraints. Let $\hat{\beta}$ be the ML estimate for the full model 3.24 and β indicate the estimate under the constraint $\mathbf{B}\beta = \varepsilon$. Then the likelihood ratio statistic is given by

© University of Venda

University of Venda
Creating Future Leaders

$LR = -2l(\beta - \hat{\beta})$ which quantifies the change of the log-likelihood l when evaluated at β and β.Under regularity requirements, LR follows a $\chi^2 - distribution$ with $s = rk(\mathbf{C})$ degrees of freedom, where $rk(\mathbf{C})$ represents the rank of the matrix $\mathbf{C}$. This involves fitting the model twice: once without constraints and once with restrictions. The log-likelihood differences are then calculated. Alternative test statistics that approximate LR include the Wald statistic and the score statistic. The Wald statistic takes the form

.

$W = (\mathbf{B}\hat{\beta} - \varepsilon)^T [\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{T}^T]_{-1}(\mathbf{C}\hat{\beta} - \varepsilon)$,

where $\mathbf{F}(\beta) = E(-\partial^2 l(\beta)/\partial\beta\partial\beta^T)$ denotes the Fisher information matrix. The Wald statistic has the benefit of just requiring the entire model to be fitted in order to produce $\hat{O}$; a fit under restrictions is not required. Keep in mind that the Wald test is comparable to the t test used in the R function glm for scalar ε. The third statistic is the score statistic

$u = s^T(\beta)\mathbf{F}^{-1}(\beta)s(\beta)$ where $s(\beta) = \partial l(\beta)/\partial\beta = (\partial l(\beta)/\partial\beta_1, ..., \partial l(\beta)/\partial\beta_p)^T$ is the score function evaluated at the fit of the constrained model. The likelihood ratio, Wald statistic, and score statistic all share the same asymptotic distribution

$\mathbf{LR}, w, u \ \chi^2(rank\mathbf{C})$

3.5.2 Residuals and Goodness-of-Fit

In the previous section, maximum likelihood estimation of discrete hazard models relied on binary representation of transitions between classifications. When analyzing residuals and goodness-of-fit, keep in mind that the original data is made up of n independent observations $(t_i, \delta_i, \mathbf{x}_i)$. Therefore, the deviations and residuals from fitting a binary model are not applicable here. We will provide methodologies for constructing valid residuals and goodness-of-fit statistics for discrete survival data.

© University of Venda

University of Venda
Creating Future Leaders

No Censoring

In a model where there is no censoring, the discrete time $T_i \in 1, ..., q$ follows a multinomial distribution. In general, a vector $\mathbf{Y}^T = (Y_1, ..., Y_k)$ is multinomially distributed, $\mathbf{Y}$ $M(n, (\pi_1, ..., \pi_k))$, if the probability mass function is given by

$$f(y_1, \ldots, y_k) = \begin{cases} \frac{n!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdots \pi_k^{y_k} & \text{if } y_i \in \{0, \ldots, n\}, \ \sum_{i=1}^k y_i = n, \\ 0 & \text{otherwise.} \end{cases}$$

where$\pi^T = (\pi_1, ..., \pi_k)$ is a probability vector, that is,$\pi_i \in [0, 1], \sum_i \pi_i = 1$ In discrete survival one considers the multinomially distributed vector $(T_{i1}, ..., T_{ik})$ $M(n, (\pi_1, ..., \pi_k))$, where T_{is} counts the number of responses in period s given that n_i observations are available at the same measurement value $\mathbf{x}_i$. If $n_i = 1$ just one of the variables $T_{i1}, ..., T_{ik}$ takes value 1. Consider the scenario with no filtering and data gathered at N measurement values $\mathbf{x}_i, i = 1, ..., N$, with n_i observations at $\mathbf{x}_i$. In other words, the covariate values (and consequently the observations) may be classified into N categories. The tests are identical to those used in categorical data analysis (Tutz 2012; Agresti 2013). With β denoting the maximum likelihood estimate, one obtains the fitted hazard $\hat{\lambda(t|\mathbf{x}_i)} = h(\gamma_{0t} + \mathbf{x}_i^{T\hat{}})$

and the fitted probabilities

$$\hat{\pi_{it}} = P(T \hat{=} t|\mathbf{x}_i) = \hat{\lambda}(t|\mathbf{x}_i) \prod_{s=1}^{t-1} (1 - \hat{\lambda}(s|\mathbf{x}_i)) \tag{3.25}$$

Furthermore, let p_{it} indicate the proportion of occurrences in subpopulation i during time t. The Pearson statistic is an adequate goodness of fit metric and is given by;

$$\chi_P^2 = \sum_{i=1}^N \sum_{t=1}^k n_i \frac{(p_{it} - \hat{\pi_{it}})^2}{\hat{\pi}_{it}}, \tag{3.26}$$

where the inner sum reflects the squared Pearson residual.

$$r_{P,i}^2 = \sum_{t=1}^k n_i \frac{(p_{it} - \hat{\pi_{it}})^2}{\hat{\pi}_{it}}. \tag{3.27}$$

© University of Venda

University of Venda
Creating Future Leaders

A smaller Pearson statistic (sum of Pearson residuals) indicates better model fit. The deviation is an alternative goodness-of-fit metric.

$$D = 2\sum_{i=1}^{N} n_i \sum_{t=1}^{k} p_{it} \log(\frac{p_{it}}{\hat{\pi}_{it}}) \tag{3.28}$$

with the corresponding quadratic deviance residuals given by

$$r_{D,i}^2 = 2n_i \sum_{t=1}^{k} p_{it} \log(\frac{p_{it}}{\hat{\pi}_{it}}) \tag{3.29}$$

When the model fits the data well, the deviation decreases, as seen using the Pearson statistic. Assuming fixed cells and regularity criteria, χ_P^2 and D are asymptotically $\chi^2 - distributed$ with $N(k-1) - p$ degrees of freedom, where p is the number of estimated parameters. A model is judged unsuitable if both or one of the test statistics exceed the $1-\alpha-quantile$ of the relevant distribution, $\chi_{1-\alpha}^2(N(k-1)-p)$. These tests are only useful for determining goodness-of-fit if n_i is not too tiny. For $n_i = 1$, with just one observation at fixed value $\mathbf{x}_i$, there is no asymptotic distribution available, making them useless. Nevertheless, residuals can be employed as diagnostic tools when $n_i = 1$. The squared deviation residual for observation i in this instance is

$$r_{D,1}^2 = 2\log(\frac{1}{\hat{\pi_{it_i}}}) = -2\log(\hat{\lambda}_{it_i}) - \sum_{s=1}^{t_i-1} 2\log(1-\hat{\lambda}_{is},) \tag{3.30}$$

where $\hat{\lambda}_{is} = \hat{\lambda}(s|\mathbf{x}_i)$. It can be re-written in a form that is familiar from response models,

$$r_{D,i}^2 = 2\sum_{s=1}^{t_i} y_{is} \log(\hat{\lambda}_{is}) + (1-y_{is})\log(1-\hat{\lambda}_{is}), \tag{3.31}$$

where the transitions over periods are indicated by the syntax $(y_{i1}, ..., y_{it_i}) = (0, ..., 0, 1)$. As a result, the residual for the binary response vector that codes the non-transitions may be expressed as the deviation residual. The depiction makes the comparison between observations and fit extremely clearer.

$$r_{D,i}^2 = 2\sum_{s=1}^{t_i} y_{is} \log(\frac{y_{is}}{\hat{\lambda}_{is}}) + (1-y_{is})\log(\frac{1-y_{is}}{1-\hat{\lambda}_{is}}), \tag{3.32}$$

© University of Venda

University of Venda
Creating Future Leaders

which reveals that at each time point the discrepancy between data and fit is measured by $\log(y_{is}/\hat{\lambda}_{is}) if y_{is} = 1$ and by $\log(1-y_{is})/(1-\hat{\lambda}_{is}) if y_{is} = 0$ For the Pearson statistic no such representation seems available. The adjusted deviance residual is a modification of the deviance residuals that, in the case of a well-fitting model, is usually closer to a normal distribution.

$$d_i = \sum_{s=1}^{t_i}\{sign()(y_{is}-\hat{\lambda}_{is})\sqrt{y_{is}\log(\frac{y_{is}}{\hat{\lambda}_{is}}) + (1-y_{is})\log(\frac{1-y_{is}}{1-\hat{\lambda}_{is}})}\}+\sum_{s=1}^{t_i}(1-2\hat{\lambda}_{is})/\sqrt{\hat{\lambda}_{is}(1-\hat{\lambda}_{is}).36} \tag{3.33}$$

In 3.33, take note that we assume $0.\log(0) \equiv 0$. Model fit may be evaluated while taking this kind of residual into account by looking at normal quantile–quantile plots.

3.5.3 Deviance in the Case of Censoring

Censoring complicates the examination of the model that has been fitted. With this situation, one models the observed time intervals as $min(T_i, C_i)$. With $q = k-1$, one can only see combined events.

$(t, \delta = 1) : \{T = t, C \geq t\}, t = 1, ..., q,$
$(t, \delta = 0) : \{T > t, C = t\}, t = 1, ..., q,$

Consequently, it is necessary to differentiate between 2_q categories. The comparable probabilities under the random censoring assumption are

$$\pi^{(\delta)_t} = \begin{cases} P(T=t)P(C>t) & \text{if } \delta = 1, \\ P(T>t)P(C=t) & \text{if } \delta = 0. \end{cases}$$

which in closed form can be written as $\pi_t^{(\delta)} = P(T=t)^{\delta}P(T>t)^{1-\delta}P(C \geq t)^{\delta}P(C=t)^{1-\delta}$

Let the probabilities for observation i be collected in $\boldsymbol{\pi}_i^T = ((\boldsymbol{\pi}_i^{(1)})^T, (\boldsymbol{\pi}_i^{(0)})^T)$ where

© University of Venda

University of Venda
Creating Future Leaders

$\boldsymbol{\pi}_i = ((\pi_{i1}^{(1)}, ..., \pi_{iq}^{(1)})^T$, and $\boldsymbol{\pi}_i^0 = ((\pi_{i1}^{(1)}, ..., \pi_{iq}^{(0)})^T$, Then the corresponding

$$D = 2\sum_{i=1}^{N} n_i\{\sum_{r=1}^{q} p_{ir}^{(1)} \log(\frac{p_{ir}^{(1)}}{\hat{\pi}_{ir}^{(1)}}) + \sum_{r=1}^{q} p_{ir}^{(0)} \log(\frac{p_{ir}^{(0)}}{\hat{\pi}_{ir}^{(0)}})\} \tag{3.34}$$

with the terms on the right-hand side given by

$\log(\frac{p_{ir}^{(1)}}{\hat{\pi}_{ir}^{(1)}}) = \log(\frac{P(T_i=r)}{\hat{P}(T_i=r)}) + \log(\frac{P(C_i\geq r)}{\hat{P}(C_i\geq r)})$,

$\log(\frac{p_{ir}^{(0)}}{\hat{\pi}_{ir}^{(0)}}) = \log(\frac{P(T_i>r)}{\hat{P}(T_i>r)}) + \log(\frac{P(C_i=r)}{\hat{P}(C_i=r)})$.

To utilize deviation as a goodness-of-fit metric, define two models: one for survival time and one for censoring procedure. Goodness-of-fit applies to both models. However, it would be preferable to assess the survival model's fit separately. Although the likelihood of code transitions does not provide a goodness-of-fit statistic, it can still be used to define residuals for censored observations. For single observations (i.e., $n_i = 1$) the contribution of the i^th observation to the corresponding "deviance" $D = -2l$ (with l denoting the log-likelihood (3.25)) is

$$r_{D,i}^2 = -2\{\delta_i \log(\hat{P}(T_i = t_i)) + (1 - 1\delta_i) \log(\hat{P}(T_i > t_i))\} \tag{3.35}$$

$$= -2\sum_{s=1}^{t_i} y_{is} \log(\hat{\lambda}_{is}) + (1 - y_{is}) \log(1 - \hat{\lambda}_{is}). \tag{3.36}$$

$$\tag{3.37}$$

which is the same as 3.31. Even in the situation of censored data, we take it into consideration as "squared deviance residuals."

3.5.4 Martingale Residuals

An alternate sort of residual that brings censoring into account and is well suited for examining the functional forms of predictor impact is the martingale residual denoted by

© University of Venda

University of Venda
Creating Future Leaders

$$m_i = \delta_i - \sum_{s=1}^{t_i} \hat{\lambda}_{is}, i = 1, ..., n, \tag{3.38}$$

where $\hat{\lambda}_{is} = \hat{\lambda}(s|\boldsymbol{x}_i)$. Here, $\Lambda(t_i) = \sum_{s=1}^{t_i} \hat{\lambda}_{is}$ measures the cumulative risk of observation i up to time t_i. In survival analysis for continuous time, this concept is comparable to the cumulative hazard function. Comparing the actual number events that occurred up to t_i (measured by δ_i) and the predicted number of occurrences up to t_i (measured by $\Lambda(t_i)$) for each person is the concept behind the martingale residual. If one uses the binary variables representation with $(y_{i1}, ..., y_{it_i})$ the residuals can be defined as

$$m_i = \sum_{s=1}^{t_i} (y_{is} - \hat{\lambda}_{is}), i = 1, ..., n. \tag{3.39}$$

The martingale residuals, unlike the deviance residuals, which use the log-transformed proportion of observation and fit, consider the disparities between transition indicators and calculated failure probability.It can be demonstrated that the martingale residuals for the logistic model add up to zero. The martingale residuals should be "random" and uncorrelated with the covariate values for a well-fitting model that incorporates all pertinent variables. According to this theory, the functional forms and significance of the variables in a discrete-time survival model may be evaluated using martingale residuals. Plotting the residuals against the relevant covariate values is how this is accomplished for each covariate. Plotting a smooth estimate, such as one produced using a P-spline estimator, adds more details on the functional shape of the corresponding covariate impact.

3.5.5 Measuring Predictive Performance

We looked at ways to evaluate parametric regression models' goodness-of-fit in the preceding section. These techniques reveal if a survival model is well calibrated, i.e., on

© University of Venda

University of Venda
Creating Future Leaders

how well the fitted survival probabilities agree with their corresponding observed proportions. It is frequently of interest to evaluate survival models' effectiveness in terms of forecasting the survival of future data in addition to their goodness-of-fit. Because a high goodness of fit in the observed data does not always indicate that the model will perform similarly on future data, this problem is significant in many applications. Furthermore, statistical hypothesis testing frequently yield scant information on the significance of a covariate (or the entire model) for future survival prediction. For instance, even if covariate effects have a minor impact on survival (as indicated, for instance, by the magnitude of the parameters $\boldsymbol{\gamma}$), they may still be important if sample sizes are high. Three methods for assessing the survival models' forecast accuracy will be discussed in the sections that follow. For all approaches we assume that the survival model was fitted on a learning data set $(t_i, \delta_i, \boldsymbol{x}_i) = 1, ..., n$ and that prediction accuracy is evaluated on an additional independent test data set $(t_j^{\mathcal{T}}, \delta_j^{\mathcal{T}}, \boldsymbol{x}_j^{\mathcal{T}})j = 1, ..., n^{\mathcal{T}}$, that follows the same distribution as the learning data. The superscript $\mathcal{T}$ indicates that the respective observations are from the test data. It should be noted that assessing predictive performance requires the use of separate sets of test and learning data. This helps to prevent overly optimistic forecasts, which would result from using the same data for both model fitting and prediction accuracy assessment. Cross-validation procedures can be used when there isn't an independent test data set available. In this instance, the goal is to repeatedly divide the data into test and learning data subsets. The model is then fitted to all learning data subsets, the model fits are assessed on the corresponding test data subsets, and the resultant estimations of prediction accuracy are averaged to produce cross-validated predictions.

Predictive Deviance and R2 Coefficients

The predictive deviance evaluated on the test data is denoted by

© University of Venda

University of Venda
Creating Future Leaders

$$D = -2\sum_{j=1}^{n^{\mathcal{T}}}\{\delta_j^{\mathcal{T}}\log(\hat{P}(T_j^{\mathcal{T}} = t_j^{\mathcal{T}})) + (1-\delta_j^{\mathcal{T}})\log((\hat{P}(T_j^{\mathcal{T}} = t_j^{\mathcal{T}}))\} \tag{3.40}$$

$$= -2\sum_{j=1}^{n^{\mathcal{T}}}\sum_{s=1}^{t_j^{\mathcal{T}}}\{y_{js}^{\mathcal{T}}\log(\hat{\lambda}_{js}) + (1-y_{is}^{\mathcal{T}})\log(1-\hat{\lambda}_{js})\} \tag{3.41}$$

$$\tag{3.42}$$

where $\hat{\lambda}_{js} = \hat{P}(T_j^{\mathcal{T}} = s|T_j^{\mathcal{T}} \geq s, \boldsymbol{x}_j^{\mathcal{T}})$ and where $T_j^{\mathcal{T}}$ are (unobserved) true survival times in the test data. As before, $(y_{j1}^{\mathcal{T}}, ..., y_{jt_j^{\mathcal{T}}}^{\mathcal{T}}) = (0, ..., 0, 1)$ if $\delta_j^{\mathcal{T}} = 1$ and $(y_{j1}^{\mathcal{T}}, ..., y_{jt_j^{\mathcal{T}}}^{\mathcal{T}}) = (0, ..., 0, 0)$ if $\delta_j^{\mathcal{T}} = 0$ denote the transitions over periods. The predictive deviance is the negative log-likelihood of a binomial regression model based on test data. Take note that the predicted deviation is an unlimited quantity. To simplify understanding, take R^2-type coefficients provided by

$$R^2 = \frac{1 - \exp((\sum_{j=1}^{n^{\mathcal{T}}} T_j^{\mathcal{T}})^{-1}(D - D_0))}{\exp((\sum_{j=1}^{n^{\mathcal{T}}} T_j^{\mathcal{T}})^{-1} D_0} \tag{3.43}$$

where D_0 is the predictive deviance obtained from a null model without covariate information. It can be shown that (4.8) is equal to 1 if prediction accuracy is perfect and equal to zero if predictions are based on the null model (Nagelkerke 1991). However, there are various issues with using R^2 coefficients in actual applications. Even in highly predictive models, R^2 coefficients frequently fall below 1, making interpretation challenging. Unlike predictive deviance, it's uncertain if R^2 coefficients are "proper" in the sense that they become maximum when computed from the real underlying model.

