Time-to-Degree: Identifying Factors for Predicting Completion of Four Year Undergraduate Degree Programmes in the Built Environment at the University of Witwatersrand.

A Masters Research Report

Submitted by:

Innocent Mamvura

E-mail: innocentmamvura9@gmail.com

TO

School of Statistics and Actuarial Science,

Faculty of Science,

University of Witwatersrand,

Supervised

By

Kevin Mc Loughlin

A research report submitted to Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

2012

DECLARATION

I declare that this research report is my own, unaided work. It is being submitted for the degree of Masters in Mathematical Statistics in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

(Innocent Mamvura)

…….day of ……

ABSTRACT

The study aims to identify the variables which best predict completion of four year undergraduate degree programmes, in the Schools of Construction Economics and Management and Architecture and Planning, at the University of Witwatersrand (Wits) in South Africa. The research is important to the University and in particular the schools under investigation, because there are only a few studies done in South African universities on this topic and it will contribute to the knowledge on variables that positively influence Time-to-Degree. Selected demographic variables such as Gender, Race, and Home Language were analysed. Other variables considered include: University Courses, First Year Scores, Matric Aggregate, Financial Aid and Residence Status.

The Binary Logistic Models, a Multinomial Logistic Model and Classification Tree Model were developed to test for the significance of the predictor variables at 5% level of significance. The Statistical packages that were used in the analysis of data are Statistical Package for Social Sciences (SPSS), Statistical Analysis System (SAS).

The logistic regression models indicated that Home Language is English and the first year university course Building Quantities 1 are the most important predictors of Time-to-Degree. The other variables that were significant are Gender is Female, Not Repeat, Theory & Practice of QS 1, Architectural Representation I, Building Quantities 1, Construction Planning and Design, Physics Building and Planning for Property Developers. Architectural Representation I, Building Quantities 1, Construction Planning and Design, Physics Building and Planning for Property Developers. Matric Aggregate is an important predictor of university first year success though it has no impact on TTD. The Classification Tree indicated that passing first year at university was significant as it increases the chances of completing the degree programme within the minimum time.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Mr Kevin Mc Loughlin, for your guidance throughout this research project. I also want to thank my work colleague, Ms Bindu Cherian, for proof reading and editing this document. Many thanks to Prof Peter Fridjhon for your contributions which made this research a success.

To my wife Matsiliso Kotelo and mom, thank you so much for your support and encouragement.

To the Almighty God we give You Glory.

TABLE OF CONTENTS

CHAPTER 4: ANALYSES AND RESULTS

CHAPTER 5: DISCUSSION AND CONCLUSIONS

LIST OF FIGURES

LIST OF TABLES

DEFINITION OF TERMS

Completion Rates:	The proportion of students who complete a given level of the education system.
Full-time equivalent (FTE):	The measure that universities use to convert module registrations into a portion of the standard annual credit workload that a full-time student would be expected to undertake at his or her level of study.
Graduation Rate:	The cumulative number of students from the baseline cohort who received the original qualification within or before the specified time.
Throughput Rate:	The number of graduates who complete their studies in prescribed time.
Retention:	The capacity of an institution to keep the students until they graduate.
Semester Elapsed:	When the student is not registered in the university system.
Semester Enrolled:	When the student is registered in the university system.
Total Moved:	The total number of students who changed programmes within or across faculties.
Time-To-Degree (TTD):	The number of academic years the student takes to complete the degree.
Specificity:	Measures the proportion of negatives which are correctly identified.
Sensitivity:	Measures the proportion of positives which are correctly identified.

ABBREVIATIONS

AIC	Akaike Information Criterion
ASCII	American Standard Code for Information Interchange
AUC	Area under the ROC curve
CHAID	Chi Square Automatic Interaction Detection
ECSA	Engineering Council of South Africa
ECSA	Engineering Council of South Africa
EM	Expectation-Maximization
FB0000	Bachelor of Architectural Studies
FF0000	Bachelor of Science in Quantity Surveying
FF0003	Bachelor of Science in Construction Management
FF0004	Bachelor of Science in Property Studies
FTE	Full-time equivalent
HG	Higher Grade
KNIME	Konstanz Information Miner
MAR	Missing at Random
ROC	Receiver Operating Curve
SAS	Statistical Analysis System
SC	Schwarz Criterion
SET	Science Engineering and Technology
SET	Science, Engineering and Technology
SG	Standard Grade
SPSS	Statistical Package for Social Sciences

TTD	Time-to-Degree
URL	Uniform Resource Locator
USA	United States of America
Wits	University of Witwatersrand
YESA	Young Engineers of South Africa

CHAPTER 1

INTRODUCTION

This chapter contains the background information on Time-to-Degree (TTD), statement of research problem, research objectives, the importance of the study and the outline of the research.

