

Modelling Time To Graduation of Durban University of Technology Students Using Event History Analysis

A dissertation presented in partial fulfilment of the
requirements for the degree of Master of Science

by

Bonginkosi Duncan Ndlovu

832838362

School of Mathematics, Statistics and Computer Science
University of Kwa-Zulu Natal

April 2015

Abstract

Tertiary institutions experienced a steady growth of students from other races after the repeal of the apartheid laws. This growth picked up pace after the promulgation of the Education White Paper of 1997 whose main thrust was to make the previously exclusive institutions accessible to the wider populace. Disturbingly, however, and contrary to the goals and the spirit of the White Paper, these institutions also experienced higher failure and lower retention rates amongst the previously disadvantaged students.

This study seeks to model time to graduation using survival analysis methods. We begin the analysis by assessing the relevance of the available variables to the exercise of modelling time to graduation using descriptive statistics and non-parametric techniques. We compared the Cox regression to its extensions in discrete time, the Discrete Time to Event Approach, with the view to find the best model to explain time to graduation given the available variables.

In light of limited availability of relevant data, we evaluated unobserved heterogeneity in both models. We closed the analysis by considering the cure models and mixture competing risks in discrete time.

Notwithstanding arguments against suitability of the Cox regression in continuous time for modelling inherently discrete data such as found in our study, we found that Cox's regression over all, provided a reasonably good fit given the available data.

We also found that in relation to the Cox proportional hazard model, there was a lesser degree of flexibility as certain variable effects were sacrificed to satisfy the proportionality

assumption by stratifying on those variables. The advantage of the Discrete Time to Event Approach is that we could assess the effects of all variables in the model and also obtain the estimates of risks to graduation which are true probabilities of graduation with fewer assumptions or conditions to satisfy.

We found that the data limitations did not compromise either the box Cox regression model or the Discrete Time to Event Approach.

The data also suggested existence of a sizable proportion of subjects that will eventually not graduate based on cure models. We also fractionated subjects censored due to closure of the observation period into those that will eventually graduate and those that will eventually dropout, using discrete mixture competing risks. We found that the mixture competing risks model explained graduation better than the cure model.

Declaration

The research work is the original work done by the author (Bonginkosi Duncan Ndlovu) and it is not a duplicate of some of the research work done by other authors. All the references that were used to refer to are duly acknowledged

Bonginkosi Duncan Ndlovu (832838362)

Date

Prof. Temesgen Zewotir

Date

Dedication

I dedicate this to my mother, Nelisiwe Ndlovu, she loved her children dearly.

Acknowledgements

I wish to thank my supervisor, Prof. Temesgen Zewotir for providing me with guidance and more importantly, his supervisory style which taught me independence and instilled self belief. I also wish to thank The Durban University of Technology for providing me with data. I cannot leave out Dr Paul Green who is first my dear friend, and then my Head of Department at the Durban University of Technology, he was my inspiration, and this work would not have been possible without him. I salute my son, Ntandoyenkosi, with whom I put in long hours at the office to complete this work whilst he was waging his own battle with matric.

Contents

<i>Abstract</i>	i
<i>Declaration</i>	i
<i>Dedication</i>	ii
<i>Acknowledgements</i>	iii
<i>1. Introduction</i>	1
<i>2. The Data</i>	6
<i>3. Non-parametric Analysis</i>	13
<i>4. The Cox Regression Model</i>	25
4.1 Model	27
4.2 Model Adequacy	30
4.2.1 The Proportionality Assumption	30
4.2.2 Outliers	37
4.2.3 Influential Observations	39
4.2.4 Overall Fit	44
4.3 Prediction	45
4.4 Summary	47
<i>5. Discrete Time to Event Approach</i>	49
5.1 Introduction	49
5.2 Model	52

5.3	Model Adequacy	58
5.3.1	Outliers	59
5.3.2	Influence on all Parameter Estimates	61
5.3.3	Influence on Individual Parameter Estimates	64
5.4	Prediction	67
5.5	Summary	69
6.	<i>Frailty Models</i>	71
6.1	Continuous Time	74
6.1.1	Parametric Methods	75
6.1.2	Semi-Parametric Methods	78
6.2	Discrete Time	81
6.3	Summary	84
7.	<i>Mixture Models</i>	86
7.1	Cure Models	86
7.1.1	Model & Estimation	89
7.2	Mixture Competing Risks	94
7.2.1	Model & Estimation	95
7.3	Summary	98
8.	<i>Conclusion and Discussion</i>	100

List of Figures

3.1	Gender Survivor Function	18
3.2	Faculty Survivor Function	19
3.3	Race Survivor Function	19
3.4	Age Survivor Function	20
4.1	Log Cumulative Hazards by Faculty	34
4.2	Female Schoenfeld Residuals	35
4.3	White Schoenfeld Residuals	36
4.4	Deviance Residuals	39
4.5	Female Delta-Beta	40
4.6	Indian Delta-Beta	41
4.7	White Delta-Beta	41
4.8	Coloured Delta-Beta	42
4.9	Other Delta-Beta	42
4.10	Cox-Snell Residuals	44
4.11	Fitted Survivor Functions	47
5.1	Deviance	60
5.2	Pearson's Residuals	61
5.3	D Statistic	62
5.4	D Statistic vs Fitted Probability	63
5.5	Other Delta-Beta	64
5.6	Health Sc. Time 4 Delta-Beta	65
5.7	Art & Design Time 4 Delta-Beta	65

5.8	Art & Design Time 5 Delta-Beta	66
5.9	Faculty Hazard Estimates	68
5.10	Race Hazard Estimates	69
5.11	Gender Hazard Estimates	70
7.1	Mixture & Sample Survivor Functions	98

List of Tables

2.1	Average Graduation Times	11
3.1	Log-Rank Test for Equality of Survivor Functions	23
4.1	Full & Reduced Model	29
4.2	Proportionality Test	33
4.3	Proportionality Test for the Stratified Model	34
4.4	Stratified Cox Model	35
4.5	Influential Subjects	43
4.6	Stratified Cox Model	45
5.1	Full Model	56
5.2	Reduced Model	57
5.3	The Outliers	61
5.4	Influential Observations	63
6.1	Parametric Frailty Model & Cox Regression Model	78
6.2	Semiparametric Frailty Model	80
6.3	Logistic Random Effects Model and Ordinary Logistic Model	83
7.1	Incidence & Latency Coefficients	93
7.2	latency & Incidence Coefficients	97

Introduction

Prior to the repeal of apartheid laws in 1991, Education in South Africa was segregated along racial lines. Under the former Population Registration Act of 1950, the people of South Africa were classified as Black, Indian, White and Coloured. The introduction of the Bantu Education Act of 1953, Coloured Persons Education Act of 1963, Indian Education Act and the Extension of Universities Act of 1959 formalised and legalized the concept of "separate but not equal" education for each ethnic group from primary school to higher education. Different "population" groups could only be accepted for study at their "own" tertiary institutions.

With the gradual repeal of apartheid laws commencing in 1991, the historically White universities and historically White technikons began to experience a steady growth of students from other races. The combined population of White students was 96% of the entire student population in these institutions in 1990 and 86% in 1993, the other races accounted for the difference. The growth rate of Black (African, Indian and Coloured) students in historically White institutions increased after the promulgation of the Education White Paper of 1997. The main thrust of the Education White Paper was the transformation of Higher Education through, *inter alia*, increased access of Black and women as well as disabled and mature students to higher education (Cloete et al., 2004).

The enrolment of Black students at the University of Witwatersrand alone, a historically White university, was 8% in 1986, 13% in 1990, 23% in 1994 and 35% in 1997. In 1997,

48% and 55% of the student population in predominantly Afrikaans and English universities, respectively were Black, and these numbers grew to 58% and 62% respectively in 2000. Also, 60% of the student population in historically White Technikons were Black in 1997 and this figure grew significantly to 77% in 2000 (Cloete et al., 2004).

Whilst these figures were commendable, a disturbing trend however began to emerge which was in conflict with the goals of the White Paper, there was a drop in retention rates of Black students compared to White students in higher education. The large number of financial exclusions explained only half the story, high failure rates particularly amongst Black students in the historically White institutions was the main contributing factor towards low retention rates (Cloete et al., 2004).

High dropout rates and low completion rates in Higher Education are a matter of national concern because they cost the country about R1.3 billion per year, (DoE, 2003). In a cohort study of students between year 2000 and 2005, it was found that respectively, 34% and 25% of the technikon and university students dropped out in the first year, 13% and 9% dropped out in second year, and finally, 11% and 7% dropped out in the third year.

Generally, 50% of students in Higher Education drop out by year three. A dropout rate of 50% is very high compared to UK and Germany where the rates are 22% and 27% respectively (Letseka, 2009)

Furthermore, only 30% of first-time entering students in higher education graduate within five years in a three year undergraduate programme, and this figure is 23% in former technikons. Universities fare marginally better as 38% of the students graduate within five years in undergraduate programmes that otherwise take a minimum period of three years to complete (Scott and Fisher, 2011).

Against this background, this study intends to investigate graduation and dropout rates at The Durban University of Technology. This institution came into being as a merger between a historically White Technikon Natal and a historically Black ML Sultan.

The institution has had its fair share of below par pass rates over the years. More recently, the yearly pass rates for the years 2007, 2008 and 2009 have been 21%, 20% and 24% respectively. (DoE, 2007, 2008, 2009). A more detailed discussion of the institution will be presented in the next chapter under *Data Description* section.

Put more formally, the objectives of this study are to:

- 1) Model time to graduation from 2004 to 2008
- 2) Investigate if there is a relationship between graduation and various explanatory variables such as race, gender, faculty etc.
- 3) Model a risk-to-graduate profile for students as a function of time to identify the periods when students are most at risk.
- 4) Investigate the possibility that there might exist a proportion of students that might eventually not graduate even if the study period was extended
- 5) Fractionate the censored students due to closure of observation period into those that will eventually graduate and those that will eventually dropout.

The institution tracks a student for a period of five years, commencing on the year of registration as a matter of procedure consistent with institutional rules (academic exclusion rule G19).

This rule stipulates that the maximum allowable period within which a student must complete his or her qualification is 5 years calculated from the year of first registration. A student, however, might still pursue his or her studies beyond the maximum allowable period of five years as the institution does not strictly enforce the exclusion rule but the institution does not keep records for such students. This condition imposes an automatic observation period of 5 years commencing from the year of first registration.

The main objective of the study is to model time to graduation of students at the Durban University of Technology. The observation period or the data collection period is fixed at 5 years, consequently some of the students will not have graduated or dropped out during this period of observation. Furthermore, some of the students will dropout during the course of observation. With regards to time to graduation, the last recorded times of these students or subjects are not complete observations and we refer to this condition as censoring.

Traditional and established statistical techniques such as linear regression fail to deal adequately with censored data, instead, *Survival Analysis* techniques will be considered as the main tool of analysis.

Survival analysis is an umbrella term which refers to statistical techniques used to model time to occurrence of an event in the presence of censoring. Although the roots of survival analysis are in natural sciences, the techniques have in recent years found wider applications in areas as far afield as sociology and economics. These methods are referred to as *event history analysis* in sociological fields such as education, whereas the same techniques are referred to as *duration analysis* in economics.

Irrespective of the field of application, the variable of interest is time to occurrence of a particular event or events of interest. One of the reasons that survival analysis techniques have found wider appeal in fields outside the customary realm of traditional

sciences is that the requirements, namely: 1) a clearly defined time of origin 2) a scale for measuring time and 3) a well-defined event, are not particularly stringent to meet.

The three requirements mentioned above are met in this study because we have a clear time of origin which is the incoming students, registered for the first time in the year 2004, we also use years as the measurement of time and our event of interest is graduation of this cohort group.

This study is organised as follows; in Chapter 2, we discuss the covariates and present descriptive statistics, Chapter 3 is devoted to the introduction of survival analysis with special attention to *non-parametric* techniques. In Chapter 4, we introduce Cox's regression model, followed by the logistic regression model in Chapter 5. We then complete the analysis with Frailty Models in Chapter 6 and Cure Models and Mixture Competing Risks in Chapter 7. We finally close the study with conclusions and recommendations in Chapter 8.

The Data

The Durban University of Technology was formed as a merger between Natal Technikon and ML Sultan Technikon in April 2002. It was initially known as the Durban Institute of Technology and the name was changed to Durban University of Technology in March 2006 when the status of all former technikons was upgraded to that of universities of technology.

The institution has six campuses in total, four in Durban - Brickfield, City, ML Sultan and Ritson Road, the other two campuses; Indumiso and Riverside are in Pietermaritzburg, about 80 kilometres from Durban. The institution has six faculties; Accounting & Informatics, Applied Sciences, Arts & Design, Engineering & The Built Environment, Health Science and Management Sciences.

The institution offers three year diploma qualifications in its various faculties which may then be followed upon by a one year Bachelor of Technology qualification. The institution also offers the Master and Doctor of Technology qualifications by dissertation.

The sample of this study consists of diploma students registered for the first time in 2004. The institution tracks a student for a period of five years only, but students can still

pursue their studies beyond the maximum allowable period of five years. The institution however does keep records which indicate if a student did register or not at the beginning of the sixth year.

Some qualifications are offered on annual bases running from January to December, others are offered on semester bases and as a result students can also graduate halfway through the year instead of year-end only.

Likewise, a student can drop out at any point in time during the course of the year and the only way that the institution can tell if a student has withdrawn during the year is if the student formally applies for de-registration or fails to register in the following year. A dropout is defined as a withdrawal from the programme for whatever reason other than graduation. A student, for example, can withdraw for health, financial, death or voluntarily, and since most of the students do not formally tender their withdrawal, the real reasons for withdrawal cannot be known.

Censoring arises in two way in this study. Firstly, the group of students who have left the study for reasons other than graduation are regarded as censored subjects. Secondly, the group of students who are still in pursuit of their studies at the close of the observation period are also regarded as censored.

Although students can graduate halfway through the year and ideally the metric of time should therefore be semesters, but dropouts can occur at any point through out the year, but the instant of their occurrence cannot be known exactly so as to apportion them to the corresponding semester. To align dropouts with graduates, we therefore assume a yearly metric of time. Thus, students who graduated halfway through the year are regarded as having graduated at the end of that year. Likewise, dropouts during the year are also assumed to have dropped out at the end of that year.

This study is focused upon students that were continuously registered at The Durban Institute of Technology from the year 2004 until they either graduated, dropped out or were still in pursuit of their studies at the close of data collection at the end of year 2008. The instances where a student de-registers in one year, skips a year or two then re-registers are excluded as their case belongs to the class of repeated events which falls outside the scope of this study.

Out of 4866 students who were registered in 2004, 325 or 6.7% are excluded from the study on the grounds of discontinuous registration. The time to graduation is calculated as the continuous time from 2004 until graduation. Thus, time to graduation can only assume the values 3, 4 or 5.

A myriad of factors in the literature are hypothesised to affect student success or failure in higher education. Race, age, gender, socio-economic background, matriculation grades, academic ability, financial support, family structure, school quality etc. are some of the variables that have been found to have sizable impact on the performance of students in higher education and eventually inform graduation in the literature (Bradley and Lenton, 2007; Desjardins and McCall, 2010).

This study is limited by the availability of variables that have been identified to explain graduation in the literature, consequently, we will also limit our discussion to the available variables namely - *race, age, gender and faculty*. We will also motivate the use of race as proxy for other variables.

It has been found in many studies that students from disadvantaged backgrounds, particularly Africans in South African context, are less likely to graduate and more likely to dropout than other racial groups (van Heerden, 1995; Strauss et al., 2003). Apart from inferior schooling system for African students which, on its own, has a significant impact on student performance later in higher education, there are other factors that

contribute to African students difficulties in higher education. Second language medium of instruction, financial strain, accommodation, transition and adaptation from African culture to a predominantly Western culture etc. are some of the factors that have been found to contribute substantially towards poor performance of African students in higher education. In general, African students tend to have the worst graduation rates and the highest dropout rates whereas the opposite is true for their White counterparts (Wilson, 1984; van Heerden, 1995; Scott and Fisher, 2011).

In the same cohort study of year 2000 DoE (2005), it was found that the graduation rates vary by faculties. The percentage of students that have completed a three year diploma at what was formerly known as technikons, within a five year period are 33%, 34%, 17% and 29% in Business/Management, Computer Science, Engineering and Social Sciences/ Public Administration respectively. Clearly, the Engineering stream has the lowest output rate with Computer Science enjoying the highest rate (Scott et al., 2007),(Scott and Fisher, 2011).

Regarding age, older students are more likely to have lower completion rates than younger students. Older students are more likely to have other commitments such as work, dependents or spouses etc. and these have a tendency to distract them away from full attention towards their academic studies. Younger students tend to have fewer commitments outside their academic studies and in many instances they are full-time students (Desjardins et al., 2002).

There are conflicting findings in the literature regarding the effect of gender on success or failure of students in higher education but women students typically tend to have higher completion rates than males (Desjardins et al., 2002).

The most commonly used measures of Socio-Economic Status (SES), are family income and the level of parental education. Access to more household income implies ability

to afford better resourced schools, investment in child health, transport costs, uniforms, private tuition etc. Furthermore, better educated parents are more likely to rank children education as one of the household priorities and therefore invest accordingly in it including choosing to settle in neighbourhoods with better schools. Better educated parents are also more likely to involve themselves directly in children education by assisting with homework and participating in school management (van den Berg and Louw, 1984).

It is commonly accepted in the literature that a student from a lower SES background tends to have lower educational aspirations and attainment. Furthermore, they are more likely to drop out than their counterparts from higher SES backgrounds. Parents from higher SES backgrounds tend to send their children to better schools which in turn improves their chances of success later at tertiary level. Parental expectations and the definition of "success" in wider society which are also inextricably linked to SES background, have also been found to have substantial impact on student persistence and attainment (van den Berg and Louw, 1984). The last two factors are referred to as the cultural capital factors according to Bourdieu's *capital cultural* model (Bourdieu and Passeron, 1977).

Driesden (2001) argues that the combination of the economic capital and the less obvious cultural capital tend to perpetuate and reproduce existing social stratification. Consequently, students from higher SES backgrounds tend to generally outperform their counterparts from poorer SES backgrounds in terms of persistence and attainment. These cultural factors are "specialized insider knowledge" which are not taught in schools but are passed down from one generation to the next.

Previously disadvantaged communities have steadily migrated to previously White suburbs and began sending their children to better schools in these suburbs since 1994. The majority of African families, however, are still trapped in townships or rural poverty where their children attend poor quality schools in these communities and about 70%

of African families are classified as lower SES where parents or guardian are poorly educated or not educated at all (Letseka and Maile, 2008).

On grounds of the above, it is not unreasonable to regard race as proxy for SES and school quality.

Table 2.1 Average Graduation Times

	Category	n	%	\bar{x}	S.D
Race	African	2075	71	3.64	0.70
	Indian	585	19	3.60	0.74
	White	222	7.6	3.27	0.56
	Coloured	45	1.5	3.64	0.73
	Other	7	0.2	3.60	0.55
	Faculty	Engineering & Built Env.	678	23	4.06
Management Sciences		973	33	3.54	0.65
Health Sciences		223	8	3.53	0.67
Applied Sciences		276	9	3.60	0.71
Arts & Design		317	11	3.22	0.52
Accounting & Informatics		467	16	3.55	0.64
Gender	Males	1414	48	3.71	0.74
	Female	1520	52	3.51	0.67
Age	19 or Younger	1583	54	3.61	0.72
	Older than 19	1351	46	3.58	0.70

In Table 2.1 above, we have computed the average graduation time after excluding dropouts and the censored subjects. The Age variable has been discretized into two categories separated by the median age of 19.

African students take the longest to graduate with Whites taking the shortest time to

graduate on average. The Engineering & The Built Environment has the longest average graduation time with Arts & Design having the shortest average graduation time.

On average, males take longer to graduate than females and students younger than 19 take a little longer than students older than 19 to graduate on average.

In this chapter we discussed the data and the variables that we intend to investigate if they explain graduation, we also conducted a descriptive analysis of these variables.

Descriptive analysis indicate that all the variables may explain graduation. We did not conduct any statistical tests, we will perform that exercise in the next chapter where we will use *non-parametric* techniques to assess if indeed these variables do explain graduation.

Non-parametric Analysis

In the previous Chapter we established that there is a possibility that the hypothesised variables namely - race, age, faculty and gender may explain graduation. We did not conduct any statistical tests to confirm these results. In this Chapter we shall use graphical methods and conduct statistical tests to verify the results we obtained in Chapter 2. These graphical methods and the tests belong to what we refer to as *non-parametric* methods that are part of broader *Survival Analysis* methods. Before we use these techniques we shall introduce survival analysis in more concrete terms than our introductory discussion in Chapter 1.

In survival analysis, the variable of interest is the time until occurrence of an event of interest. Sometimes this variable might be referred to as *lifetime*, *survival time* or *failure time* (even though the event of interest might be a positive outcome such as graduation) and so forth, depending on the field of application.

In engineering, for example, interest might centre upon endurance or failure time of a component and a sample of items might be tested and observed until they fail so as to obtain their failure times. A group of patients who are diagnosed with a particular disease might have been subjected to a certain treatment, and in this instance, the

variable of interest might be the time until recovery.

Despite the differences in the units of analysis and the events of interest between the two situations described above, the variable of interest is *time until occurrence of an event* in both instances. Survival analysis techniques can be applied in both examples as long as the *time of origin*, *time metric* and the *event* are clearly defined, the actual nature of the event of interest is immaterial.

Ideally, to observe exact failure times, all subjects or units of analysis that are entering a study should be observed until they all fail, however due to constraints such as time, budget, impracticability etc. the data collection period is often fixed before hand. This period between the commencement and the close of observation is often referred to as the *observation period* or *follow-up* period. The latter term is inherited from medical research.

It is often the case that some of the subjects under study do not experience the event of interest during this observation period and we refer to the observations associated with these subjects as censored observations.

Thus, in the examples described above, some of the components may have not failed, likewise, some of the patients may have not recovered at the end of the observation period. In such instances, the true survival times i.e. the time between entry into observation and the event time, can not be known. The censored subjects only provide partial information about their true survival times. This type of censoring is referred to as *right censoring* where the missing information is to the right after entry into observation. There are other forms of censoring in survival analysis such as *left* and *interval* censoring, but we will restrict our attention to right censoring because this is the form of censoring that arises in this study.

Putting these concepts firmly within the context of our study, some of the students will not have graduated at the close of observation. These subjects would still be in pursuit of their studies at the end of the fifth year. Furthermore, some will withdraw or leave observation for various reasons that are not related to the event of interest such as financial affordability, death, voluntary withdrawal etc. All the observations associated with all the cases described above will not be complete observations, instead they will be referred to as right censored observations.

The time until occurrence of some specified event can be characterised by many functions in survival analysis. The most frequently used functions are the survival function, $S(t)$, which represents the probability that a subject survives until time t . The other function is the hazard function $h(t)$, which represents the rate of occurrence of the event of interest.

The estimation of the survival and the hazard functions using *non-parametric* techniques is the usual starting point in survival analysis. The appeal of these techniques lies in the fact that it is not necessary to assume any distributional form underlying the functions to be estimated, hence the name *distribution free* techniques.

We will also use non-parametric methods to build on the results obtained using descriptive statistics we presented in the previous chapter as well to provide a bridge to semi-parametric methods that we will consider in the next chapter. Before we apply non-parametric methods we need to introduce a few basic concepts that underpin survival analysis methods.