Prediction Error Curves

Prediction Error(PE) curves are a time-dependent metric of prediction error based on the squared distance between expected survival functions
$\hat{S}_j(t) = \prod_{s=1}^{t}(1-\hat{\lambda}_{js})$, $j = 1, ..., n^{\mathcal{T}}$, and the corresponding *observed* survival functions

© University of Venda

University of Venda
Creating Future Leaders

$\tilde{S}_j(t) = I(t < t_j^{\mathcal{T}})$, where $I(.)$denotes the indicator function. PE curves measure the difference between observed (i.e., $\tilde{S}_j(t)$) and predicted (i.e., $t_j^{\mathcal{T}}$) survival functions. These functions are defined as $I(t < t_j^{\mathcal{T}}) = 1 if t < t_j^{\mathcal{T}}$ and 0 otherwise. For fixed t, the PE curve (which includes *inverse* probability of censoring *weights* $w_j^{\mathcal{T}}(t)$) is given by

$$\widehat{PE}(t) = \frac{1}{n^{\mathcal{T}}} \sum_{j=1}^{n^{\mathcal{T}}} [\frac{\delta_j^{\mathcal{T}}(1 - 1\tilde{S}_j(t))}{\hat{G}_j(t_j^{\mathcal{T}})} + \frac{\tilde{S}_j(t)}{\hat{G}_j(t)}](\hat{S}_j(t) - \tilde{S}_j(t))^2 \tag{3.44}$$

$$= \frac{1}{n^{\mathcal{T}}} \sum_{j=1}^{n^{\mathcal{T}}} w_j^{\mathcal{T}}(t)(\hat{S}_j(t) - \tilde{S}_j(t))^2 \tag{3.45}$$

$$\tag{3.46}$$

The weights $w_j^{\mathcal{T}}(t)$ include the statement $\hat{G_j}(t) = \hat{P}(C_j > t|\boldsymbol{x}_j^{\mathcal{T}}$, which represents the estimated survival function of the censoring process acquired from the learning data. Their role will be addressed later. The second item in 3.44 reflects the fundamental measure of disagreement, which is the squared deviation between observed and expected survival for fixed t. One limitation of this squared distance is that it can only be calculated using test data.

- (a) if an event was seen at or before t (implying that $\tilde{S}_j(t) = 0$)
- (b) if the person is still living at t (implying that the event happened after t, that is, $\tilde{S}_j(t) = 1$).

The two summands in the first component of 3.44 reflect the situations (a) and (b). The first summand's numerator $\delta_j^{\mathcal{T}}(1 - \tilde{S}_j(t))$ becomes 1 in situation (a) and 0 otherwise, while the second summand's numerator $\tilde{S}_j(t)$ becomes 1 in situation (b) and zero otherwise, yielding the weights $w_j^{\mathcal{T}} = 1/\hat{G}_j(t_j^{\mathcal{T}-1}$ For (a), $w_j^{\mathcal{T}} = 1/\hat{G}_j(t)$, while for (b).

© University of Venda

University of Venda
Creating Future Leaders

Censoring data before t makes it difficult to determine the squared deviation in 3.44. These observations are not included in the computation of the PE curve at t. Ignoring observations that were suppressed before t might result in bias. The inclusion of the inverse probability of censoring weights $w_j^{\mathcal{T}}, j = 1, ..., n^{\mathcal{T}}$ in 3.44 is due to this. The concept behind inverse probability of censoring weighting is as follows: Assume all survival times were completely observed (without censorship). To calculate $\widehat{PE}(t)$, use the whole data sample with all observations and compute the squared deviation $(\hat{S}_j(t) - \tilde{S}_j(t))^2$ for each t. Censoring prevents this computation, thus we must rely on a smaller sample of uncensored data for a given time point (t). The weights $w_j^{\mathcal{T}}$ are used to "reconstruct" the unknown entire data sample's properties. Weights determine the likelihood of censoring an observation in test data after t. Assume $\hat{G}_j$ is close to 1 at a specific time point.The censoring event is expected to occur after this time point, suggesting that $w_j^{\mathcal{T}}$ is also close to 1. The value of $w_j^{\mathcal{T}}$ indicates confidence that the observed data is roughly representative of the complete sample without censoring. If $\hat{G}_j$ is tiny, observations up to the time point of interest may only represent a small portion of the whole data sample. As a result, they are weighted to recreate the overall structure of the sample.

In surveys, for instance, under-represented groups of observations are frequently up-weighted by their inverse probabilities of being included in the survey to obtain unbiased estimates, and groups with missing data are frequently up-weighted by their inverse probabilities of being completely observed. In mathematical terms, inverse probability of censoring weighting ensures the consistency of the estimator for the mean of the random variable $(\hat{S}_j(t) - S_j(t))^2$. This approach is comparable to strategies for avoiding selection bias and for handling missing data in survey research.

Obtaining Estimates of the Censoring Process

The survival function of the censoring process, $G_j(t) = P(C_j > t|\boldsymbol{x}_j^{\mathcal{T}})$, must be

© University of Venda

University of Venda
Creating Future Leaders

calculated in order to calculate the inverse probability of censoring weights. Similar to estimates of the survival function S_j, estimates of $\hat{G}_j$ may be derived from the learning data by modeling the censoring model as a binary regression model. However, censorship is now the event of interest, and the censoring process is influenced by survival observations. Take, for instance, the case when the measured survival time is t_i and $\delta_j = 1$. Since $T_i \leq C_i$ in this instance, it follows that the censoring event must occur after $t_i - 1$ and that the unobserved censoring time C_i has been more than $t_i - 1$. Consequently, the binary observations and design variables for the estimation of the censoring process are given by

Binary Observations	Design Variables
0	$1\ x_i^T$
0	$2\ x_i^T$
0	$3\ x_i^T$
$\vdots$	
0	$t_i - 1\ x_i^T$

where the running time t has to be considered as a factor when using appropriate software. Similarly, the censoring event occurred at t_i, and the censoring time C_i is equal to t_i if the observed survival time is t_i and $\delta_i = 0$. In this instance, the design variables and binary data used to estimate the censoring process are provided by

Binary Observations	Design Variables
0	$1\ x_i^T$
0	$2\ x_i^T$
0	$3\ x_i^T$
$\vdots$	
0	$t_i\ x_i^T$

© University of Venda

University of Venda
Creating Future Leaders

The value of the PE curve in 3.44 is frequently insensitive to the set of design variables that are part of the censoring model as it is an average over the $n^{\mathcal{T}}$ observations in the test sample. To estimate the censoring process, it is common practice to fit a covariate-free censoring model (i.e., to utilize columns 1 and 2 of the aforementioned design matrix but not the covariates). For the censoring process, this basically translates to a life table estimate.

Properties and Integrated Prediction Error Curve

If the anticipated survival functions closely match the observed survival functions, the PE curve is by definition tiny. Furthermore, it can be demonstrated that the PE curve is a "proper scoring rule" for every t in that it becomes minimum if $\hat{S}$ equals the genuine survival function $P(t_j^{\mathcal{T}} > t|\boldsymbol{x})$.A time-independent coefficient of prediction error is given by the *integrated* PE curve, which is defined as

$$\widehat{PE}_{int} = \hat{E}_T[\widehat{PE}(T)] = \sum_t \widehat{PE}(t).\hat{P}(T = t) \tag{3.47}$$

PE values for all t should, in general, be less than 0.25 in a well-predicting survival model. For every t, the value of the PE curve derived from a non-informative model with $\hat{S}(t) = 0.5$ is 0.25, which explains this.

Brier Score The Brier score is a frequently used metric to assess probabilistic predictions in discrete response models. For one observation it has the form

$$b_i = \delta_i(1 - \hat{\pi}_{it_i})^2 + \sum_{s \neq t_i} \hat{\pi}_{is}^2. \tag{3.48}$$

Discrimination Measures

Calculating discrimination measures for the predictor $\eta : \boldsymbol{x}^T\boldsymbol{\gamma}$ is an alternate method for assessing prediction accuracy. Considering survival outcomes as time-dependent binary variables with levels "event at t" and "event after t" is the concept behind dis-

© University of Venda

University of Venda
Creating Future Leaders

criminating measures. As a result, recognized notions for assessing binary classification methods may be used to gauge prediction accuracy. This is carried out in this manner: Observations with an event at t are regarded as "cases" for each fixed time point t, whereas observations with an event after t are regarded as "controls."

According to this definition, if *eta* has *highdiscriminativepower*—that is, the capacity to distinguish between cases and controls in the test data—it has a high prediction accuracy. Specificities and "time-dependent sensitivities" can be used to quantify discriminative power., which are defined as

$$sens(c,t) := P(\eta > c|T = t) \tag{3.49}$$

and

$$sens(c,t) := P(\eta \leq c|T > t), \tag{3.50}$$

respectively.Here, c is a threshold of the predictor . A high sensitivity $P(\eta > c|T = t)$ indicates that η and hence the hazard $\lambda(t|\boldsymbol{x}) = h(\eta)$ are likely to be considerable. This is equivalent to detecting "cases" or "hit rate" in signal detection theory. A high specificity $P(\eta \leq c|T > t)$ indicates that the dangers for $T > t$ are minimal, implying that "controls" are accurately recognized. In signal detection theory, this is referred to as "correct rejection," as the "cases" signal is not present. Sensitivity and specificity rates vary among thresholds. As $c \to -\infty$, sensitivity decreases to one and specificity decreases to zero. Conversely, as $c \to \infty$, sensitivity increases to zero and specificity decreases to one. Summarizing $sens(c,t) and spec(c,t)$ for fixed t yields the $time - dependent ROC curve$, which is defined as

© University of Venda

University of Venda
Creating Future Leaders

$$\text{ROC}(c,t) \coloneqq \{(1-\text{spec}(c,t), \text{sens}(c,t))\}_{c\in\mathbb{R}}$$

The ROC curve compares the hit rate (sensitivity) and false positive rate (1-specificity) at different thresholds. It typically has a concave shape connecting the points (0,0) and (1,1). If η has a high discriminative power the curve is strongly concave and has a large area below the curve. Analogously to ROC analysis for binary outcomes, it is possible to calculate the areas under the time-dependent ROC curves for each time point t. This results in the $time-dependentAUCcurve$, denoted by $AUC(t)$. The AUC curve should exceed 0.5 at all time points, as this is the result achieved by a null model without covariate information. The area under the time-dependent AUC curve, like the integrated prediction error curve, may be used to quantify discriminative power independent of time. Following the approach by Heagerty and Zheng (2005), we consider the index

$$C^* = \sum_t AUC(t).w(t) \tag{3.51}$$

with weights $w(t) = P(T=t)P(T>t)/\sum_t P(T=t)P(T>t)$. The probability $P(\eta_{j1}^{\mathcal{T}} > \eta_{j2}^{\mathcal{T}} | T_{j1}^{\mathcal{T}} < T_{j2}^{\mathcal{T}})$, which is a global concordance index measuring the probability that observations with large values of η have shorter survival times than observations with small values of η, is equal to C^*. The variables $\eta_{j1}^{\mathcal{T}}, \eta_{j2}^{\mathcal{T}}, T_{j1}^{\mathcal{T}}, and T_{j2}^{\mathcal{T}}$ represent the predictors and survival times of two randomly chosen observations j_1 and j_2 from the test sample. For a predictor η to outperform chance, C^* should be greater than 0.5, similar to the time-dependent AUC curve. Following Uno et al. (2007), we estimate by

$$\widehat{sens}(c,t) = \frac{\sum_j {\delta_j}^{\mathcal{T}} I(\hat{\eta}_j^{\mathcal{T}} > c \cap t_j^{\mathcal{T}} = t)/\hat{G}_j(t_j^{\mathcal{T}} - 1)}{\sum_j {\delta_j}^{\mathcal{T}} I(t_j^{\mathcal{T}} = t)/\hat{G}_j(t_j^{\mathcal{T}} - 1)} \tag{3.52}$$

$$\tag{3.53}$$

respectively, where $\hat{\eta}_j$, $j = 1, ..., n^{\mathcal{T}}$ denote the estimates of $\boldsymbol{x}^T\boldsymbol{\gamma}$ in the test data. Similar to the estimator of the prediction error curve in 3.44, the weights $1/\hat{G}_j(t_j^{\mathcal{T}} -$

© University of Venda

University of Venda
Creating Future Leaders

1) ensure the consistency of 3.52. Estimates of $AUC(t)$ can be obtained by using numerical integration of the estimated ROC curve $\{1 - \widehat{spec}(c,t), \widehat{sens}(c,t)\}_{c\in R, t\geq 0}$. The concordance index C^* can be estimated by

$$C* = \sum_t \widehat{AUC}(t)\hat{P}(T=t)\hat{P}(T>t) / \sum_t \hat{P}(T=t)\hat{P}(T=t), \tag{3.54}$$

where $\widehat{AUC}(t)$ denotes the estimated time-dependent AUC curve.

3.6 Discrete-time competing risk model

3.6.1 Introduction

In this section, we will first look into parametric competing risk models for discrete time-to-event data, and then illustrate how the estimation may be integrated into the framework of GLMs. In this study, we will construct a discrete temporal cause-specific hazard model to investigate the influence of certain factors on the length of graduating or dropping out.

3.6.2 Parametric Models

In basic time-to-event models with one goal event, the process dynamics were characterized by a single hazard function. When there are several target events, it is useful to construct separate hazard functions for each type of event. Let $R \in 1; ...; m$ represent different target events. For discrete time $T \in 1; ...; q+1$, the cause-specific hazard function resulting from cause or risk r is defined by

$$\lambda_r(t|\mathbf{x}) = P(T=t, R=r|T\geq t, \mathbf{x}), \tag{3.55}$$

where $\mathbf{x}$ is a vector of time-independent covariates. The m hazard functions $\lambda_1(t|\mathbf{x}), ..., \lambda_m(t|\mathbf{x})$ can be combined into one hazard function that describes the process regardless of the type of target event. The resulting overall hazard function is defined by

© University of Venda

University of Venda
Creating Future Leaders

$$\lambda(t|\mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = P(T = t, R = r|T \geq t, \mathbf{x}). \tag{3.56}$$

The survival function and unconditional probability of an occurrence in period t are identical to those for a single target event, i.e.,

$$S(t|x) = P(T > t|\mathbf{x}) = \prod_{i=1}^{t}(1 - \lambda(i|\mathbf{x})) \tag{3.57}$$

and

$$P(T = t|\mathbf{x}) = \lambda(t|\mathbf{x})S(T - 1|\mathbf{x}) \tag{3.58}$$

When a person reaches interval $[a_{t-1}, a_t)$, there are m alternative outcomes, including completing one of the m objective events or surviving beyond $[a_{t-1}, a_t)$. The corresponding conditional response probabilities are given by

$\lambda_1(t|\mathbf{x}), ..., \lambda_,m(t|\mathbf{x}), 1 - \lambda(t|\mathbf{x}),$

where $1 - \lambda(t|\mathbf{x})$ is the probability of survival. Natural models for $m + 1$ occurrences are multi-category models used to model categorical data.

We'll look at the popular multinomial model and a discretized version of a continuous-time competing risks model.

Cause-specific discrete hazard

$$\lambda_r(t|\mathbf{x}) = P(T = t, R = r|T \geq t, \mathbf{x}), \tag{3.59}$$

Overall hazard

$$\lambda(t|\mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = P(T = t, R = r|T \geq t, \mathbf{x}). \tag{3.60}$$

Survival function

$$S(t|x) = P(T > t|\mathbf{x}) = \prod_{i=1}^{t}(1 - \lambda(i|\mathbf{x})) \tag{3.61}$$

© University of Venda

University of Venda
Creating Future Leaders

Event probability

$$P(T = t, R = r|\mathbf{x}) = \lambda_r(t|\mathbf{x})S(t-1|\mathbf{x}) \tag{3.62}$$

Multinomial Logit Model

The multinomial logit model is commonly used to analyze categorical answers (e.g., Tutz and Petry (2012) or Agresti (2012)). Discrete survival involves reactions to either target events or survival. The corresponding model is given by

$$\lambda_r(t|\mathbf{x}) = \frac{\exp(\gamma_{0tr} + \mathbf{x}^T\boldsymbol{\gamma}_r)}{1 + \sum_{i=1}^{m}\exp(\gamma_{0ti} + \mathbf{x}^T\boldsymbol{\gamma}_i).} \tag{3.63}$$

for $r = 1, ..., m, t = 1, ..., q$. The parameters $\gamma_{01j}, ..., \gamma_{0qj}$ represent the cause-specific baseline hazard function, and $\boldsymbol{\gamma}_r$ is the cause-specific vector of coefficients. Although $m + 1$ response categories are possible, specifying the conditional probability of the target events $1, ..., m$ is sufficient. The reference category is conditional survival, and its probability is simply stated by

$$P(T > t|T \geq t, \mathbf{x}) = 1 - \sum_{r=1}^{m}\lambda_r(t|\mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{m}\exp(\gamma_{0tj} + \mathbf{x}^T\boldsymbol{\gamma}_j)}. \tag{3.64}$$

Thus the conditional model that is used is the multinomial logit model for $m+1$ categories. With $R \in 0, 1, ..., m, where$R=0 denotes conditional survival, the conditional probabilities that sum up to 1 are given by $\lambda_0(t|\mathbf{x}) = P(T > t|T \geq t, \mathbf{x}), \lambda_1(t|\mathbf{x}), ..., \lambda_m(t|\mathbf{x})$.