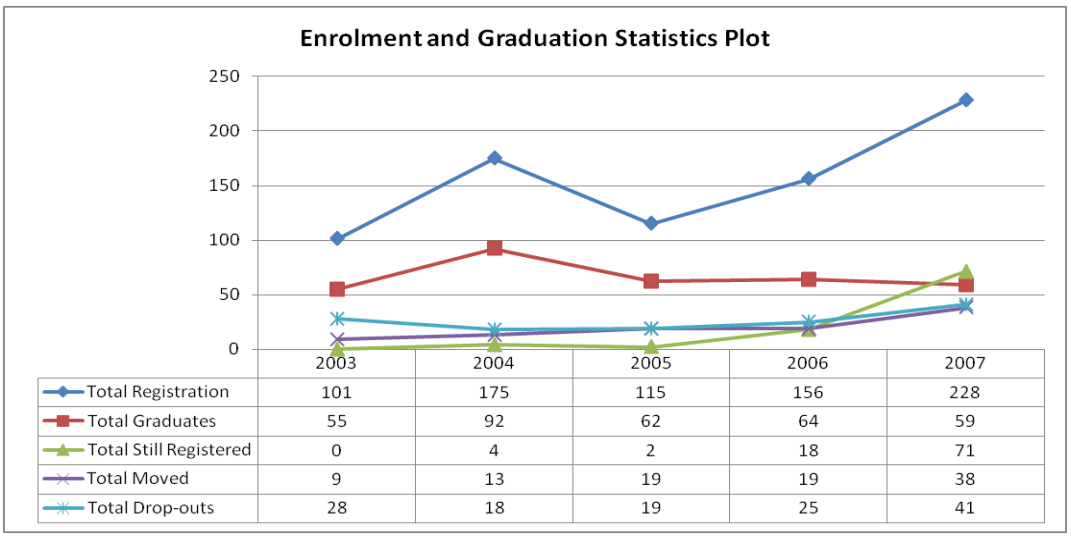
1.1 Introduction

Completion of a university degree programme within the stipulated minimum number of years is a goal for many students and an expectation of many parents. The TTD for a student is the number of academic years the student takes to complete the degree. TTD relates to student success, institutional success, accountability, education expenditure, time investment and graduates entering the job market (Reeves and Haynes, 2008).

The selection process of students who may enrol at the University of Witwatersrand (Wits) identifies students whom the university regards as having a higher probability of completing their programmes. Some of these students are awarded bursaries from companies, sponsorships from various donors, loans from financial institutions and financial assistance from the university. Bursaries are forfeited if the students do not meet the minimum credits required to register for the next year of study. The enrolled students are expected to pass all courses they register for and progress to the next year and complete the degree within the minimum time.

The statistics from Oracle Business Intelligence dashboard (2011) at Wits shows that the cohort of 2005 had 115 undergraduate students enrolled for Bachelor of Science in Construction Management (FF0003), Bachelor of Science in Property Studies (FF0004), Bachelor of Science in Quantity Surveying (FF0000), and Bachelor of Architectural Studies (FB0000). From this cohort, 53.9% completed their degree programmes and 16.5% dropped out. The cohort of 2006 had 156 undergraduate students, of which 41% students completed their degree programmes and 16% dropped out. The cohort of 2007 had 228 students registered of which 25.9% completed their degree programmes and 31.1% are still registered. These statistics are plotted in Figure 1 below.

Figure 1: Enrolment and Graduation Statistics



	2003	2004	2005	2006	2007
Total Registration	101	175	115	156	228
Total Graduates	55	92	62	64	59
Total Still Registered	0	4	2	18	71
Total Moved	9	13	19	19	38
Total Drop-outs	28	18	19	25	41

Figure 1 indicates that total graduates increased steadily between 2003 and 2004 as total registrations also increased. Between 2004 and 2005 there was a sharp decrease in enrolments and in that same period the total number of students who dropped out increased. After 2005 there was a sharp increase in the total number of undergraduate students who registered in the above mentioned programmes.

The success of a university is partly measured by the number of graduates the university produces each academic year, and its revenue (subsidy) is affected by the TTD. A student who takes more than the minimum time to complete a degree programme prejudices the university in that he/she occupies space which would have otherwise been occupied by other potential students who could register, generate input subsidies for the university and upon completion of their programmes, generate output subsidies. The main purpose of this study is to examine the cohorts of 2003, 2004, 2005, 2006 and 2007 (to ensure a sizable data set) undergraduate students against a number of possible academic and demographic factors that impact on TTD at Wits for students enrolling in the Schools of Construction and Management; and of Architecture and Planning. This study will give the school's management and administrators a better understanding of the factors that will positively reduce the time taken in completing degree programmes.

The State of Texas estimated that the cost to students (or parents) for a degree, completed in the prescribed four years, is $41,636, while the cost jumps to $60,264 if the degree takes six years. In the same scenario the cost to the state jumps from $24,948 to $31,752 per student (Texas Higher Education Coordinating Board, 1996). A study by Astin, Tsui and Avalos (1996) revealed that 39.9 % of the 1993 undergraduate cohort in an American University managed to complete their degrees within four years of entering college and the remaining portion of the students completed their studies within nine years after enrolment.