Suppose that T , a non-negative random variable, represents survival times of subjects in a hypothetical population with respect to some specified event. Let $f(t)$ and $F(t)$ represent both the corresponding probability density function and the distribution function respectively. The familiar relationship between $f(t)$ and $F(t)$ is given by

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

On the other hand, the *survival function* is given by

$$S(t) = P(T > t) = \int_t^\infty f(u) du = 1 - F(t).$$

The other function that we need an expression for is the *hazard rate*. Intuitively, consider a time interval $[t, t+dt)$ of length dt . The conditional probability that a subject fails in this interval given that the subject has survived up to the beginning of the interval is given by

$$P(t < T < t + dt | T > t) = \frac{F(t + dt) - F(t)}{S(t)}.$$

Dividing the above equation by dt we obtain the rate of failure over this infinitely small interval $[t, t+dt)$ and in the limit we obtain the hazard rate, which is an instantaneous rate, given by

$$h(t) = \lim_{dt \rightarrow 0} \frac{F(t + dt) - F(t)}{dtS(t)} = \frac{f(t)}{S(t)}.$$

The estimate of the probability that an event occurs in the infinitesimal interval $[t, t+dt)$ described above, conditional on not having occurred prior to time t is approximately $h(t)dt$. Whereas the unconditional probability that an event occurs in the same interval is approximately $f(t)dt$.

In the context of this study, $f(t)dt$, is the estimate of the proportion of the total sample that graduate by time t , whereas $h(t)dt$ is the estimate of the proportion of the total number in the sample that sat for the examination that graduate.

The term "risk" is more widely used than the term "hazard rate". Thus, a specialist surgeon would say, for example, the risk of dying from surgical complications after undergoing an operation is high immediately after surgery but drops a few weeks later. More formally, the surgeon is saying that the hazard rate of experiencing event "death" is high initially but it improves with time. That is; after surviving the first few critical days immediately or few days after surgery, the instantaneous rate of dying drops.

The hazard rate is not a true probability in continuous time because it can exceed one as opposed to discrete time setting where $0 \leq h(t) \leq 1$ and therefore a true probability.

In the previous chapter we introduced four covariates; *race*, *gender*, *faculty* and *age*, and based on preliminary results, there is a case for regarding these variables as possible covariates that explain graduation. To strengthen our argument concerning the four variables even further, we will move beyond descriptive methods and use The Kaplan-Meier techniques.

Our intention is to establish whether the four variables; gender, race, faculty and age do explain the graduation rate. We will compare the survivor functions of the categories within each variable and also conduct tests to confirm graphical findings.

Let $0 < t_1 < t_2, \dots < t_j$, be distinct ordered times at which events occur. Let d_k be the number of subjects that fail at time, t_k , out of a total of n_k subjects that are at risk at time t_k . Define $[t_{k-1}, t_k)$ as the k^{th} interval and also define the probability of surviving through the k^{th} interval conditional on not having experienced the event at the beginning of the interval as p_k . Also define, q_k , as the probability of experiencing the event during the k^{th} interval conditional on not having experienced the event at the beginning of the interval where $q_k = 1 - p_k$. Then the probability of surviving beyond

time t is given by

$$\hat{S}(t) = \prod_{t_k \leq t} p_k.$$

Since $\hat{q}_k = \frac{d_k}{n_k}$ then $\hat{p}_k = \left(\frac{n_k - d_k}{n_k}\right)$, the Kaplan-Meier survivor function estimate of $S(t)$ is given by

$$\hat{S}(t) = \prod_{t_k \leq t} \left(\frac{n_k - d_k}{n_k}\right).$$

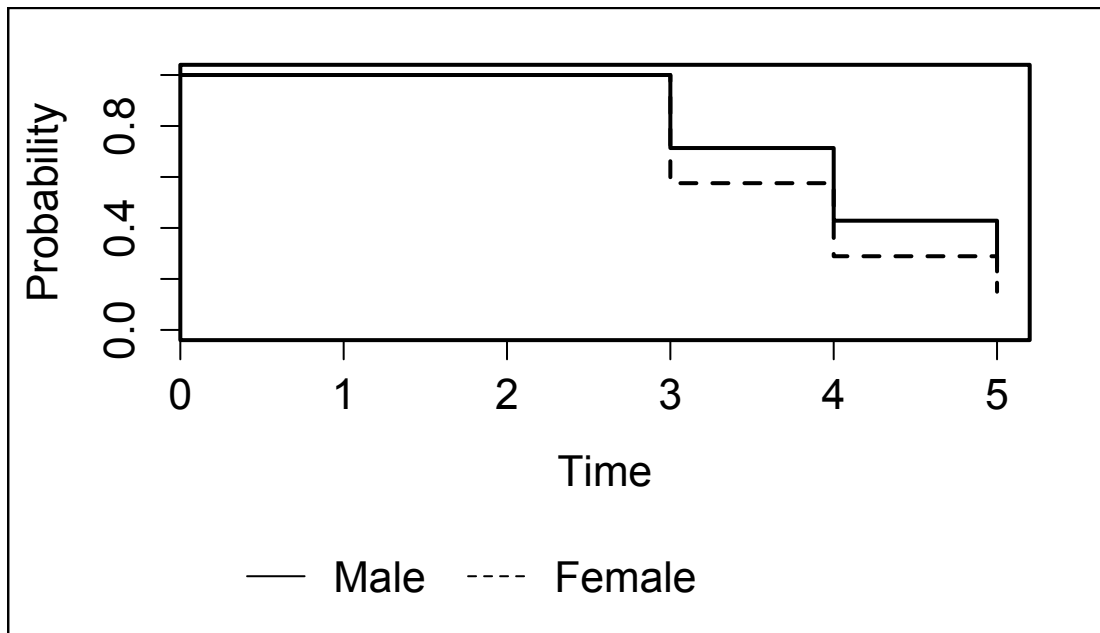


Figure 3.1 Gender Survivor Function

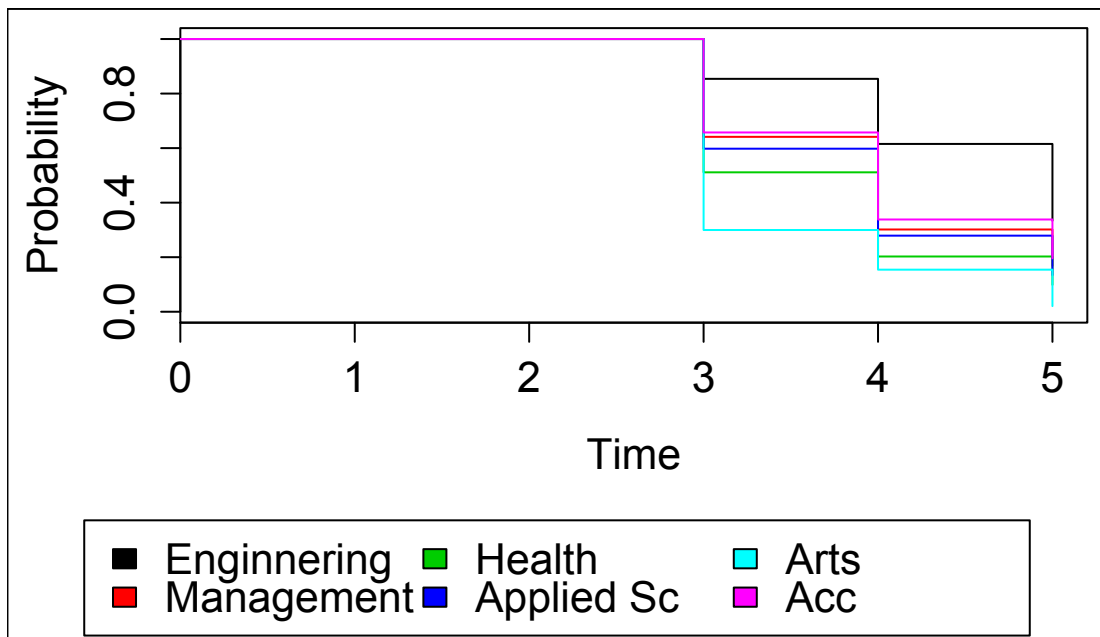


Figure 3.2 Faculty Survivor Function

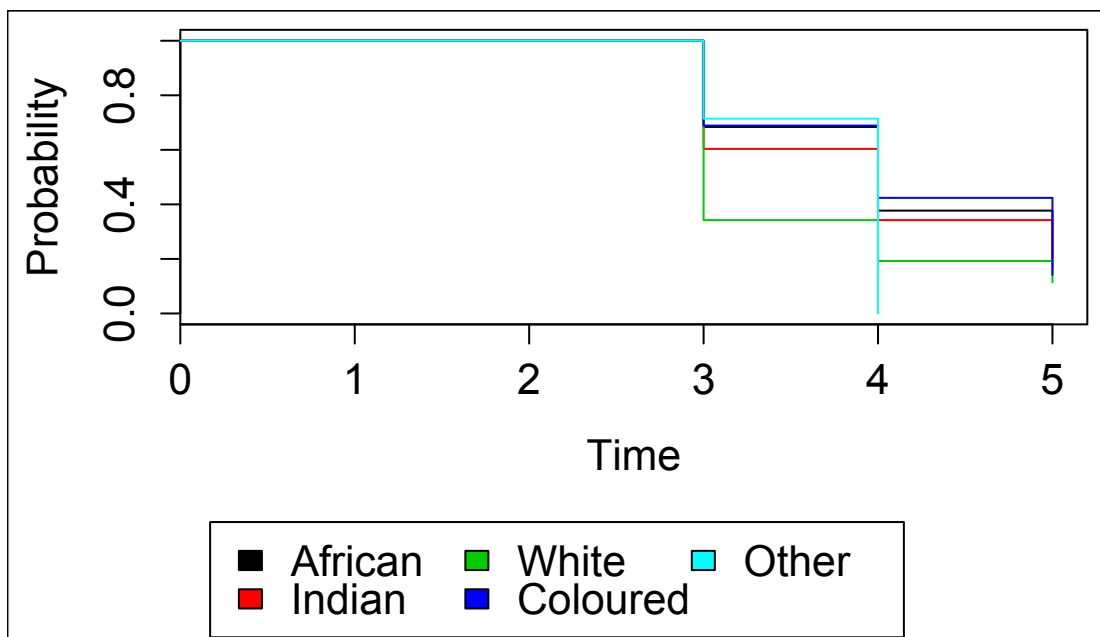


Figure 3.3 Race Survivor Function

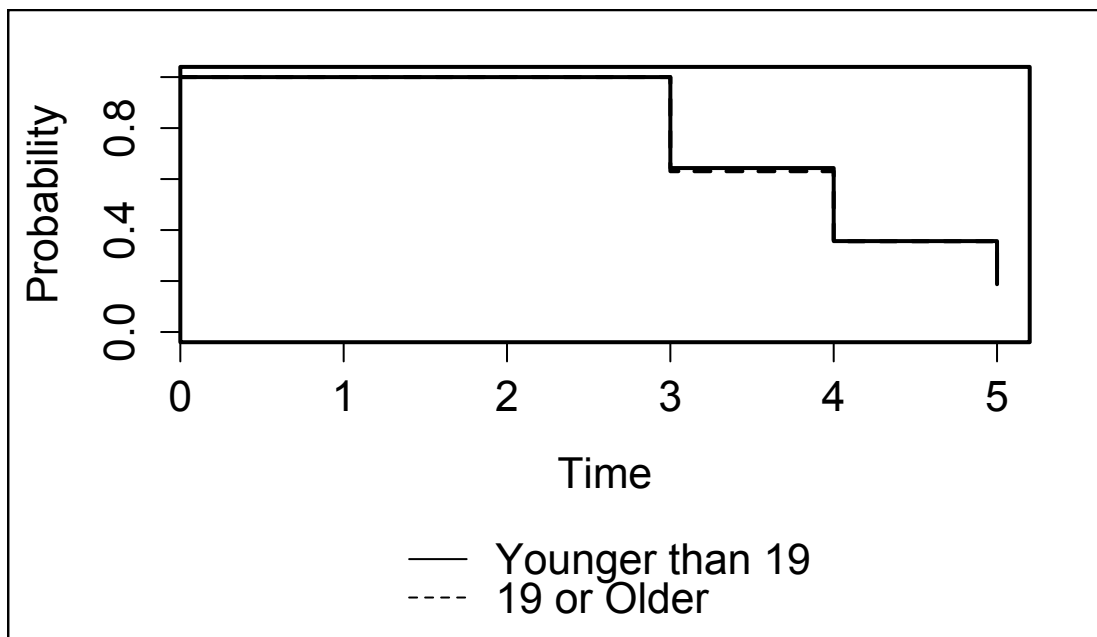


Figure 3.4 Age Survivor Function

We have plotted the survivor function in Figure 3.1 to Figure 3.4. We observe that the Engineering & The Built Environment faculty has the worst graduation survival rate and Arts & Design faculty has the best one. This implies that students in the Engineering & The Built Environment faculty take the longest to graduate compared to student in the Arts faculty who take shortest time. Females have a better survival rate than males, subjects older than the median age, 19, have a slightly better graduation survival rate than younger ones. Coloured and the Other race seem to have the worst graduation survival rate and Whites have the best one.

Suppose we have variable $Z = [z_1, z_2, \dots, z_J]'$ with J categories. To test the equality of survivor functions, $H_0 : S_{z_1} = S_{z_2} = \dots = S_{z_J} \Leftrightarrow h_{z_1} = h_{z_2} = \dots = h_{z_J}$.

Denote the distinct times of observed failures as $0 < t_1 < t_2, \dots < t_K$

Let:

- d_{kj} be the number of observed events from group j at time k
- Y_{kj} be the number of subjects in group k that are at risk at time k
- $d_k = \sum_{j=1}^J d_{kj}$
- $Y_k = \sum_{j=1}^J Y_{kj}$, and
- $W(t_k)$ be the weight at time k

To test the hypothesis above, a vector Z is computed with the following components

$$z_j = \sum_{k=1}^K W(t_k) [d_{kj} - Y_{kj} \frac{d_k}{Y_k}]$$

where K is the number of event times and $W(t_k)$ are the weights which depend on the test used. Define the variance of z_j , V_{jj} as

$$\hat{V}_{jj} = \sum_{k=1}^K W(t_k)^2 \frac{Y_{kj}}{Y_k} \left(1 - \frac{Y_{kj}}{Y_k}\right) \left(\frac{Y_k - d_k}{Y_k - 1}\right) d_k$$

for $j = 1, \dots, J$, and the covariance of z_j and z_g , V_{jg} as

$$\hat{V}_{jg} = \sum_{k=1}^K W(t_k)^2 \frac{Y_{kj}}{Y_k} \frac{Y_{kg}}{Y_k} \left(\frac{Y_k - d_k}{Y_k - 1}\right) d_k.$$

For $g \neq j$.

\hat{V}_{jj} and \hat{V}_{jg} form the components of V , the variance-covariance matrix of $Z = (z_1, z_2, \dots, z_{J-1})$. $ZV^{-1}Z^t$, the test statistic follows a χ^2 follows a chi-squared distribution when the null hypothesis is true with $J - 1$ degrees of freedom (Klein and Moeschberger, 2003). When $W(t_k) = 1$, we have what is referred to as the *Log-rank* test.

The proportion that experiences the event at time k from group j is $\frac{d_{kj}}{Y_{kj}}$, whereas it is $\frac{d_k}{Y_k}$ from the study population and these should not be very different when the null hypothesis is true. i.e.

$$\frac{d_{kj}}{Y_{kj}} = \frac{d_k}{Y_k},$$

or

$$d_{kj} = Y_{kj} \frac{d_k}{Y_k}.$$

Thus, the test statistic should not be significantly different from zero if the null hypothesis is true. We have presented the results of our tests in Table 3.1 below:

In all cases, the survivor functions are not equal except for age, therefore, all the hypothesised variables explain graduation except for age. We must, however, bear in mind

Table 3.1 Log-Rank Test for Equality of Survivor Functions

Variable	Graduation		
	χ^2	DF	<i>P</i> -value
Gender	60.2	1	8.33×10^{-15}
Faculty	307	5	0.00
Race	93.3	4	0.000
Age	1.6	3	0.671

that we have discretized a continuous variable and the choice of cut points will definitely have effect on the test with regard to "age" variable.

We have noted that different race, faculty and gender groups have different graduation survival rate and that there is no difference in survival rate between the two age groups.

Although non-parametric methods can indicate whether the survival potential of two or more groups are equal or not by visual inspection, as well as by conducting tests, it must be noted their usefulness is limited by their inability to deal adequately with certain types of covariates. In our study, for example, the method worked very well with gender, race and faculty because they each had fewer categories, which in turn gave rise to fewer survivor functions to compare. The factors partition the sample into fewer subpopulations, but this does not hold true for a continuous variable such as age, where we resorted to discretizing age into two groups (younger than 19; 19 and older) and thereby discarding information by collapsing the continuous scale to a dichotomous variable.

The other obvious disadvantage of non-parametric methods is that even when all covariates are factors and with even fewer categories, we still have to consider each factor

separately. Another shortcoming is that the tests associated with the methods can indicate whether survival functions are the same or different but if the conclusion is that they are different the methods do not quantify the difference.

Ideally, we would want a model that regresses the time to event with all the covariates simultaneously as well as be able to quantify the differences between survival experiences if they are found to exist. The models that will be introduced in the following chapters will attempt to overcome all these shortcomings of non-parametric techniques. The first model to be considered is the Cox's regression model.

The Cox Regression Model

The advantage of non-parametric methods is that in modelling time to event it is not necessary to assume any distribution and thereby the costly risk of misspecification is easily avoided. However, and as it is often the case in practice, there might be risk factors or covariates that have influence on time to event and it might be of interest to quantify their effect. Ideally, that objective would be achieved by regressing the time to event on these covariates as in classical regression methods. The piecemeal fashion of estimation in non-parametric methods can easily become cumbersome and even more so when some of the variables are continuous.

Cox (1972) proposed one such model that regresses the hazard function on all covariates simultaneously, and it is given by

$$h(t, \mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}).$$

The above equation has since become famously known as Cox's regression or semiparametric regression. The covariates enter the model through the vector $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ with a corresponding vector of coefficients $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_p)$. In essence, the hazard is

broken down into two parts - (a) the baseline hazard $h_0(t)$, a function of time that is left unspecified (*non-parametric*) (b) the covariate effect $\exp(\boldsymbol{\beta}'\mathbf{x})$ (*parametric*), which in its simplest term does not include term t . The equation therefore consists of a parametric and a non-parametric part, and hence the name *semi-parametric*.

Let us suppose that there are two individuals at some time t with the following fixed covariates \mathbf{x} and \mathbf{x}^* respectively, then the ratio of their hazards is given by

$$\frac{h(t, \mathbf{x})}{h(t, \mathbf{x}^*)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}^*)} = \exp[\boldsymbol{\beta}'(\mathbf{x} - \mathbf{x}^*)].$$

Since the covariates are fixed, $\exp[\boldsymbol{\beta}'(\mathbf{x} - \mathbf{x}^*)]$ is also fixed, implying that the ratio of the two hazards is constant regardless of time t . The hazard functions are proportional with $\exp[\boldsymbol{\beta}'(\mathbf{x} - \mathbf{x}^*)]$ as the proportionality constant, hence the nomenclature "Cox's *proportional hazard model*" or Cox's PH model.

The hazard rate of an individual with $\mathbf{x} = \mathbf{0}$ is $h_0(t)$, which is a hazard rate of an individual for whom all the covariates that make up the vector \mathbf{x} are set at zero. The individuals with $\mathbf{x} = \mathbf{0}$ are often referred to as the "reference group". In general, the hazard rate of the i^{th} individual is given by

$$h_i(t) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i).$$

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

$$\frac{h_i(t)}{h_0(t)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{h_0(t)} = \exp(\boldsymbol{\beta}'\mathbf{x}_i).$$

Taking logarithms of both sides in the above equation we obtain

$$\log \frac{h_i(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

and thus the logarithm of the relative risk is linear in β 's.

4.1 Model

Cox (1972) showed that the likelihood function to be maximised in order to obtain the estimates of β is given by

$$L(\beta) = \prod_{j=1}^r \frac{\exp(\beta' \mathbf{x}_j)}{\sum_{\ell \in R(t_j)} \exp(\beta' \mathbf{x}_\ell)}. \quad (4.1.1)$$

$R(t_j)$, is the risk set at time t_j , i.e. the number of subjects that are "alive" immediately before "death" time t_j , and r is the total number of events. Censored times do not contribute to Equation 4.1.1. Even event times themselves do not contribute to Equation 4.1.1, directly as in a conventional likelihood function because the baseline hazards, the function of time, cancel out. In a sample of size n , with survival times t_1, t_2, \dots, t_n , and an indicator variable δ_i which is one when t_i is an event time or zero otherwise, Equation 4.1.1 can be written as (Collet, 2003)

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta' \mathbf{x}_i)}{\sum_{\ell \in R(t_i)} \exp(\beta' \mathbf{x}_\ell)} \right\}^{\delta_i}.$$

The above equation is maximised using iterative methods to obtain the estimate of β .

The partial likelihood described in Equation 4.1.1 is based on the assumption that there are no ties, a situation that rarely obtains in practice. There are quite a few approximations to the partial likelihood when there are ties and the most widely used is the Breslow estimation given by

$$L(\beta) = \prod_{j=1}^r \frac{\exp(\beta' \mathbf{s}_j)}{\{\sum_{\ell \in R(t_j)} \exp(\beta' \mathbf{x}_\ell)\}^{d_j}}$$

where \mathbf{s}_j is the sum of the p covariates for the subjects that experience the event at time t_j , and d_j is the number of those subjects.

To conduct hypothesis tests using the maximum likelihood estimates of the effect of *race*, *gender*, *faculty* and *age*, we require reference categories. We observed from the previous chapter that Africans, males and Engineering & the Built Environment are the worst performing categories within race, gender and faculty variables respectively. It is common practice to treat the worst performing group i.e. the "African males in Engineering & The Built Environment" as the reference group (Klein and Moeschbecker, 2003).

In order to build the best reduced model, we shall use the *backward elimination* as explained in (Collet, 2003), which entails fitting all the variables and then excluding one variable at a time where the variable excluded is the one that increases $-2 \log L(\hat{\beta})$ by the least amount, this process ends when the next variable excluded increases $-2 \log L(\hat{\beta})$ by more than say 5% or some other specified value.

We have fitted all variables, including 2-way interaction terms, in Model A. Using the backward elimination, we moved from Model A to a reduced model, Model B. The results

are listed in Table 4.1.

Table 4.1 Full & Reduced Model

Variable	Model A		Model B		
	$\hat{\beta}$	Sig.	$\hat{\beta}$	Sig.	$\exp \hat{\beta}$
Female	-0.378	0.268	0.088	0.097	1.092
Management Sciences	0.680	0.000	0.573	0.000	1.773
Health Sciences	0.872	0.000	0.691	0.000	1.996
Applied Sciences	0.849	0.015	0.591	0.000	1.806
Accounting & Informatics	0.941	0.100	0.539	0.000	1.714
Arts & Design	1.342	0.004	1.001	0.000	2.720
Coloured	-0.152	0.561	0.056	0.771	1.057
Indian	0.111	0.214	0.022	0.791	1.022
Other	0.389	0.409	0.588	0.190	1.801
White	0.287	0.096	0.390	0.000	1.477
Age	-0.009	0.620	-	-	-
Female \times Faculty	0.005	0.871	-	-	-
Female \times Race	0.182	0.564	-	-	-
Female \times Age	0.012	0.433	-	-	-
Faculty \times Race	-0.011	0.592	-	-	-
Faculty \times Age	-0.003	0.529	-	-	-

We have included $\exp \hat{\beta}$ for Model B which are essentially ratios of hazards.

Since the ratio of hazards are given as follows

$$\frac{h(t, \mathbf{x})}{h(t, \mathbf{x}^*)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}^*)} = \exp[\boldsymbol{\beta}'(\mathbf{x} - \mathbf{x}^*)],$$

and $\mathbf{x}^* = 0$ for African males in Engineering & the Built Environment, the reference

category, then the ratio of their hazards is $\exp \beta'(\boldsymbol{x}) = 1.09$.