For the interpretation of parameters it is useful to consider the model in the form

$$\log(\frac{\lambda_r(t|\mathbf{x})}{\lambda_0(t|\mathbf{x})}) = \gamma_{0tr} + \mathbf{x}^T\boldsymbol{\gamma}_r. \tag{3.65}$$

It is seen that the linear predictor determines the cause-specific log-odds, that means the logarithm of the proportion $\lambda_r(t|\mathbf{x})/\lambda_0(t|\mathbf{x})$, which compares the conditional probability of the target event $R = r$ to conditional survival. With parameter vector $\boldsymbol{\gamma}_r^T = \gamma_{r1}, ..., \gamma_{rp}$ one obtains

© University of Venda

University of Venda
Creating Future Leaders

$$\frac{\lambda_r(t|\mathbf{x})}{\lambda_0(t|\mathbf{x})} = \exp(\gamma_{0tr}) \exp(\gamma_{r1})^{x_1} ... \exp(\gamma_{rp})^{x_p}. \tag{3.66}$$

Increasing x_j by one unit improves the cause-specific probability by $\exp(\gamma_{rj})$. γ_{0tr} represents the additive influence of variable x_j on log-odds, whereas $\exp(\gamma_{rj})$ demonstrates the multiplicative effect, which is typically more clear. To simplify the baseline hazard function, assume a smooth function and vary the weight $\boldsymbol{\gamma}_r$ based on time, similar to the single-event scenario.

Ordered Target Events

The multinomial logit model is applicable to several risk categories, but has numerous parameters. To estimate the baseline hazard γ_{0tr} for each risk category, use a function in t and the parameter vector $\boldsymbol{\gamma}_r$. Ordering the causes or dangers results in sparser parameterizations. In a research on unemployment, the goal events $R \in 1, ..., m$ can represent "part-time job" or "full-time job" as alternatives to staying jobless, with the latter represented by $R = 0.R$ represents an ordered categorical answer at a specified time point. McCullagh's (1980) classical ordinal response models can be used to simulate cause-specific risks if R follows a natural ordering. The cumulative-type model calculates cumulative probability for discrete hazards as follows:

$$\lambda_r(t|\mathbf{x}) = \sum_{j=1}^{r} \lambda_j(t|\mathbf{x}) = P(T = t, R \leq r|T \geq t, \mathbf{x}). \tag{3.67}$$

The so-called cumulative model for ordered responses assumes

$$\lambda_r(t|\mathbf{x}) = F(\gamma_{0tr} + \mathbf{x}^T \boldsymbol{\gamma}_. ...). \tag{3.68}$$

where $F(.)$ s a fixed cumulative distribution function, and the intercepts must satisfy $\gamma_{0tr} \leq \gamma_{0t,r+1}$ for all r. Specifically, if $F(.)$ is the logistic distribution function, one obtains McCullagh's proportional odds model for responses R given by

$$\log(\frac{\lambda_r(t|\mathbf{x})}{1 - \lambda_r(t|\mathbf{x})}) = \gamma_{0tr} + \mathbf{x}^T \boldsymbol{\gamma}. \tag{3.69}$$

© University of Venda

University of Venda
Creating Future Leaders

The model's benefit over the multinomial logit model is the necessity for just one parameter vector $\boldsymbol{\gamma}$, rather than one for each category.

Separate Modeling of Single Targets

Competing risk models simulate numerous target events at the same time. To simplify modeling, consider focusing on a single target event and treating other data as censored. Then, for fixed r_0 one would use the hazard function

$$\lambda_{r0}(t|\mathbf{x}) = \lambda(t|\mathbf{x}) = P(T = t, R = r_0|T \geq t, \mathbf{x}). \tag{3.70}$$

Let the data for the competing risks again be given by $(t_i, r_i, \delta_i, \mathbf{x})$, where $r_i \in 1, ..., m$ indicates the target event. When focussing on this event, one considers the transformed data
$(t_i, r_i, {\delta_i}^{(r_0)}, \mathbf{x})$, where ${\delta_i}^{(r_0)} = 0 if \delta_i = 0 or r_i \neq r_0$.

The indicator function ${\delta_i}^{(r_0)}$ denotes censoring in the single-cause model for the r_0^{th} target event.This strategy appears appealing due to its use of simple binary models, but has significant drawbacks. Estimation, in particular, is susceptible to significant bias. Simple discrete survival models require independence between censoring and survival (known as random censoring). When survival times for competing events are associated, the simplified model's censoring process for individual targets correlates with survival time. Ignoring the link between survival time and censoring leads to erroneous estimates of dangers. If survival times are connected, it is not advisable to model individual targets separately. In practice, when the correlation structure is unclear, competing risks modeling is a preferable option.

3.6.3 Maximum Likelihood Estimation

Consider the data as $(t_i, r_i, \delta_i,)\mathbf{x}_i$, with $r_i \in 1, ..., m$ indicating the target event. We assume random censoring at the end of the period with $t_i = minT_i, C_i$, where events are determined by the indicator function

© University of Venda

University of Venda
Creating Future Leaders

$$\delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i \\ 0, & \text{otherwise.} \end{cases} \tag{3.71}$$

The following demonstrates how maximum likelihood estimation in multivariate GLMs may be used to derive estimates of competing risk models. For the generic model, the ith observation's probability contribution is

$$L_i = P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$$

If random censoring is assumed, the element $P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ may be left out, then the probability drops to

$$L_i = P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} = \lambda_{r_i}(t_i|\mathbf{x}_i)^{\delta_i}(1 - \lambda(t_i|\mathbf{x}_i))^{1-\delta_i} \prod_{t=1}^{t_i-1}(1 - \lambda(i|\mathbf{x}_i)) \tag{3.72}$$

The following is how one specifies indicator variables: Every observation has a definition for $t < t_i$.

$$\mathbf{y}_{it}^T = (y_{it0}, y_i t1, ..., y_{itm}) = (1, 0, ..., 0)$$

which encodes survival for all time points before $t - i$. For $t = t_i$ one defines for $\delta_i = 1$ $\mathbf{y}_{it_i}^T = (y_{it_i0}, y_i t_i 1, ..., y_{it_i m}) = (0, ...1, ..., 0)$

with $y_{it_i ri} = 1$and all other entries zero, and for $\delta_i = 0$

$$\mathbf{y}_{it_i}^T = (y_{it_i0}, y_i t_i 1, ..., y_{it_i m}) = (1, 0, ..., 0)$$

With these indicator variables the likelihood for the $i_t h$ observation can be written in the form

© University of Venda

University of Venda
Creating Future Leaders

$$L_i = \prod_{t=1}^{t_i}\prod_{r=1}^{m}\lambda_r(t|\mathbf{x}_i)^{y_{itr}}(1-\lambda(t|\mathbf{x}_i)^{y_{it0}}) = \prod_{t=1}^{t_i}\prod_{r=1}^{m}\lambda_r(t|\mathbf{x}_i)^{y_{itr}}(1-\sum_{r=1}^{m}\lambda_r(t|\mathbf{x}_i)^{y_{it0}}) \quad (3.73)$$

In a multinomial response model, the likelihood for the i_th observation equals the likelihood for the ti observations $\mathbf{y}_{i1}, ..., \mathbf{y}_{it_i}$. The indicator variables reflect distributions at a certain interval. Given that an individual reaches interval $[a_{t-1}, a_t)$, the response is multinomially distributed with $\mathbf{y}_{it}^T = (y_{it0}, y_it1, ..., y_{itm})\ M(1, 1-\lambda(t|\mathbf{x}_i), \lambda_1(t|\mathbf{x}_i), ..., \lambda_m(t|\mathbf{x}_i)$ The dummy variable $y_{it0} = 1 - y_{it1} - ... - y_{itm}$has value 1 if individual i does not fail in interval $[a_{t-1}, a_t)$ and $y_{it0} = 0$ if individual i fails in $[a_{t-1}, a_t)$ As a consequence, the likelihood is that of the multi-categorical model $P(Y_{it} = r) = h_r(\mathbf{X}_t\beta)$, where $Y_{it} = r$ if $y_{itr} = 1$. Maximum likelihood estimates may be produced in multivariate generalized linear models by augmenting design matrices, similar to single-cause models. The response matrix for the ith observation is as follows:

$$\begin{bmatrix} \mathbf{y}_{i1} \\ . \\ . \\ . \\ \mathbf{y}_{it_i} \end{bmatrix}$$

and design matrix is given by

$$\begin{bmatrix} \mathbf{X}_1 \\ . \\ . \\ . \\ \mathbf{X}_{ti} \end{bmatrix}$$

and the total log-likelihood is given by

$$\begin{aligned} l &= \sum_{i=1}^{n}\sum_{t=1}^{t_i}(\sum_{r=1}^{m} y_{itr} log\lambda_r(t|\mathbf{x}_i) + y_{it0} log(1-\sum_{r=1}^{m}\lambda_r(t|\mathbf{x}_i))) \\ &= \sum_{t=1}^{q}\sum_{i\in R_t}^{t_i}(\sum_{r=1}^{m} y_{itr} log\lambda_r(t|\mathbf{x}_i) + y_{it0} log(1-\sum_{r=1}^{m}\lambda_r(t|\mathbf{x}_i))), \end{aligned} \quad (3.74)$$

© University of Venda

University of Venda
Creating Future Leaders

where in the latter form $R_t = i|t_i \geq t$ is the number of individuals at risk in the interval $[a_{t-1}, a_t]$.

3.6.4 Variable Selection

The cause-specific hazard's linear predictor in the multinomial model $\lambda_r(t|\mathbf{x})$ possesses the form

$$\eta_r = \gamma_{0tr} + \mathbf{x}^T \boldsymbol{\gamma}_r, \tag{3.75}$$

where $\mathbf{x}^T = (x_1, ..., x_p)$ and $\boldsymbol{\gamma}_r{}^T = (\gamma_{r1}, ..., \gamma_{rp})$. Let all of the parameters be gathered in the vector $\boldsymbol{\gamma}_r$ as follows: In single-cause models, variable selection can be achieved by employing strategies that optimize the penalized log-likelihood. Rather than using the standard log-likelihood $l(\boldsymbol{\gamma})$, $l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{\lambda}{2} J(\boldsymbol{\gamma})$. If there are a lot of relevant predictors available, the lasso can be utilized to find them. However, the direct application of the lasso has significant drawbacks for the cause-specific model as the basic lasso penalty

$$J(\boldsymbol{\gamma}) = \sum_{r-1}^{m} \sum_{j=1}^{p} \gamma_{rj} \tag{3.76}$$

enforces parameter selection but not variable selection. A portion of the tuning parameters will be set to zero with proper parameter selection, while the remaining parameters can be connected to any of the variables. As a result, it's possible that the model still has to include every variable. Alternatively, a penalty that identifies groupings of parameters and associates them with a single covariate each is used to directly impose variable selection as opposed to parameter selection. Let all the effects on the jth variable be collected in $\boldsymbol{\gamma}_{.j}{}^T = (\gamma_{1j}, ..., \gamma_{mj})$.

and consider the grouped penalty

$$J(\boldsymbol{\gamma}) = \sum_{j=1}^{p} \boldsymbol{\gamma}_{.j} = \sum_{j=1}^{p} (\gamma_{1j}{}^2 + ... + \gamma_{mj}{}^2)^{\frac{1}{2}} \tag{3.77}$$

where $\mathbf{u} = \mathbf{u} = \sqrt{\mathbf{u}^T \mathbf{u}}$ denotes the L_2 norm. Variable selection is enforced by the penalty, meaning that every parameter in $\boldsymbol{\gamma}_{.j}$ is concurrently decreased toward zero. It

© University of Venda

University of Venda
Creating Future Leaders

has a close connection to the group lasso. In contrast, grouping in the current model originates from the multivariate response structure, whereas in the group lasso grouping relates to the parameters associated with a categorical predictor inside a univariate regression model. Tutz et al. (2016) was the first to propose it for multinomial replies. To guarantee that the baseline hazard is sufficiently smooth across time, the penalty in discrete survival should be modified by a term. A penalty that smoothes the baseline dangers across time and enforces organized and efficient variable selection is provided by

$$J_{\zeta 1 \zeta 2}(\boldsymbol{\gamma}) = \zeta 1 \sum_{r=1}^{m} \sum_{t=2}^{q} (\gamma_{0tr} - \gamma_{0,t-1,r})^2 + \zeta 2 \sum_{j=1}^{p} \phi_j \boldsymbol{\gamma}_{.j},$$

where the penalty level for parameter vectors $\boldsymbol{\gamma}_{.j}$ is adjusted for their dimension by a weight $\phi_j = \sqrt{m}$. The tuning parameters $\zeta 1$ and $\zeta 2$ influence the significance of the penalty terms. Ordinary maximum likelihood estimation is achieved using $\zeta 1 = \zeta 2 = 0$, which is penalty-free. In general, it is not the best option to use the same penalty level $\zeta 2$ for every $\boldsymbol{\gamma}_{.j}$. Penalties are inconsistent if used with a common penalty parameter. The proposed remedy are so-called adaptive weights, which are obtained by replacing the weights $\phi$by

$$\phi_j{}^a = \frac{\sqrt{m}}{\hat{\boldsymbol{\gamma}}_{.j}{}^{Init}}, \tag{3.78}$$

When a good starting estimator is indicated by $\hat{\boldsymbol{\gamma}}_{.j}{}^{Init}$ respectively. For instance, the initial estimator can be the penalized estimator that emerges from applying a penalty with $\zeta 2 = 0$. Here, the first estimator employs an active smoothing penalty on the baseline effects but unpenalized covariate effects. It is necessary to select the tuning parameters oneself, for instance, using cross-validation.

© University of Venda

University of Venda
Creating Future Leaders

3.7 Machine Learning Methods

Machine learning works well when there are a high number of examples in a decent dimensional feature space, however this is not true for several survival analysis issues. Machine learning approaches have the benefit of being able to model nonlinear interactions as well as the quality of their overall predictions.

3.8 Recursive partitioning in discrete competing risks

Competing risks occur when an individual is subject to multiple possible failure events, where the occurrence of one event prevents the occurrence of others (Fine and Gray, 1999). Traditional statistical methods such as cause-specific hazard models or subdistribution hazard models have been used to analyze competing risks, but these approaches rely on strong assumptions about the hazard functions. Recursive partitioning, particularly classification trees, offers a nonparametric alternative for handling competing risks in discrete time settings.

Decision trees recursively split the feature space at different thresholds. Each split defines a region where a simple rule applies. These regions can be highly irregular and adaptive, which allows for modeling nonlinear patterns effectively.

Unlike linear models, tree-based methods do not assume a linear relationship between predictors and the outcome, making them ideal for educational or behavioral datasets where interactions are often complex and non-additive. According to Friedman (2009) , tree-based models excel at uncovering structure in high-dimensional data through recursive partitioning, which naturally incorporates variable interactions and nonlinearities. Studies like Couronné et al. (2018) have shown that tree-based models often outperform linear models, especially in datasets where nonlinear and interactive effects dominate. Moreover, Breiman (2001) highlights that random forests, in particular, provide robustness against overfitting and can handle large numbers of input

© University of Venda

University of Venda
Creating Future Leaders

variables without variable deletion, while still capturing intricate interaction effects. These characteristics make tree-based methods a powerful choice for modeling dropout and graduation events, where the underlying mechanisms are likely influenced by a complex interplay of demographic, academic, and institutional factors.

3.8.1 Classification Trees for discrete competing risks

The tree-based strategy by (Schmid et al., 2016), developed for time-to-event models with a single kind of event ($J = 1$), involves fitting a discrete hazard model of the form

$$\lambda(t|x_i) = f(t, \mathbf{x}_i),$$

, where the function $f(\cdot)$ is represented by a classification tree with binary outcome. The explanatory variables $x_1, \ldots, x_p$ and the time t (expressed by an ordinal variable) are taken into consideration as splitting possibilities for the tree's construction. (Schmid et al., 2016) suggested using the CART method with minimal node size pruning, which is based on the Gini impurity measure.This approach produces a collection of terminal nodes, each associated with a set of binary outcome values, which serve as the basis for estimating the hazard function. Constructing a tree involves iteratively partitioning the predictor space defined by the explanatory variables. The root of the tree represents the top node, encompassing the entire predictor space, while the terminal nodes correspond to a disjoint partition of this space.

When constructing a tree, each node M is iteratively divided into two subsets, $\mathbf{M_1}$ and $\mathbf{M_2}$, Berger et al. (2019). In each step, an explanatory variable x_s or the time t is chosen along with a corresponding splitting rule. The splitting rule is applied based on the type of variable. For a metric or ordinal variable x_s, the partitioning follows the form: $M_1 = M \cap \{x_s \leq c\}, \quad M_2 = M \cap \{x_s > c\}$

where c represents the splitting point. For a categorical variable without an inherent order, where $x_s \in \{1, \ldots, r\}$, the partitioning is defined as:

$M_1 = M \cap C_1, \quad M_2 = M \cap C_2$

© University of Venda

University of Venda
Creating Future Leaders

where C_1 and C_2 are disjoint, non-empty subsets such that $C_1 \subset \{1, \ldots, r\}$ and $C_2 = \{1, \ldots, r\} \setminus C_1$.

In the case of competing events ($J > 1$), (Berger et al., 2019) propose an extension of the survival tree method by (Schmid et al., 2016) by introducing a cause-specific hazard model given by

$$\lambda_j(t \mid x_i) = f_j(t, \mathbf{x}_i),$$

where the functions $f_j(\cdot)$ are represented by a classification tree with a multi-categorical outcome.