According to Lawless (2005), the ratio of registered engineers to the total population in South Africa is 1:3166, compared to 1:543 in Malaysia, 1:389 in the USA and 1:130 in China. These ratios show that South Africa has a lower proportion of engineers compared to other countries. The figures indicating the lower proportion of engineers is supported by the South African Council on Higher Education (2009) which indicated that the graduation rate for Science, Engineering and Technology was 17% in 2007. The Wits 2013 strategy document (2011) reveals that the undergraduate throughput rate was 15% in 2009 and the target is 30% by 2013. These low graduation rates are of serious concern to the Department of Education and Training (DHET) as they results in shortages of qualified engineers and surveyors. Sunjka (2010) in his article stated that on average, only a third of engineering registered students graduate. For this reason the government of South Africa funds the Young Engineers of South Africa (YESA) programme that was established by the Mereka Institute. The role of YESA is to contribute to the pipeline of Science, Engineering and Technology (SET) graduates and postgraduates through providing learning programmes.

The article on Future Engineering in South Africa by Sunjka (2010) also stated that the scarcity of engineering professionals has two problems namely: there will be shortage of practitioners for ongoing work and the engineering projects are being done without skilled engineering input. In order to reduce shortage of engineers, the Engineering Council of South Africa (ECSA) is in the process of drafting the Identification of Engineering Work Act which will require all engineering personnel to register with the board. Not having enough practitioners for perpetual work means that there are no qualified artisans, qualified technicians and technologists to do the engineering work.

1.2 Statement of the Research Problem

When students enrol for a degree programme any of the following scenarios can occur: they finish within the minimum time, take longer than the minimum time, transfer to other faculties or drop out without completing the programme. The Schools of Construction Economics and Management; and of Architecture and Planning have not done a study that probes factors which affect TTD and this research aims to provide insight and a better understanding of these factors.

A longer average TTD increases the financial burden on students, parents, institutions, state and tax-payers (Scott, Brown and Yang, 2007). Delays in completion of degree programmes affect the university in the sense that generation of output subsidy is delayed and the objective of achieving higher throughput is compromised. Throughput is one of the factors that the government uses for funding a university (Department of Higher Education, 2001). Stock, Finegan and Siegfried (2009) noted that those students, who fail completely to earn a degree, are affected by costs in terms of psychological costs and delayed entry into alternative careers that better match their skills.

1.3 Research Objectives

The main objective of this research is to investigate, in the cohorts of 2003, 2004, 2005, 2006 and 2007 (to ensure a sizeable sample) undergraduate students, in the Schools of Construction and Management; and of Architecture and Planning at Wits, the factors (Gender, Race, Home Language, University Courses, First Year Marks, Matric Aggregate, Financial Aid, Residence Status) that impact on TTD.

The following questions will be answered in this research:

- Do Gender, Race, Home Language, Financial Aid, Residence Status, Matric Aggregate, and University Courses affect the time taken to complete degree programmes?
- Which of these variables are the most important in predicting TTD? The outcome variable of the first model had the following categories; *Pass* (all students who passed their first year with an average mark of 50% and above) and *Fail* (all students who got an average first year mark of 49% and below). The first model was run to analyse the impact of Matric Aggregate, Gender and Race on predicting university first year success. The outcome variable of the second model had two categories: *Completed* (all students who

completed in minimum time) and *Not Completed* (all students who did not complete). The Third Binary Logistic Model also had two categories namely: *Completed* (all students who completed their degree programmes within the minimum time and after) and *Not Completed* (all students who did not complete). A Multinomial Logistic Model was also developed where the outcome variable was a polytomous outcome variable with more than two categories. The values of the outcome variable are *Completed* (all students who graduated within the minimum time of their degree programmes), *Not Completed* (all students who completed after the minimum time and those who are still registered) and *Dropped Out* (all students who were excluded from the programmes and drop outs).

1.4 Significance of the Study

This research is important to the university because currently there are only a few similar studies done in South African universities and the researchers have not focused on investigating the effects of course pass rates on completion-which will be analysed in this study. This research will contribute to the knowledge on factors that influence the TTD rate. Some of the studies which were done in South Africa focused on demographic variables such as Gender, Age, Race, Home Language, and Marital Status (Hall 1999, Zhu 2003, and Lam 1999).

Improving the TTD results in more students graduating and this potentially increases the number of graduates enrolling for postgraduate studies; this would be in line with the Wits' vision found in the Teaching and Learning Plan 2010- 2014 (2010) which states that by 2014, at least 40% of the students should be registered in postgraduate programmes.

1.5 Outline of the Study

The following topics will be covered in each chapter:

Chapter 2: the views of other researchers on the factors that affect TTD.
Chapter 3: the statistical methods, how the data was collected and the methodology used in this research.
Chapter 4: the findings of the analysis. The results of the Binary Logistic model, Multinomial Logistic Model and Classification Trees are discussed.
Chapter 5: discussions of the models built and conclusion of the findings.