The risk of graduation for African females in Engineering & the Built Environment is therefore about 9% higher than for African males in the same faculty provided the *proportionality assumption* holds true.

4.2 Model Adequacy

Assessment of the survival model adequacy can be divided into five steps. The first step is the verification of the statistical significance of covariates, which is already conducted in the previous section. The Cox proportional hazard model specific assumptions are then examined, namely; the linear effect of covariates on the logarithm of hazards and the proportionality assumption. The other two steps are the identification of poor fit and influential observations, which is then followed by the assessment of the overall goodness-of-fit.

4.2.1 The Proportionality Assumption

There are number of ways in which the proportionality assumption can be verified, the techniques roughly fall into either graphical or formal hypothesis testing. The graphical methods are for example plots based on *Schoenfeld residuals* and *Stratified Cox regression*. The test attributed to Grambsch and Therneau (1994) is one of the most widely used "objective" technique to assess the proportionality of hazards assumption.

In the previous chapter we established the following:

$$h(t) = \frac{f(t)}{S(t)}$$

which can also be written as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t).$$

Therefore

$$S(t) = \exp\{-H(t)\}$$

where

$$H(t) = \int_0^t h(s) ds.$$

$H(t)$ is referred to as the cumulative hazard.

Briefly, stratification entails modelling $h_{ij}(t)$ instead of $h_i(t)$, which is the hazard rate of the i^{th} subject in the j^{th} stratum and the base line hazard rate of the j^{th} stratum is now $h_{0j}(t)$. The relationship is given by

$$h_{ij}(t) = h_{0j}(t) \exp(\boldsymbol{\beta}' \mathbf{x}_{ij}).$$

All the other variables are assumed to satisfy the proportional hazard assumption and the regression coefficients are the same for all strata with only the baseline hazards varying from one stratum to the next.

If the proportionality assumption is valid then the baseline cumulative hazards, $H_0(t)$, within the strata should be constant multiple of each other at each time point. If the variable under consideration has say K , strata, the plots of $\log[\hat{H}_{i0}(t)]$, for $i = 1, \dots, K$, against time should be approximately parallel when the proportionality assumption

holds. If the variable under consideration is a continuous variables, it must be discretized into some suitable K categories (Klein and Moeschbeger, 2003).

An objective test by Grambsch and Therneau (1994) is based on residuals proposed by Schoenfeld (1982). Schoenfeld residuals are the differences between the covariates of those subjects that experienced the event and their expected values given by

$$r_i = x_{ji} - \hat{a}_{ji}$$

where

$$\hat{a}_{ji} = \frac{\sum_{k \in \mathcal{R}(t_i)} x_{jk} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_k)}{\sum_{\ell \in \mathcal{R}(t_i)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_\ell)}$$

is the expected value of the j^{th} covariate over all the subjects in the risk set $R(t_i)$, at time t_i . These residuals are calculated for all covariates. The vector \mathbf{r}_i of Schoenfeld residuals is then scaled to obtain

$$\mathbf{r}_i^* = r \text{ var}(\hat{\boldsymbol{\beta}}) \mathbf{r}_i$$

which is referred to as the scaled Schoenfeld residuals, where $\text{var}(\hat{\boldsymbol{\beta}})$ is a variance-covariance matrix and r is the number of uncensored subjects . If the j^{th} covariate has time-varying effect, then the coefficient of the j^{th} covariate can be expressed as

$$\beta_j(t_i) = \beta_j + \rho g(t_i)$$

where $g(t_i)$ is some function of time. Grambsch and Therneau (1994) showed that $E(r_{ji}^*) \approx \beta_j(t_i) - \hat{\beta}_j$, and therefore a plot of the values of the residuals against time or a function thereof will reveal if the j^{th} covariate has time-varying effect or not. A horizontal

line will suggest that the coefficient of the j^{th} covariate is constant and therefore not time-varying. Furthermore, fitting a least squares regression will provide an estimate of ρ which is used to evaluate the hypothesis: $H_0 : \rho = 0$, (implying that the proportionality assumption holds when the null hypothesis is not rejected) (Grambsch and Therneau, 1994).

We have listed the results of Grambsch and Therneau (1994) test for Model B in Table 4.2.

Table 4.2 Proportionality Test

	Variable	$\hat{\rho}$	Sig.
	Female	-0.0092	< .001
	Management Sciences	-0.1020	< .001
	Health Sciences	-0.0810	< .001
	Applied Sciences	-0.0764	< .001
	Accounting & Informatics	-0.0959	< .001
	Arts & Design	-0.0964	< .001
	Indian	-0.0226	< .001
	White	-0.0521	< .001
	Coloured	-0.0076	< .001
	Other	0.0115	< .001

Clearly, none of the variables satisfy the proportionality assumption and on the bases of the above results we proceed to fit a *stratified Cox proportional hazards* model where we stratify on the faculty variable.

In Figure 4.1 we have plotted log cumulative baseline hazards where we have stratified by faculty.

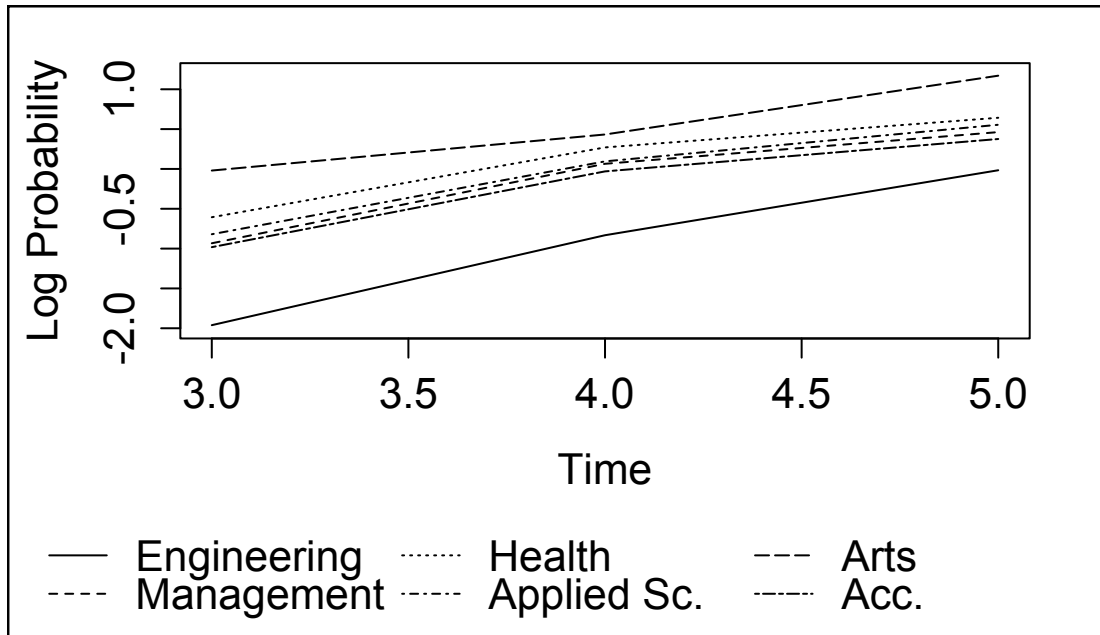


Figure 4.1 Log Cumulative Hazards by Faculty

There is a suggestion that the log cumulative baseline hazards are parallel. We confirm these graphical findings by performing the Grambsch and Therneau (1994) test on this model where we have stratified on the faculty variable, with results as listed in Table 4.3. All variables in the stratified model satisfy the proportionality assumption except for the "White" race category.

Table 4.3 Proportionality Test for the Stratified Model

	Variable	$\hat{\rho}$	Sig.
	Female	-0.03064	0.1679
	Indian	-0.03068	0.1726
	White	-0.0572	0.0116
	Coloured	0.00621	0.7832
	Other	0.0148	0.5089

Based on the above results, our final model will therefore be Model B stratified on the

Table 4.4 Stratified Cox Model

	Variable	$\hat{\beta}$	S.E	Sig.
	Female	0.1471	0.0478	0.0021
	Indian	0.1718	0.0575	0.0028
	White	0.3473	0.0869	< 0.001
	Coloured	0.0439	0.1918	0.8189
	Other	0.6541	0.4489	0.1451

faculty variable. The results of our stratified and final model are presented in Table 4.4. Only the "Female" variable and "Indian" factor level satisfy the proportionality assumption. The "Coloured" and "Other" factor levels are insignificant, whereas the "White" factor level, though significant, but it fails the proportionality test.

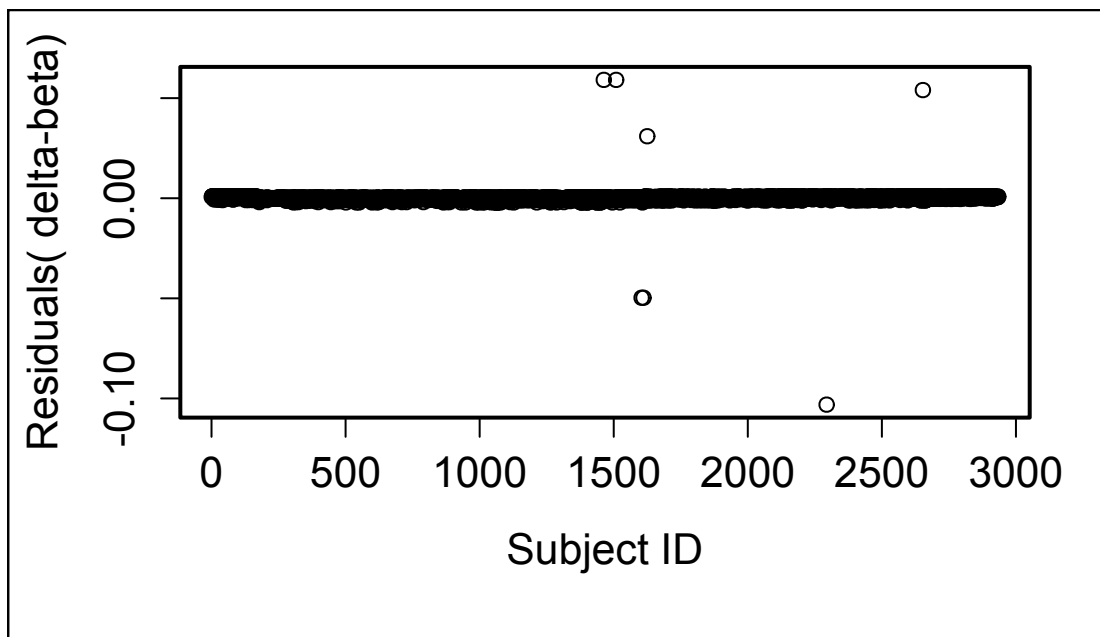


Figure 4.2 Female Schoenfeld Residuals

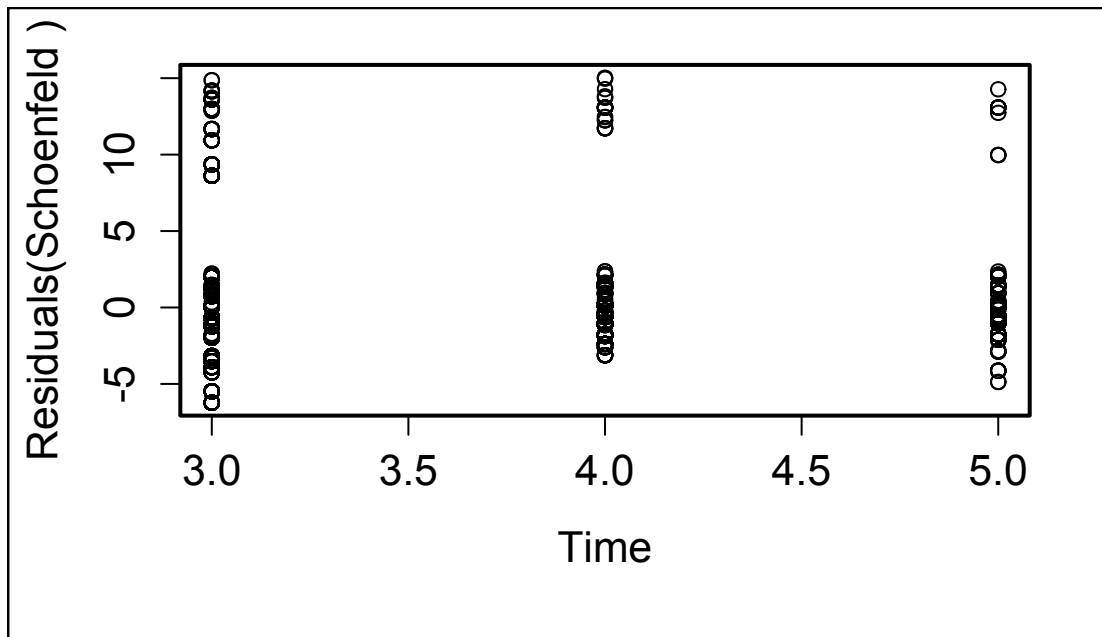


Figure 4.3 White Schoenfeld Residuals

As proposed by Grambsch and Therneau (1994), to confirm our findings in Table 4.3 and to save space, we have plotted the Schoenfeld (1982) *residuals* against time only for the "White" factor level in Figure 4.3 and "Female" factor level in Figure 4.2 . We note that the residuals have a pattern with regards to the "White" factor level, whereas, there seems to be no pronounced pattern in relation to the "Females" factor level. The "Females" factor level satisfies the proportionality assumption whereas the "White" factor level does not.

We will now proceed to assess the other aspects of model adequacy in relation to this final model. We will only consider the question of outliers and influential observations because all the variables in the model are dichotomous and therefore linearity is not a concern.

4.2.2 Outliers

We will use the *deviance residuals* to detect the outliers. Deviance residuals are the modification of *martingale residuals* which in turn depend on *Cox-Snell residuals* . Cox-Snell residuals are given by

$$r_{Ci} = \exp(\hat{\beta}' \mathbf{x}_i) \hat{H}_0(t_i) = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i).$$

The martingale residuals are given by

$$r_{Mi} = \delta_i - r_{Ci} = \delta_i - \hat{H}_i(t_i).$$

Martingale residuals can be viewed as the difference between the observed number of

deaths δ_i , for the i^{th} subject in the interval $(0, t_i)$ and the expected number of deaths $\hat{H}_i(t_i)$ according to the model. These residuals are analogous to the residuals we encounter in other areas such as classical regression.

Martingale residuals are not symmetrical, Therneau et al. (1990) introduced *deviance* residuals, based on martingale residuals, which are more symmetrically distributed around zero and given by

$$r_{Di} = \text{sign}(r_{Mi})[-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}]^{\frac{1}{2}}$$

The deviance residuals are related to the more familiar maximum likelihood based deviance: $D = -2\{\log \hat{L}_c - \log \hat{L}_N\}$, the relationship is given by $D = \sum r_{Di}^2$.

Since these residuals are approximately *Gaussian*, $N(0, 1)$, we can think of outliers as values outside say $(-3.5, 3.5)$ or even $(-2.5, 2.5)$, if we are more stringent. An index plot will reveal poorly fitting subjects, whereas plotting the residuals against time or covariates highlights poorly fitting times or covariates respectively. We will not consider the plot of deviance *vs* risk scores because, as Singer and Willet (2003) argue, these plots are more suited to continuous data. The authors also caution about symmetry of the residuals when there exist heavy censoring (more than 40%), our data has a rate of about 30%. Figure 4.4 depicts an index plot.

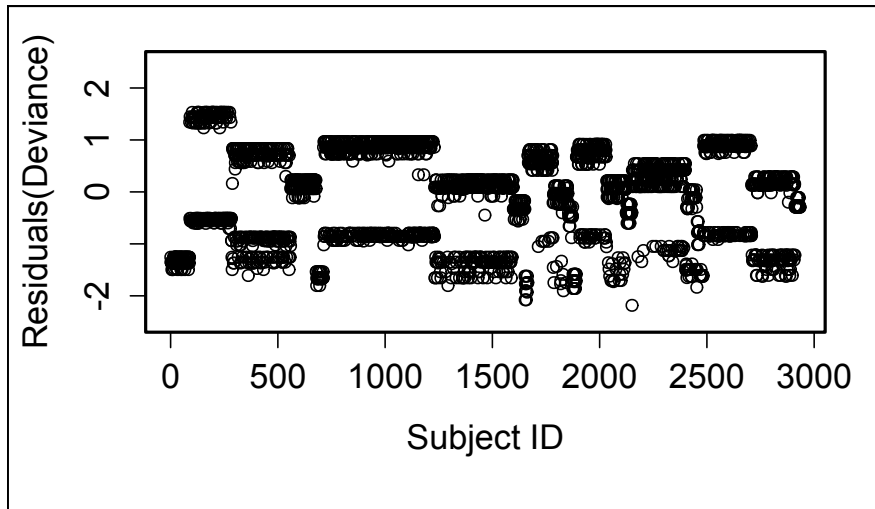


Figure 4.4 Deviance Residuals

The plot reveals that all residuals are within ± 2.5 standard deviations limits which implies that there are no outliers.

4.2.3 Influential Observations

This exercise is to identify the observation(s) that have extraordinary effect on the coefficients. This can occur either because the observation of the j^{th} covariate is an outlier, or because the observation is unusual, in the sense that it does not seem to fit in with other observations of that covariate. The observation of the j^{th} covariate is said to be influential if omission or inclusion thereof has a significant impact on $\hat{\beta}_j$. One way of identifying the observation is to fit all the n observations to obtain $\hat{\beta}_j$ and then fit the same model without the observation to obtain $\hat{\beta}_{j(i)}$ and then assess $\hat{\beta}_j - \hat{\beta}_{j(i)}$. This exercise has to be repeated for $i = 1, 2, \dots, n$ and for all $\hat{\beta}_j$, $j = 1, 2, \dots, p$, which then becomes computationally prohibitive.

The estimate of $\hat{\beta}_j - \hat{\beta}_{j(i)}$ is based on *score residuals*. The score residuals are related to

Schoenfeld residuals, for the i^{th} subject and the j^{th} covariate, they are given by

$$r_{Sji} = \delta_i(x_{ji} - \hat{a}_{ji}) + \exp(\hat{\beta}' \mathbf{x}_i) \sum_{t_r \leq t_i} \frac{(\hat{a}_{jr} - x_{ji}) \delta_r}{\sum_{\ell \in R(t_r)} \exp(\hat{\beta}' \mathbf{x}_i \ell)}. \quad (4.2.1)$$

The first term of Equation 4.2.1 is the Schoenfeld residual. letting $\mathbf{r}'_{Si} = (r_{S1i}, r_{S1i}, \dots, r_{Spi})$ where r_{Sji} is given by Equation 4.2.1, then the approximation of $\hat{\beta}_j - \hat{\beta}_{j(i)}$, also referred to as *delta-beta*, is given by

$$\hat{\beta}_j - \hat{\beta}_{j(i)} \approx \mathbf{r}'_{Si} \text{var}(\hat{\beta}).$$