Similar to the single-event scenario, the fundamental components of the tree-based method proposed by (Berger et al., 2019), which will be described in the following sections, include: (i) defining the data structure, (ii) selecting an appropriate splitting criterion, (iii) determining suitable tuning parameters for splitting and pruning the constructed tree, and (iv) estimating the cause-specific hazard functions.

Choice of the splitting criterion

For constructing a classification tree with a multi-categorical outcome variable, widely used splitting strategies rely on impurity measures, see the Gini impurity by (Breiman, 2017). Given a multi-categorical outcome variable $\mathbf{y}_i$, the Gini impurity measure for a node M in a constructed tree is defined as:

$$\mathrm{GI}_M = \sum_{j=0}^{J} \sum_{k \neq j}^{J} \pi_j(M)\pi_k(M), \tag{3.79}$$

where $\pi_j(M)$ represents the proportion of observations with outcome category j in node M (Breiman, 1996). At each step of the tree-growing process, the algorithm selects the split (among all explanatory variables and corresponding splitting rules) that minimizes the pooled Gini impurity:

© University of Venda

University of Venda
Creating Future Leaders

$$\mathrm{GI}(M_1, M_2) = |M_1|\mathrm{GIM}_1 + |M_2|\mathrm{GIM}_2, \tag{3.80}$$

where $|\cdot|$ denotes the number of observations in the node. The proportions of outcome values in each node M provide estimates for the cause-specific hazards, given by:

$$\hat{\lambda}_1(M) = \pi_1(M), \quad \hat{\lambda}_2(M) = \pi_2(M), \quad \ldots, \quad \hat{\lambda}_J(M) = \pi_J(M),$$

for the subset of individuals within a time interval defined by the node. The overall hazard estimate is determined by the proportion of non-zero values, expressed as:

$$\hat{\lambda}(M) = 1 - \pi_0(M).$$

An alternative criterion for splitting in decision trees is the Hellinger distance(HD), which is utilized in the Hellinger Distance Decision Tree (HDDT) algorithm. This algorithm was initially proposed by Cieslak et al. for binary classification and further developed in subsequent work, focusing on tree-based methods tailored for imbalanced datasets where one class is significantly underrepresented (Cieslak and Chawla, 2008). The research demonstrated that HDDT surpasses traditional CART algorithms in terms of Area Under the Curve (AUC) performance, particularly in scenarios with severe class imbalance. For the proposed tree-based method, the Hellinger distance is an attractive option, given that the augmented data matrix used in tree construction contains a disproportionately large number of zero values. Given a binary outcome pair, such as $\{0, 1\}$, and a node M that is split into two subsets M_1 and M_2, the Hellinger distance is defined as:

$$HD(M_1, M_2) = \sqrt{\left(\sqrt{\pi_1(M_1)} - \sqrt{\pi_0(M_1)}\right)^2 + \left(\sqrt{\pi_1(M_2)} - \sqrt{\pi_0(M_2)}\right)^2},$$

where $\pi_0(\cdot)$ and $\pi_1(\cdot)$ represent the proportions of zeros and ones in the respective nodes.

To extend this concept to a multi-categorical outcome variable, consider pairs of outcome subsets $j_1 \subset \{0, \ldots, J\}$ and $j_2 = \{0, \ldots, J\} \setminus j_1$. The categories in j_1 are

© University of Venda

University of Venda
Creating Future Leaders

assigned a value of zero, while those in j_2 are assigned a value of one. During each step of tree construction, $HD(M_1, M_2)$ is computed for all possible pairs of subsets. The optimal split is then determined by selecting the one that minimizes the Hellinger distance across all pairs of outcome values, explanatory variables, and corresponding splitting rules.

For instance, imagine you're building a decision tree to predict whether students graduate on time or drop out (plus those still studying) based on early academic performance and background data. At any potential split, by "attendance rate", the HD computes the proportion of graduates and dropouts in each branch (left/right). Suppose after the split there is a High-attendance node with 95% graduate, 5% dropout and a Low-attendance node with 65% graduate, 35% dropout. The Hellinger distance formula calculates how different these two probability distributions are. A larger distance means the split creates child nodes with distinct outcome patterns. The algorithm chooses the split with the highest Hellinger distance, focusing on separating dropout from graduate probabilities.

Specification of the tuning parameters

When constructing trees, the primary tuning parameter is the number of splits, which controls the depth and thus the size of the tree. Various strategies exist to determine the optimal size of classification trees.In discrete time-to-event analysis, pruning is crucial to ensure reasonable estimates of the cause-specific hazards $\hat{\lambda}_1, \ldots, \hat{\lambda}_J$ (discussed in detail in the next subsection), as the variance of these estimators is inversely related to the size of the terminal nodes. Therefore, the accuracy of these estimates is highly dependent on the number of observations in the terminal nodes.

In conventional methods, tuning is typically achieved through cost-complexity pruning. This process starts at the terminal nodes of a fully grown tree and iteratively collapses nodes that result in the least reduction in classification accuracy. The out-

© University of Venda

University of Venda
Creating Future Leaders

come is a series of nested subtrees, each of which minimizes a cost-complexity criterion among all subtrees of the same size. The optimal tree in this sequence is the one that achieves the lowest value of the cost-complexity criterion, effectively balancing the trade-off between classification accuracy and tree size Mingers (1989). However, numerous studies have indicated that this pruning strategy is not optimal for estimating probabilities from terminal nodes, as trees optimized for classification accuracy often end up being too small Provost and Domingos (2003).

To determine the tree with the optimal minimal node size, we propose using the log-likelihood of the model. One approach is to minimize an information criterion (such as AIC or BIC) or maximize the predictive log-likelihood. In a manner similar to (Schmid et al., 2016), we define the information criteria as $-2\ell + \xi(Q-1)$, where ℓ represents the log-likelihood, Q is the number of terminal nodes (which serves as a measure of the tree's complexity), and $\xi \in \{2, \log(\tilde{n})\}$. When the predictive log-likelihood is used for tuning, the algorithm performs five-fold cross-validation using subsamples without replacement, each of size 80% (drawn from the original non-augmented data.

To ensure that each sub-sample includes a sufficient number of observations for each observed event time, the sub-samples are stratified by the variable $\tilde{T}_i$. The pruning strategy described aims to balance between creating trees that are too small, which may optimize classification accuracy but are less effective for probability estimation, and trees that are too large, where terminal nodes have too few observations, leading to cause-specific hazard estimates with high variance. This approach generates a sequence of nested trees, similar to cost-complexity pruning. However, unlike cost-complexity pruning, minimal node size pruning does not guarantee that the resulting tree will always be the optimal choice among trees of the same size for probability estimation purposes.

© University of Venda

University of Venda
Creating Future Leaders

3.9 Performance Measures

This section examines the evaluation of machine learning techniques and statistical models in survival analysis, focusing on their ability to predict outcomes. The assessment of model performance is based on both discrimination and calibration.

3.9.1 Discrimination Measures

To effectively compare different models, it is crucial to use evaluation metrics specifically suited for survival analysis. These include time-dependent area under the curve (AUC), the concordance index (C-index), and the log-rank test p-value.

Concordance Index

A popular metric for evaluating the prediction performance of statistical and machine learning models is the Concordance Index (C-index). It assesses how well the model ranks anticipated survival periods. Comparing the C-index readings for several models allows one to identify the model with the best predicted accuracy. Better discriminating capacity is indicated by higher C-index values, which imply that the model is more accurate at predicting each person's relative survival time.

The concordance index is a pivotal measure for assessing the discriminative ability of prognostic models in the presence of competing risks, as it evaluates the model's capacity to correctly rank individuals by their risk while rigorously handling censoring and multiple event types. Unlike simpler metrics, the C-index accommodates right-censored data by excluding or weighting ambiguous pairs, preserving the validity of the assessment. And when there are competing risks, it makes sure only comparisons that make sense(where one event happens before another, and of a specific type) are counted, so you don't get misleading scores from ignoring those other outcomes.

The C-index is defined as:

© University of Venda

University of Venda
Creating Future Leaders

$$C\text{-index} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} \cdot I(T_i < T_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}} \tag{3.81}$$

where:

- n is the total number of subjects in the dataset,
- C_{ij} represents the pairwise comparison of predicted survival outcomes or risk scores for individuals i and j, If $C_{ij} = 1$, it indicates that the predicted outcome for individual i is concordant with the actual outcome of individual j. If $C_{ij} = 0$, it means the outcomes are discordant or tied,
- $I(T_i < T_j)$ is an indicator function that returns 1 if the survival time of individual i is less than that of individual j, and 0 otherwise.
- T_i represents the survival time or time-to-event for individual i.

Time-dependent Area Under the Curve (AUC)

The time-dependent Area Under the Curve (AUC) is a performance metric that evaluates the discriminatory ability of survival models across different time points. It is commonly used to compare the ability of machine learning and statistical models to distinguish between survival outcomes at various times. The AUC provides insight into how well a model differentiates between individuals based on their survival probabilities.

The time-dependent AUC is defined as:

$$AUC(t) = \frac{[C_i(t) - D_i(t)]}{C_i(t) \cdot D_i(t) + 0.5 \cdot C_i(t) \cdot (C_i(t) - 1)} \tag{3.82}$$

where:

- $C_i(t)$ represents the number of correctly predicted pairs at time t,
- $D_i(t)$ represents the number of discordant pairs at time t.

© University of Venda

University of Venda
Creating Future Leaders

The numerator in this calculation determines the difference between concordant and discordant pairs at a specific time point. The denominator, on the other hand, represents the product of these concordant and discordant pairs, adjusted by a correction factor. The Area Under the Curve (AUC) ranges from 0.5, which signifies no discriminatory power, to 1, which indicates perfect discrimination. A higher AUC value implies that the model is more effective at distinguishing between different survival times at a given time point.

3.9.2 Calibration Measures

When comparing statistical models with machine learning algorithms, calibration models provide a useful approach to assess how well predicted probabilities correspond to observed outcomes. Commonly used calibration methods include the Reality Plot, Calibration Curve, Brier Score, and the Hosmer-Lemeshow test.

Calibration Curve

Like a reliability plot, a calibration curve depicts the average expected probability for various quantiles instead of using predetermined intervals to show the link between anticipated probabilities and actual outcomes. An ideal calibration curve closely follows the diagonal line, indicating well-calibrated predictions.

Brier Score

One of the most popular metrics for assessing probabilistic predictions in a variety of fields, such as survival analysis, statistical modeling, and forecasting, is the Brier Score. It provides a measure of predictive accuracy by quantifying the discrepancy between expected probabilities and actual outcomes. The most popular score function is probably the Brier score. Beyond evaluating whether the model correctly predicts that one observation is more likely to be a "success" than another, the Brier Score also evaluates the accuracy of the model's predicted probabilities. The Brier Score is

© University of Venda

University of Venda
Creating Future Leaders

a metric designed to assess the quality of predicted probability scores. Similar to the mean squared error, it specifically applies to probability predictions that range from 0 to 1, where a score of 1.0 indicates the worst possible model and a score of 0.0 indicates the best.

The Brier Score $BS(t^*)$ at a specific time t^* is defined as:

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t^*) - \hat{\pi}(t^*|\hat{X}_i)\right)^2 \tag{3.83}$$

where $BS(t^*)$ represents the Brier Score at time t^*, quantifying the squared deviation between the predicted probability and the observed outcome. Here, n denotes the total number of subjects or observations, $\sum_{i=1}^{n}$ represents the summation over all observations, and $I(T_i > t^*)$ is an indicator function that equals 1 if the event for the i-th subject occurs after t^*, and 0 otherwise. Additionally, $\hat{\pi}(t^*|\hat{X}_i)$ is the predicted probability of the event occurring at time t^* for the i-th subject, given the estimated predictor variables $\hat{X}_i$.

Integrated Brier Score

To evaluate the overall performance of a model across all time points $t_1 \leq t \leq t_{\max}$, the Integrated Brier Score (IBS) is used. It is defined as:

$$IBS = \int_{t_1}^{t_{\max}} BS^c(t)dw(t) \tag{3.84}$$

where $BS_c(t)$ represents the time-dependent Brier Score, and the weighting function is given by $w(t) = \frac{t}{t_{\max}}$. The IBS provides a summary measure of predictive performance over the specified time interval $[t_1, t_{\max}]$.

© University of Venda

University of Venda
Creating Future Leaders

3.10 Procedure about data processing and manipulation

3.10.1 Introduction

This section provides an explanation of how the student enrollment, graduation, and subject data was cleaned, filtered, merged, and processed in order to account for a competing risks survival analysis.

We initially had four different subsets, namely,Subjects, Residence, Enrolment and Graduation list. Subjects dataset had matric subjects for each student, Residence only had information about on campus residence for students who resided at campus during their studies. Enrolment data showed academic information like qualification name, year of enrolment, formal time of a course and any information related to the course for each student at each year of registration. Graduation list only contained graduation information about students who had graduated between the years 2014 and 2021. There were information like the course name which the person had graduated from, decision date of the course completion, graduation date and the Year of graduation.

3.10.2 Enrolment dataset

From this dataset, we only extracted students who started their university studies between 2014 and 2021, using the First Time Entering status column that was analysed for each year at each student. This is because we only wanted a cohort of 6 years between 2014 and 2021. Students who were doing a course that exceeded 3 years were removed from this subset, including three programmes that had an extension. Every student who first enrolled to a foundation programme before going to the respective three year programmes were removed as well, this is because their formal time for the course would have ended up summing to four years where as we only want three years. We also removed rows where "ACADEMIC BLOCK" is "FOR GRADUATION

© University of Venda

University of Venda
Creating Future Leaders

ONLY",since academic block shows the reason for registration for that academic year, ensuring that only active enrollment records are kept. For graduation only is for a case where a students just registers so that they can get their examination results (especially for those writing special exams). This was done so that the time that these students took to graduate does not add an unnecessary year. For students who did two or more courses between 2014 and 2021, we only kept the years were they were doing the first course. The faculties were three in total, since the health sciences faculty had no courses which were offered for a period of three years.

3.10.3 Other datasets

Before merging the graduation list with the enrolment dataset, we first made sure that the course name on the enrolment were matching with the course name on the graduation list, this means that if the course that the person graduated for on the graduation list does not match with the one from the enrolment dataset, we would remove it(since it could be the second or third course which the student had registered and we are only dealing with the first course registered for in our enrolment dataset). After merging those datasets, we then calculated the duration of studying for each student as follows: $Duration = "Lastacademicyear" - "Firstyearofregistration" + 1$. Now that we had graduation information merged to the enrolment dataset, we were now able to compute the exiting status for each student. A student was known to have graduated if he or she had a Decision Year, dropped out if there was no Decision Year on his or her decision year column but last academic year is before 2021 and still enrolled if there is no Decision Year, but last academic year is 2021. It was at this stage that we could now check the time it took for a student to finally graduate. Because of some missing information on this dataset, there were cases were a student would have graduated within one or two years, so we removed such students as it did not make sense to graduate from a three year course in two years. The subjects dataset was used to compute the average matric mark for each student, but before doing that we

© University of Venda

University of Venda
Creating Future Leaders

first had to take the mean mark for those subjects that were duplicated. The columns, average mark, English mark and Math mark for each student were then merged into the enrolment dataset. We applied mean imputation to handle missing values in the average mark, English mark, and Math mark columns.This approach helped preserve a large number of observations in the model. Finally students who exceeded 6 years were censored at 6 years and considered still enrolled even if they ended up being graduated or dropped out at the 7th or 8th year.

3.11 Model fitting and R packages

Competing risks models require data to be on long format in order to account for discrete survival analysis and right censoring. The "Data Long" format has been extended to encompass all time periods up until each individual's most recent observation. Data used in this study was already on a long format as every students' information was available from first year of registration until the exiting time or the last observation time. To transform survival data containing competing events to a multinomial responses, one uses the "dataLongcompRisk" function in the R package "discSurv" (Welchowski and Schmid, 2015).

To fit the discrete survival models for competing risks in R, we used the vglm function for the VGAM package. For the evaluation of the model, we used the Concordance Index and the Brier score. The **cIndexCompRisks** function which is part of the discSurv package in R was used to compute the C-Index. The reason for choosing the discSurv package is that it is designed for discrete-time models. It builds on discrete-time event probabilities, integrating them into concordance calculations. This leverages the competing-risks structure directly, leading to more accurate evaluation.

The proposed tree-based method is implemented in the R program that is available from GitHub(https://github.com/jmober/CompetingRisksTreeDiscSurvival.git) with help

© University of Venda

University of Venda
Creating Future Leaders

from the paper by Berger et al. (2019). The package used was the "CRTree". To improve tree size, we suggest utilizing the minimum number of observations in nodes (minimalnodesize) as the key pruning option during tree creation. We suggest using log-likelihood of the model to get the optimum tree with smallest node size. One can minimize an information criteria (e.g., AIC and BIC), we used the BIC information criteria and the Gini impurity for the splitting criteria.

Based on the BIC, the model will choose the optimal tree. By penalizing model complexity, BIC promotes smaller trees with superior generalizability, hence preventing overfitting. It computes the BIC for each tree. The best model is determined by choosing the tree with the lowest BIC. A range for the bare minimum of observations per split (node size) is defined by minimal node size. Using this range, the model will assess several tree sizes and select the optimal one using the BIC criteria. The splitting criteria is the Gini impurity. Although this dictates the tree's growth, the BIC tuning actually directs the choice of the finished tree.

© University of Venda

University of Venda
Creating Future Leaders

Chapter 4

Data analysis and Results

4.1 Introduction

This chapter provides an in-depth examination of the study's data and its key findings. It commences with an overview of the data utilized, which is then followed by a detailed presentation of the results aligned with the study's primary objectives. The overarching goal of the research was to investigate the temporal dynamics of student progression using discrete-time survival analysis techniques.