CHAPTER 2

LITERATURE REVIEW

This chapter contains a theoretical discussion of related studies done by other researchers. Emphasis is given to those factors that are relevant to this study such as Gender, Matric Aggregate, Financial Aid, University Courses and Home Language.

The paper presented to the Minister of Higher Education and Training, by the South African Council of Higher Education (2010), on the research done by the Working Group on Retention and Throughput at Wits, showed that in the Faculty of Engineering and Built Environment, less than 50% of the graduates qualified within the minimum time of four years and a further 25% completed after five years. Gender and Race were found to be non-significant in predicting throughput. The Working Group requested comments from all faculties; the Faculty of Engineering and Built Environment suggested that poor high school grades made it difficult for students to keep up with Mathematics and Science requirements, and heavy workloads and financial problems led to students dropping out. A number of factors which caused delays in completion were identified and these included student-related factors such as:

- Under preparedness (students not being academically strong enough);
- Students' approach to learning, and their attitude and expectations;
- Students' taking less responsibility for their learning;
- Issues of the students' life and other pressures such as personal, social, financial or family matters.

The group also identified staff related factors such as the attitudes of the staff, the skills of the staff, and staff being demotivated by changes at the university. The research by the Working Group on Retention and Throughput at Wits is very relevant as this group also investigated the Faculty of Engineering and Built Environment and used some of the demographic variables (Gender, Race and Matric Aggregate) that were used in this study.

Research conducted at the University of Western Cape in South Africa by Latief (2005), investigated the throughput of students who did at least one semester of third year statistics. The researcher defined throughput as the completion of undergraduate studies by a student in three

consecutive years. The following factors were explored: Gender, Race, Home Language, and Grade 12 Aggregate, Grade 12 Mathematics results, students entering university directly after school and student registration before and after the 1994 elections in South Africa (first democratic election). Logistic Regression and Decision Trees were used to identify the factors that predict successful throughput. The results on race indicated that more Non-African students finish their degrees in prescribed time as compared to African students. Home language and Grade 12 Aggregate were also found to be significant predictors at the 5% level of significance. The research by Latief (2005) is very relevant as this author used some of the demographic variables (Gender, Race, Home language and Matric Aggregate) and the same statistical techniques (Logistic Regression Model) that were used in this study.

Zhu (2003) examined the TTD of students with respect to their College Preparation, Academic Performance, Time Management, Financial Support and Demographics such as Gender. Financial Support, Gender, and High School Average (Matric Aggregate) are some of the variables that were investigated in this research report. The purpose of the research was to identify the factors that significantly related to the degree completion within four, five, or six years in a public four year college. The TTD had two values: *Graduated* or *Not Graduated*. The researcher used Logistic Regression technique to identify factors which had an impact on the TTD. The Logistic Regression Technique used by Zhu was also used in this research report since the dependent variable under investigation has two outcomes (*Completed* and *Not Completed*). The results of Zhu indicated that the percentage of females who graduated within four years and five years was significantly higher than that of their male counterparts. The results also indicated that High School Average had a positive and statistically significant impact on the TTD. The higher the High School Average the better the chances for students to earn their bachelor degrees within four years. The number of hours spent on study per week, was also significantly related to the TTD. The researchers grouped Financial Aid into family support, grants, part-time income, loans and savings, but found out that there was no significant relationship with the TTD.

Spoerre (2010) investigated the factors that affect graduation in Construction Management education programs at Southern Illinois University, Carbondale. The results showed that students with a high grade average and higher course completion rates, graduated faster than those with

low course completion rates. The impact of the variable higher grade average is relevant in this research report as it is analysed as matric aggregate. The average graduation rate for students enrolled during the study was 41% and 52.6% required more than the prescribed four semesters to complete the program of study. The research showed that student academic factors such as Grade Point Average and course completion Rates are significant predictors in student retention and graduation.

Shulruf, Tumen and Hattie (2010) investigated variables such as Student Participation, Achievements and Completion, Gender, Age, Secondary School Achievements and Courses taken. A series of Regression Models were used to identify the predicting factors for the student pathways each year. The results showed that only 15.2% of the students, who started their programme in 2002, completed their studies within the minimum time, at the end of 2004. Demographic characteristics of students had little to no effect on completion. High pass rates in the first year were associated with completion in the third year, and high pass rates in the third year was the most significant factor for completion of the degree.

Scott (2005) explored three aspects of tertiary study: duration, attrition and completion. The paper investigated the length of time taken to complete degree programmes, if study was adjusted to part-time or part-year. The researcher used a cohort of domestic students, starting at any tertiary institution in 1998 and tracked their equivalent full-time enrolment over a six year period. The results showed a strong relationship between study load and completion (which will also be investigated in this research). Scott mentioned some of the reasons why students take longer than the prescribed time to complete, which include: failing, re-sitting of particular courses, papers, units or modules within the qualification; or a change of qualification during study.