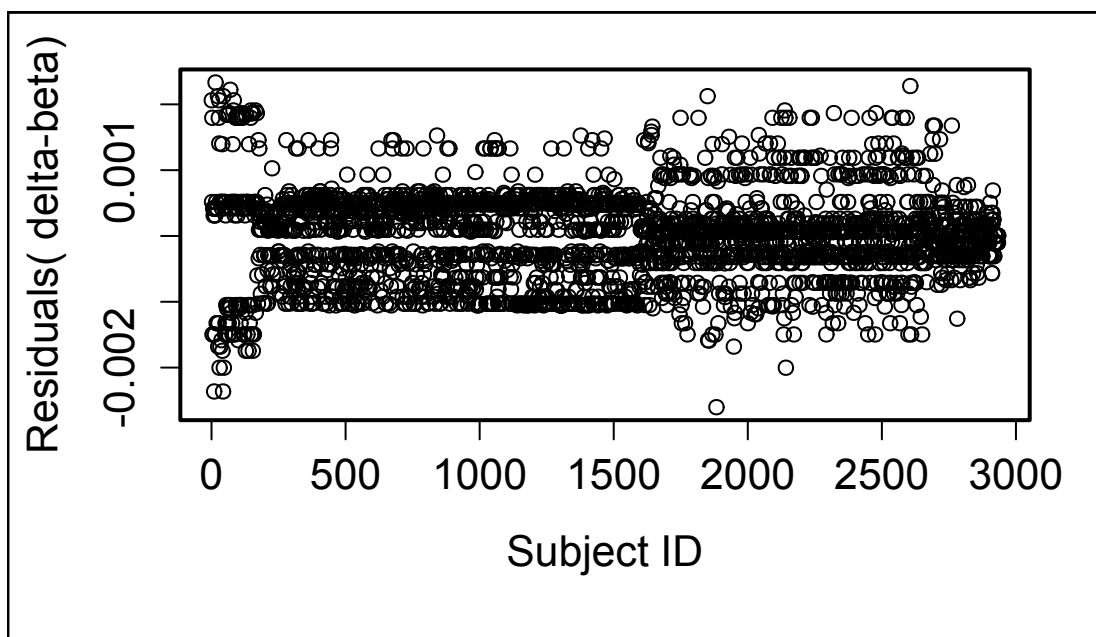


Figure 4.5 Female Delta-Beta

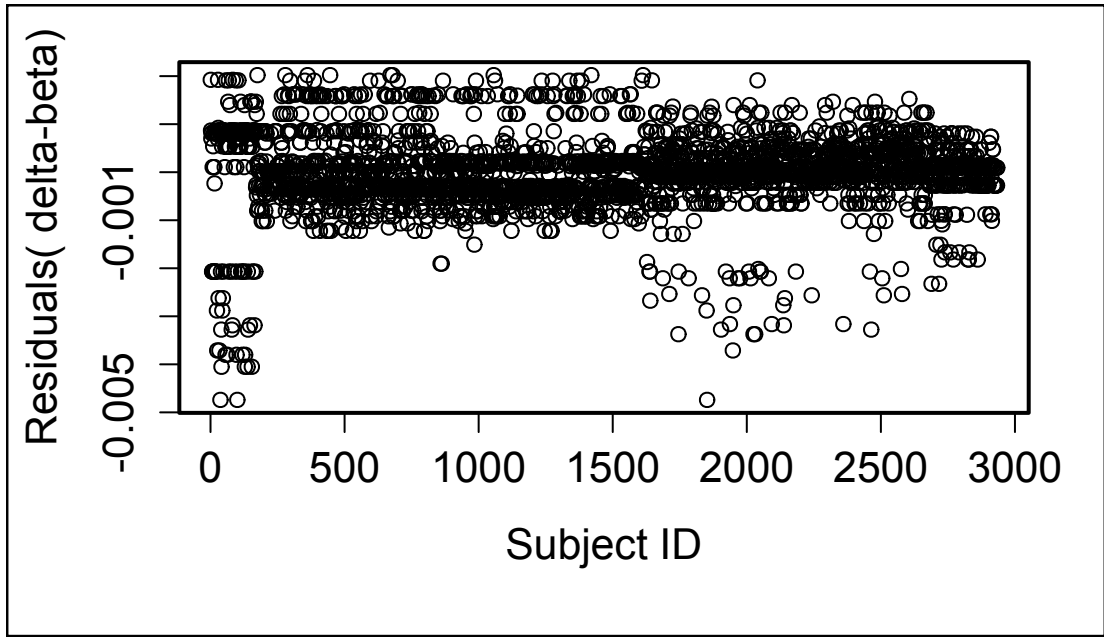


Figure 4.6 Indian Delta-Beta

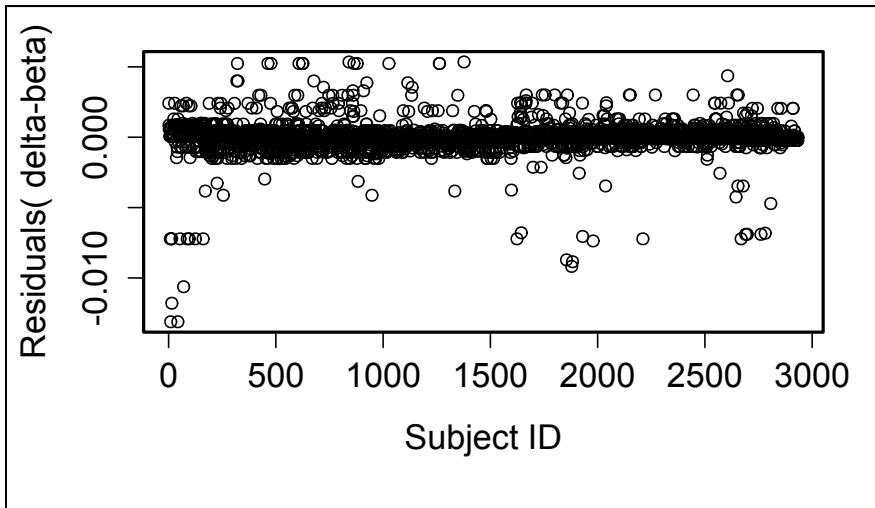


Figure 4.7 White Delta-Beta

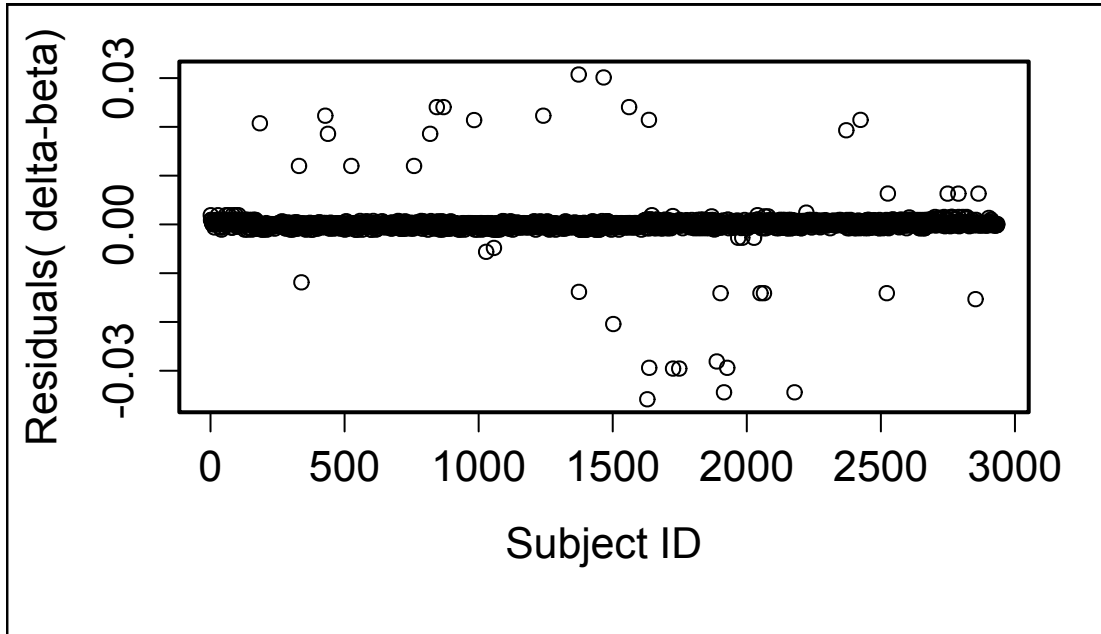


Figure 4.8 Coloured Delta-Beta

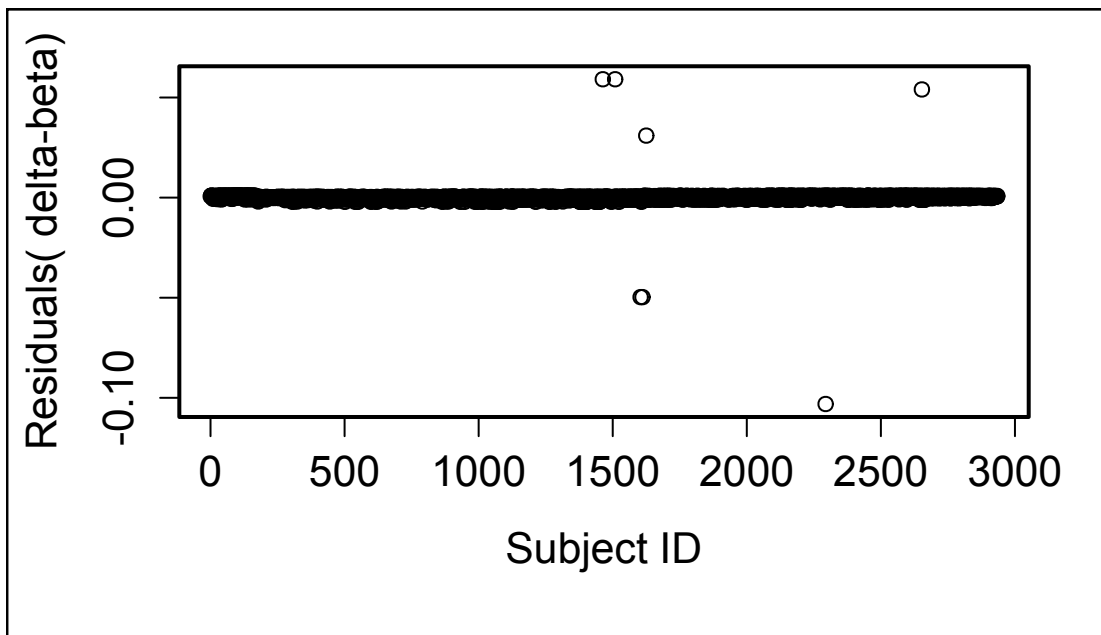


Figure 4.9 Other Delta-Beta

In Figure 4.5 to Figure 4.9, we have plotted delta-beta's for "Female", "Indian" and "White" factor levels, as well as delta-beta's for "Coloured" and "Other" factor levels. There are no noticeable outliers for all these factor levels, with the exception of the "White" and the "Other" plots. There are four *delta-beta*'s that do not seem to belong with the rest of the residuals in relation to the "White" factor level and 3 for the "Other" factor level, these delta-beta's are smaller than -0.01 . These subjects are listed in Table 4.5. For the "White" factor level, subjects ID=10 and ID=43 share a similar delta-beta which is also the largest in absolute terms. The actual difference between the coefficients including and omitting these two subjects is about 15% of $\text{var}(\hat{\beta})$ corresponding to the factor level "White". For the "Other" factor level, the largest difference corresponds to ID=2294, which is about 31% of $\text{var}(\hat{\beta})$. ID=2294 delta-beta is quite far from the rest of the other subjects, and since there are only 7 subjects it is not surprising that it has such inordinate influence on the coefficient estimate. Both variables are insignificant in the final stratified model.

Table 4.5 Influential Subjects

ID	Faculty	Gender	Race	Time	<i>delta-beta</i>	$\hat{\beta}_j$	$\hat{\beta}_{j(i)}$	$\hat{\beta}_j - \hat{\beta}_{j(i)}$
10	Health	Female	White	Year 5	-0.013114	0.347259	0.360719	-0.013460
15	Arts	Male	White	Year 5	-0.011794	0.347259	0.359593	-0.012334
43	Health	Female	White	Year 5	-0.013114	0.347259	0.360719	-0.013460
70	Health	Male	White	Year 5	-0.010624	0.347259	0.358124	-0.010865
1604	Engineering	Male	Other	Year 3	-0.049700	0.654091	0.705172	-0.051081
1611	Engineering	Male	Other	Year 3	-0.051204	0.654091	0.705172	-0.051081
2294	Engineering	Male	Other	Year 4	-0.103091	0.654091	0.792570	-0.138479

4.2.4 Overall Fit

The *Cox-Snell residuals* are used to assess the overall goodness-of-fit for the Cox regression model. These residuals were discussed when the *martingale residuals* were introduced i.e. $r_{Ci} = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i)$. It is shown in Collet (2003) that $-\log \hat{S}_i(t_i)$ follows an exponential distribution irrespective of the form of $S(t)$. Furthermore, if the model fits the data well, then the plot of the estimated cumulative hazard against the residuals should be a straight line through the origin with a slope of 1. A plot of estimated cumulative hazard against the Cox-Snell residuals is given in Figure 4.10.

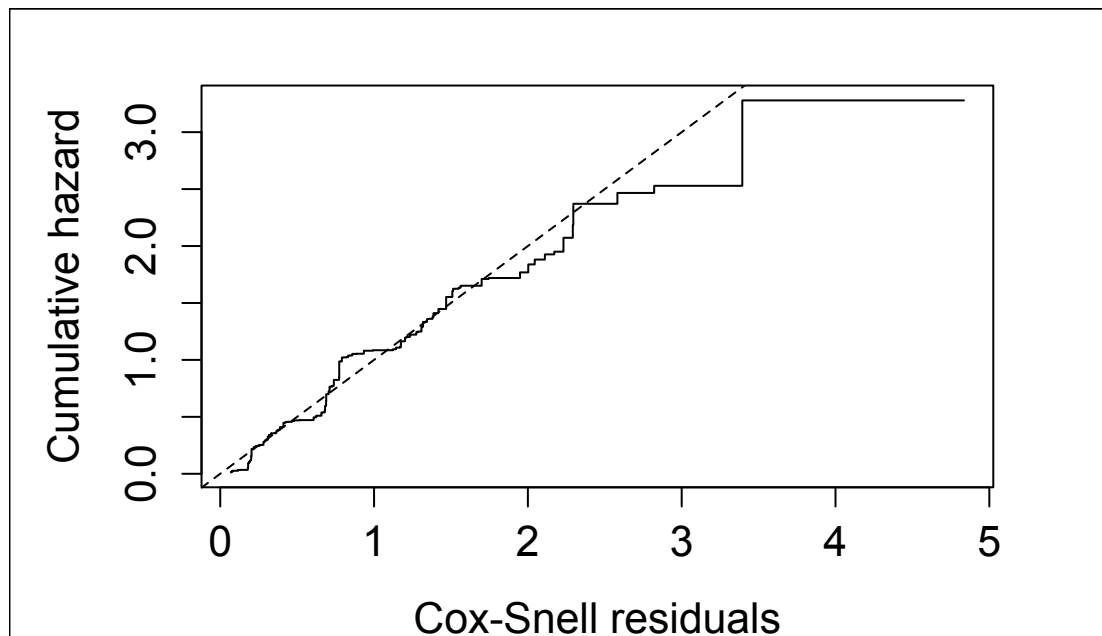


Figure 4.10 Cox-Snell Residuals

There is very little departure of residuals from a straight line through the origin with a slope of 1. Even if there was readily obvious departure, as Collet (2003) argues, this fact alone should not be the bases for discarding a model on grounds of poor lack-of-fit.

4.3 Prediction

Table 4.6 Stratified Cox Model

Variable	$\hat{\beta}$	S.E	Sig.
Female	0.1471	0.0478	0.0021
Indian	0.1718	0.0575	0.0028
White	0.3473	0.0869	< 0.001
Coloured	0.0439	0.1918	0.8189
Other	0.6541	0.4489	0.1451

We have re-printed the results of the stratified Cox model in Table 4.6. Race was found to be significant, however, the "White" factor level did not satisfy the proportionality test and it also insignificant in the final stratified model. The "Other" factor level will also be left out from interpretation of the results due to unstable estimates as suggested by delta-beta's. Gender was found to be significant and the variable also satisfied the proportionality test.

On the bases of the above model, the "risk" of graduation is about 16% ($\exp(0.147) = 1.158$) higher for females compared to males for all faculties and identical racial group. Also, the "risk" is 19% ($\exp(0.1718) = 1.187$) higher for Indians compared to "Africans" for identical gender category and all faculties .

We now proceed to plot faculty survivor functions on the bases of the stratified Cox regression model at average values of the variables. The relationship between the survivor function and the coefficients in Equation 4.3.1

$$\hat{h}_i = \hat{h}_0 \exp\{0.1471\text{Female}_i + 0.1718\text{Indian}_i + 0.3473\text{White}_i + 0.0439\text{Coloured}_i + 0.6541\text{Other}_i\}$$

(4.3.1)

,

is given by

$$\hat{S}_i(t) = \{\hat{S}_0(t)\}^{\exp\{\hat{h}_i\}},$$

The approximation of $\hat{S}_0(t)$ when there are ties is given by

$$\hat{S}_0(t) = \prod_{j=1}^k \left\{ 1 - \frac{d_j}{\sum_{\ell \in R(t_j)} \exp(\hat{\beta}' \mathbf{x}_\ell)} \right\}$$

where d_j is number of events at the j^{th} ordered event time $t_{(j)}$.

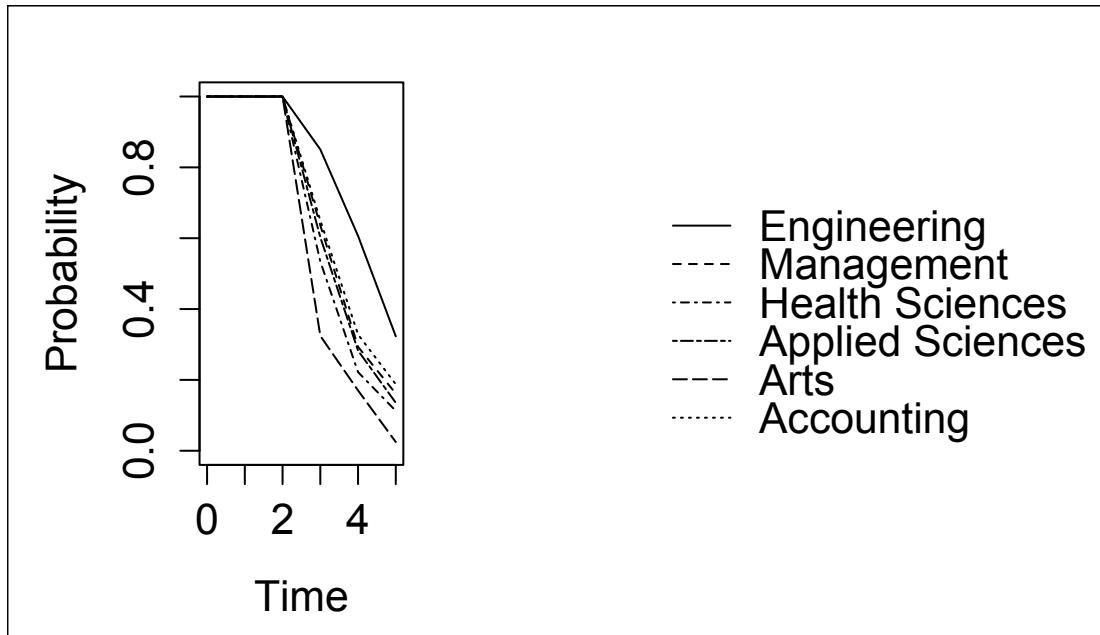


Figure 4.11 Fitted Survivor Functions

Figure 4.11 confirms that the subjects in the Engineering & The Built Environment take longest to graduate with the subjects in the Arts & Design taking the shortest time to graduate.

4.4 Summary

In this chapter we assessed the effect of race, gender, faculty and age on graduation at the given time. We fitted a Cox regression model in an attempt to model the relationship between graduation and these covariates. We found that age was insignificant.

We then performed a proportionality test and found that none of the variables passed the test, which then led us to fit a stratified regression model where we stratified on the faculty variable. We found that, upon stratifying on the faculty variable, the "Race"

and "Gender" variables satisfied the proportionality assumption.

We performed the other diagnostic tests namely; influential variable and outlier test and our model passed these tests. In the end we also assessed the overall fit of the model where we found that our fitted model fits the data reasonably well.

On the bases of our regression model, and with regard gender, we have found that this variable explains graduation and that females are more likely to graduate than males. Regarding race, we have found that Indians are more likely to graduate than African.

In summary, there are reservations in the literature about the suitability of the Cox regression model to model data that is inherently discrete because the coefficient estimates are biased "downwards" (Singer and Willet, 1993, 2003). Notwithstanding its limitations, we have nevertheless fitted a Cox regression model to serve as bridge between non-parametric survival analysis of the previous chapter and the logistic regression that we will present in the next chapter which is also premised on the Cox regression model albeit in discrete context.

Discrete Time to Event Approach

5.1 Introduction

In the previous chapter we used Cox's regression model to estimate the hazard functions and the results that we obtained were not only consistent with our findings in Chapter 3, but there was improvement on *non-parametric* techniques. Firstly, we were able to regress the hazard function on all hypothesised variables simultaneously, secondly, we could also compare hazards and quantify the differences between them. The latter benefit accrues provided a somewhat onerous condition of proportionality is met, failing which, a time varying alternative should be considered.

Whereas the Cox model considered in the previous chapter was in continuous time context, we will now present an extension of the same model in discrete time setting. In the same seminal publication in which Cox (1972) presented the *proportional hazards* model, he proposed that the hazard function in discrete time, a true probability, can be modelled to have a logistic dependence on time and explanatory variables. However, he did not elaborate any further on the model, but authors such as Brown (1975) and Efron (1988) formalized logistic regression hazard as the standard alternative to continuous

hazard formulation.

Traditionally, survival analysis has its roots in the "hard" sciences, particularly in medical research, and its application has been restricted to that area until, Allison (1982) extended the methods to the sociological fields. He showed that the "survival" likelihood function, in discrete setting, is equivalent to the familiar Bernoulli likelihood function. Much later, the work of Singer and Willet (1993) firmly grounded the application of survival analysis in sociological fields and since then, survival analysis in discrete time has become popularly known as *Event History Analysis*.

In discrete time, the random variable T indicates the time of occurrence of an event such that if $T = t_l$ then an event occurred at time t_l . Implicit in the above statement is that t_{l-1} indicates the duration of non-occurrence of the event which leads to the definition of the discrete hazard given by

$$h(t_l) = P(T = t_l | T \geq t_l).$$

Thus, the definition of the discrete hazard is that it expresses the probability that an event occurs at time t_l , given that it has not occurred up to time t_{l-1} . Equivalently to continuous time hazard, the discrete hazard can then be expressed as

$$h(t_l) = P[T = t_l | T \geq t_l] = \frac{P(T = t_l)}{P(T \geq t_l)} = \frac{f(t_l)}{S(t_{l-1})}.$$

Now, $f(t_l) = S(t_{l-1}) - S(t_l)$, therefore

$$h(t_l) = \frac{S(t_{l-1}) - S(t_l)}{S(t_{l-1})} = 1 - \frac{S(t_l)}{S(t_{l-1})}.$$

After some algebra, the above expression becomes

$$S(t_l) = S(t_{l-1})[1 - h(t_l)].$$

Since $S(0)=1$ i.e. at or before time zero no event occurs, recursively using the above we obtain

$$S(t_l) = \prod_{k=1}^l [1 - h(t_k)] \tag{5.1.1}$$

and

$$f(t_l) = h(t_l)S(t_{l-1}) = h(t_l) \prod_{k=1}^{l-1} [1 - h(t_k)]. \tag{5.1.2}$$

We now introduce a different notation for the hazard of a specific subject. If the time interval is split into contiguous subintervals $(0, t_1], (t_1, t_2], \dots$, then $T = t_l$ can also be written as $T = l$, where $(0, t_1)$, refers to $T = 1$ and likewise, $T = l$ refers to the interval $(t_{l-1}, t_l]$ i.e. the event occurred during the l^{th} interval. The hazard of the i^{th} subject with \mathbf{x}_{il} vector of explanatory variables at time l becomes

$$h_{il} = P[T_i = l | T_i \geq l, \mathbf{x}_{il}].$$

5.2 Model

Suppose that we observe n subjects ($i = 1, 2, \dots, n$) from time $t = 1$ until t_i , where for the i^{th} subject an event occurs or the subject is censored at time t_i . Again, we introduce an indicator variable δ_i which is one when the i^{th} subject experiences the event, or zero otherwise.

The probability that the i^{th} subject experiences the event at time t_i is $P(T_i = l_i)$, whereas the probability that the subject is censored is, $P(T_i > l_i)$. When censoring is *non-informative* then the likelihood of the sample is given by

$$L = \prod_i^n [P(T_i = l_i)]^{\delta_i} [P(T_i > l_i)]^{1-\delta_i}$$

Substituting Equation 5.1.1 and Equation 5.1.2 above we obtain

$$L = \prod_{i=1}^n \left(\left[h_{il_i} \prod_{j=1}^{l_i-1} (1 - h_{ij}) \right]^{\delta_i} \left[\prod_{j=1}^{l_i} (1 - h_{ij}) \right]^{1-\delta_i} \right).$$

Taking logarithms of the above, we obtain

$$\log L = \sum_{i=1}^n \delta_i \log \left[\frac{h_{il_i}}{(1 - h_{il_i})} \right] + \sum_{i=1}^n \sum_{j=1}^{l_i} (1 - \delta_i) \log(1 - h_{ij}).$$

Define y_{it} , an indicator variable to be 1 when the i^{th} subject experiences the event and zero otherwise. The above equation leads to

$$\log L = \sum_{i=1}^n \sum_{j=1}^{l_i} y_{ij} \log \left[\frac{h_{ij}}{(1-h_{ij})} \right] + \sum_{i=1}^n \sum_{j=1}^{l_i} \log(1-h_{ij}).$$

Upon taking anti-logarithm of the above we obtain

$$\prod_{i=1}^n \prod_{j=1}^{l_i} h_{ij}^{y_{ij}} (1-h_{ij})^{1-y_{ij}}. \quad (5.2.1)$$

Allison (1982) and Singer and Willet (1993) argue that the above equation reduces to a familiar *bernoulli* likelihood function where y_{ij} is the observations with probability h_{ij} . A logistic parameterization of h_{ij} is given by

$$h_{ij} = \frac{1}{1 + \exp[-\eta_{ij}]} \quad (5.2.2)$$

where $\eta_{ij} = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \dots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_p x_{pij}$.

J is the length of the observation period i.e. the maximum time period, $[D_{1ij}, D_{2ij} \dots D_{Jij}]$ are dummy variables indexing the time periods, with $[d_{1ij}, d_{2ij} \dots d_{Jij}]$ as observations.

The focus now shifts away from the subjects to time periods which become the new unit of analysis. For a given subject, a record is created for each time period that the subject remains in the follow up. If the subject experiences the event or is censored at time l , then l records will be created for that subject together with y_{ij} , where $y_{ij} = 0$ for the time periods 1 to $l-1$ and $y_{il} = 1$ if the subject experiences the event, or 0 if the subject is censored. A new record is therefore, a time period together with the original covariates of the subject which may be time varying or fixed and then the y_{ij} . In total there will be $N = l_1 + l_2 \dots + l_n$ records.