To analyse the incident of dropout. Baseline survival functions are reported in the later subsections. The number of dropouts in each year is also reported

The study reports the estimated coefficients, which represent the log odds of dropout and graduation, along with their standard errors and p-values, for both single and competing risks models. The primary goal is to compare the dropout risk profiles across different student groups. Initially, the risk of dropout is visually compared using the baseline hazard function to provide a graphical overview of how dropout risks vary among these groups. We used the baseline hazard of the simple discrete survival model and the time points 1 and 2 only in order to check the effect of time on dropout without checking for graduation simultaneously.

© University of Venda

University of Venda
Creating Future Leaders

4.2 Data description and exploratory data analysis

4.2.1 Data description

The study utilized data sourced from the University of Venda's Integrated Tertiary System (ITS). The dataset was filtered to include only students enrolled in three-year undergraduate programs during the period from 2014 to 2021. For each student, we only considered the first course which the student was registered for (the student will be a dropout of the first course, if they jumped into the second one before completing the first one).

The data covers students enrolled for the first time between the years 2014 and 2021. Data was analysed using R statistical analysis software . The cohort was followed for 8 years from 2014 through 2021, inclusive. However, students who took seven years and beyond to their exit status were censored at year six(they were treated as students who were still enrolled at the sixth year). At the end of the observation time, a student would either have dropped out, graduated or still be enrolled for the programme.

4.2.2 Exploratory data analysis

The marital status variable was collapsed into two categories(Single and been married) due to the small numbers of those who were married, divorced or widowed. The age at the first year of registration for each student was categorised into 4 groups, i.e, less than 20, 20-24, 25-29 and 30+. The faculty variable had three categories, namely, faculty of management, commerce and law(FMCL), faculty of science, engineering and agriculture(FSEA) and the faculty of human and social sciences(FHHS). The average mark was also categorised into 4 groups, i.e, less than 50, 50-59, 60-69 and 70+. The categorisation was necessary for these variables in order to accommodate the decision tree in Figure 4.5 as it appeared to be too messy with continuous variables.

© University of Venda

University of Venda
Creating Future Leaders

Demographic and academic variables that are known to be associated with student academic progression that were available on the provided datasets were used as explanatory variables. Some potentially important explanatory variables , e.g., financial support and school quintile, could not be obtained from the database. APS, English and Maths score could not be used since there were a lot of missing values which could potentially reduce the sample size when fitting the models. The full list of the variables used in the study is presented in the table below:

Table 4.1: Variables used in the study.

Variable	**Description**	**Variable type**
Gender	Male	Categorical
	Female	
Faculty	FMCL- Faculty of Management, Commerce and Law	Categorical
	FSEA- Faculty of Science, Engineering and Agriculture	
	FHSS- Faculty of Human and Social Sciences	
Marital status	Single	Categorical
	Been married	
Average mark	Used as a measure of matric performance	Categorical
Age	Age between 16 and 51	Categorical

Frequency tables and summary statistics

Table 4.2 shows the descriptive statistics of some of the covariates included in the study. The number of students used was 7131, of which 51.7% were females and 48.3.7% were males. Females constituted a higher percentage than males. While the majority (93.8%) of individuals are single, about 5.93% of individuals have missing or unknown marital status and only a very small fraction (0.27%) of individuals have been married before or still are. The faculty distribution shows that Engineering and Applied Sciences (FSEA) is the most popular faculty,(representing 39.2% of the population), followed

© University of Venda

University of Venda
Creating Future Leaders

Table 4.2: Frequency distribution of qualitative explanatory variables

Category	Subcategory	Frequency	Percent	Cumulative Percent
Gender	F	3690	51.7	51.7
	M	3441	48.3	100.0
Marital Status	Single	6691	93.8	93.8
	NA	423	5.93	99.8
	Been Married	17	0.27	100.0
Age Group	15-19	3542	49.7	49.7
	20-24	3360	47.1	96.8
	25-29	176	2.47	99.3
	30+	53	0.743	100.0
Average Mark Group	<50	217	3.04	3.04
	50-59	1123	15.7	18.74
	60-69	5402	75.8	94.54
	70+	389	5.46	100.0
Faculty	FSEA	2796	39.2	39.2
	FMCL	2386	33.5	72.7
	HSS	1949	27.3	100.0

by Management and Law(33.5%). Humanities and Social Sciences (HSS) make up the smallest segment(27.3%). Most students fall into the 60-69% average mark range (75.8%), with only 3.04% scoring below 50% and 5.46% achieving 70% or higher. The majority of students in the dataset are between 15-19 years old (49.7%) and 20-24 years old (47.1%). Only 2.47% of students are between 25-29 years old, and an even smaller percentage (0.74%) are 30 years or older. Given this distribution, the impact of age on graduation or dropout rates may be more pronounced for younger students, as they make up the vast majority of the sample. The small sample size in the 30+ age group

© University of Venda

University of Venda
Creating Future Leaders

may lead to less precise estimates for this category.

Table 4.3: Summary Statistics of `Age at first year`

Statistic	Minimum	1st Quarter.	Median	Mean	3rd Quarter.	Maximum
Value	16.00	19.00	20.00	20.01	21.00	51.00

Table 4.3 shows the summary statistics of students' age. The ages of students ranged from 16 to 51 years, with a mean age of 20.01 and a median of 20, this suggests a fairly symmetric distribution. The median age is 20 years, indicating that half of the individuals are younger than or equal to 20 years, and the other half are older. The third quartile is 21. This means 75% of the individuals are aged 21 or younger, while 25% are older than 21 years.. However, the presence of older individuals, with ages in the 50s, suggests some diversity, potentially representing adult learners. Overall, the dataset captures a primarily young demographic with a small proportion of older participants.

Violin plot

A violin plot is a combination of a boxplot (represented by the black rectangle and white dot) and a kernel density plot (the blue area). It shows the central tendency (median and quartiles) and the distribution shape (density) of the data simultaneously. The violin plot 4.1 below shows that most individuals are aged around 19–21 years, (the widest section of the plot). The density decreases sharply for ages beyond 21, showing fewer older students. The black box in the center represents the interquartile range (IQR), with the white dot inside indicating the median age (20). The distribution appears unimodal and symmetric, aligning with the summary statistics. This makes it a useful visualization for understanding the spread and shape of age data in the dataset.

© University of Venda

University of Venda
Creating Future Leaders

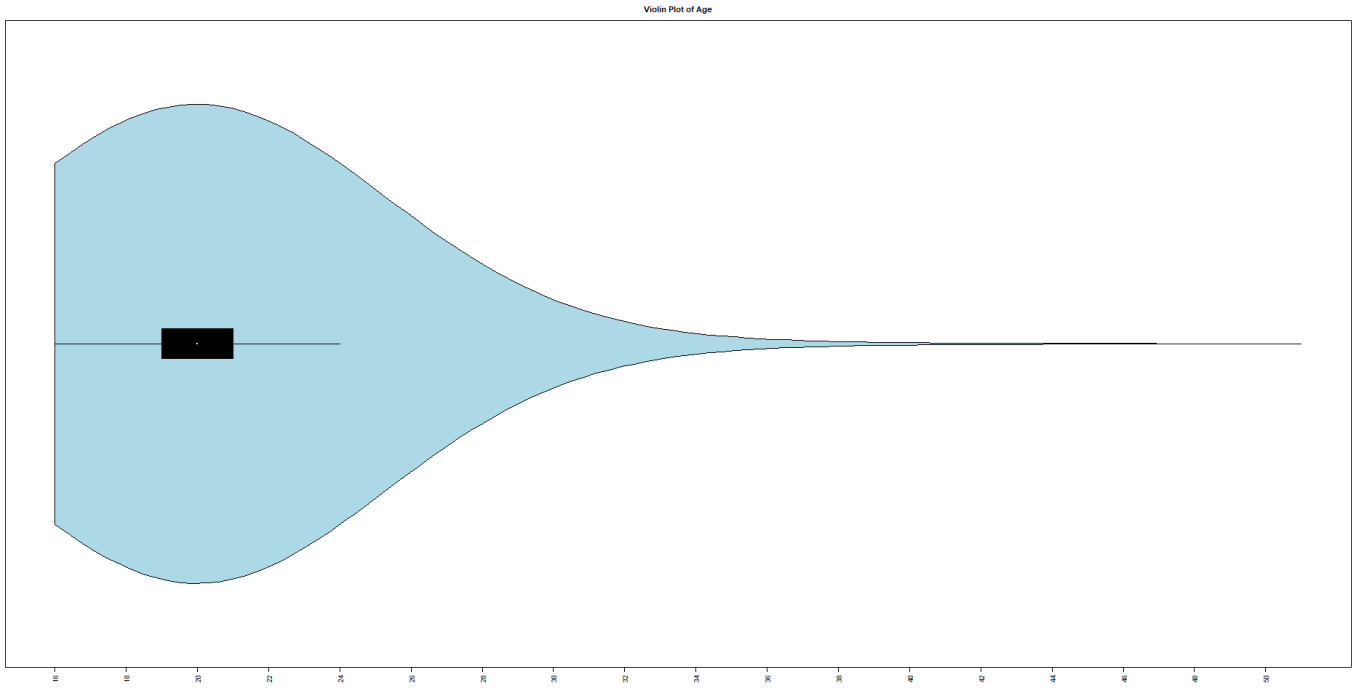


Figure 4.1: Violin plot of Age at first year

4.3 Single risk model

In this section, a single risk discrete-time hazard model is applied using a logit link function. This approach enables the assessment of how various covariates influence student dropout, facilitating the prediction of dropout risk based on a specified set of explanatory variables. The model was fitted using a binomial logit link function to assess the relationship between student characteristics (age, gender, faculty, and matric academic performance) and the dependent variable(dropout). The odds ratios and their 95% confidence intervals were computed to evaluate the effect sizes. Statistical significance was assessed using p-values, with a threshold of $p < 0.05$. This model was applied to the first two years after registration since there is no possibility of graduation in those two years.

© University of Venda

University of Venda
Creating Future Leaders

4.3.1 Baseline hazard

The baseline hazard graph 4.2 shows the effect of time on dropout. It is evident that the risk of dropping out was the highest at the first year of registration and decreased drastically at the second year.

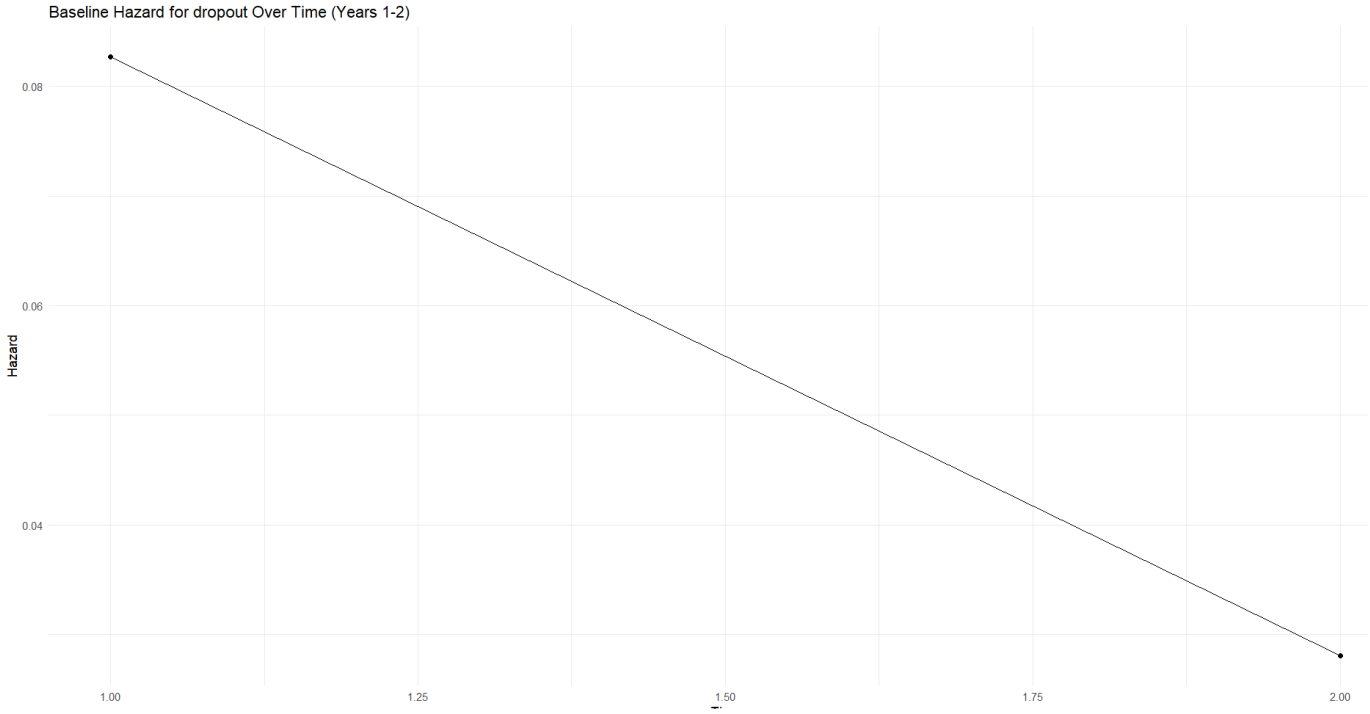


Figure 4.2: Baseline hazard for dropout

Table 4.4 shows the estimates of the coefficients, the corresponding estimated standard errors (SE), Odds ratios and their 95% lower and upper confidence intervals and the p-values of the covariate effects obtained from the single risk model

4.3.2 Effect of Age

Keeping gender, average mark, and faculty constant, it is observable from 4.4 that older students had significantly higher odds of dropout (OR = 1.44, 95% CI: 1.10–1.89, p = 0.008). We are 95% confident that the odds of dropping out for this group are between 14.3% and 81.8% higher than those under 20. The other groups do not differ

© University of Venda

University of Venda
Creating Future Leaders

Table 4.4: Logistic Regression Results

Variable	Reference	Estimate	SE	OR	Lower CI	Upper CI	p-value
Age	< 20						
20-24	-	0.366	0.118	1.442	1.143	1.818	0.002
25-29	-	0.545	0.307	1.725	0.944	3.151	0.076
30+	-	0.858	0.480	2.357	0.920	6.036	0.074
Average Mark	< 50						
50-59	-	-0.138	0.348	0.871	0.440	1.724	0.692
60-69	-	0.122	0.318	1.130	0.606	2.106	0.701
70+	-	-0.333	0.453	0.716	0.295	1.742	0.462
Faculty	FSEA						
HSS	-	-0.303	0.143	0.739	0.558	0.978	0.034
FMCL	-	0.107	0.123	1.113	0.875	1.415	0.383
Gender	Female						
Male	-	0.678	0.117	1.970	1.567	2.475	0.001

significantly from the reference category in terms of the odds of dropout. These findings suggest that while students in the 20-24 age group are significantly more likely to drop out than younger students, the effects for older age groups (25-29 and 30+) are not conclusive due to the wide confidence intervals and lack of statistical significance. The confidence intervals for both age groups were relatively wide, indicating a level of uncertainty around these estimates.

4.3.3 Effect of Average Mark

The results on table 4.4 suggest that average mark does not have a statistically significant effect on dropout odds in this sample.

© University of Venda

University of Venda
Creating Future Leaders

4.3.4 Effect of Faculty

Controlling for gender,age and faculty, students in the Faculty of Humanities and Social Sciences (HSS) have 26.1% lower odds of dropping out compared to those in the Faculty of Science, Engineering, and Agriculture (FSEA). This suggests that students in the HSS faculty are less likely to dropout as compared to students in the Faculty of Science, Engineering and Agriculture. We can conclude with 95% confidence that the odds of dropout for students in HSS are between 2.2% and 44.2% lower than those in FSEA.

4.3.5 Effect of Gender

Keeping age, average mark, and faculty constant, the odds of dropping out for male students are 97.0% higher compared to female students. This effect is highly significant ($p < 0.001$), and we can conclude with 95% confidence that the odds of dropping out for male students are between 56.7% and 147.5% higher than for female students. This finding suggests that male students are significantly more likely to drop out than female students, even after controlling for other factors such as age, academic performance, and faculty affiliation.

4.3.6 Conclusion

The analysis of the factors influencing dropout odds among students revealed the following significant findings: Students aged between 20 and 24 years have higher odds of dropping out compared to those under 20 years. However, no significant effect was observed for students in the older age groups (25-29 and 30+ years), as the wide confidence intervals for these groups indicate inconclusive results.The analysis suggests that average mark does not have a statistically significant impact on the odds of dropout in this sample. Therefore, we cannot conclude that academic performance plays a major role in predicting dropout odds in this context. Students in the Faculty of Humanities and Social Sciences (HSS) are found to have lower odds of dropping out compared to

© University of Venda

University of Venda
Creating Future Leaders

students in the Faculty of Science, Engineering, and Agriculture (FSEA). This suggests that students in the HSS faculty are significantly less likely to drop out. The analysis indicates that male students are more likely to drop out than female students.

4.4 Discrete time cause specific for competing risks

In this section, dropout and graduation are jointly analyzed to account for the potential interdependence between these two outcomes. The model is applied to the third, fourth, fifth and sixth years, as graduation only occurs starting from the third year.

4.4.1 Baseline hazard

The baseline hazard Graph 4.3 shows the effect of time on graduation and dropout.The rate of dropout of graduation is higher than that of dropout and it kept increasing with time. Although the rate of dropout was lower than that of graduation, it was also increasing with time but at a slower rate.