Lam (1999) examined the relationship between the TTD and the type of Financial Aid received by students. The purpose of the study was to provide financial aid researchers with empirical evidence of how different types of financial aid and employment, impact the TTD. The undergraduate students were classified into eight groups based upon sources of financial support. The categories included “loans, gifts and work”, “gifts and work”, “loans and gifts”, ”loans and

work", "work only", "gifts only", "loans only", and "unknown". Students in the "unknown" category had neither financial aid nor payroll records kept by the university. Students in the "loans, gifts, and work" and "loans and work" categories took a longer time to complete their degrees in both elapsed time and registered time. The students in the "loans only" took the least registered time to complete their programmes. The Stepwise Regression Models showed that financial aid variables and percent of loans were both significant. The results also showed that variables related to academic performance and enrolment behaviour, remained the significant variables in determining TTD. The academic variables included admission test scores and cumulative grade point averages, while the enrolment variables included transfer hours from other institutions, number of major changes, number of summer sessions enrolled, and number of semesters enrolled as part-time.

Knight (2004) researched the effect of student participation, demographics, pre-college characteristics, enrolment behaviour variables, academic outcomes, financial aid, parent's educational level, and program accreditation status on TTD. The results of the Regression Model, with semesters elapsed prior to degree attainment as the dependent variable, had a $r^2 = 0.48$. The significant predictors included participation in the Summer Success Challenge Program, average Student Credit Hours Earned per semester, participation in the President's Leadership Academy and Student Credit Hours Earned at the time of graduation. The Regression Model, with the semesters enrolled prior to degree attainment as the dependent variable, had a $r^2 = 0.50$. The significant predictors included Student Credit Hours Transferred, graduation in the Arts disciplines, need-based loan dollars received, students enrolling in the College Reading and Learning Skills class, and Student Credit Hours Earned at graduation.

Yathavan (2008) used Multinomial Logistic Regression and the Chi Square Automatic Interaction Detection (CHAID) analysis to identify factors which affect the students' performance during the first year in the Faculty of Commerce at Wits University. Selected variables such as Previous Institution Type, Gender, Age, Matriculation Aggregate, First Year Performance and Matriculation Courses (Accountancy, Biology, English, History, Mathematics and Physical Science) were used as predictor variables. The CHAID analyses indicated that Matriculation Aggregate was the most important predictor; however Previous Institution Type,

Age, Accountancy, English and Physical Science were also significant predictors. The results of Multinomial Logistic Regression analysis showed that Age, Aggregate, Accountancy, English, Mathematics and Physical Science were significant predictors.

CHAPTER 3

METHODOLGY

This chapter describes the background information of the statistical methods and the data analysis used in this study. The main purpose of this study is to examine the academic and demographic factors that impact on TTD at Wits, for students enrolling in the Schools of Construction and Management; and of Architecture and Planning.

3.1 Statistical Methods

The Statistical methods that were used in this analysis are Binary Logistic Regression, Multinomial Logistic Regression and Classification Tree Analysis. The section below describes these statistical methods.

3.1.1 Binary Logistic Regression

Binary Logistic Regression is a technique used when the outcome variable is a dichotomous variable (has two values). Logistic Regression uses Binomial Probability Theory in which there are only two outcome categories. The technique forms a function using the Maximum Likelihood Method, which maximizes the chances of grouping the observed data into the suitable category given the regression coefficients. One good reason why Logistic Regression is used over a linear Regression model is that the outcome variable is dichotomous and for linear Regression to be valid model the observed data should contain a linear relationship and when the outcome variable is dichotomous, this assumption is usually violated (Berry, 1993). Logistic Regression expresses the Multiple Linear Regression equation in logarithmic terms and thus overcomes the problem of violating the assumption of linearity.

3.1.1.1 Assumptions of Logistic Regression

- No Linear Relationship between the outcome and predictor variables is required;
- The outcome variable must have two categories;
- The predictor variables do not follow a normal distribution, or linear relationship;
- Maximum Likelihood coefficients are large sample estimates.

3.1.1.2 The Model

The Logistic Regression Model equates the Logit Transform, the Log-Odds of the probability of success, to the linear component as defined by Czepiel (2002):

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{ik}\beta_k \qquad i = 1,2,\dots N \qquad (1)$$

The matrix of independent variables, X, is composed of N rows and $K+1$ columns, where K is the number of independent variables specified in the model. The parameter vector, β, is a column vector of length $K+1$.

The Joint Probability Density Function of Y as defined by Czepiel (2002) is:

$$f(\mathrm{y}|\beta)\ ) = \prod_{i=1}^{N} \frac{n_i!}{y_i!\,(n_i - y_i)!}.\pi_i^{y_i}(1-\pi_i)^{n_i - y_i} \qquad (2)$$

The Joint Probability Density Function expresses the values of y as a function of known, fixed values for β.