Equation 5.2.2 can also be expressed as

$$\left\{ \frac{h_{ij}}{1 - h_{ij}} \right\} = \exp\{\alpha_1 D_{1ij} + \dots + \alpha_J D_{Jij}\} \times \exp\{\beta_1 x_{1ij} + \dots + \beta_p x_{pij}\}. \quad (5.2.3)$$

In much the same way as the hazard is the product of unspecified $h_0(t)$, the base line hazard which is a function of time and $\exp(\beta \mathbf{x}_i)$ in the Cox regression model, we can also think of $\exp\{\alpha_1 D_{1ij} + \dots + \alpha_J D_{Jij}\}$, the function of time, as the base line hazard which is also left unspecified. This expression, as Singer and Willet (2003) argue, invokes Cox's proportional hazards of the previous chapter albeit in discrete survival time. By taking logarithms of 5.2.3 we obtain the familiar $\logit(h_{ij})$ given by

$$\log \left\{ \frac{h_{ij}}{1 - h_{ij}} \right\} = \alpha_1 D_{1ij} + \dots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \dots + \beta_p x_{pij}. \quad (5.2.4)$$

We already have a likelihood function in Equation 5.2 which leads to

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^{l_1} \left\{ y_{ij} \log \left[\frac{h_{ij}}{(1 - h_{ij})} \right] + \log(1 - h_{ij}) \right\} \quad (5.2.5)$$

where $\boldsymbol{\theta}' = [\alpha_1, \alpha_2, \dots, \alpha_J, \beta_1, \beta_2, \dots, \beta_p]$ and $h_{ij} = \frac{1}{1 + \exp(-\eta_{ij})}$.

The partial derivatives of Equation 5.2.5 are taken and then equated to zero to solve for $\boldsymbol{\theta}$ which maximizes the likelihood function. The estimates of $\boldsymbol{\theta}$ are obtained by using iterative methods.

Since the minimum time to graduation is three years for all subjects, we have only considered the subjects that have survived up to the beginning of year 3 as our initial "risk set". We have fitted a full model (Model A) with interaction terms in Table 5.1,

and then moved to a reduced (Model B) in Table 5.2, by using backward elimination method.

Table 5.1 Full Model

Model A			
Variable	$\hat{\beta}$	S.E	Sig.
Time 3	-1.460	0.452	0.001
Time 4	0.2330	0.435	0.593
Time 5	1.4710	0.627	0.019
Female	-0.229	0.495	0.644
Management Sciences	0.3770	0.196	0.000
Health Sciences	1.9610	0.375	0.000
Applied Sciences	2.0960	0.537	0.000
Accounting & Informatics	2.3450	0.705	0.001
Arts & Design	3.6110	0.896	0.000
Coloured	-0.967	1.094	0.377
Indian	-0.049	0.358	0.890
Other	-0.371	1.635	0.820
White	0.1540	0.717	0.830
Age	0.0010	0.041	0.972
Females \times Time	-0.233	0.094	0.013
Female \times Race	0.3990	0.099	0.000
Female \times Faculty	-0.003	0.041	0.940
Female \times Age	0.0150	0.020	0.472
Race \times Time	-0.126	0.074	0.088
Faculty \times Time	-0.130	0.029	0.000
Age \times Time	-0.021	0.014	0.144
Race \times Faculty	0.0040	0.029	0.885
Race \times Age	0.0170	0.015	0.245
Faculty \times Age	-0.008	0.008	0.299

Table 5.2 Reduced Model

Model B			
Variable	$\hat{\beta}$	S.E	Sig.
Time 3	-1.905	0.117	< 0.0001
Time 4	-1.058	0.109	< 0.0001
Time 5	-0.268	0.097	< 0.0001
Female	0.103	0.073	0.560000
Management Sciences \times Time 3	1.178	0.131	< 0.0001
Management Sciences \times Time 4	1.073	0.142	< 0.0001
Health Sciences \times Time 3	1.477	0.178	< 0.0001
Health Sciences \times Time 4	1.230	0.227	< 0.0001
Applied Sciences \times Time 3	1.231	0.168	< 0.0001
Applied Sciences \times Time 4	1.001	0.203	< 0.0001
Accounting & Informatics \times Time 3	1.160	0.149	< 0.0001
Accounting & Informatics \times Time 4	0.925	0.166	< 0.0001
Arts & Design \times Time 3	2.378	0.169	< 0.0001
Arts & Design \times Time 4	0.806	0.266	< 0.0001
Arts & Design \times Time 5	1.972	0.767	< 0.0001
Indian	0.009	0.107	0.935000
White	0.530	0.169	0.002000
Coloured	0.066	0.169	0.795200
Other	1.151	0.253	0.002000
Female \times Indian	0.679	0.158	< 0.0001
Female \times White	0.687	0.286	< 0.0001

Before we proceed with the interpretation of the model, we assessed the overall goodness-of-fit by using Hosmer and Lemeshow (2000) test. Briefly, the data is split into 10 groups according to their estimated probabilities i.e. $0.0- < 0.1, 0.1- < 0.2, \dots, 0.9- < 1$. Let y_{ij} denote the binary outcome for observation j in group i where $i = 1, 2, \dots, 10$. Let π_{ij} denote the corresponding fitted probability for the model fitted for ungrouped data. The statistic is given by

$$C = \sum_{i=1}^{10} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{\sum_j \hat{\pi}_{ij} (1 - (\sum_j \hat{\pi}_{ij})/n_i)}.$$

This statistic does not however have a limiting chi-square with $df=(10-2)$ as noted by Agresti (2002), but the authors of the test, Hosmer and Lemeshow (2000) maintain that if the number of covariate patterns (subjects with identical covariate values) is equal or approximately equal to the sample size, the statistic follows a chi-square distribution.

The test statistic is in the same spirit as the usual, $\sum \frac{(O-E)^2}{E}$, such that the numerator of the statistic should be very small when the model fits the data very well. $C = 11.563$, $df.= 8$, for the reduced model compared to 15.507 at 5% significance level, suggesting a good fit.

5.3 Model Adequacy

Having already assessed overall goodness-of-fit, we will now consider outliers and influential variables.

5.3.1 Outliers

In general, graphical methods are used to detect outliers and a plot of either *Pearson's* or the *deviance* residuals, will reveal outliers. The pearson residuals are given by

$$r = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}} \quad (5.3.1)$$

where $\hat{\pi}_i$ are the estimated probabilities and y_i are the observations which are either 0 or 1. Dividing Equation 5.3.1 by $\sqrt{(1 - h_i)}$, we obtain what is referred to as the *standardized Pearson residuals*. It must be stressed that the unit of analysis is the time period and no longer the subject, therefore the i^{th} subject hereafter refers to the i^{th} time period. Note that there are will be N subjects, where $N = l_1 + l_2 + \dots + l_n$ instead of n which corresponds to the original number of subjects or the sample size.

$$r_{pi} = \frac{r}{\sqrt{1 - h_i}}$$

where h_i is the i^{th} diagonal element of $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$, the hat matrix of Pregibon (1981). The matrix X is a $n \times p$ matrix of covariates. W contains weights, with the diagonal elements equal to $\hat{\pi}_i(1 - \hat{\pi}_i)$.

The term h_i is referred to as the leverage, it measures the extent to which the i^{th} subject is distant from the others in term of explanatory variables, such that the larger the value of the leverage, the more distant the subject is from others. On the other hand, the deviance residuals are given by

$$d_i = \text{sign}(y_i - \pi_i) \sqrt{-2[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]}$$

and dividing by $\sqrt{(1 - h_i)}$ again, we obtain *standardized deviance residuals*

$$r_{Di} = \frac{d_i}{\sqrt{1 - h_i}}$$

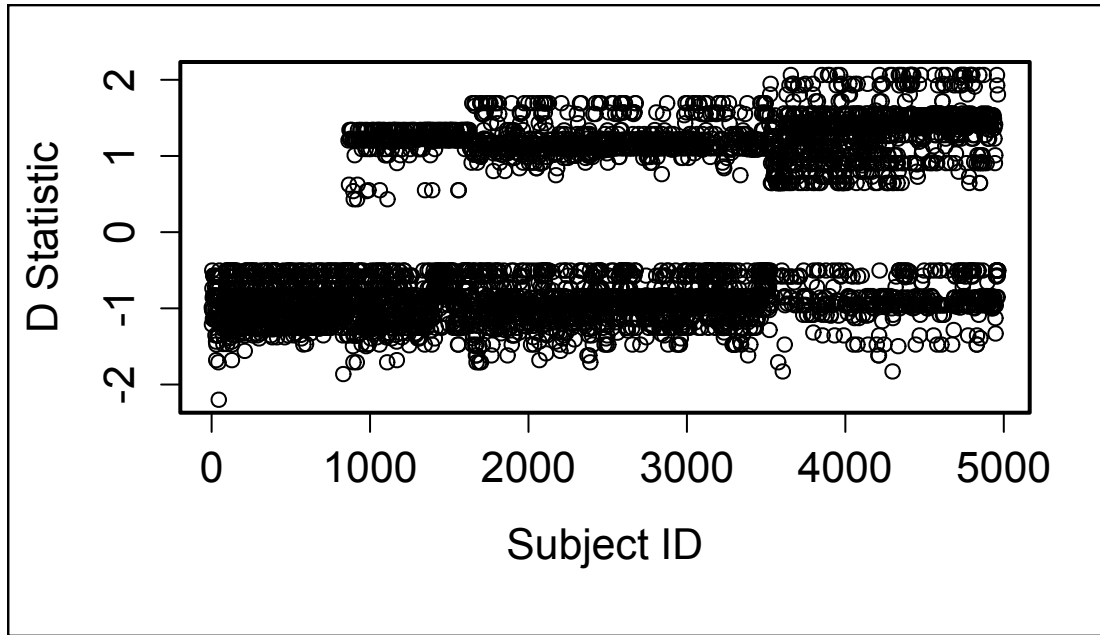


Figure 5.1 Deviance

The general idea is to identify those observations that do not belong with the majority of the other observations, by examining the plot of Pearson or deviance residuals. From Figure 5.1 and Figure 5.2, there is one observation which seems to be an outlier and its details are given in Table 5.3. Outliers occur when the observed y value is 0 and the estimated probability is near 1 or when observed y value is 1 and estimated probability is near 0. In this instance we have $y = 0$ and a high estimated probability value of 0.91067. We will take up this point about the outlier when we assess the impact of influential values.

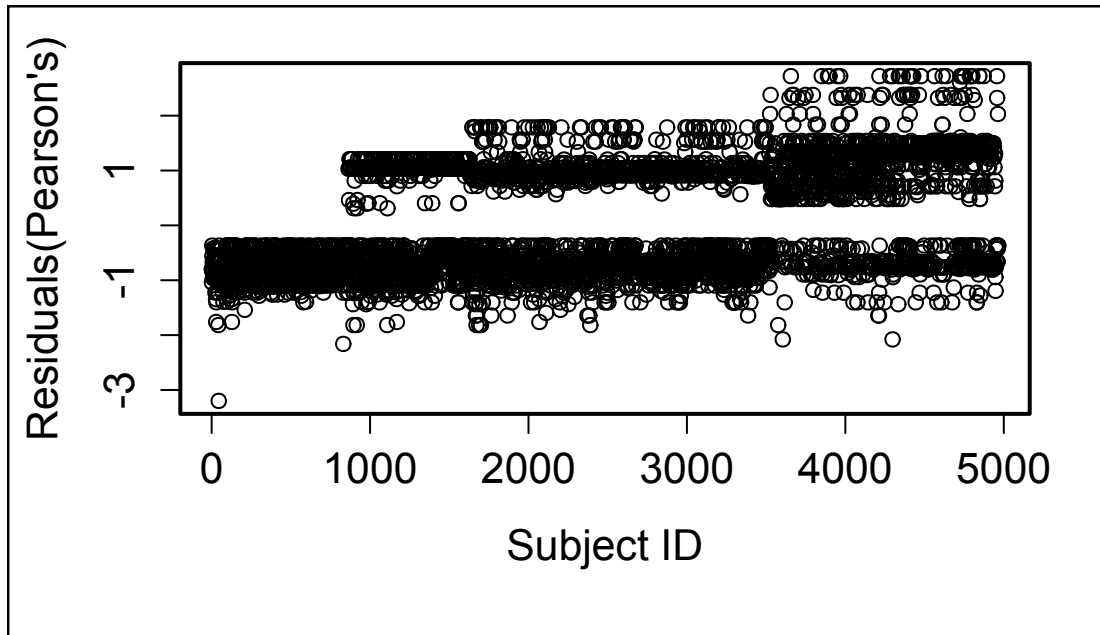


Figure 5.2 Pearson's Residuals

Table 5.3 The Outliers

ID	Gender	Race	Faculty	Age	Status	y	$\hat{\pi}$
15	Male	White	Art & Design	19	censored	0	0.91052

5.3.2 Influence on all Parameter Estimates

As in the previous chapter, here we assess the effect of deleting the i^{th} subject on the parameter estimate $\hat{\beta}$ for the full model compared a to $\hat{\beta}_{(i)}$ based on the same model, but without the i^{th} subject. To avoid the tedious exercise of repeating this for all n subjects, the extent to which the set of parameter estimates is affected by the exclusion of the i^{th} subject is given by

$$D_i = 1/\pi \sum_{j=1}^n \left\{ \text{logit}(\hat{\pi}_j) - \text{logit}(\hat{\pi}_{j(i)}) \right\}^2 w_i.$$

Intuitively, it is the squared distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ approximated (Collet, 1991) by

$$D_i \approx \frac{h_i r_{pi}^2}{\pi(1 - h_i)}.$$

The plot of the D against subjects is in Figure 5.3 and the estimated probability is in Figure 5.4 .

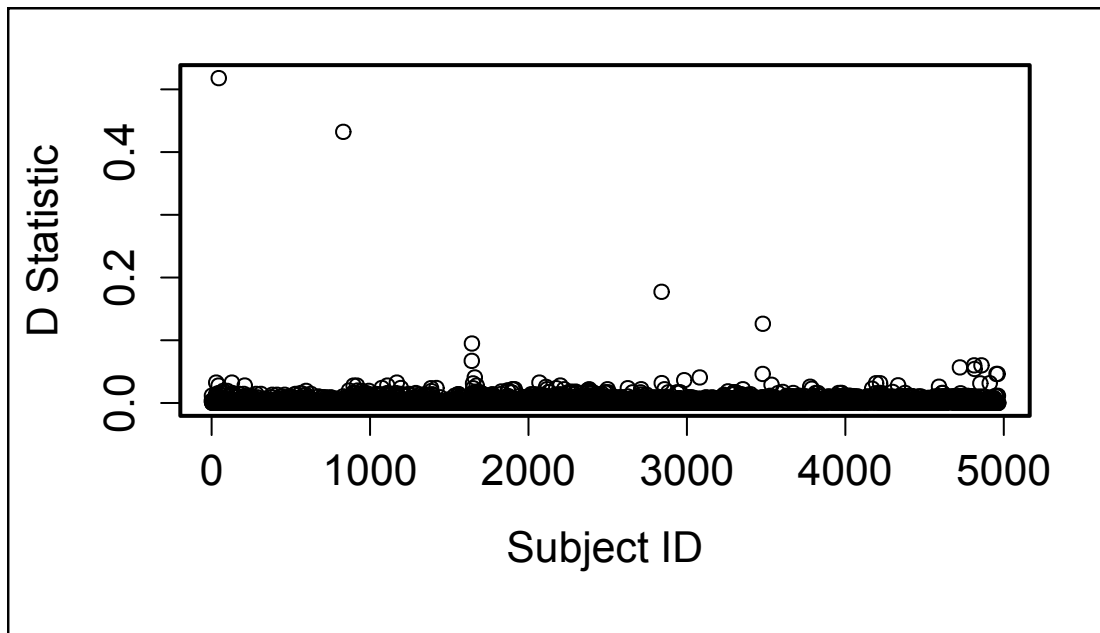


Figure 5.3 D Statistic

There are two subjects that do not belong with others, they are listed in Table 5.4. The plot of the D statistic against the estimated probability in 5.4 indicates that the two subjects have relatively large estimated probability values, possibly compared to observed outcome of 0. Noteworthy is that subject ID=15, the outlier in terms of

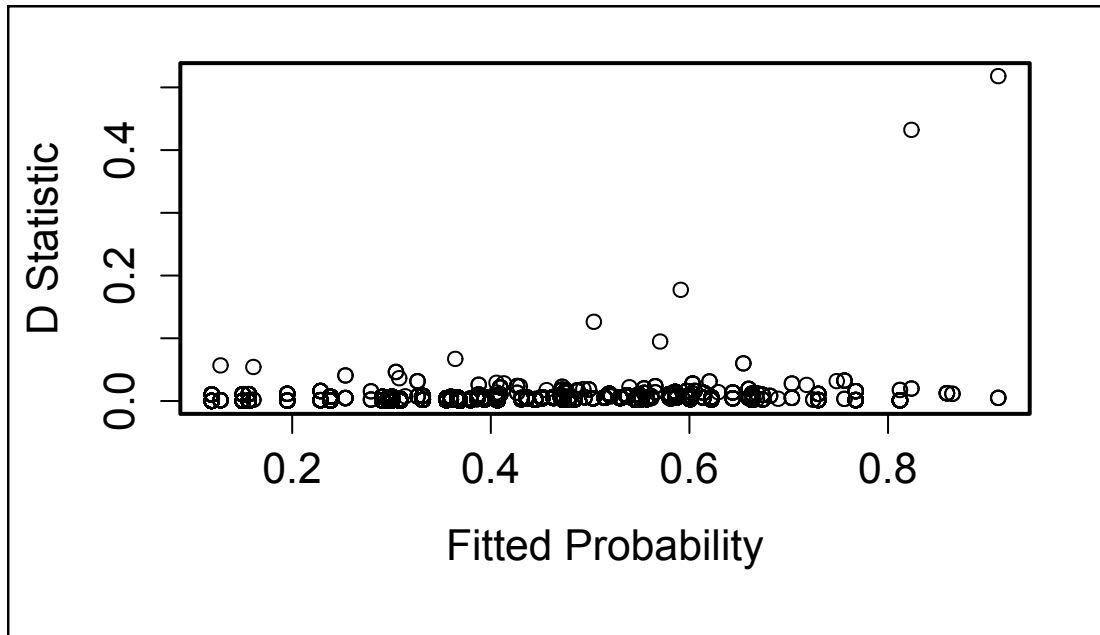


Figure 5.4 D Statistic vs Fitted Probability

residuals, reappears again as having an effect on the parameter estimates. Despite the appearance of these two subjects as extreme, as noted by Hosmer and Lemeshow (2000), the value of the D statistic should exceed 1 for a subject to deserve special attention. We will, however, assess their exact effect later in this section. The other subjects which do not belong with the rest, but not as extremely as the first two, belong to the "Other" factor level. We will however not pay special attention to them as the variable is insignificant in the final model anyway.

Table 5.4 Influential Observations

ID	Gender	Race	Faculty	Age	Status	y	$\hat{\pi}$
15	Male	White	Art & Design	19	censored	0	0.91052
277	Male	African	Art & Design	21	censored	0	0.82395

The D statistic is a summary statistic, which assesses the change over all parameter

estimates, we will assess the effect on individual parameter estimates in the next section.

5.3.3 Influence on Individual Parameter Estimates

The effect of excluding the i^{th} subject on the value of β_j is given by

$$\frac{(X'WX)_{j+1}^{-1}x_i(y_i - \hat{y}_i)}{(1 - h_i)\text{s.e}(\hat{\beta}_j)}.$$

where $(X'WX)_{j+1}^{-1}$, is the $(j + 1)$ th row of the variance-covariance matrix of $\hat{\beta}$. The above statistic is referred to as the *delta-beta*.

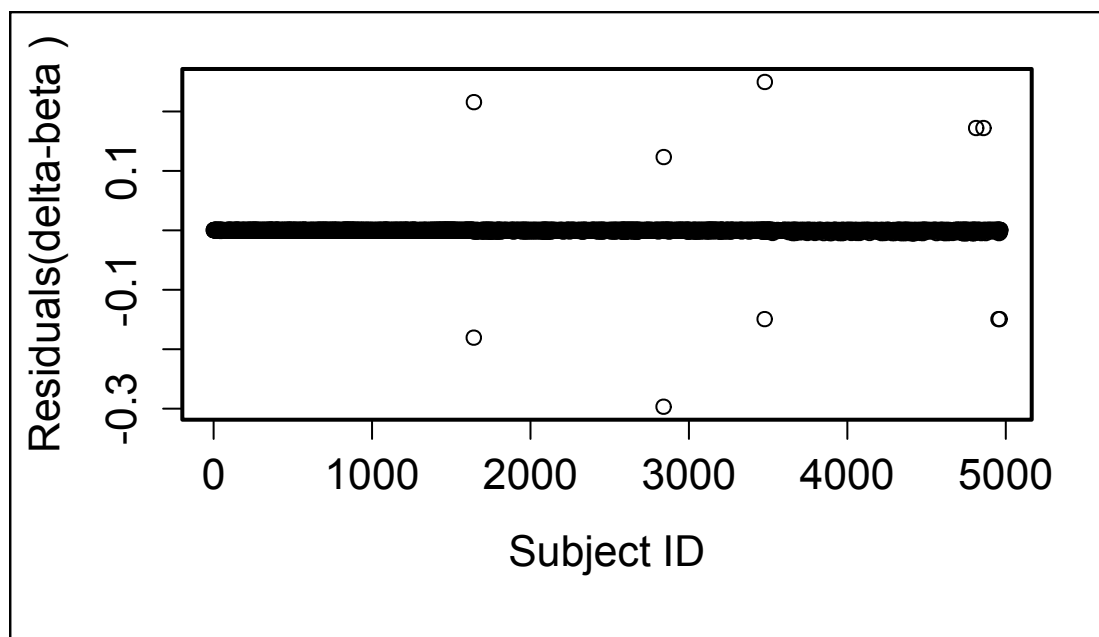


Figure 5.5 Other Delta-Beta

Whilst the D statistic of the previous section raises a flag if there are subjects that have influence on overall parameters estimates without indicating the actual parameters that