© University of Venda

University of Venda
Creating Future Leaders

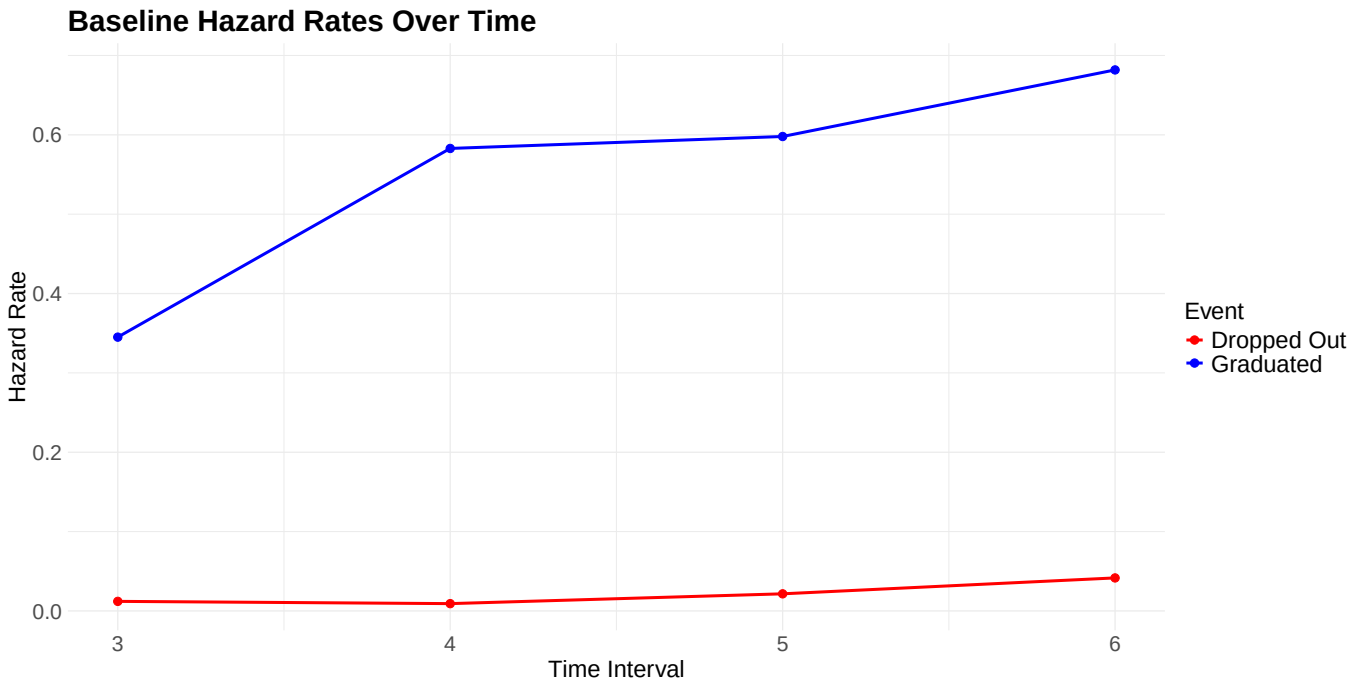


Figure 4.3: Baseline hazard for graduation and dropout

Table 4.5 shows the estimates of the coefficients, the corresponding estimated standard errors (S.E), Odds ratios and their 95% lower and upper confidence intervals and the p-values of the covariate effects obtained from the discrete cause-specific model for competing risks.

© University of Venda

University of Venda
Creating Future Leaders

4.4.2 Effect of Gender

Keeping age, average mark, and faculty constant, being male increases the odds of dropping out by 87.4% compared to females. We can conclude with 95% confidence that the odds of dropping out for males are between 40.5% and 149.8% higher than for females. Conversely, being male decreases the odds of graduating by 29.1% compared to females. We can conclude with 95% confidence that the odds of graduating for male students are between 20.5% and 36.7% lower than those for female students. This means that male students are more likely to dropout and less likely to graduate as compared to female students.

4.4.3 Effect of Age

Keeping gender, average mark, and faculty constant, students aged between 20 and 24 years have 50.5% higher odds of dropping out compared to those under 20. We are therefore 95% confident that the odds of dropping out for this group are between 12.5% and 101.3% higher than those under 20. However, the effect on graduation is not statistically significant. Students aged 25 to 29 years and 30 years plus do not have a significant effect on either graduation or dropout compared to those under 20.

4.4.4 Effect of Faculty

The odds of students in the Faculty of Humanities and Social Sciences (HSS) dropping out are 34.4% lower than those in the Faculty of Science, Engineering, and Agriculture (FSEA), meaning that HSS students are less likely to dropout as compared to FSEA students. We conclude with 95% confidence that, keeping gender, age, and average mark constant, the odds of dropping out for HSS students are between 6.9% and 53.7% lower than those of FSEA students. However, HSS students do not significantly differ from FSEA students in graduation odds. The Faculty of Commerce, Management, and Law (FMCL) does not significantly differ from FSEA in either dropout or graduation.

© University of Venda

University of Venda
Creating Future Leaders

4.4.5 Effect of Average Mark

The effect of average mark on graduation is not significant for all categories of students except the 70 plus group. After controlling for age, gender, and faculty, students with an average mark of 70+ have 48.2% higher odds of graduating compared to those with an average below 50. We conclude with 95% confidence that the odds of graduation for students in this category are between 0.9% and 121.5% higher than those with an average below 50. The effect of average mark on dropout is not statistically significant for any mark group.

4.4.6 Conclusion

Male students have significantly higher odds of dropping out compared to female students, to be specific, being male increases the odds of dropping out by 87.4% compared to females. Additionally, male students are less likely to graduate (being male decreases the odds of graduating by 29.1% compared to females), indicating that male students are at a higher risk of both dropping out and not graduating compared to female students, even after controlling for other factors.

Students aged between 20 and 24 years have significantly higher odds of dropping out compared to students under 20. However, age does not have a statistically significant effect on graduation for this group. Students aged 25-29 years and 30 years and older do not show significant differences in either dropout or graduation rates compared to those under 20.

Students in the Faculty of Humanities and Social Sciences (HSS) are significantly less likely to drop out, with 34.4% lower odds of dropping out compared to students in the Faculty of Science, Engineering, and Agriculture (FSEA). However, HSS students do not show significant differences in graduation odds. Similarly, students in the Faculty of Commerce, Management, and Law (FMCL) do not significantly differ from FSEA students in terms of either dropout or graduation.

© University of Venda

University of Venda
Creating Future Leaders

Average mark has a significant impact on graduation odds, particularly for students with an average mark of 70+ are more likely to graduate compared to students with an average below 50. However, average mark does not significantly affect dropout odds for any group, indicating that academic performance does not appear to be a key determinant of dropout in this dataset.

In summary, gender, age (for 20-24 years), and faculty (HSS) are significant predictors of dropout risk, with male students, students aged 20-24, and those in FSEA being at higher risk. Average mark plays a significant role in graduation, especially for students with higher academic performance (70+).

© University of Venda

University of Venda
Creating Future Leaders

Variable	Ref Category	Estimate	S.E	OR	Lower CI	Upper CI	p-value
(Intercept):1	Still Enrolled	-3.976	0.491	0.019	0.007	0.049	< 0.001
(Intercept):2	Still Enrolled	-0.622	0.170	0.537	0.385	0.749	< 0.001
Gender	Female						
Male:1	-	0.628	0.147	1.874	1.405	2.498	< 0.001
Male:2	-	-0.344	0.058	0.709	0.633	0.795	< 0.001
Age Group	< 20						
20-24:1	-	0.409	0.148	1.505	1.125	2.013	0.006
20-24:2	-	-0.092	0.061	0.912	0.809	1.028	0.131
25-29:1	-	0.355	0.421	1.427	0.625	3.258	0.399
25-29:2	-	0.043	0.191	1.044	0.718	1.517	0.822
30+:1	-	0.925	0.554	2.522	0.851	7.468	0.095
30+:2	-	-0.139	0.313	0.870	0.471	1.606	0.656
Faculty	FSEA						
HSS:1	-	-0.421	0.178	0.656	0.463	0.931	0.018
HSS:2	-	-0.101	0.073	0.904	0.784	1.042	0.165
FMCL:1	-	-0.046	0.153	0.955	0.708	1.289	0.766
FMCL:2	-	-0.121	0.067	0.886	0.776	1.011	0.072
Average mark	< 50						
(50-59):1	-	0.198	0.504	1.220	0.454	3.275	0.694
50-59:2	-	-0.147	0.175	0.863	0.613	1.217	0.402
60-69:1	-	0.665	0.468	1.944	0.777	4.865	0.156
60-69:2	-	0.040	0.162	1.041	0.758	1.429	0.805
70+:1	-	0.246	0.630	1.279	0.372	4.397	0.696
70+:2	-	0.393	0.205	1.482	0.991	2.215	<0.001

Table 4.5: Multinomial Logistic Regression Results

© University of Venda

University of Venda
Creating Future Leaders

4.5 Decision tree for competing risks

We finally fitted a recursive partitioning model. When constructing trees, an appropriate criterion for splitting must be selected. One of the earliest approaches in tree construction involves using impurity measures to determine splits, we used the Gini impurity while building the tree. To determine the ideal tree size, one must select an appropriate tuning parameter that regulates its complexity. In the case of a discrete survival tree, this parameter is the minimum number of observations required in each terminal node (minimal node size). Optimizing this value helps prevent overfitting by ensuring that the number of terminal nodes does not become excessively large while also avoiding nodes that are too small (Berger and Schmid, 2018). Splitting stops when any further division would create a node with fewer observations than the specified minimal node size. A sequence of tree models is generated based on different minimal node sizes, and the optimal tree—defined by the best minimal node size is selected using an information criterion, such as AIC or BIC which in our model we have used the BIC.

© University of Venda

University of Venda
Creating Future Leaders

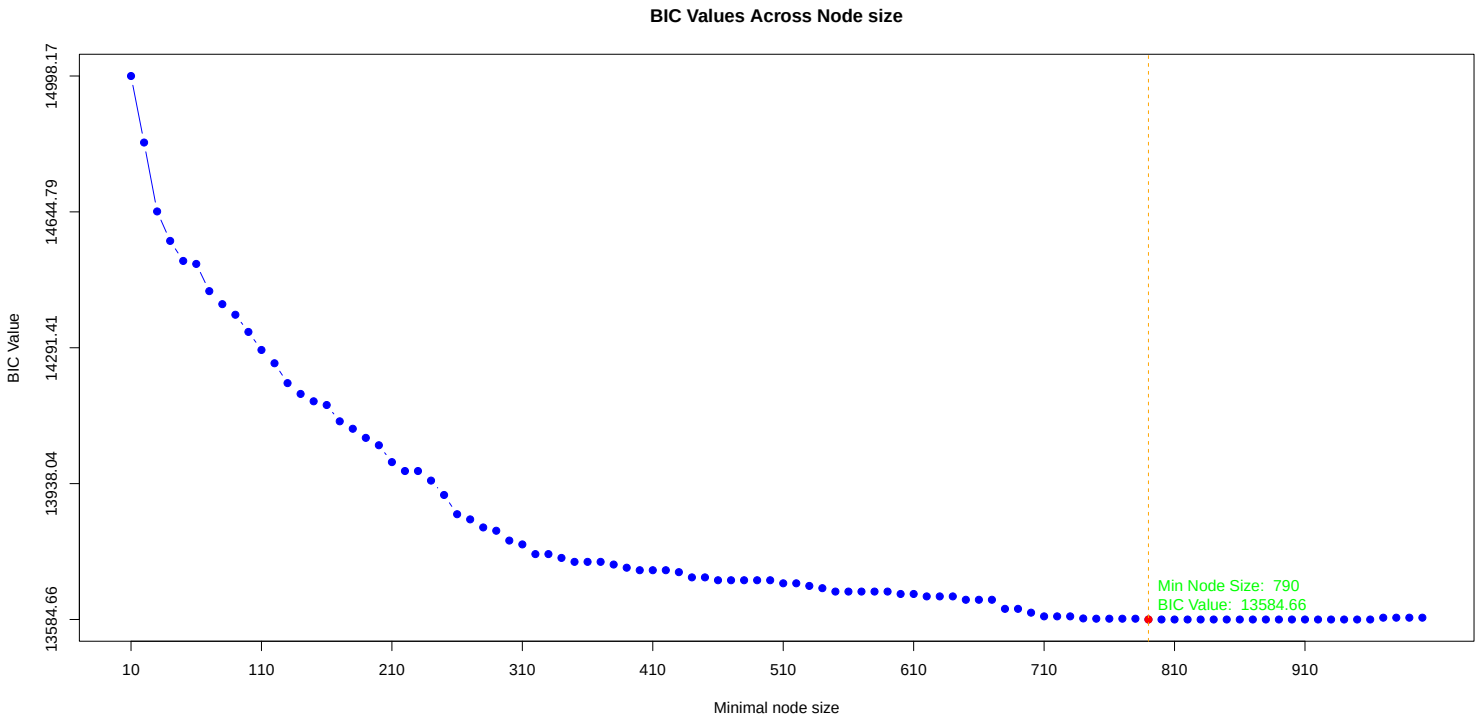


Figure 4.4: BIC plot

Figure 4.4 is the plot that shows the BIC values for the sequence of trees that were obtained by fitting a model with minimal node sizes ranging from 10 to 1000. The minimal BIC value (obtained for node size 790) is marked by the vertical dashed line. The tuning parameter of the model was the BIC with the Gini impurity splitting criterion.

© University of Venda

University of Venda
Creating Future Leaders

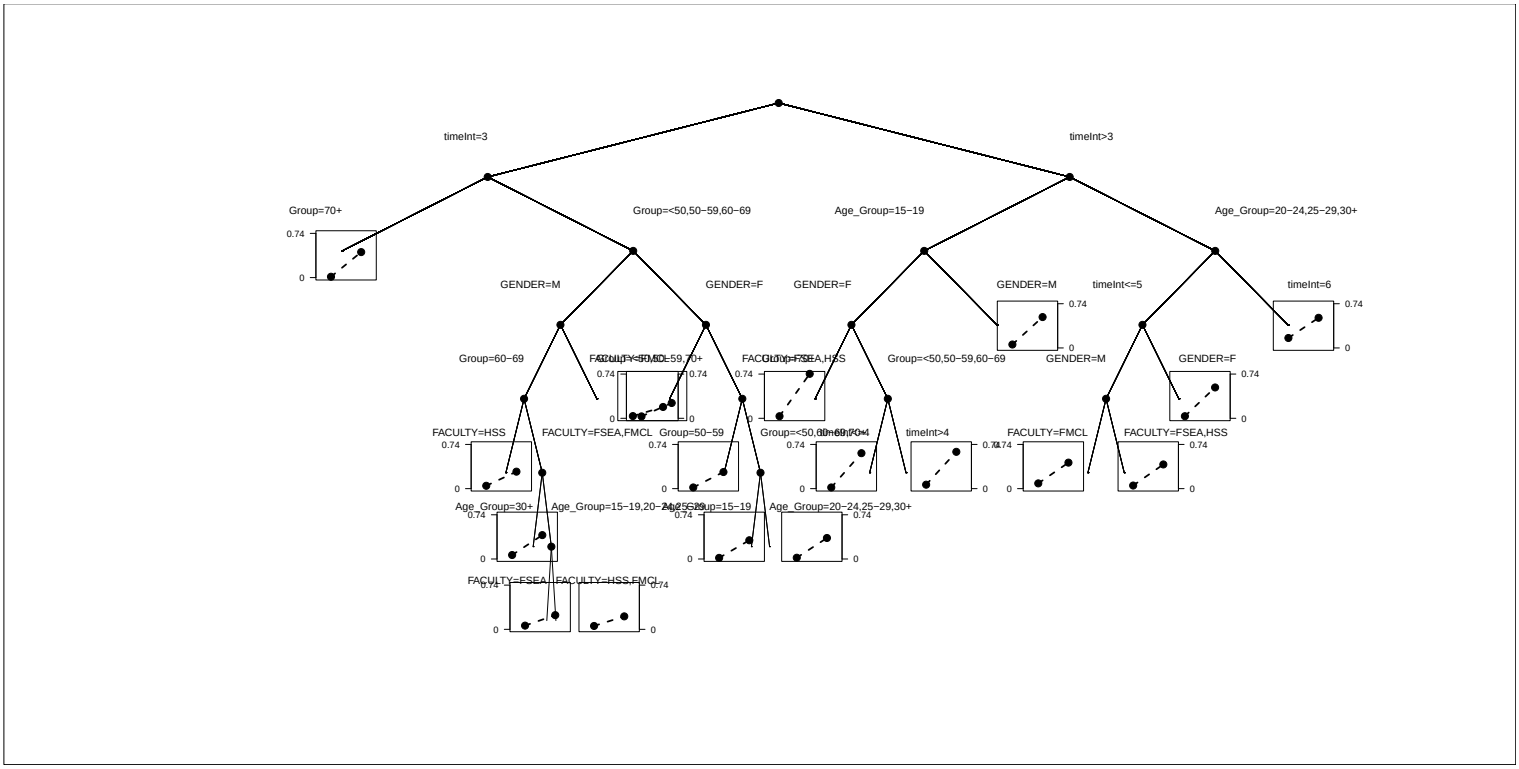


Figure 4.5: Decision tree

According to the BIC plot in Figure 4.4 , the optimal tree model has minimal node size 790 which is marked by the dashed line. This results in a decision tree with eighteen. nodes as shown in Figure 4.5. The tree first splits on timeInt, meaning the time interval (year) is the primary factor determining outcomes graduation and dropout. If timeInt=3, the tree splits again based on average mark(Group). When time >3, the tree will split according to age groups. Further splits occur based on timeInt(years) Group (average mark group), Age Group, Gender, and Faculty.If time is ≤ 3 and Group < 70 then further split based on gender.

Table 4.6 shows the Node probabilities of dropping out P(D), graduatingP(G) and still enrollingP(SE) and the variables that affect these outomes at each node.