The Likelihood Function expresses the values of β in terms of known, fixed values for y, as defined by Czepiel (2002)

$$L(\beta|\mathrm{y})\ ) = \prod_{i=1}^{N} \frac{n_i!}{y_i!\,(n_i - y_i)!}.\pi_i^{y_i}(1-\pi_i)^{n_i - y_i} \qquad (3)$$

Rearranging the terms in Eq. 3 gives:

$$\prod_{i=1}^{N} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)^{n_i} \qquad (4)$$

Taking e to both sides of Eq. 1 gives:

$$\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\sum_{k=0}^{K} x_{ik}\beta_k} \qquad (5)$$

Solving Eq.5 for π_i gives the following result:

$$\pi_i = \left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right) \qquad (6)$$

Substituting Eq. 5 for the first term and Eq. 6 for the second term, as defined by Czepiel (2002), Eq. 4 becomes:

$$\prod_{i=1}^{N}\left(e^{y_i\sum_{k=0}^{K}x_{ik}\beta_k}\right)\left(1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right)^{-n_i} \quad (7)$$

Taking the natural log of Eq. 7 yields the log likelihood function

$$l(\beta)=\sum_{i=1}^{N}y_i\left(\sum_{k=0}^{K}x_{ik}\beta_k\right)-n_i log\left(1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right) \quad (8)$$

Setting the first partial derivative with respect to each β equal to zero to find the critical points of the log likelihood function,

$$\frac{\partial}{\partial\beta_k}\sum_{k=0}^{K}x_{ik}\beta_k=x_{ik} \quad (9)$$

$$\frac{\partial l(\beta)}{\partial\beta_k}=\sum_{i=1}^{N}y_i x_{ik}-n_i.\frac{1}{1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}}.\frac{\partial}{\partial\beta_k}\left(1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right) \quad (10)$$

$$=\sum_{i=1}^{N}y_i x_{ik}-n_i.\frac{1}{1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}}.e^{\sum_{k=0}^{K}x_{ik}\beta_k}\frac{\partial}{\partial\beta_k}\sum_{k=0}^{K}x_{ik}\beta_k \quad (11)$$

$$=\sum_{i=1}^{N}y_i x_{ik}-n_i\pi_i\,x_{ik} \quad (12)$$

The maximum likelihood estimates for β is found by setting each of the $K+1$ equations in Eqn. 12 equal to zero and solving for each β_k (Czepiel, 2002) .

3.1.2 Multinomial Logistic Regression

The idea of a Multinomial Logistic Regression Model was generalised from the Binary Logistic Regression (Aldrich and Nelson 1984, Hosmer and Lemeshow 2000). A Multinomial Logistic Model provides several equations for classifying individuals into one of many categories, and is similar to Binary Logistic Regression, but it is more general because the outcome variable is not restricted to only two categories. For example, this analysis involves a dependent variable with three categories: *Completed*, *Not Completed* and *Dropped Out*.

3.1.2.1 Baseline-Category Logistic Model

Consider $J > 2$ categories and the response is assumed to have a Multinomial Distribution Function, taking J as the baseline-category the model according to Czepiel (2002) is:

$$log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj} \quad j = 1,2,\dots,J-1 \;, i = 1,2,\dots N \quad (13)$$

N= number of observations

K= number of independent variables

β= matrix with K+1 rows and J-1 columns, such that each element β_{kj} contains the parameter estimate for the k^{th} covariate and the j^{th} value of the dependent variable.

We want to model the probability π_{ij} that observation i in each j^{th} class of the J-1 categories. The first category class j = 1 is taken as the base class; so the base probability π_{i1} is computed as the residual probability.

Solving Eq. 13 for π_{ij} we have:

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}{1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}} \qquad j < J \qquad (14)$$

$$\pi_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}} \qquad (15)$$

3.1.2.2 Parameter Estimation

The Joint Probability Density Function of a Multinomial Distribution as defined by Czepiel (2002) is:

$$f(\mathrm{y}|\beta)\,) = \prod_{i=1}^{N} \frac{n_i!}{\prod_{j=1}^{J} y_{ij!}} . \prod_{j=1}^{J} \pi_{ij}{}^{y_{ij}} \qquad (16)$$

The likelihood function expresses the unknown values of β in terms of known fixed constant values for y. Maximizing equation (16) with respect to β to obtain the log likelihood function gives equation (17).

The Log Likelihood Function for the Multinomial Logistic Regression Model is:

$$L(\beta|\mathrm{y})\,) \approx \prod_{i=1}^{N} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \qquad (17)$$

Replacing the J^{th} terms in Eq.16 becomes

$$\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}.\pi_{iJ}^{n_i-\sum_{j-1}^{J-1}y_{ij}}$$

$$=\prod_{i=1}^{N}\prod_{j=1}^{J-1}\pi_{ij}^{y_{ij}}.\frac{\pi_{iJ}^{n_i}}{\prod_{j=1}^{J-1}\pi_{iJ}^{y_{ij}}} \quad (18)$$

Grouping the terms that are raised to the y_{ij}

$$\prod_{i=1}^{N}\prod_{j=1}^{J-1}\frac{\pi_{ij}}{\pi_{iJ}}^{y_{ij}}.\pi_{iJ}^{n_i}$$