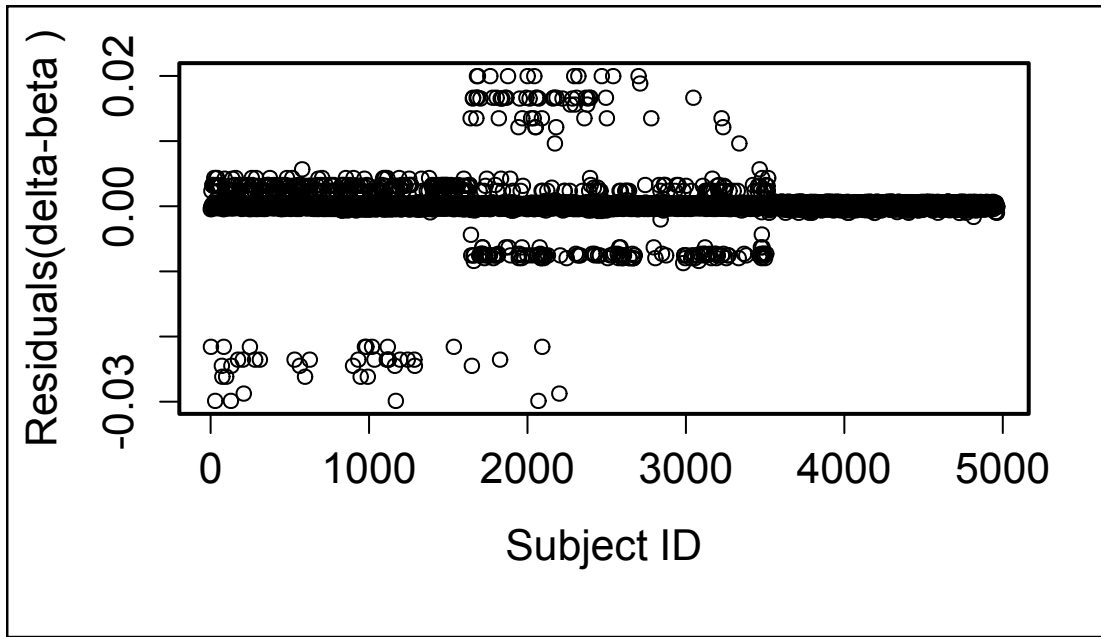


Figure 5.6 Health Sc. Time 4 Delta-Beta

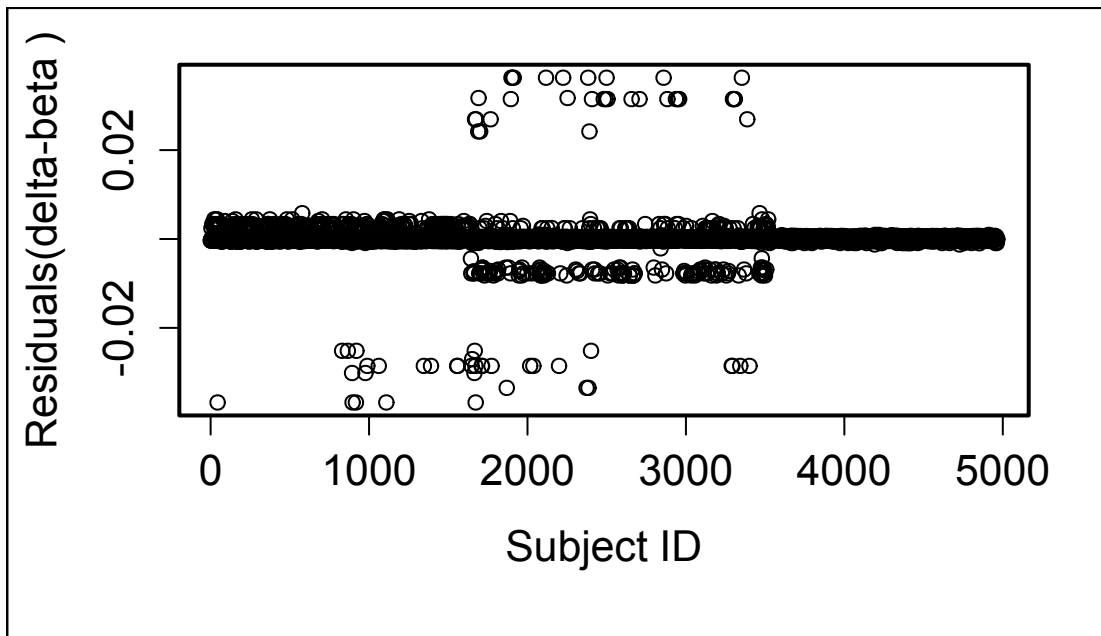


Figure 5.7 Art & Design Time 4 Delta-Beta

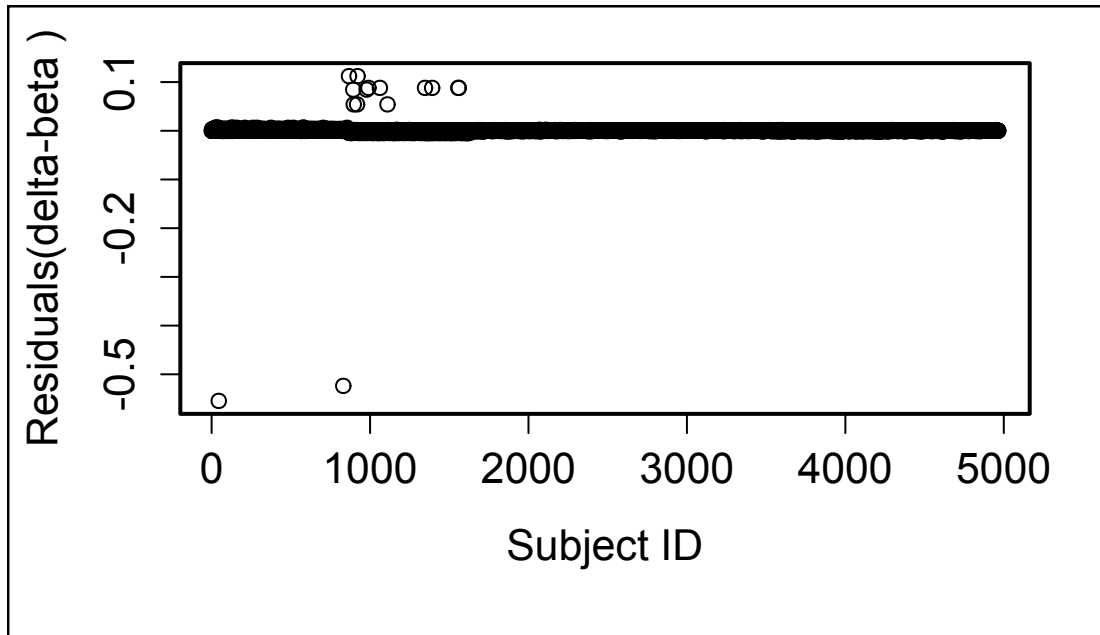


Figure 5.8 Art & Design Time 5 Delta-Beta

are affected, the *delta-betas* goes a step further by pinpointing the actual parameter estimates affected.

We have plotted the *delta-betas* only for the factor levels where there was a suggestion that outliers might exist in Figure 5.5 to Figure 5.8 to save space. Even though there seems to be outliers in the plots of Health Sciences Time 4 and Art & Design Time 4, they are in the order of 0.03 and 0.04 in absolute terms, respectively. On the other hand, the Art & Design Time 3 plot reveals subject ID=15 and ID=277 again as producing the largest changes in on the coefficient of Art & Design Time 3 when they are included and excluded. Also, The same applies to the "Other" factor level. Incidentally, the first two subjects have large leverage values (h_i), with subject ID=276 having the largest leverage value. It is therefore not surprising that they should be influential subjects, and was a question of finding which amongst the three Art & Design interaction terms do they have effect on.

We have established the two subjects, or, rather the two time periods with subject ID=15 and 277 explanatory variables that have undue influence on the Arts & Design Time 3 coefficient estimate. The estimate of the coefficient of Arts & Design Time 5 with both subjects included in the model is 1.972 with S.E = 0.767 and when we exclude subject ID=15, the coefficient estimate is 2.688, a coefficient estimate difference of 93% relative to S.E. Excluding subject ID=277, the coefficient estimate is 2.656, resulting in a coefficient estimate difference of 89%. In both cases, the difference is very high.

The two subjects in question are censored in period 3 out 15 that are at risk. The other 13 graduate, which may explain the reason the two have excessively high probabilities, otherwise there is no other reason such as transcription error etc. which could explain the disproportionately high estimated probabilities.

The other coefficient estimates do not differ substantially whether we include or exclude the subjects except for the Arts & Design time 5 coefficient, therefore the hazards estimates will not differ markedly, whichever model we use, unless the estimation is in relation to Arts & Design time 5. We will leave both subjects in the model but bear this fact in mind when we interpret the concerned coefficient.

5.4 Prediction

In Figure 5.9, we have plotted the hazard estimates for the Faculty variable. The largest hazard estimates occur in the fourth year for all faculties except for the Arts & Design and the Engineering & The Built Environment faculties. The largest hazard estimate occurs in year five for the two faculties, bearing in mind that the estimate for the Arts & Design faculty is unstable in that time period. These results suggest that students are more likely to graduate in the fourth year, than any other period for all other faculties

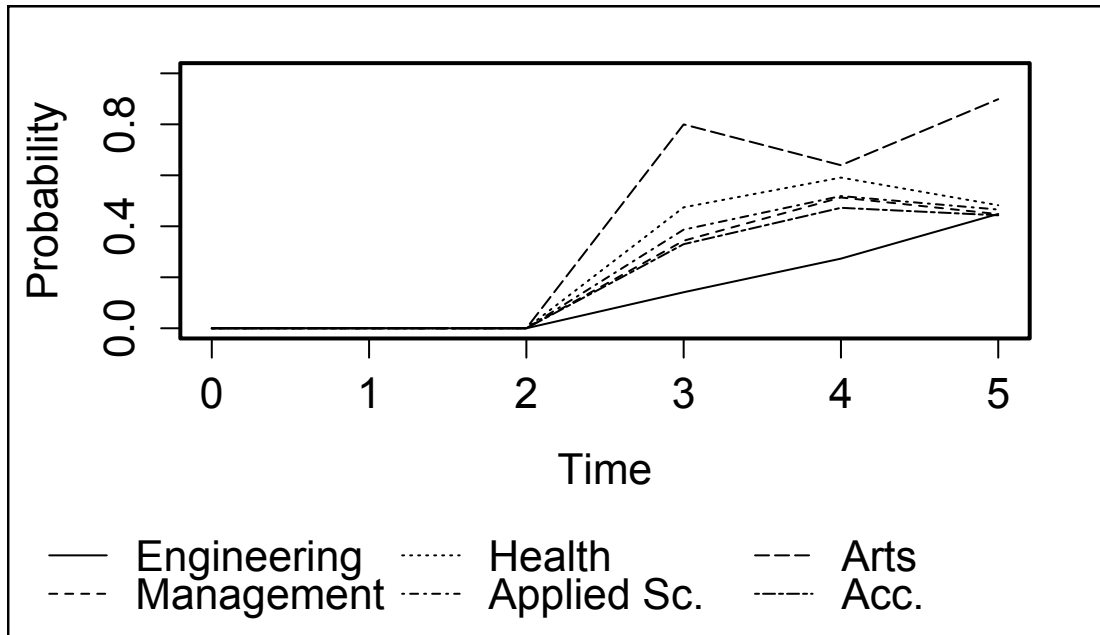


Figure 5.9 Faculty Hazard Estimates

except the Arts & Design and the Engineering & The Built Environment faculties where the highest "risk" of graduation is in the fifth year. Over all, the Arts & Design faculty has the best graduation profile with the Engineering & The Built Environment faculty having the poorest profile.

We have plotted the race hazard estimates in Figure 5.10. All racial groupings are most likely to graduate in year 5 than in any other period. White subjects have the best graduation experience and African subjects have the weakest experience. There is very little difference between the African and the Indian subjects and the marginal edge that the Indian subjects have over the African subjects, is due to Indian female subjects. Coloureds and the Other subjects share the same graduation experience as African subjects.

The gender hazard estimates are plotted in Figure 5.11. Female subjects have a better graduation profile compared to male subjects. Females are most likely to graduate in

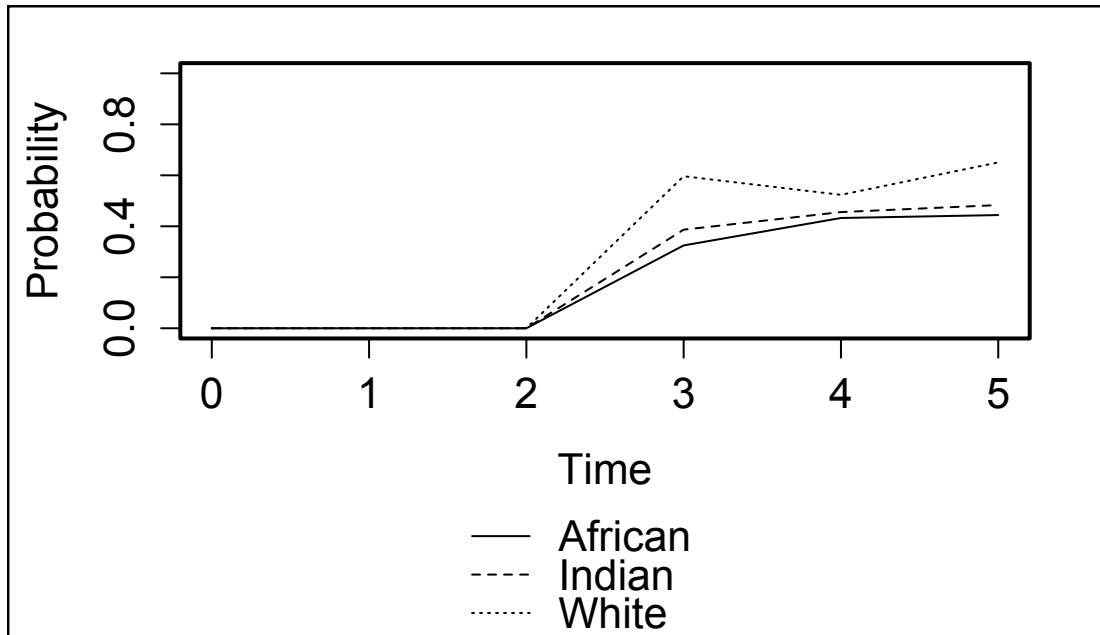


Figure 5.10 Race Hazard Estimates

year 4 than any other period, whereas the male subjects are most likely to graduate in year 5.

5.5 Summary

In the previous chapter we fitted a Cox regression model by stratifying on the faculty variable to obtain proportional hazards between Africans and Indians as well as between males and females. In this chapter we fitted a logistic regression, a model that is not restricted by the stringent condition of the proportionality assumption.

We could not determine the effect of the "White" factor level with Cox's regression model because the factor level failed the proportionality test. This model provides with the estimates for the "White" factor level. Secondly, we lost the effect of "faculty" in

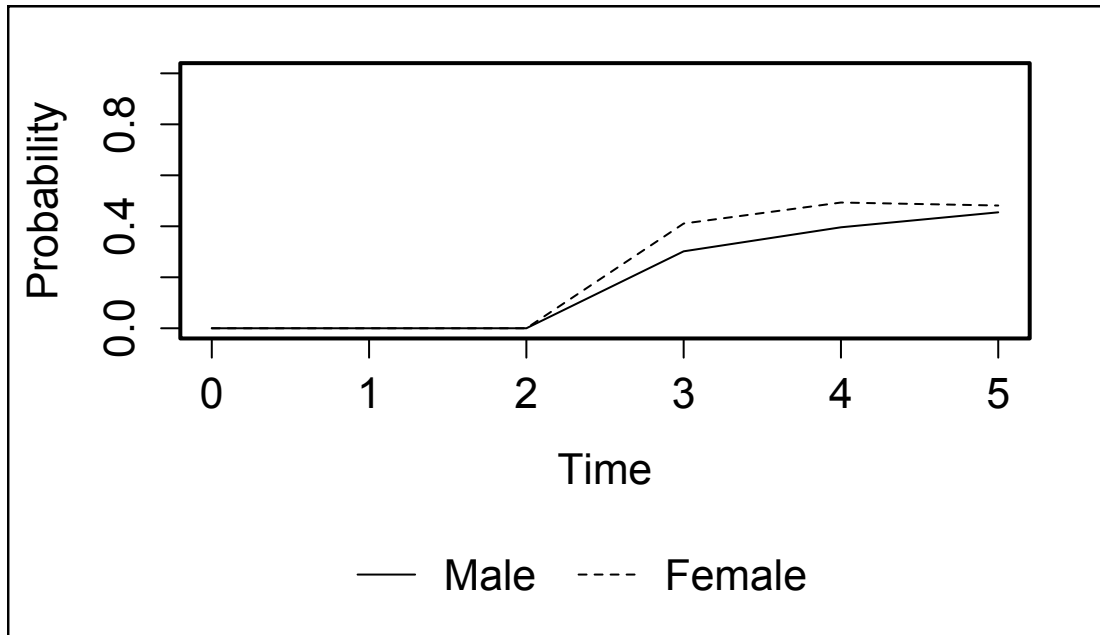


Figure 5.11 Gender Hazard Estimates

the Cox's regression because we used it as the stratifying variable. In this chapter we have the effect of the faculty variable. Thirdly, logistic regression provides the estimates of true probabilities of graduation in contrast to Cox regression where hazards are not true probabilities.

The Cox regression however quantifies the difference between the risk of graduation for male and female subjects as well as between Indian and African subjects.

We raised a concern in Chapter 2 that we have limited data at our disposal, that we did not have access to variables that have been found to explain graduation in the literature. The immediate consequence of leaving out relevant data is overdispersion. We will address this question in the next chapter.

Frailty Models

The Cox (1972) regression model of Chapter 4 is premised on the following two assumptions, namely - (a) The survival times of the subjects in the sample are independent and (b) The survival times of the subjects in the sample are identically distributed.

There are situations, however, where these two assumptions may be found to be untenable. Regarding the first assumption, it is inconceivable, for example, to regard the survival times of siblings or rats from the same litter to be independent of each other because the subjects come from the same genetic 'stock' or background. Two rats from the same litter are more prone to exhibit similar behaviour than rats from different litters and their survival times should not be an exception. The siblings or the rats are said to belong to the same *cluster*.

The second assumption that the survival times are identically distributed underpins the notion that the population under study is homogenous. In principle, this implies that all the subjects under study face the same risk of experiencing the event of interest. This assumption may be suspect at times because certain subjects in a sample may be more prone to experience the event of interest than others.

Vaupel et al. (1979) refer to this "susceptibility" as *frailty*. In life expectancy studies, they argued that individuals differ in certain unobservable attributes or "endowment", which renders some subjects to be more susceptible, or more frail, heightened to experience the event of interest than others. Thus, over and above the heterogeneity that is explained by observed explanatory variables, there are other attributes or "unobserved frailties" unique to the subjects in a given sample.

There are two distinguishable broad classes of frailty models namely; models describing the *univariate* survival times and *multivariate* models. The example of modelling survival times of siblings or "clustered" subjects falls under the multivariate methods. These methods were introduced by Clayton (1978) in his seminal paper where he questioned the validity of the independence assumption in respect of survival times amongst relatives in relation to chronic diseases. On the other hand, Vaupel et al. (1979) are credited with the introduction of frailty in biostatistics from demography and specifically to account for unobserved heterogeneity in univariate models.

More closer to our study, when we discussed the variables in Chapter 2, we highlighted the limitation imposed by unavailability of data such as matriculation results, IQ etc. Over and above these measurable variables, albeit not available, there are other personal characteristics or attributes that may have bearing on time to graduation. Invariably, these attributes are near impossible to quantify, and as such they cannot be directly included with other explanatory variables in a model, because they are *unobservable* and yet they may have sizable influence on graduation. They are what constitutes the different "endowment" or what is loosely referred to as *unobserved heterogeneity* in the literature.

One example is that of the "Big Five" traits (Openness to Experience, Extraversion, Conscientiousness, Neuroticism, Agreeableness), which have been found to have substantial impact on academic performance (Chamorro-Premuzic and Furnham, 2003; Furnham

et al., 2009). The "Big Five" traits are therefore some of the unobserved "frailties" that are unique to the subjects in this study, such that some subjects are more susceptible to graduate than other subjects, as a consequence of these personal attributes.

The often cited example in the literature to illustrate the pitfalls of ignoring frailty is that of two sub-populations, with different, but constant hazards.

Let $\theta_i(t)$ and $S(t|\theta_i)$ be the proportion of subjects in the i^{th} subpopulation and the corresponding conditional survival probability at time t , for $i = 1, 2$, respectively

The unconditional survival probability of the population at time t is given by

$$S(t) = \theta_1(0)S(t|\theta_1) + \theta_2(0)S(t|\theta_2) \quad (6.0.1)$$

whereas, the corresponding unconditional hazard is given by

$$h(t) = \theta_1(t)h(t|\theta_1) + \theta_2(t)h(t|\theta_2). \quad (6.0.2)$$

If say, $h(t|\theta_1) = \lambda_1 > h(t|\theta_2) = \lambda_2$, then $h(t)$ will decline as $\theta_2(t)$ increasingly outweighs $\theta_1(t)$, because more of subjects in subpopulation 1 experience the event earlier. The population hazard will decline in later periods as it increasingly reflects the lesser frailty of the second group, which eventually predominates the remaining risk set. Thus, at population level, the hazard will reflect a declining risk, which is at variance with subject level hazard experience of two constant hazards.

Against the background of data limitations as discussed, we will limit our analysis to univariate methods in this chapter, where we will focus on accounting for possible unobserved frailty which can also be viewed as adjustment for overdispersion due to

omission of material variables (Klein and Moeschbeger, 2003).

In medicine, where the term *frailty* originates, more specifically in gerontology, more frail individuals have a higher risk for "death" as opposed to less frail subjects. Likewise, in our study, a combination of unobservable intrinsic factors and measurable but unmeasured variables for a given subject, reduce or magnify the risk of graduation.

6.1 Continuous Time

In general, frailty models are the equivalent of *random effects* or *mixed models* in survival analysis. They are essentially an extension of Cox's regression model in that a frailty term is introduced which acts multiplicatively to reduce or magnify the "risk" or the hazard as follows

$$h_i(t|w) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i + w_i) \quad (6.1.1)$$

This model can be re-written as

$$h_i(t|u) = h_0(t) u_i \exp(\boldsymbol{\beta}' \mathbf{x}_i) \quad (6.1.2)$$

where $u_i = \exp(w_i)$ is the frailty term that accounts for the unobserved heterogeneity. Vaupel et al. (1979) proposed a gamma distribution for the $\{u_i\}$, which has since become the standard choice for the distribution of the frailty term. The p.d.f is given by

$$g(u) = \frac{u^{1/\theta-1} \exp(-u/\theta)}{\theta^{1/\theta} \Gamma(1/\theta)} \quad (6.1.3)$$

with $E(U) = 1$ and $Var(U) = \theta$. This has the interpretation that if $u_i > 1$, then the i^{th} subject is more frail and more susceptible to experience the event than the k^{th} subject with $u_k < 1$.

The choice of the distribution for the frailty term, u_i is not limited to the gamma distribution. The *Lognormal*, *Inverse Gaussian* and *positive stable* are some of the other choices (Duchateau and Janssen, 2008; Munda et al., 2012).

There are two approaches of estimation in frailty models namely; *parametric* and *semi parametric*. In the former case, the distribution of the baseline hazard is specified i.e. a distributional form of the baseline hazard is assumed such as the Weibull or the exponential distributions and the estimates of the relevant parameters are obtained by maximizing the marginal log-likelihood function. Naturally, the baseline distribution is not assumed in the case of semi parametric approach. The derivation of the parameter estimates when semi-parametric formulation is assumed is far more involved and complex.

6.1.1 Parametric Methods

Equation 6.0.1 illustrates the idea that the population survival function is the average of survival functions of two groups where subjects in each group have identical survival functions. We can extend the idea to a situation more closer to reality where each subjects will have a unique conditional survival function, $S(t|u_i)$ which, however, is unobservable. Instead of the weights used in Equation 6.0.1, we introduce $G(u)$, the distribution function of the frailty term u to obtain the population survival function,

$S(t)$, by integrating out the frailty term if u is continuous.

The population survival function is obtained as follows.

$$S(t) = \int_0^\infty S(t|u)g(u)du. \quad (6.1.4)$$

Since

$$S(t|u_i) = \exp - \int_0^t u_i(h(s))ds = \exp -[H(t)u_i].$$

Substituting the above result in Equation 6.1.4 and assuming that u has a c.d.f, $G(u)$ and a p.d.f $g(u)$, then

$$S(t) = \int_0^\infty \exp -[H(t)u]g(u)du = M_g[-H(t)],$$

where $M_g()$ is the moment generating function of the frailty distribution evaluated at $-H(t)$, and $H(t) = H_0(t) \exp(\beta' \mathbf{x})$, is the cumulative hazard function.

In writing out the likelihood function for n subjects, we typically consider the observed (event or censoring) times $t_1, t_2 \dots t_n$, together with $\delta_1, \delta_2 \dots \delta_n$ indicator variables, such that $\delta_i = 1$ when the i^{th} subject experiences the event, or zero when it is censored, as well as unobserved $u_1, u_2 \dots u_n$. The conditional likelihood is given by

$$\begin{aligned}
L(\boldsymbol{\xi}, \boldsymbol{\beta}|\mathbf{u}) &= \prod_{i=1}^n [h_i(t_i)]^{\delta_i} S_i(t_i) \\
&= \prod_{i=1}^n (h_0(t_i)u_i \exp(\boldsymbol{\beta}' \mathbf{x}_i))^{\delta_i} \exp(-H_0(t_i)u_i \exp(\boldsymbol{\beta}' \mathbf{x}_i))
\end{aligned}$$

where $\boldsymbol{\xi}$ is the vector of baseline hazard parameters (Duchateau and Janssen, 2008). To obtain the unconditional likelihood function we integrate out the frailty term as follows

$$L_{\text{marg}}(\zeta) = \int_0^\infty L(\boldsymbol{\xi}, \boldsymbol{\beta}|u)g(u)du$$

where $\zeta = (\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\beta})$. This expression is then maximised to obtain the estimates for $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Larger/smaller values of $\hat{\theta}$ suggests a greater/lesser variability in the u_i 's and therefore a larger/smaller degree of heterogeneity amongst the subjects. We can verify the above by conducting a test based on the Wald statistic:- $\hat{\theta}/SE(\hat{\theta})$, to test if $\hat{\theta}$ is significantly different from zero.