© University of Venda

University of Venda
Creating Future Leaders

4.5.1 Nodes with Higher Risk of Dropout

Node 18 Students in their 6th year with ages 20+ have the highest probability of dropping out which is16.7—This suggests that prolonged enrollment beyond the typical graduation period increases the likelihood of dropout by 16.7%.

Node 15 Male students aged from 20 years and above who are spending 5 years and less in faculty of Management, Commerce and Law also have a high dropout risk of 8.8%, suggesting that these students are more likely to dropout.

Node 13 Male students who spend 4 years and more in the programs marks lower than 70 have a moderate risk of dropout. This indicates that longer study duration combined with lower performance increases dropout chances by 6.6%.

Node 3 Male students with marks between 60 and 69 who spend three years in the HSS faculty also tend to dropout more likely as well. These students have a dropout rate of 6.7%

4.5.2 Nodes with Higher Probability of Graduation

Node 11 Female students in the faculty of Science, Engineering and Agriculture and HSS who have average marks less than 50 and 60-69 are most likely to graduate. This suggests that these faculties may provide better support or have students with stronger academic persistence.

Node 12 Female students who spend 4 years and less and have average marks less than 50 and between 60 and 69 are more likely to graduate with a probability of 59.4%.

Node 13 Students who spent 4 years and more in their programs and had an average mark that is less than 70 also have a high risk of graduation which is 61.6% This means

© University of Venda

University of Venda
Creating Future Leaders

that even with longer study period, many students still manage to graduate indicating that persistence can lead to eventual success, even if delayed.

Node 17 Female students who spend 5 years and less are also likely to graduate with a rate of 51.6% This suggests a gender-related difference in academic persistence or success rates.

Node 14 Male students aged between 15 and 19 years who manage to reach three years in the program have a higher graduation probability of 51.7%

4.5.3 Conclusion

Time in the programme is a major factor—longer enrollment increases dropout risk, but persistence can also lead to graduation. Students with marks below 70% are more likely to drop out, but those who persist beyond four years still have a high chance of graduating. Female students, even with lower marks, show stronger graduation trends. Male students with lower academic performance are more prone to dropout, particularly in the Faculty of Management, Commerce, and Law. Younger students who manage to spend three years on their projects are more likely to graduate.

© University of Venda

University of Venda
Creating Future Leaders

Node	P(SE)	P(D)	P(G)	Variables Associated
1	0.559	0.014	0.427	time=3, marks=70+
2	0.667	0.048	0.284	males, HSS faculty, time=3, marks=60-69
3	0.533	0.067	0.400	time=3, males, FSEA, FMCL faculty, age=30+
4	0.700	0.062	0.238	time=3, male, faculty=FSEA
5	0.726	0.057	0.217	time=3, males, faculty=FSEA, FMCL, age=15-19, mark=60-69
6	0.773	0.037	0.190	time=3, males, marks=70+
7	0.712	0.032	0.256	time=3, females, marks=70+
8	0.702	0.020	0.278	time=3, females, faculty=FSEA, HSS, marks=50-59
9	0.668	0.019	0.314	time=3, females, age=15-19
10	0.627	0.021	0.352	time=3, females, marks<50, 60-69, age=20-24, 25-29
11	0.221	0.035	0.744	time=3, females, faculty=HSS, FSEA
12	0.386	0.020	0.594	time≤4, mark≤50, 60-69
13	0.317	0.066	0.616	time>4, mark<50, 50-59, 60-69
14	0.425	0.058	0.517	time>3, males, age=15-19
15	0.476	0.088	0.436	time≤5, male, faculty=FMCL, age=20-24, 25-29, 30+
16	0.543	0.052	0.405	time≤5, male, faculty=FSEA, HSS, age=20-24, 25-29, 30+
17	0.449	0.035	0.516	time≤5, female
18	0.329	0.167	0.505	time=6, age=20-24, 25-29, 30+

Table 4.6: Node probabilities of dropping out(P(D)), graduating(P(G)) and still enrolled(P(SE))

© University of Venda

University of Venda
Creating Future Leaders

4.6 Comparison of Decision tree results with the Discrete time Competing risk survival model results

Table 4.7 shows the comparison of the key results from the Discrete time Cause-specific hazard model and the Decision Tree model. From both these models, male students seem to be more likely to dropout and female students seem to have higher chances of graduation. Academic performance have a huge impact on both dropout and graduation. Students with average mark of 70 are more likely to graduate.

While both the discrete-time cause specific model and the decision tree highlight similar risk patterns (increased dropout likelihood for males and higher graduation rates for students with high average marks), the decision tree model provides greater insight due to its interpretability and ability to detect interactions between variables. Unlike the the discrete time cause specific hazard model, which assesses independent effects of each variable, the decision tree allows one to explore how combinations of student characteristics jointly influence outcomes.

© University of Venda

University of Venda
Creating Future Leaders

Variable	Cause specific model	Decision Tree Nodes
Gender	Males have significantly higher odds of dropout (OR = 1.87, $p < 0.001$) and lower odds of graduation (OR = 0.71, $p < 0.001$)	Nodes 2, 3, 4, 5, 6: Males consistently show higher dropout probabilities (e.g., Node 5 = 5.7%) and lower graduation rates
Average Marks	Only the 70+ group shows marginally significant higher odds of graduation (OR = 1.48, $p < 0.001$)	Node 1 (marks 70+): highest graduation probability (42.7%), Node 6 (males, 70+): only 19% graduation
Age Group	only the 20-24 group is more likely to dropout. Others are not statistically significant on both graduation and dropout	Multiple nodes (e.g., 3, 14, 15, 18) show how age interacts with faculty, gender, and marks. Node 18 (older students at time=6): 16.7% dropout
Faculty	Students in HSS have lower odds of dropout (OR = 0.66, $p = 0.018$), but significant on graduation; FMCL faculty is not significant on both graduation and dropout	Node 2 (males in HSS): dropout = 4.8%, Node 11 (females in HSS/FSEA): graduation = 74.4%
Interactions	Model assesses independent effects only	Tree shows multi-way interactions: Node 5 (young males, FSEA/FMCL, marks 60–69): 5.7% dropout vs. <2% in similar females

Table 4.7: Comparison of Key Results Between the discrete time Cause-specific hazard model and the Decision Tree model

© University of Venda

University of Venda
Creating Future Leaders

4.7 Single model with age and average mark as continuous variables

Table 4.8: Single Risk model Results

Variable	Ref Category	Estimate	SE	OR	Lower CI	Upper CI	p-value
Age	-	0.070	0.019	1.072	1.034	1.112	0.000
Average mark	-	0.004	0.008	1.004	0.987	1.020	0.672
Faculty	FSEA						
HSS	-	-0.278	0.143	0.758	0.573	1.002	0.052
FMCL	-	0.124	0.122	1.132	0.890	1.438	0.312
Gender	Female						
Male	-	0.704	0.115	2.021	1.612	2.534	0.000

4.7.1 Effect of explanatory variables(Age, Gender, Average mark and Faculty) on dropout

The results in Table 4.8 show that, keeping gender, faculty and average mark constant, each additional year of age increases the odds of dropout by 7.2%. We can therefore conclude with 95% confidence that the odds of dropping out for older students increase by between 3.4% and 11.2%. This finding suggests that older students are significantly more likely to drop out than younger students.

Keeping age, average mark, and faculty constant, male students have 102.1% higher odds of dropping out compared to female students. Therefore, we can conclude with 95% confidence that the true increase in dropout odds for male students is between 61.2% and 153.4%. This finding strongly suggests that male students are significantly more likely to drop out than female students. However, the effect of average mark and

© University of Venda

University of Venda
Creating Future Leaders

faculty on dropout is not significant.

4.8 Cause specific hazard model with age and average mark as continuous variables

Table 4.9: Cause specific model Results

Variable	Ref Category	Estimate	SE	OR	Lower CI	Upper CI	p-value
Intercept	-	-5.077	0.670	0.006	0.002	0.023	<0.001
	-	-1.196	0.333	0.302	0.158	0.581	<0.001
Gender	Female						
Male(D)	-	0.548	0.117	1.730	1.374	2.177	<0.001
Male(G)	-	-0.363	0.047	0.695	0.634	0.763	<0.001
Age	-						
Age(D)	-	0.064	0.019	1.067	1.027	1.107	<0.001
Age(G)	-	-0.015	0.011	0.985	0.963	1.007	0.190
Faculty	FSEA						
HSS(D)	-	-0.318	0.145	0.728	0.548	0.967	0.028
HSS(G)	-	-0.096	0.059	0.909	0.810	1.020	0.104
FMCL(D)	-	0.063	0.125	1.065	0.834	1.359	0.614
FMCL(G)	-	-0.146	0.055	0.864	0.776	0.963	0.008
Average Mark	-						
Average mark(D)	-	0.009	0.008	1.009	0.992	1.026	0.304
Average mark(G)	-	0.014	0.004	1.014	1.006	1.021	<0.001

Table 4.9 shows the results of the casue specific hazard model where the variables age and average mark are treated as continuous variables. This model examines the likelihood of dropout (D) and graduation (G) compared to the reference category (still

© University of Venda

University of Venda
Creating Future Leaders

enrolled). The results show log-odds estimates, standard errors (SE), odds ratios (OR), 95% confidence intervals (CI), and p-values for each predictor. Below follows an interpretation about those results.

4.8.1 Effect of Age

Although age has no significant effect on graduation, the results in Table 4.9 show that, keeping gender, faculty, and average mark constant, each additional year of age increases the odds of dropout by 6.7%. We can therefore conclude with 95% confidence that the odds of dropping out for older students increase by between 2.7% and 10.7%, suggesting that older students are significantly more likely to drop out than younger students.

4.8.2 Effect of Average mark

Although the effect of average mark on dropout is not statistically significant, Each one-unit increase in the average mark increases the odds of graduation by 1.4%. We can conclude with 95% confidence that, keeping age, gender and faculty constant, the true increase in graduation odds per unit increase in average mark is between 0.6% and 2.1%, suggesting that higher academic performance slightly increases the likelihood of graduation.

4.8.3 Effect of Faculty

The results in Table 4.9 indicate that students in the Humanities and Social Sciences (HSS) faculty have 27.2% lower odds of dropping out compared to students in the Science, Engineering, and Agriculture (FSEA) faculty. With 95% confidence, we can conclude that the true reduction in dropout odds for HSS students lies between 3.3% and 45.2%. However the odds of this faculty have no effect on the incidence of graduation. Although the FMCL group has no significant effect on dropout, students in the FMCL faculty have 13.6% lower odds of graduating compared to FSEA students.

© University of Venda

University of Venda
Creating Future Leaders

We can conclude with 95% confidence that, keeping age, average mark and gender, the true reduction in the odds of graduation for students in the faculty of Management, Commerce and Law lies between 3.7% and 22.4%

4.8.4 Effect of Gender

Keeping age, faculty, and average mark constant, male students have 73% higher odds of dropping out compared to female students. With 95% confidence, the true increase in dropout odds for male students lies between 37.4% and 117.7%. This finding strongly suggests that male students are significantly more likely to drop out than female students. Male students have 30.5% lower odds of graduating compared to female students. With 95% confidence, the true reduction in graduation odds for male students lies between 23.7% and 36.6%, indicating that female students are significantly more likely to graduate than male students.

4.9 Model evaluation

Since we fitted a total of six models, we will first rename the models as follows; the discrete survival model with categorical age and average mark will be called model1, the one with continuous variables will be called model2. The discrete cause specific model with categorical variables will be model, the continuous one will be model4. The decision tree model with categorical variable will be model5 while the last one, which is a decision tree with continuous variables only, will be called model6.

4.9.1 Concordance Index

The C-index (Concordance Index) measures the discriminatory ability of a survival model. It is the likelihood that, out of the pair of comparable samples selected at random, the sample with the greater risk prediction will either belong to a higher binary class or experience an event before the other sample. It ranges from zero to

© University of Venda

University of Venda
Creating Future Leaders

one where C-index > 0.7 indicates a model with good predictive ability. C-index < 0.5 suggests poor discrimination, meaning the model's predictions are worse than random guessing. From Table 4.10, all models have C-index values far below 0.5, ranging from 0.105 to 0.253. This indicates very poor predictive performance, meaning these models struggle to distinguish between individuals with different risks.The model with the highest C-index is single risk model with categorical age and average mark, followed by the single risk model with continuous age and average mark,these models which were only checking the effect of certain variables on dropout only. However, the classification tree models(with both categorical continuous age and average mark) seemed to perform better than the discrete cause specific models. The model with the highest C-index is the Decision tree for competing risks with continuous age and average, although the best tree obtained from this appeared to be too messy.

A possible explanation for this suboptimal performance could be the omission of key variables such as financial aid status, residence status, school quintile, coupled with a significant amount of missing data in many of the explanatory variables. It is recommended that the University explore more effective methods of capturing student information to improve the quality and completeness of this dataset.

Model name	C Index
Discrete survival model with categorical age and average mark	0.253
Discrete survival model with continuous age and average mark	0.245
Discrete cause specific model for competing risks with categorical age and average mark	0.108
Discrete cause specific model for competing risks with continuous age and average mark	0.105
Decision tree for competing risks with categorical age and average mark	0.174
Decision tree for competing risks with continuous age and average mark	0.195

Table 4.10: C indexes for the models

© University of Venda

University of Venda
Creating Future Leaders

4.10 Discussion

4.11 Determinants of student progression

Gender

In South Africa, females constitute a higher proportion of higher education enrollments compared to males.

These results are consistent with findings from other studies. For instance, Bhorat et al. (2012) found that female students generally outperform their male counterparts in terms of higher education retention and completion rates. These findings are supported by Van Broekhuizen and Spaull (2017), who identified significant gender disparities in four-year completion rates among the 2006 first-time entering undergraduate cohort in the Western Cape. Their results indicate that over half (52%) of female students completed their undergraduate qualifications by the end of 2009, a significantly higher rate than that of their male counterparts. Regarding dropout, Van Broekhuizen and Spaull (2017) found that females were slightly less likely to drop out within the first three years of enrollment compared to males.

Matric performance(Average mark)

South African higher education institutions (HEIs) place significant reliance on standardized school-leaving examination results. These marks are considered key indicators of a student's current knowledge and potential for success in higher education. Since these examinations are nationally standardized and quality-controlled, they are regarded as reliable measures of academic ability, allowing for fair comparisons of students across different years. Research on the predictive value of Matric performance for future academic success has yielded inconsistent results. Some studies suggest that Matric results are strong predictors of academic performance, particularly for students who attended high-quality high schools and whose first language is English (Essack et al.,

© University of Venda

University of Venda
Creating Future Leaders

2012). On the other hand, Marnewick (2012) found no correlation between first-year university performance among Information Technology (IT) students and their Matric results.

Similar to the study by Marnewick (2012), the current study found that the effect of average mark on dropout is not statistically significant for any mark group, on both the single risk and the competing risk model, if we using the discrete time model with average mark as a categorical variable. The effect of average mark on graduation is not significant for all categories of students except the 70 plus group. Students with an average mark of 70+ have 48.2% higher odds of graduating compared to those with an average below 50.

Faculty

Looking at the competing risk model, keeping gender, age, and average mark constant, the odds of dropping out for students in the Faculty of Human and Social Sciences are between 6.9% and 53.7% lower than those of FSEA students, meaning that they are less likely to dropout as compared to those in the faculty of Science, Engineering and Agriculture. However, HSS students do not significantly differ from FSEA students in graduation odds. The Faculty of Commerce, Management, and Law (FMCL) does not significantly differ from FSEA in either dropout or graduation.

Age

This study found that older students are more likely to dropout than younger ones, from the the single risk model in both two spheres(categorical and continuous). One of the reasons for this variation might be that older learners have more family or personal commitments. These results are similar to those of Zewotir et al. (2015) who discovered that older students seem to take longer to successfully complete a master's degree compared with otherwise identical students who are younger. However, age has no effect on graduation on both spheres of the competing risk model.

© University of Venda

University of Venda
Creating Future Leaders

4.12 Multiple imputation

Table 4.11 presents results from a discrete-time cause-specific hazard model where missing data were addressed using multiple imputation. Multiple imputation is a robust statistical technique that replaces each missing value with a set of plausible values based on the observed data, creating multiple complete datasets. These datasets are analyzed separately, and the results are then combined or "pooled" to produce final estimates that account for the uncertainty caused by the missing data. This method has a major advantage over simpler techniques—such as mean imputation—because it preserves the relationships between variables and reflects the variability that naturally arises due to missing values. While simple to implement, mean imputation distorts the distribution of the data, underestimates variability, and can lead to biased estimates—especially in models that rely on the relationships among variables, such as regression or hazard models. By using multiple imputation in this analysis, the risk of biased coefficient estimates and incorrect standard errors was reduced, resulting in more reliable inferences. Thus, the pooled outcomes shown in Table 4.11 more accurately reflect the underlying data patterns, improving both the validity and generalizability of the findings.

Table 4.11 shows that older students, particularly those aged 30 plus, are significantly more likely to drop out compared to those under 20 years old (OR = 2.82, 95% CI: 1.81–4.40, $p < 0.001$), while age has no significant effect on the likelihood of graduation. Male students have more than twice the odds of dropping out (OR = 2.12, 95% CI: 1.92–2.33, $p < 0.001$) and significantly lower odds of graduating (OR = 0.83, 95% CI: 0.80–0.86, $p < 0.001$) than the females. Students in the HSS faculty have lower odds of dropping out (OR = 0.88, 95% CI: 0.78–0.99, $p = 0.032$) compared to those in the FSEA faculty, while students in the FMCL faculty are slightly less likely to graduate (OR = 0.95, 95% CI: 0.90–0.99, $p = 0.017$). Students who achieved 70% or higher had significantly higher odds of graduating (OR = 1.28, 95% CI: 1.14–1.44, $p < 0.001$), however, academic performance did not significantly affect the odds of dropping out.