The likelihood function for the Multinomial Logistic Regression Model

$$l(\beta)=\sum_{i=1}^{N}\sum_{j=1}^{J-1}\left(y_{ij}\sum_{k=0}^{K}x_{ik}\beta_{kj}\right)-n_i log\left(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}\right) \quad (19)$$

We take the first partial derivatives

$$\frac{\partial l(\beta)}{\partial\beta_{kj}}=\sum_{i=1}^{N}y_{ij}x_{ik}-n_i.\frac{1}{1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}.\frac{\partial}{\partial\beta_{kj}}\left(1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}\right)$$

$$=\sum_{i=1}^{N}y_{ij}x_{ik}-n_i.\frac{1}{1+\sum_{j=1}^{J-1}e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}}.e^{\sum_{k=0}^{K}x_{ik}\beta_{kj}}.\frac{\partial}{\partial\beta_{kj}}\sum_{k=0}^{K}x_{ik}\beta_{kj}$$

$$=\sum_{i=1}^{N}y_{ij}x_{ik}-n_i\pi_{ij}x_{ik} \quad (20)$$

3.1.2.3 Goodness of Fit

The goodness of fit of a model describes how well the model fits a set of observations. The deviance of a fitted model compares the log likelihood of the fitted model to the log likelihood of a model with n parameters that fits the n observations perfectly (Saturated Model). The deviance for the fitted model is:

$$G^2 = 2\sum_{i=1}^{N}\sum_{j=1}^{J} y_{ij}\, log\frac{y_{ij}}{\mu_{ij}}$$

The smaller the deviance, the closer the fitted value is to the Saturated Model and the larger the deviance, the poorer the fit.

3.1.3 Classification Tree

Classification is the process of forming groups from a large set of cases based on their characteristics (Fletcher, Lyon, Barnes, Stuebing, Francis, Olson, Shaywitz and Shaywitz, 2001). The Classification Tree procedure creates a Tree-based Classification Model which predicts values of the outcome variable based on values of independent variables. Classification researchers evaluate the consistency and validity of a hypothetical grouping of interest (Fletcher, Francis, Rourke, Shaywitz and Shaywitz, 1993; Morris and Fletcher, 1988; Skinner, 1981). The technique of Binary Recursive Partitioning is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. The process begins with a training set consisting of pre-classified records (outcome variable has a known class for example, *Completed* and *Not Completed*). Then, every possible split is tried and considered, and the best split is the one which gives the largest increase in homogeneity (Frontline Systems Inc, 2010). The process ranks all the best splits and selects the variable that achieves the highest purity at root and classes are asigned to the nodes according to a rule that minimizes the misclassification costs.The process is continued at the next node until a full tree is generated (Breiman, Friedman, Olshen and Stone, 1984).

3.2 Statistical Software Packages

The Statistical packages that were used in the analysis of data are Statistical Package for Social Sciences (SPSS), Statistical Analysis System (SAS).

3.2.1 Statistical Package for Social Sciences (SPSS)

SPSS is a statistical package developed by Norman H. Nie and C. Hadlai Hull in 1968 (Levesque, 2007)**.** SPSS is a powerful program which provides many ways to rapidly examine data and it can produce basic descriptive statistics, such as averages and frequencies, as well as

advanced tests such as Binary Logistic Regression analysis and Multivariate analysis. The program is also capable of producing high-quality graphs and tables.

3.2.2 Statistical Analysis System (SAS)

SAS is a widely used and powerful computer package for analyzing statistical data. It was developed in the early 1970s at North Carolina State University. SAS is currently the most commonly used statistical package when large databases have to be managed, but is also easy to use for small or medium-sized data sets. SAS has an Enterprise Guide which performs statistical tests, estimate statistical parameters and compute significant values (Dilorio, 1991).

3.2.3 Konstanz Information Miner (KNIME)

KNIME is a modern data analytics platform that performs data mining and statistical analysis. Its workbench combines data access, data transformation, predictive analytics and visualization. This package was developed by the Chair for Bioinformatics and Information Mining at the University of Konstanz in Germany.

3.3 Measurements

3.3.1 Output Variable

Three Binary Logistic Regression Models were developed. The outcome variable of the first model had the following categories; *Pass* (all students who passed their first year with 50% and above) and *Fail* (all students who got 49% and below). *Pass* was coded with a value of 1 and *Fail* a value of 0. The outcome variable of the second model had two categories: *Completed* (all students who completed within the minimum time) and *Not Completed* (all students who did not complete). *Completed* was coded a value of 1 and *Not Completed* a value of 0. The third model had two categories: namely *Completed* (all students who completed their degree programmes within the minimum time and after) and *Not Completed* (all students who did not complete). A Multinomial Logistic Model was developed where the outcome variable was a polytomous outcome variable with more than two categories. The values of the output variable are *Completed* (all students who graduated within the minimum time of their degree programmes), *Not Completed* (all students who completed after the minimum time and those who are still registered), *Dropped Out* (all students who were excluded from the programmes and drop outs).