The disadvantage of parametric models is that if the assumed baseline distribution does not provide a reasonable fit, overdispersion is captured in the frailty term. The larger the difference between the fitted baseline hazard and the observed hazard, the larger the frailty effect (Gutierrez, 2002). Thus a large value of $\hat{\theta}$, which is only due to a misspecified baseline distribution, may lead us erroneously to conclude that there exists heterogeneity when in fact it does not exist.

In Table 6.1, we have fitted a frailty model contrasted with an ordinary Cox regression

Table 6.1 Parametric Frailty Model & Cox Regression Model

	Frailty Model			Cox regression Model		
	$\hat{\beta}$	S.E	P-value	$\hat{\beta}$	S.E	P-value
Intercept	0.7961	0.0284	< 0.001			
female	0.2199	0.034	< 0.001	0.1471	0.0478	< 0.001
Indian	-0.0835	0.0419	< 0.001	0.1718	0.0575	< 0.001
White	-0.3003	0.0602	< 0.001	0.3473	0.0869	< 0.001
Coloured	-0.0293	0.1441	< 0.001	0.0439	0.1918	< 0.001
Other	-0.1161	0.3353	< 0.001	0.6541	0.4490	< 0.001
$\hat{\theta}=0.55$						

of Chapter 4. The assumed baseline distribution for the frailty model is the lognormal distribution, with a gamma frailty distribution. We note that the estimate of the variance for the frailty term is 0.55, which is relatively large and therefore suggesting existence of heterogeneity. Further evidence is the differences of the coefficient estimates between the frailty and the ordinary Cox regression models.

To isolated the possible confounding effect of misspecified baseline distribution, a semi-parametric model is fitted such that the variance of the frailty term only captures the extent of heterogeneity. In the next section we introduce the Cox regression with random effects to account for heterogeneity.

6.1.2 Semi-Parametric Methods

To recall from Chapter 4, The partial likelihood to be maximized to obtain the β when $h_0(\cdot)$, the baseline hazard is left unspecified is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^n \left\{ \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_j)}{\sum_{\ell \in R(t_j)} \exp(\boldsymbol{\beta}' \mathbf{x}_\ell)} \right\}^{\delta_j}. \quad (6.1.5)$$

When

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i) \quad (6.1.6)$$

In the presence of frailty where $w_i = \log u_i$ and $V(w) = \gamma$ we have the following equation

$$h_i(t|w) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i + w_i). \quad (6.1.7)$$

if w_i 's were fixed in Equation 6.1.7, they would be treated as extra parameters to be estimated together with $\boldsymbol{\beta}$. Therneau et al. (2000) suggested maximizing the partial likelihood given in Equation 6.1.5 augmented with frailty, l_{part} , subject to penalty term l_{pen} as follows

$$l_{ppl}(\gamma, \boldsymbol{\beta}, \mathbf{w}) = l_{part}(\boldsymbol{\beta}, \mathbf{w}) - l_{pen}(\gamma, \mathbf{w}).$$

$$l_{part}(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^n \delta_i \left\{ \eta_i - \log \left(\sum_{\ell \in R(t_i)} \exp \eta_\ell \right) \right\}$$

where $\eta_i = \boldsymbol{\beta}' \mathbf{x}_i + w_i$, and

$$l_{pen} = - \sum_{i=1}^n \log g(w_i). \quad (6.1.8)$$

Equation 6.1.8 is referred to as a penalty term because if the actual value of the frailty term $w_i = \log u_i$ deviates from its mean(zero) substantially, then l_{pen} has a large negative contribution towards l_{ppl} .

Maximization consists of an *inner* loop and an *outer* loop. First an initial value γ is suggested and then $l_{ppl}(\gamma, \boldsymbol{\beta}, \mathbf{w})$ is maximized to obtain estimates (BLUPs) of $\boldsymbol{\beta}$ and γ , using Newton-Raphson procedure, then in the outer loop, RELM estimator for \mathbf{w} is obtained. This process is repeated until convergence (Duchateau and Janssen, 2008).

We have fitted a stratified Cox regression model with faculty as the stratifying variable, and using the gamma and the Gaussian distributions for the frailty. We have used *R-package: Survival* (Therneau, 2012) and the results are listed in Table 6.2.

Table 6.2 Semiparametric Frailty Model

	Cox Regression		Gamma Frailty		Gaussian frailty	
Variable	$\hat{\beta}$	Sig.	$\hat{\beta}$	Sig.	$\hat{\beta}$	Sig.
Female	0.1471	0.00211	0.1471	< 0.001	0.1471	< 0.001
Indian	0.1718	0.00281	0.1718	< 0.001	0.1718	< 0.001
White	0.3472	< 0.001	0.3473	< 0.001	0.3474	< 0.001
Coloured	0.0439	0.81887	0.0439	< 0.001	0.0434	< 0.001
Other	0.6541	0.1451	0.6541	< 0.001	0.6541	< 0.001
			$\gamma = 5 \times 10^{-7}$		$\gamma = 0.0013$	

We note that in both frailty models the variance (γ) of the random term is very close to zero, and the coefficients have not changed substantially.

The suggestion that there exist random effects when the baseline hazard is specified in parametric formulation, is due to the misspecification of the baseline hazard. In actual

fact, there is no random effect as suggested by fitting a semiparametric model above.

6.2 Discrete Time

In the continuous case, we fitted a frailty term, u_i , which corresponds to the i^{th} subject. This has a meaningful interpretation in terms of unobserved heterogeneity of a subject, in that it accounts for omitted covariates of a subject. This is not the case in discrete time, since the sampling unit is no longer a subject, but a time period. Instead, we fit a random effects term u_{ij} , which corresponds to the j^{th} time period, with covariates of the i^{th} subject to account for overdispersion.

We recall from Chapter 5 that the hazard h_{ij} in discrete time is modelled to have logistic dependence on covariates and time, as follows;

$$h_{ij} = \frac{1}{1 + \exp[-\eta_{ij}]} \quad (6.2.1)$$

where $\eta_{ij} = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij}$

To account for overdispersion occasioned by possible omission of relevant covariates, we introduce a random effects term u_{ij} , for the j^{th} time period which enters as follows;

$\eta_{ij} = \alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + u_{ij}$, where

$\{u_{ij}\}$ are independent and are assumed to follow a normal distribution with mean zero and variance θ .

The likelihood function of Chapter 5 becomes the conditional likelihood on the u_{ij} as follows;

$$L(\boldsymbol{\xi}|u_{ij}) = \prod_{i=1}^n \prod_{j=1}^{l_i} h_{ij}^{y_{ij}} (1 - h_{ij})^{1-y_{ij}}$$

where $\boldsymbol{\xi}' = [\alpha_1, \alpha_2, \dots, \alpha_J, \beta_1, \beta_2 \dots \beta_p]$ and

$$\text{logit } h_{ij} = [\mathbf{D}, \mathbf{X}]\boldsymbol{\xi} + u_{ij}.$$

Let $N = t_1 + t_2 \dots + t_n$, then \mathbf{D} is a $(N \times J)$ matrix, \mathbf{X} a $(N \times p)$ matrix and $\boldsymbol{\xi}$ a $((J + p) \times 1)$ vector of parameters. The marginal likelihood is given by

$$L(\boldsymbol{\xi}, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \prod_{i=1}^n \prod_{j=1}^{l_i} h_{ij}^{y_{ij}} (1 - h_{ij})^{1-y_{ij}} g(u) du.$$

This likelihood is integrated using Gauss-Hermite or quadrature (Collet, 1991) and then maximized to obtain both $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\theta}}$. If $\hat{\boldsymbol{\theta}}$ is very large, then there is a strong suggestion of omitted explanatory variables. The results of fitting the above model are displayed in Table 6.3. We note that $\hat{\boldsymbol{\theta}}=0.0048$ is very small and that there is very little difference between the coefficients of the models including and excluding random effects.

Table 6.3 Logistic Random Effects Model and Ordinary Logistic Model

Variable	Model C		Model D	
	$\hat{\xi}$	Sig.	$\hat{\xi}$	Sig.
Time 3	-1.905	< 0.001	-1.906	< 0.001
Time 4	-1.058	< 0.001	-1.059	< 0.001
Time 5	-0.268	0.0058	-0.268	0.00058
Female	0.103	0.1565	0.268	0.1563
Management Sciences \times Time 3	1.178	< 0.001	1.178	< 0.001
Management Sciences \times Time 4	1.074	< 0.001	1.070	< 0.001
Health Sciences \times Time 3	1.477	< 0.001	1.478	< 0.001
Health Sciences \times Time 4	1.230	< 0.001	1.231	< 0.001
Applied Sciences \times Time 3	1.231	< 0.001	1.231	< 0.001
Applied Sciences \times Time 4	1.001	< 0.001	1.002	< 0.001
Accounting & Informatics \times Time 3	1.160	< 0.001	1.161	< 0.001
Accounting & Informatics \times Time 4	0.925	< 0.001	0.925	< 0.001
Arts & Design \times Time 3	2.378	< 0.001	2.380	< 0.001
Arts & Design \times Time 4	0.806	0.00250	0.807	0.00250
Arts & Design \times Time 5	1.972	0.01000	1.974	< 0.001
Indian	0.009	0.93500	0.009	0.93463
White	0.530	0.00180	0.530	0.00180
Coloured	0.066	0.79520	0.066	0.79522
Other	1.151	0.08453	1.152	0.08450
Indian \times Female	0.669	< 0.001	0.670	< 0.001
White \times Female	0.687	0.01632	0.687	0.01630
$\hat{\theta} =$	0.0048			
$SE(\hat{\theta}) =$	0.06951			

6.3 Summary

When we discussed the variables at our disposal in Chapter 2, we noted that we did not have access to some variables that have been found to inform both graduation and dropouts in the literature. The objective of this chapter was to assess the effect of omitting measurable and unmeasurable variables in our modelling exercise with reference to the Cox's regression model of Chapter 4 and the discrete model of Chapter 5.

We began by fitting a parametric stratified model with random effects. This was an extension of the stratified Cox regression model of Chapter 4 with the faculty variable as the stratifying variable. The results of the exercise suggested that there exist significant unobserved heterogeneity.

We then fitted a semiparametric stratified Cox regression model with random effects. The general idea was to isolate the impact of the random component without the confounding interference that may arise due to a misspecified baseline hazard. The results indicated that there was no significant unobserved heterogeneity, the unobserved heterogeneity suggested by the parametric model was merely due to a misspecified baseline hazard. We also fitted a discrete model of the previous chapter with random effects and also found that there was no overdispersion.

Our models passed the heterogeneity test that they are both not compromised by limited access to relevant variables. We could not investigate the possibility of *clustering* for lack of appropriate markers in our data to distinguish possible clusters.

Both models, the Cox regression model and its equivalent in discrete time, are also premised on the assumption that all subjects will eventually graduate, had it not been for the 5 year maximum allowable completion period. In the next chapter we evaluate

this assumption.

Mixture Models

7.1 Cure Models

We have addressed the two assumptions upon which Cox's regression model is premised in the previous chapter namely:- the independence and identical distribution of survival times. Yet another assumption that is often unstated is that all subjects under study will eventually experience the event of interest provided the observation period is long enough.

There are instances, however, where some subjects do not eventually experience the event of interest, and we refer to such subjects as "cured", a term inherited from clinical trials. This is usually evidenced by a Kaplan-Meier curve which ultimately levels off into a plateau instead of approaching zero. Standard survival models are based on the assumption that $\lim_{t \rightarrow \infty} S(t) = 0$, a possibility that does not obtain in the presence of cured subjects. (Boag, 1949; J. Berkson and Gage, 1952) are credited with the earliest discussion of this subject. The topic only received renewed attention towards the close of the last century after the seminal work of Kuk and Chen (1992), and the associated modelling techniques have since been referred to as *mixture* or *cure* models in the survival

analysis literature.

In this study, the event of interest is graduation and there might be a possibility that some of the subjects might never graduate (cured), even if the observation period is long enough. Thus, the application of Cox's regression or any other estimation procedures that does not take into account the possibility of cure, might not be entirely appropriate. We therefore investigate the possibility that there might exist a non-ignorable proportion of cured subjects with the ultimate objective to estimate this proportion should there be evidence that it exists and also adjust the survivor function of the subjects that will eventually graduate (uncured) accordingly.

Typically in clinical trials, subjects would undergo a treatment, say, against cancer and cured subjects are those that would not experience relapse if the observation period was long enough (Peng and Dear, 2000; Sy and Taylor, 2000). In our study, we can also think of the three year period as the treatment period towards experiencing the event (graduation), as opposed to not experiencing the event (relapse) in clinical trials. Whilst a cured subject in clinical trials is a subject who does not experience relapse, with regards to our study, a cured subject here, refers to one who eventually does not graduate.

Subjects can either experience the event of interest, or are censored during the observation period. The focus centres on the censored subjects with the view to splitting them into those who would experience the event, and those who would not. Typically in clinical trials, subjects who are lost to follow up, may still experience the event of interest even though their eventual status may not be known i.e. subjects who have received treatment for some ailment may recover or relapse even though they are no longer under observation.

In this study, censored subjects, that is subjects who have left the institution will cer-

tainly not graduate from the institution. Perhaps we should then view our data as a sample coming from a population of all universities of technology so that a dropout can be regarded as censored in the sense that the subject may eventually graduate albeit from another institution. Having made this assumption, we should therefore bear in mind however that censored subject are more likely not to graduate at all, even from another institution and this will tend to overstate the proportion that will eventually graduate. We will re-visit this point later.

Armed with above assumptions and more formally, letting T be the event time and U the indicator variable of uncured subjects, such that $U = 1$ if a subject is uncured and zero otherwise, then the mixture model is defined as:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S_u(t|\mathbf{x}) + 1 - \pi(\mathbf{z}) \quad (7.1.1)$$

$S_{pop}(t|\mathbf{x}, \mathbf{z})$ is the marginal survival function and $S_u(t|\mathbf{x}) = P(T > t|U = 1, \mathbf{x})$ is conditional survival function of uncured subjects with a covariate vector \mathbf{x} . It then follows that $\pi(\mathbf{z}) = P(U = 1|\mathbf{z})$ is the proportion of uncured subjects given a covariate vector \mathbf{z} . We thus have two survival functions:-one for the uncured subjects with survival function $S_u(t|\mathbf{x})$, and $1 - \pi(\mathbf{z})$, for cured subjects that does not depend on time.

Classically, $\pi(\mathbf{z})$ is often referred to as the "incidence" and $S_u(t|\mathbf{x})$ the "latency". The incidence is commonly modelled to have a logistic dependence on \mathbf{z} , i.e

$$\text{logit}(\pi(\mathbf{z})) = \boldsymbol{\gamma}'\mathbf{z}$$

Another variation of mixture models is *non-mixture* models or *promotion time* models.(Chen et al., 1999; Tsodikov et al., 2003; Y.GU et al., 2011). We however consider

the mixture model in this study, and specifically, the discrete model of Chapter 5 and the grouped survival methods.

We begin by developing the necessary theoretical background which is then followed by estimation.

7.1.1 Model & Estimation

We start by re-writing Equation 7.1.1 as follows;

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S_{u0}(t)^{\exp(\boldsymbol{\beta}'\mathbf{x})} + 1 - \pi(\mathbf{z}).$$

In continuous time, and when the distribution of the baseline, $S_{u0}(t)$, is specified, we have the parametric formulation and the parameters to be estimated are $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ (Farewell, 1982). If the baseline is unspecified (semi-parametric), the parameters to be estimated, using the EM algorithm, are $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, as well as $S_{u0}(t)$ (Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000).

Most of the advances in cure models are restricted to continuous time framework, the literature on cure models in discrete time is less developed. We will extend the discrete methods of Chapter 5 to account for cured subjects (Steele, 2003). We will also introduce grouped survival methods (Prentice and L.A.Gloecker, 1978; Allison, 1982) as an alternative to Chapter 5 methods.

To the best of our knowledge, there is one program in **R** to fit mixture models; the *smcure* package (C.Cai et al., 2012) and the *semicure* package in **S-plus** (Peng, 2003) which were both designed for continuous data with fewer ties. There is also a macro in

SAS which we have not investigated (F. Corbie're and P. Joy, 2007). We have had to write our own macros to specifically fit the two the discrete models.

As in Chapter 5, we again consider contiguous intervals $(0, t_1], (t_2, t_3], \dots$ of time. Equation 7.1.1 still obtains:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S_u(t|\mathbf{x}) + 1 - \pi(\mathbf{z}),$$

however, the latency part is now: $S_u(t|\mathbf{x}) = \prod_{k=1}^t (1 - h_u(k|\mathbf{x}))$, see (Steele, 2003)

The contributions to the likelihood function:

$$\text{Contribution} = \begin{cases} \pi(\mathbf{z}_i)h_u(t_i|\mathbf{x}_i)S_u(t_i - 1|\mathbf{x}_i) & \text{if } (\delta_i, u_i) = (1, 1) \\ \pi(\mathbf{z}_i)S_u(t_i|\mathbf{x}_i) & \text{if } (\delta_i, u_i) = (0, 1) \\ 1 - \pi(\mathbf{z}_i) & \text{if } (\delta_i, u_i) = (0, 0) \end{cases}$$

Suppressing the covariates, and assuming the u_i 's the observation of the random variable U are known, the likelihood is given by:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^n \pi_i^{u_i} (1 - \pi_i)^{1-u_i} h_u(t_i)^{\delta_i} (1 - h_u(t_i))^{u_i - \delta_i} S_u(t_i - 1)^{u_i}$$

where $\text{logit}(\pi(\mathbf{z})) = \boldsymbol{\gamma}'\mathbf{z}$.

In Chapter 5 we parametrized the hazard as $\text{logit}(h(t|\mathbf{x})) = \boldsymbol{\beta}'\mathbf{x}$. We will also consider the following parametrization of the hazard, given by

$$h(t|\mathbf{x}) = 1 - \left[\frac{S_0(t|\mathbf{x})}{S_0(t-1|\mathbf{x})} \right]^{\exp(\boldsymbol{\beta}'\mathbf{x})}$$

,

which leads to $\log[-\log(1 - h(t|\mathbf{x}))] = \alpha + \boldsymbol{\beta}x$, the cloglog link function (Prentice and L.A.Gloecker, 1978).

Since u_i 's are not known, the estimates of γ and $\boldsymbol{\beta}$ are obtained by using the Expectation-Maximization(EM) algorithm. In the E-step, u_i 's in $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are replaced by their expectation w_i at the r^{th} step given by

$$w_i^{(r)} = E(u_i|\boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}) = \begin{cases} 1 & \text{if } \delta_i = 1 \\ \frac{\pi_i^{(r)} S_u^{(r)}(t_i)}{\pi_i^{(r)} S_u^{(r)}(t_i) + 1 - \pi_i^{(r)}} & \text{if } \delta_i = 0 \end{cases} . \quad (7.1.2)$$

Thereafter, $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is maximized with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the M-step to obtain $\boldsymbol{\gamma}^{(r)}$ and $\boldsymbol{\beta}^{(r)}$, the estimates at the r^{th} iteration. These estimates in turn are used in E-step of the $(r + 1)^{\text{th}}$ iteration, this process is repeated until convergence.

The likelihood can be split into $L_1(\boldsymbol{\gamma})$ and $L_2(\boldsymbol{\beta})$:

$$L_1(\boldsymbol{\gamma}) = \prod_{i=1}^n \pi_i^{u_i} (1 - \pi_i)^{1-u_i}$$

and

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^n h_u(t_i)^{\delta_i} (1 - h_u(t_i))^{u_i - \delta_i} S_u(t_i - 1)^{u_i} .$$

$L_1(\gamma)$, is a straightfoward logistic regression, $L_2(\beta)$ can be re-written as

$$\begin{aligned}
L_2(\beta) &= \prod_{i=1}^n h_u(t_i)^{\delta_i} (1 - h_u(t_i))^{u_i - \delta_i} \left(\prod_{k=1}^{t_i-1} (1 - h_u(k)) \right)^{u_i} \\
&= \prod_{i=1}^n \left(\frac{h_u(t_i)}{1 - h_u(t_i)} \right)^{\delta_i} \prod_{k=1}^{t_i} (1 - h_u(k))^{u_{ik}} \\
&= \prod_{i=1}^n \prod_{k=1}^{t_i} h_u(t_i)^{y_{ik}} (1 - h_u(k))^{u_{ik} - y_{ik}}.
\end{aligned}$$

where $y_{ik} = 0$ for $k = 1, \dots, t_i - 1$, $y_{ik} = \delta_i$ for $k = t_i$ and $u_{ik} = u_i$ for $k = t_i$. Thus $L_2(\beta)$ is a likelihood of binomial data with y_{ik} successes out of u_{ik} trials. This likelihood is similar to the likelihood in Chapter 5, the only difference is that $u_{ik} = 1$ in Chapter 5, hence a binary data.

The estimates of β and γ can be obtained by using standard programmes that accomodate fractional success/failures and the number of trials. w_i is fractional number of trials in $L_2(\beta)$, and the outcome variable in $L_1(\gamma)$.

Estimation could be initialized by setting Δ with elements $w_i = \delta_i$ as the response variable in $L_1(\gamma)$ to obtain $\mathbf{w}^{(0)}$ with elements $w_i = \hat{\pi}_i$. $\mathbf{w}^{(0)}$ is passed on to $L_2(\beta)$ as the number of trials for $\delta_i = 0$ and 1 otherwise. Therefater, using $L_2(\beta)$ we determine $S_u^{(0)}(t_i)$ which together with Equation 7.1.2 and, $\pi_i^{(0)}$, are used to update to $\mathbf{w}^{(1)}$. This is repeated until covergence.

We have fitted both logistic and cloglog as link functions in the latency and the results as displayed in 7.1 are very close in exponential scale. The logistic incidence suggests that about 86% of students in the study will eventually graduate, compared to about 85% according to the cloglog incidence.

Table 7.1 Incidence & Latency Coefficients

Latency Coefficients				
Logistic			Cloglog	
Variable	$\hat{\gamma}$	Sig.	$\hat{\gamma}$	Sig.
Time 3	-0.3432	< .001	-0.8143	< .001
Time 4	0.2671	< .001	-0.5300	< .001
Time 5	1.1988	< .001	-0.4500	< .001

Incidence Coefficients				
Variable	$\hat{\gamma}$	Sig.	$\hat{\gamma}$	Sig.
Intercept	1.8340	< .001	1.7661	< .001

Before we proceed to interpret the result, we need to attend to the concern we raised in the introduction concerning the censored subjects. Out of 2934 subjects, 1968 graduated during the observation period, 797 dropped out(censored) and 169 were censored due to the closure of the observation period. About 67% graduated during the observation period, thus this model implies that a further 18% or about 528 will eventually graduate from the censored subjects, resulting in 85% "eventual graduation rate".

All in all, the cure model is premised on the assumption that censored subjects may still experience the event of interest. In clinical trials, censored subjects may still experience the event of interest, but experience in higher education suggests that dropouts are less likely to ever graduate. The remaining 169 subjects at the close of observation, fall far short of the required 528, even if they all eventually graduate, to justify the figures of 85% and 86%, as suggested by the model.