© University of Venda

University of Venda
Creating Future Leaders

In comparison with the results from the mean imputation, these results are similar.

© University of Venda

University of Venda
Creating Future Leaders

Variable	Ref Category	Estimate	SE	OR	Lower CI	Upper CI	p-value
(Intercept):1	Still enrolled	-3.210	0.179	0.040	0.027	0.060	<0.001
(Intercept):2	Still enrolled	-0.319	0.052	0.727	0.656	0.806	<0.001
Age	<20						
20–24:1	-	0.385	0.052	1.470	1.330	1.630	<0.001
20–24:2	-	-0.020	0.022	0.981	0.940	1.020	0.369
25–29:1	-	0.306	0.148	1.360	1.020	1.820	0.039
25–29:2	-	-0.081	0.069	0.922	0.806	1.060	0.239
30+:1	-	1.040	0.225	2.820	1.810	4.400	<0.001
30+:2	-	-0.149	0.131	0.862	0.667	1.110	0.256
Faculty	FSEA						
FMCL:1	-	0.049	0.053	1.050	0.947	1.170	0.354
FMCL:2	-	-0.057	0.024	0.945	0.902	0.990	0.017
HSS:1	-	-0.131	0.061	0.877	0.778	0.989	0.032
HSS:2	-	0.003	0.026	1.000	0.953	1.060	0.902
Gender	Female						
Male:1	-	0.749	0.050	2.120	1.920	2.330	<0.001
Male:2	-	-0.186	0.021	0.830	0.798	0.864	<0.001
Average mark	<50%						
50–59:1	-	-0.056	0.148	0.945	0.683	1.310	0.711
50–59:2	-	0.032	0.050	1.030	0.935	1.140	0.522
60–69:1	-	0.026	0.176	1.030	0.680	1.550	0.886
60–69:2	-	0.077	0.048	1.080	0.982	1.190	0.112
70+:1	-	-0.077	0.303	0.925	0.440	1.950	0.807
70+:2	-	0.246	0.060	1.280	1.140	1.440	<0.001

Table 4.11: Cause specific model results from Multiple imputation

© University of Venda

University of Venda
Creating Future Leaders

Chapter 5

Conclusion

5.1 Introduction

This chapter provides a summary of the study's findings discussed in the previous chapter. It also presents the conclusions derived from the results, along with study recommendations and identified limitations. The study aimed to examine the patterns of student progression at the University of venda using discrete-time survival analysis methods in a statistical and machine learning approach. In particular, the main objectives of the study were to: (1)Build a discrete time competing risk model to identify factors which determine the time to graduation or dropout, (2) Develop discrete time machine learning model(decision tree) to identify the factors which determine the time to graduation or dropout, (3)Compare the discrimination and calibration performance of the machine learning and statistical models for time to graduation or dropout, (4) Make recommendations on the strategies that universities can use to reduce the time from registration to graduation and to reduce dropout rates among their students. The findings of the study are discussed in accordance with the research objectives.

© University of Venda

University of Venda
Creating Future Leaders

5.2 Conclusion

This comparison was done on models that were fitted while considering the categorical case of the explanatory variables. The current study found that the effect of average mark on dropout is not statistically significant for any mark group, on both the single risk and the competing risk model(age and average mark categorical). Students with an average mark of 70+ have 48.2% higher odds of graduating compared to those with an average below 50. Students in the faculty of Human and Social Sciences are less likely to dropout as compared to those in the faculty of Science, Engineering and Agriculture. However, HSS students do not significantly differ from FSEA students in graduation odds. The Faculty of Commerce, Management, and Law (FMCL) does not significantly differ from FSEA in either dropout or graduation. This study found that older students are more likely to dropout than younger ones, from the the single risk model in both two spheres(categorical and continuous). One of the reasons for this variation might be that older learners have more family or personal commitments. The results from both models, i.e, Single risk model and the competing risk model show that females are more likely to graduate than males and males a more likely to dropout as compared to females.

However, when age and average mark are considered as continuous variables, there is a slight change in the results in that FMCL are less likely to graduate on the discrete time competing risk survival model.

The study's findings demonstrate the kinds of extra insights that may be obtained when analyzing student dropout using discrete-time competing risk survival analysis techniques as opposed to more traditional methods like logistic regression. We see this from the set of two models namely discrete cause-specific model and decision tree for competing risks, these two models gained more insight about both the competing events(dropout and graduation) of students progression. However, the decision tree model seemed to have gained more insight than the cause-specific model. This is evident

© University of Venda

University of Venda
Creating Future Leaders

on the C-index table where the values are 0.174(Decision tree model for competing risks with categorical age and average mark) and 0.195 (Decision tree model for competing risks with continuous age and average mark).

5.3 Recommendation

This study suggests that using a decision tree model is more efficient than standard approaches for analyzing student dropout and academic results and recommends that it should therefore be used for analysis of academic outcomes. Interventions strategies like care and support programmes for reducing dropout rates and shortening the time from first registration to graduation should target the identified high risk groups such as male and older students.

5.4 Limitation of the study

The primary limitation of this study stems from the use of secondary data, restricting model estimation to the variables available in the database. Information on financial support residence status, school quintile and the marital status of students was not available on the database.

5.5 Future Research

Future research should focus on fitting more competing risk models like the Sub-distribution hazard model, Random Survival Forest and the Neural Networks. Since a lot of time spent on data processing and cleaning, there were a lot of work that should have been done to evaluate these models. One could also look at other evaluation metrics like the Log-Rank Tests and Prediction error curves in order to evaluate the

© University of Venda

University of Venda
Creating Future Leaders

discrimination and the calibration of the models. More research has to be done on the factors(gender, faculty, age and average mark) since they have a significant effect on student progression. To examine the underlying problems connected to these factors, more study is required.

© University of Venda

University of Venda
Creating Future Leaders

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2012.

P. D. Allison. Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13:61–98, 1982.

P. D. Allison. *Event History Analysis: Regression for Longitudinal Event Data*, volume 46. Sage, 1984.

P. D. Allison. *Event History and Survival Analysis: Regression for Longitudinal Event Data*, volume 46. SAGE Publications, 2014.

S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 903–912. ACM, 2016.

A. Andersen. Discrete models for the analysis of time series. *Journal of the American Statistical Association*, 65(332):9–33, 1970.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Science & Business Media, 2012.

E. Arias Ortiz and C. Dehon. Roads to success in the belgian french community's higher education system: Predictors of dropout and degree completion at the université libre de bruxelles. *Research in Higher Education*, 54:693–723, 2013.

© University of Venda

University of Venda
Creating Future Leaders

M. Berger and M. Schmid. Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 18(3-4):322–345, 2018.

M. Berger, T. Welchowski, S. Schmitz-Valckenberg, and M. Schmid. A classification tree approach for the modeling of competing risks in discrete time. *Advances in Data Analysis and Classification*, 13:965–990, 2019.

M. Berger, M. Schmid, T. Welchowski, S. Schmitz-Valckenberg, and J. Beyersmann. Subdistribution hazard models for competing risks in discrete time. *Biostatistics*, 21 (3):449–466, 2020.

H. Bhorat, N. Mayet, and M. Visser. *Student Graduation, Labour Market Destinations and Employment Earnings*. HSRC Press, 2012.

E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186, 1998.

E. M. Biganzoli, P. Boracchi, F. Ambrogi, and E. Marubini. Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial Intelligence in Medicine*, 37(2):119–130, 2006.

K. Bokana and D. Tewari. Determinants of student success at a south african university: An econometric analysis. *The Anthropologist*, 17(1):259–277, 2014.

J. Borucka. Methods of handling tied events in the cox proportional hazard model. *Studia Oeconomica Posnaniensia*, 2(2):91–106, 2014.

L. Breiman. Some properties of splitting criteria. *Machine Learning*, 24:41–47, 1996.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

L. Breiman. *Classification and Regression Trees*. Routledge, 2017.

© University of Venda

University of Venda
Creating Future Leaders

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

Y. Chen, A. Johri, and H. Rangwala. Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 270–279. ACM, 2018.

J. R. Chimka, T. Reed-Rhoads, and K. Barker. Proportional hazards models of graduation. *Journal of College Student Retention: Research, Theory & Practice*, 9(2): 221–232, 2007.

D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2008, Antwerp, Belgium, Proceedings, Part I*, volume 5211 of *Lecture Notes in Computer Science*, pages 241–256. Springer, 2008.

D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24:136–158, 2012.

N. Cloete, P. Maassen, and T. Bailey. *Knowledge Production and Contradictory Functions in African Higher Education*. African Minds, 2015.

F. J. d. Costa, M. d. S. Bispo, and R. d. C. d. F. Pereira. Dropout and retention of undergraduate students in management: A study at a brazilian federal university. *RAUSP Management Journal*, 53:74–85, 2018.

R. Couronné, P. Probst, and A.-L. Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19:1–14, 2018.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

© University of Venda

University of Venda
Creating Future Leaders

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

R. Csalódi and J. Abonyi. Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout. *Mathematics*, 9(5):463, 2021.

S. L. DesJardins. Event history methods: Conceptual issues and an application to student departure from college. In J. C. Smart, editor, *Higher Education: Handbook of Theory and Research*, pages 421–471. Springer, 2003.

S. L. DesJardins, D. A. Ahlburg, and B. P. McCall. A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education*, 73(5):555–581, 2002.

L. L. Doove, E. Dusseldorp, K. Van Deun, and I. Van Mechelen. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification*, 8 (4):403–425, 2014.

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

B. Efron. Logistic regression, survival analysis, and the kaplan–meier curve. *Journal of the American Statistical Association*, 83(402):414–425, 1988.

F. Emmert-Streib and M. Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

S. Y. Essack, V. Wedekind, and I. Naidoo. Selection tools predictive of success in the health sciences—nated vs. nsc students. *South African Journal of Higher Education*, 26(3):472–486, 2012.

I. Eugene, I. Jeremiah, and E. Atombosoba. Teenage pregnancy and obstetric outcome: A comparative study in urban nigeria. *Nepal Journal of Obstetrics & Gynaecology*, 11(1), 2016.

© University of Venda

University of Venda
Creating Future Leaders

L. Fahrmeir. Dynamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika*, 81(2):317–330, 1994.

L. Fahrmeir. Discrete survival-time models. *Wiley StatsRef: Statistics Reference Online*, 2014.

L. Fahrmeir and S. Wagenpfeil. Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, 91(436):1584–1594, 1996.

R. J. Fehring, M. Schneider, K. Raviele, D. Rodriguez, and J. Pruszynski. Randomized comparison of two internet-supported fertility-awareness-based methods of family planning. *Contraception*, 88(1):24–30, 2013.

J. P. Fine and R. J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*, 2009.

H. Frydman and A. Matuszyk. Random survival forest for competing credit risks. *Journal of the Operational Research Society*, 73(1):15–25, 2022.

N. Gitinabard, F. Khoshnevisan, C. F. Lynch, and E. Y. Wang. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. *arXiv preprint arXiv:1809.00052*, 2018.

P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

A. Groll and G. Tutz. Variable selection in discrete survival models including heterogeneity. *Lifetime Data Analysis*, 23:305–338, 2017.

© University of Venda

University of Venda
Creating Future Leaders

J. Heckman and B. Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984a.

J. J. Heckman and B. Singer. Econometric duration analysis. *Journal of Econometrics*, 24(1-2):63–132, 1984b.

I. Hertz-Picciotto and B. Rockhill. Validity and efficiency of approximation methods for tied survival times in cox regression. *Biometrics*, pages 1151–1156, 1997.

P. Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

P. Hougaard. *Analysis of multivariate survival data*, volume 564. Springer, 2000.

T. T. Ishitani and S. L. DesJardins. A longitudinal investigation of dropout from college in the united states. air 2002 forum paper. *ERIC*, 2002.

H. Ishwaran and U. B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80(13-14):1056–1064, 2010.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2008.

H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

S. Janitza and G. Tutz. Prediction models for time discrete competing risks. *Statistical Papers*, 2015.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2002.

© University of Venda

University of Venda
Creating Future Leaders

G. Kantidakis, H. Putter, S. Litière, and M. Fiocco. Statistical models versus machine learning for competing risks: Development and validation of prognostic models. *BMC Medical Research Methodology*, 23(1):51, 2023.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*, volume 1230. Springer, 2003.

T. Lancaster. *The Econometric Analysis of Transition Data*. Cambridge University Press, 1990.

M. Letseka and M. Cosser. Student retention & graduate destination: Higher education & labour market access & success. *HSRC Review*, 2010.

P. J. Lisboa, T. A. Etchells, I. H. Jarman, C. T. Arsene, M. H. Aung, A. Eleuteri, A. F. Taktak, F. Ambrogi, P. Boracchi, and E. Biganzoli. Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE Transactions on Neural Networks*, 20(9):1403–1416, 2009.

C. Marnewick. The mystery of student selection: Are there any selection criteria? *Educational Studies*, 38(2):123–137, 2012.

M. Mills. Introducing survival and event history analysis. 2010.

J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4:227–243, 1989.

K. Monterrubio-Gómez, N. Constantine-Cooke, and C. A. Vallejos. A review on competing risks methods for survival analysis. *arXiv preprint arXiv:2212.05157*, 2022.

K. Monterrubio-Gómez, N. Constantine-Cooke, and C. A. Vallejos. A review on statistical and machine learning competing risks methods. *Biometrical Journal*, 66(2): 2300060, 2024.

P. Moodley and R. J. Singh. Addressing student dropout rates at south african universities. *Alternation (Durban)*, 2015.

© University of Venda

University of Venda
Creating Future Leaders

P. M. Mphekgwana, T. E. Mabila, H. M. Tirivangasi, and H. M. Makgopa. Analysis of survival rates among postgraduate students at a historically disadvantaged university in south africa. *Gender & Behaviour*, 18(3), 2020.

B. Muthén and K. Masyn. Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics*, 30(1):27–58, 2005.

W. Narendranathan and M. B. Stewart. Modelling the probability of leaving unemployment: Competing risks models with flexible base-line hazards. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(1):63–83, 1993.

B. D. Ndlovu. *Modelling Time to Graduation of Durban University of Technology Students Using Event History Analysis*. PhD thesis, Durban University of Technology, 2015.

O. M. Odeniya, A. A. Abiodun, and B. A. Oyejola. Discrete-time survival analysis of age at first pregnancy among nigerian women. *Research Journal of Applied Science*, 8:215–220, 2013.

E. T. Pascarella and P. T. Terenzini. *How College Affects Students: A Third Decade of Research. Volume 2*. ERIC, 2005.

R. L. Prentice and L. A. Gloeckler. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, pages 57–67, 1978.

F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52:199–215, 2003.

P. L. Ramokolo. *Application of Discrete-Time Survival Analysis Techniques in Modelling Student Dropout: A Case of Engineering Students at Tshwane University of Technology, South Africa,*. PhD thesis.

M. Restaino. Dropping out of university of salerno: A survival approach. *Working Papers; 3.193*, 2008.

© University of Venda

University of Venda
Creating Future Leaders

S. L. Rohr. How well does the sat and gpa predict the retention of science, technology, engineering, mathematics, and business students. *Journal of College Student Retention: Research, Theory & Practice*, 14(2):195–208, 2012.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

T. H. Scheike and T. K. Jensen. A discrete survival model with random effects: An application to time to pregnancy. *Biometrics*, pages 318–329, 1997.

M. Schmid and M. Berger. Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1529, 2021a.

M. Schmid and M. Berger. *Modelling Discrete Time-to-Event Data*. Springer, 2021b.

M. Schmid, H. Küchenhoff, A. Hoerauf, and G. Tutz. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, 35(5):734–751, 2016.

D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

J. D. Singer and J. B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2): 155–195, 1993.

J. D. Singer and J. B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, 2003.

N. Spaull and N. Makaluza. Girls do better: The pro-female gender gap in learning outcomes in south africa 1995–2018. *Agenda*, 33(4):11–28, 2019.

S. Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, 24(1): 12–18, 2014.

© University of Venda

University of Venda
Creating Future Leaders

F. Steele, H. Goldstein, and W. Browne. A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, 4(2):145–159, 2004.

K. Suresh, C. Severn, and D. Ghosh. Survival prediction models: An introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(1):207, 2022.

D. Teferra. *Scientific Communication in African Universities: External Assistance and National Needs*. Routledge, 2004.

W. A. Thompson Jr. On the treatment of grouped observations in life studies. *Biometrics*, pages 463–470, 1977.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

G. Tutz and S. Petry. Nonparametric estimation of the link function including variable selection. *Statistics and Computing*, 22:545–561, 2012.

G. Tutz, M. Schmid, et al. *Modeling Discrete Time-to-Event Data*. Springer, 2016.

C. A. Vallejos and M. F. J. Steel. Bayesian survival modelling of university outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pages 613–631, 2017.

H. Van Broekhuizen and N. Spaull. The 'martha effect': The compounding female advantage in south african higher education. *Unpublished manuscript*, 2017.

J. W. Vaupel, K. G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

© University of Venda

University of Venda
Creating Future Leaders

P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

T. Welchowski and M. Schmid. discsurv: Discrete time survival analysis. https://cran.r-project.org/package=discSurv, 2015. R package version 1.1.

D. Yang, T. Sinha, D. Adamson, and C. P. Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, volume 11, page 14. Lake Tahoe, NV, 2013.

T. Zewotir, D. North, and M. Murray. Student success in entry level modules at the university of kwazulu-natal. *South African Journal of Higher Education*, 25(6):1233–1244, 2011.

T. Zewotir, D. North, and M. Murray. The time to degree or dropout amongst full-time master's students at university of kwazulu-natal. *South African Journal of Science*, 111(9–10):01–06, 2015.

B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artificial Intelligence in Medicine*, 20(1):59–75, 2000.

© University of Venda