To address the shortcomings of the above cure model, we introduce an alternative model within the competing risks framework in the next section.

7.2 Mixture Competing Risks

The most widely used competing risks model in discrete time survival analysis is the multinomial distribution (Scott and Kennedy, 2005; Ambrogi et al., 2009). In this study we have three events namely; graduation, dropouts, and censored subjects due to closure of the study. The shortcoming of the multinomial model is that it regards the censored subjects as an event, whereas we would, ideally, want to split these censored subjects into graduation and dropouts.

Larson and Dinse (1985) introduced *mixture models* which essentially entails splitting the censored subjects amongst the competing risks by using EM algorithm. In introducing the model, Larson and Dinse (1985) specified a distribution for the baseline hazard function, Ng and McLachan (2003); Escarala and Bowater (2008) extended it by assuming a non-parametric baseline hazard function.

The literature on mixture competing risks models in discrete time is very limited, consequently, we have extended the cure model of the previous section to the mixture competing risks (Steele, 2003).

Since we have only two events (graduation and dropouts), the survival function under the mixture models is given by:

$$S(t) = \pi S_1(t) + (1 - \pi) S_2(t)$$

Let us assume that we have $J = 2$ competing risks, as is the case in this study. Let $c_{ij} = 1$ when the i^{th} a subject exits due to the j^{th} cause and zero otherwise. Define another indicator variable z_{ij} , such that $z_{ij} = 1$ when the i^{th} censored subject eventually exits

due to the j^{th} cause and zero otherwise. Define $h_j(t_i)$ as the hazard of the i^{th} subject due to the j^{th} cause which is parametrized as follows

$$\text{logit}(h_j(t, \mathbf{x})) = \boldsymbol{\beta}'_j \mathbf{x}$$

Also define π_j as the probability that a subject is in the j^{th} sub-population which is parametrized as follows:

$$\text{logit}(\pi_j(\boldsymbol{\gamma}, \mathbf{z})) = \boldsymbol{\gamma}' \mathbf{z}_i.$$

Define $\Phi = \{\boldsymbol{\gamma}_j, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ and $c_i = \sum_{j=1}^2 c_{ij}$

7.2.1 Model & Estimation

Adapting the cure model likelihood of the previous section and suppressing parameters and covariates, the mixture competing risks likelihood when there are only two outcomes, is given by

$$\begin{aligned} L(\Phi) &= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij} f_j(t_i)]^{c_{ij}} [\pi_{ij} S_j(t_i)]^{(1-c_{i.})z_{ij}} \\ &= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij} h_j(t_i) S_j(t_i-1)]^{c_{ij}} [\pi_{ij} S_j(t_i)]^{(1-c_{i.})z_{ij}} \\ &= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij}]^{c_{ij} + (1-c_{i.})z_{ij}} [h_j(t_i)]^{c_{ij}} [1-h_j(t_i)]^{(1-c_{i.})z_{ij}} [S_j(t_i-1)]^{c_{ij} + (1-c_{i.})z_{ij}} \end{aligned}$$

Let $g_{ij} = c_{ij} + (1 - c_i)z_{ij}$. Replacing $(1 - c_i)z_{ij}$ with $g_{ij} - c_{ij}$ and a little algebra yields

$$\begin{aligned}
L(\Phi) &= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij}]^{g_{ij}} \left(\frac{h_j(t_i)}{1 - h_j(t_i)} \right)^{c_{ij}} [S_j(t_i)]^{g_{ij}} \\
&= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij}]^{g_{ij}} \left(\frac{h_j(t_i)}{1 - h_j(t_i)} \right)^{c_{ij}} \prod_{s=1}^{t_i} (1 - h_j(s))^{g_{ij}} \\
&= \prod_{i=1}^n \prod_{j=1}^2 [\pi_{ij}]^{g_{ij}} \prod_{i=1}^n \prod_{j=1}^J \prod_{s=1}^{t_i} \left(\frac{h_j(t_i)}{1 - h_j(t_i)} \right)^{c_{ijk}} (1 - h_j(s))^{g_{ijk}} \\
&= L_M \times L_B
\end{aligned}$$

$c_{ijk} = 0$ for $k = 1, 2 \dots t_i - 1$, and $c_{ijk} = c_{ij}$ for $k = t_i$, $g_{ijk} = g_{ij}$ for all k .

L_M is recognizable as the likelihood of a multinomial distribution. L_B is a product of 2 binomial likelihoods $L(\beta_1)$ and $L(\beta_2)$ where there are c_{ijk} successes in g_{ij} trials for the j^{th} binomial likelihood.

L_M turns out to be a Bernoulli likelihood i.e. $\pi_{i2} = 1 - \pi_{i1}$ and $z_{i2} = 1 - z_{i1}$. If the graduations are viewed as successes, we then have the number of 1's equal to the number graduations, the number of 0's equal to the number dropouts and z_{i1} 's to indicate the number of censored subjects due to the closure of the observation period.

Clearly, this is where the mixture model has advantage over the cure model, because only the subjects that are censored due to closure of the study are split between graduation and dropouts. On the other hand, the cure model splits all censored (including dropouts) subjects into graduations and subjects that will not graduate.

Since z_{ij} 's are not known and therefore g_{ij} 's are also unknown, the estimates of Φ are obtained by using the Expectation-Maximization(EM) algorithm. In the E-step, g_{ij} 's in $L(\Phi)$ are replaced by their expectation \bar{g}_{ij} at the n^{th} step, where $\bar{g}_{ij} = c_{ij} + (1 - c_i)w_{ij}$ and w_{ij} is given by

$$w_{ij}^{(n)} = E(w_{ij}|\gamma^{(n)}, \beta^{(n)}) = \frac{\pi_{i1}^{(n)} S_1^{(n)}(t_i)}{\pi_{i1}^{(n)} S_1^{(n)}(t_i) + (1 - \pi_{i1}^{(n)}) S_2^{(n)}(t_i)}.$$

Maximization can also be carried out with standard statistical packages that can accommodate fractional binomial outcomes as in the cure model and the only difference is that we now have an extra binomial likelihood to consider.

We have fitted a model with cloglog link function in the latency without explanatory variables and the results are listed in Table 7.2. We are not aware of a standard statistical package to fit mixture models in discrete time and specialized programming is required to fit the model, even more so if explanatory variables are included in the model. At any rate, we were interested in comparing the cure model and mixture competing risks at global level.

Table 7.2 latency & Incidence Coefficients

Cloglog						
Latency						
Graduation				Dropouts		
Variable	$\hat{\beta}$	S.E	Sig.	$\hat{\beta}$	S.E	Sig.
Time 3	-0.33081	0.03040	< 0.001	-0.52215	0.04751	< 0.001
Time 4	0.08308	0.03932	0.03470	-0.14555	0.05713	0.01090
Time 5	0.35448	0.06545	< 0.001	-0.16172	0.08838	0.0674
Incidence						
Intercept	0.83955	0.03883	< 0.001			

We plotted both fitted survivor function and the sample survivor function in Figure 7.1. The intention was to assess the fit of our model because we did not conduct any simulations to validate the discrete mixture competing risks, as we have not found a similar model in the literature.

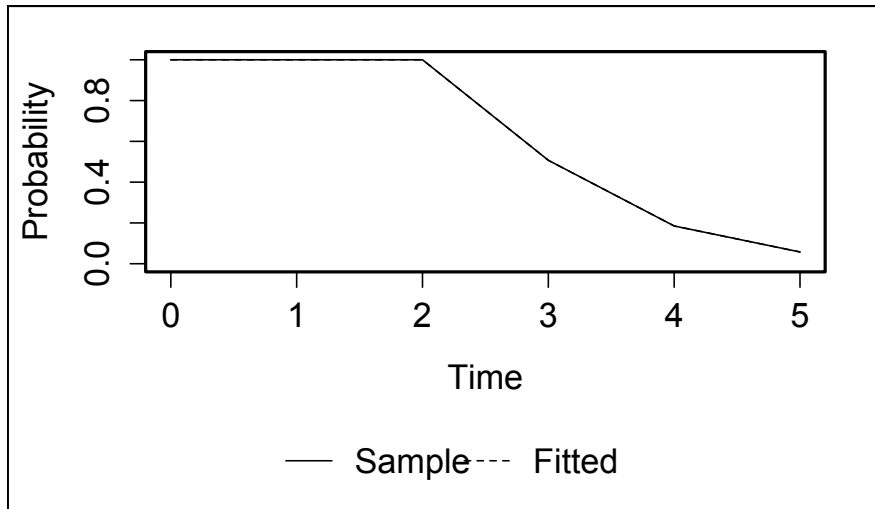


Figure 7.1 Mixture & Sample Survivor Functions

We observe that our model fits the sample results, which then increases our confidence in the model, in the absence of a simulation exercise. On the basis of these results, the estimate of the proportion of subjects that will eventually graduate is about 70% and the other 30% will eventually dropout.

7.3 Summary

Our initial objective in this chapter was to determine if there was evidence that there existed a substantial proportion of subjects that would not experience graduation, even if the observation period was long enough. If there was evidence that some subjects were cured (would graduate), our next objective was to estimate this proportion.

We extended the discrete methods of Chapter 5 into the context of cure models by considering the logistic as well as the cloglog link functions in the latency and the logistic link function in the incidence.

Assuming that dropouts may still possibly eventually graduate, we found that there existed a substantial proportion that will eventually graduate. The reason is that the cure model regards censored subjects as if they would still graduate, a possibility that is very unlikely in higher education literature. A more likely possibility is that censored subjects dropout eventually.

This limitation of the cure model led us to consider posing our modelling exercise within the competing risks realm where dropouts are regarded as competing risks to graduation.

We considered the mixture competing risks model in continuous time and adapted it by extending the univariate approach of Steele (2003), as presented in cure models, to mixture competing risks in discrete time. We noted that we have not found a discrete mixture competing risks model in the literature.

We fitted a cloglog link function in the latency as well as the logistic link function in the incidence. We gauged our modelling exercise against sample results and the survivor function plots suggested that the model provided a very good fit to sample results.

We observed that about 70% of the subjects will eventually graduate. Thus, extending the allowable study period any further from the 5 year maximum allowable period for completion, will only improve the graduation rate marginally from 67% to 70%.

Conclusion and Discussion

In light of the bleak picture around the pass rates and retention rates in the South African higher education landscape, the objective of this study was to investigate the factors that explained graduation pattern at The Durban University of Technology. More specifically, our objective was to find a statistical model that will best explain time to graduation. Furthermore, to derive a likelihood profile from the model that will indicate the periods in which the students are most likely to graduate. We also wanted to investigate if there existed a significant proportion of students who would eventually graduate if the allowable period to complete their studies was extended reasonably longer. Finally, we wished to estimate the proportion that will eventually graduate or dropout, amongst the subjects that are censored due to closure of study.

The most onerous limitation of the study was unavailability of variables that have been found to inform graduation in the literature. We could only access four variables namely - gender, race, faculty and age. Despite this limitation, we nevertheless proceeded with our attempt to find a statistical model that would best explain graduation, albeit with reservations that this limitation might have negative bearing on our modelling exercise.

Descriptive analysis and later non-parametric techniques indicated that these variables

do explain graduation with the exception of age. Non-parametric techniques went a step further by providing us with means to validate these findings that all variables, with the exception of age, explained graduation.

We then explored possible regression models to overcome the limitations of non-parametric techniques, in that they do not provide us with the means to directly regress time to graduation, or a function thereof, on these variables simultaneously.

The first regression technique that we considered was the Cox regression model. Cox's regression indirectly regresses the time to event, by expressing the hazard of the event of interest, as a function of the hypothesised variables. Because Cox's regression methods is premised on the proportionality of hazards assumption, we found ourselves having to resort to fitting a stratified model with faculty as the stratifying variable. Thus, in satisfying the proportionality assumption, we were compelled to sacrifice the faculty effect in our final model.

This approach left us with race and gender as the only significant variables. Cox's regression model suggested that Indians are 19% more likely to graduate than Africans, moreover, females are about 16% more likely to graduate than males.

We then considered a discrete time approach with all the four variables and found that age does not affect the graduation rate. Likewise, our results confirmed that Africans tend to graduate later than other racial groups and females also tend graduate sooner than males. The other advantage of the discrete time regression model is that it provided us with actual probabilities from which we could construct the graduation profile as set out in the objectives. Thus, the second objective of compiling a likelihood profile could be achieved using the discrete model.

We found that the Engineering & The Built Environment faculty had the worst gradu-

ation record compared to any other faculty, with best graduation record attributable to the Arts & Design faculty . Furthermore, we found that females have a better graduation record than males and we also found that Africans have the worst graduation record compared to all other race groups, with Whites having the best graduation record.

We then introduced the frailty model to the Cox regression model in such a way that the latent variable is a proxy for missing data.

We began by fitting a parametric model in continuous time, where we specified a distribution for the baseline hazard. The results suggested existence of unobserved individual effects although this could be the consequence of a miss-specified baseline distribution. To isolate the possible confounding effect of a miss-specified baseline distribution we fitted a semi-parametric model. We found that there was no significant unobserved heterogeneity. We obtained similar results when we considered the discrete model. These results gave us the assurance that both models were not compromised by limited access to relevant variables.

Lastly, we fitted cure models to determine if there existed a substantial proportion of cured subjects i.e. subjects that would not eventually graduate, had the observation period been long enough. We observed that cure models are not suited to our data as it is premised on that all censored subjects will eventually graduate, a possibility which is very unlikely as censored subjects are more likely to dropout permanently.

To overcome the limitation of the cure models, we considered mixture competing risks models in discrete time. We noted that even though the multinomial distribution is the standard competing risk model in discrete time, but its limitation is that it treats censored subjects as an event.

The results of fitting a mixture model in discrete time suggested that about 70% of

the subjects eventually graduate. The advantage of mixture competing risks over cure models is that it regards dropouts as competing risks and only fractionates the subjects that are censored due to closure of the observation period as either eventual graduates or dropouts.

On the other hand, the cure model fractionates all censored subjects into eventual graduates and those who will eventually not graduate. "Not to graduate" has a less meaningful interpretation in higher education literature than dropping out because subjects cannot pursue their studies indefinitely. We noted that extending the allowable study period from the existing 5 year period will only improve the graduation insignificantly from 67% to 70%.

The literature on cure models is in continuous time in most instances. Modeling both frailty and cure models simultaneously in continuous time is the new frontier in the literature (Lai and Yau, 2010; Lopes and Bolfarine, 2012; Calsavara et al., 2013). Therefore, as a possible direction of future research, this study can be extended by applying the already established techniques in the literature on continuous mixture models incorporating frailty by translating them to discrete time methods (Steele, 2003; Chi and Chen, 2011).

The advances in competing risks have been in continuous time as well. Some of the new modelling techniques are; *vertical modelling* (Nicolaie et al., 2010), *multistate models* (Putter et al., 2007), *pseudo observation models*, (Anderson and Perne, 2010). These new methods could also be investigated with a view to extend them to discrete time

We also did not include explanatory variables in our mixture competing risks model because it required specialized programming and this could also be considered.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Wiley and Sons.

Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, 13:61–98.

Ambrogi, F., Biganzoli, E., and Boracchi, P. (2009). Estimating Crude Cumulative incidences through Multinomial Logit Regression on Discrete Cause Specific Hazard . *Computational Statistics and Data Analysis*, 53:2767–2779.

Anderson, P. K. and Perne, M. P. (2010). Pseudo-observations in survival analysis. *Statistical Methods*, 19:71–99.

Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical, Society B*, 11:15–53.

Bourdieu, P. and Passeron, J. C. (1977). *Reproduction in Education, Society, Culture*. . Beverly Hills, CA:Sage .

Bradley, S. and Lenton, P. (2007). Dropping out of Post-Compulsory Education in the UK: an Analysis of Determinants and Outcomes. *Journal of Population Economics*, 20:299–328.

Brown, C. C. (1975). On the Use of Indicator Variables for Studying the Time-Dependence of Parameters in a Response-Time Model. *Biometrics*, 31:863–871.

Calsavara, V. F., Tomazella, V., and Fogo, J. (2013). The effect of frailty term in the standard mixture model. *Chilean Journal of Statistics*, 4:95–109.

- C.Cai, Y.Zou, Peng, Y., and J.Zhang (2012). smcure: An R-package for Estimating Semi-parametric Mixture Cure Models. *Comput Methods Programs Biomed*, 108:12551260.
- Chamorro-Premuzic, T. and Furnham, A. (2003). Personality Traits and Academic Examination Performance. *European Journal of Personality*, 17:237250.
- Chen, M., Ibrahim, J., and Sinha, D. (1999). A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association*, 94:909–919.
- Chi, Y. and Chen, C. (2011). Frailty models with a cure fraction for modeling clustered discrete survival data. In *Int. Statistical Inst.: Proc. 58th World Statistical Congress*.
- Clayton, D. (1978). A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65:141–151.
- Cloete, N., Massen, P., R.Fehnel, T.Moja, Perold, H., and Gibbon, T. (2004). *Transformation in Higher Education*. Kluwer Academic Publishers, Netherlands.
- Collet, D. (1991). *Modelling Binary Data*. Chapman and Hall/CRC, New York.
- Collet, D. (2003). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, Boca Raton.
- Cox, D. R. (1972). Regression Models and Life Tables. *Journal of Royal Statistical Society B*, 34:187–220.
- Desjardins, S. L. and McCall, B. P. (2010). Simulating the Effects of Financial Aid Packages on College Student Stopout, Re-enrollment Spells, and Graduation Chances. *The Review of Higher Education*, 33:513–541.
- Desjardins, S. L., McCall, B. P., and Ahlburg, D. A. (2002). A Temporal Investigation of Factors Related to Timely Degree Completion. *The Journal of Higher Education*, 73:555–581.
- DoE (2003). *Annual Report*. Department of Education, Pretoria.

- DoE (2005). *Annual Report*. Department of Education, Pretoria.
- DoE (2007). *Education Statistics*. Department of Education, Pretoria.
- DoE (2008). *Education Statistics*. Department of Education, Pretoria.
- DoE (2009). *Education Statistics*. Department of Education, Pretoria.
- Driesden, G. W. J. M. (2001). Ethnicity, Forms of Capital, and Educational Achievement . *International Review of Education*, 47:513–538.
- Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer.
- Efron, B. (1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. *Journal of the American Statistical Association*, 83:414–8424.
- Escarala, G. and Bowater, R. J. (2008). Fitting a Semi-Parametric Mixture Model for Competing Risks in Survival Data. *Communications in Statistics - Theory and Methods*, 37:277–293.
- F. Corbière and P. Joy. A SAS macro for parametric and semiparametric mixture cure models.
- F. Corbière and P. Joy (2007). A SAS macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 85:173–180.
- Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046.
- Furnham, A., Monsen, J., and Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Psychology*, 79:769–782.
- Grambsch, P. M. and Therneau, T. M. (1994). Proportional Hazards Tests in Diagnostics based on Weighted Residuals. *Biometrika*, 81:515–526.

- Gutierrez, R. (2002). Parametric frailty and shared frailty survival models. *Stata Journal*, 2:22–44.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. New York.
- J. Berkson, J. and Gage, R. (1952). Survival Curve for Cancer Patients Following Treatment. *Journal of the American Statistical Association*, 47:501–515.
- Klein, J. and Moeschbeger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.
- Kuk, A. Y. C. and Chen, C. (1992). A Mixture Model Combining Logistic Regression with Proportional Hazards Regression. *Biometrika Trust*, 79:531–541.
- Lai, X. and Yau, K. (2010). Extending the long-term survivor mixture model with random effects for clustered survival data. *Computational Statistics and Data Analysis*, 54:2013–2112.
- Larson, M. G. and Dinse, G. E. (1985). A Mixture Model for the Regression Analysis of Competing Risks Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34:201–211.
- Letseka, M. (2009). University Drop-out and researching(lifelong) learning and work.
- Letseka, M. and Maile, M. (2008). High university dropout rates: a threat to South Africa’s future. [http://www.hsrc.ac.za/uploads/pageContent/1088/Dropout %20 rates.pdf](http://www.hsrc.ac.za/uploads/pageContent/1088/Dropout%20rates.pdf). Human Sciences Research Council.
- Lopes, C. and Bolfarine, H. (2012). Random effects in promotion time cure rate models. *Computational Statistics and Data Analysis*, 56:75–87.
- Munda, M., Rotolo, F., and Legrand, C. (2012). parfm: Parametric Frailty Models in R. *Journal of Statistical Software*, 51.

- Ng, S. K. and McLachan, G. J. (2003). An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data. *STATISTICS IN MEDICINE*, 22:1097–1111.
- Nicolaie, M. A., van Houwelingen, H. C., and Putter, H. (2010). Vertical modeling: a pattern mixture approach for competing risks modeling. *Statistical Medicine*, 29:1190–1205.
- Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41:481–490.
- Peng, Y. and Dear, K. (2000). A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics*, 56:237–243.
- Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, 9:705–724.
- Prentice, R. and L.A.Gloecker (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics*, 34:57–67.
- Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26:2389–2430.
- Schoenfeld, D. A. (1982). Partial Residuals for the Proportional Hazards Regression Model. *Biometrika*, 69:239–241.
- Scott, I. and Fisher, G. (2011). The Role of Higher Education in Closing The Skills gap in South Africa. http://www.ched.uct.ac.za/usr/ched/docs/Fisher_Higher%20Education%20role.pdf. World Bank.
- Scott, I., Yeld, N., and Hendry, J. (2007). A Case for Improving Teaching and Learning in South African Higher Education. Higher Education Monitor No.6. http://www.chet.ac.za/media_and_publications/publications/higher_education_monitors. Council for Higher Education.
- Scott, M. A. and Kennedy, B. (2005). Pitfalls in Pathways: Some Perspectives on Competing Risks Event History Analysis in Education Research. *Journal of Educational and Behavioral Statistics*, 30:413–442.

- Singer, D. and Willet, J. B. (1993). It's about Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18:155–195.
- Singer, J. and Willet, J. (2003). *Applied Longitudinal Data analysis*. Oxford University Press, New York.
- Steele, F. (2003). A Discrete-Time Multilevel Mixture Model for Event History Data with Long-term Survivors, with an Application to Analysis of Contraceptive Sterilization in Banglades. *Lifetime Data Analysis*, 9:155–174.
- Strauss, R., Sennet, J., Finchelescu, G., and Gibson, K. (2003). Adjustment of Black Students at a Historically White South African University. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 23:107–116.
- Sy, J. and Taylor, J. (2000). Estimation in a Cox Proportional Hazard Cure Models. *Biometrics*, 56:227–236.
- Therneau, T. (2012). **Survival**: A Package for Survival Analysis in R package version 2.3612.
- Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based Residuals for Survival Models. *Biometrika*, 77:147–160.
- Therneau, T. M., Grambsch, P. M., and Pankratz., V. S. (2000). Penalized survival models and frailty. Technical report, Mayo Foundation.
- Tsodikov, A., Ibrahim, J., and Yakolev, A. (2003). Estimating cure rates from survival data. *Journal of the American Statistical Association*, 98:1063–1078.
- van den Berg, S. and Louw, M. (1984). Problems of University Adjustment Experienced by undergraduates in Developing Countries . *Higher Education*,, 13:1–22.
- van Heerden, E. (1995). Black University Students in South Africa: The Influence of Sociocultural Factors on Study and Performance. *Anthropology & Education Quarterly*, 26:50–80.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography*, 16:439–454.

Wilson, B. (1984). Problems of university adjustment experienced by undergraduates in a developing countries. *Higher Education*, 13:1–22.

Y.GU, Sinha, D., and Banerjee, S. (2011). Analysis of cure rate survival data under proportional odds model. *Lifetime Data Anal*, 17:123–134.