Completed was coded a value of 1, *Not Completed* a value of 2 and *Dropped Out* was coded a value of 3. The outcome variable *Timetodegree* is a nominal scale measurement.

3.3.2 Input Variable

The following variables describe the demographics of the students: Gender grouped as 'Male' and 'Female', and the variable Race with values: 'Black' and 'White'. The variable HLanguage was coded as: 1='English' and 2= 'Non-English'. The 'Non English' category consists of all other languages except English. The categorical random variable HLanguage is a nominal scaled measurement. The matric subjects had different grades: Higher Grade and Standard Grade. The Matriculation Aggregate variable was calculated by averaging the marks received for the subjects passed in matric. To convert the Standard Grade to a same scale with the Higher Grade, a mark of 20 was subtracted for each Standard Grade score. This is consistent with the admission policy of the Faculty of Engineering and Built Environment where Standard Grade A is equivalent to a Higher Grade C (admission requirements, www.wits.ac.za). The other variable included was Financial Aid coded as follows 1='No Financial' and 2='Financialaid'. The variable Repeat was coded as follows 1='Not Repeat' and 2='Repeat'. Residence Description had the following values: 'In Residence' and 'Not In Residence'.

The minimum admission requirement for the programmes of Bachelor of Science in Construction Management (FF0003), Bachelor of Property Studies (FF0004), Bachelor of Science in Quantity Surveying (FF0000) and Bachelor of Architectural Studies (FB0000), are Mathematics Pass Higher Grade (HG), or minimum 60% at Standard Grade (SG) and English Pass HG. The Minimum Admission Points score was 23 points. Applicants with between 18 and 22 points were accepted on the basis of an exercise and an interview for either the ordinary degree or an extended curriculum programme. The Firstyear variable was created by averaging the university first year courses done by each student and the variable was coded as 1= '60 and above' and 2='below 60'. Admissionpoints was coded as 1= '23 and above' and 2= 'below23'. Both Aggregate and Admissionpoints are nominal scale variables. The variables for the first year courses are shown in Appendix 2 (see Table A2.2). The courses are continuous variables. The courses used in the Multinomial Logistic Regression Model and Classification Tree were

categorical and small letters were used with a ‘c’ at the end to differentiate them between the original variables for example ‘buqs110c’ is the same variable with ‘BUQS110’.

3.4 Data Analysis

The data set was extracted from the cohorts of undergraduate students enrolled in 2003, 2004, 2005, 2006 and 2007 (to ensure a sizeable sample) for programmes in FF0003, FF0004, FF0000 and FB0000. The student’s data is stored in the Oracle Data Warehouse managed by the Business Intelligence Services Unit at Wits. The Faculty of Engineering and the Built Environment at Wits, has seven schools and offers qualifications that address the social, spatial, cultural and infrastructural needs of a transforming South Africa. The primary aim of Engineering and Built Environment education at Wits is to produce graduates competent to create and develop policies, devices and systems in many areas including buildings, transportation and communication systems generation, the distribution of electrical energy, extracting and processing of naturally occurring minerals and materials. This study considered the School of Construction Economics and Management; and of Architecture and Planning which, at undergraduate level, offer the following four year programmes:

- Bachelor of Science in Construction Management (FF0003);
- Bachelor of Property Studies (FF0004);
- Bachelor of Science in Quantity Surveying (FF0000);
- Bachelor of Architectural Studies (FB0000).

Each student has a unique student number that enabled the tracking of his/her academic progress. The data included the programme name in which the student was enrolled, the duration of the programme, Matric Scores, Matric Province (the region where the students completed their matric education), Residence Status, Gender, whether student had Financial Aid, and Course Components. This study also investigated the relationship between matric scores and first year scores. Roux, Bothma and Botha (2004) discovered that a very small percentage of those students with a high school result of below 70%, obtained a first-year University average performance of 50% or more.

The statistical methods were implemented using the SPSS statistical package, Konstanz Information Miner (KNIME) package and SAS. The data sets from the University's Oracle Data Warehouse, in comma delimited (csv) format, were loaded into KNIME and merged into one file using the joiner node, pivoting node and the group node. The file node read data from an ASCII file or URL location and can be configured to read various formats. The group node, groups the rows of a table, by the unique values in the selected columns. A row is created, for each unique value group, of the selected column(s). The remaining rows are aggregated by the defined method. The output table, therefore, contains one row for each existing value combination of the selected group column(s). The pivoting node, counts the co-occurrences of all value pairs between the group and pivot column. If an aggregation column is selected, the value between the co-occurrences is computed, based on the selected aggregation method. The joiner node joins two tables in a database-like way. The join is based on the joining columns of both tables. The comma delimited writer, executes the data table coming through its input port, into a file. The node provides many options, to customize the output format.

The data was then exported as an excel file to SPSS, where categorical variables were created and data cleaning was performed. Variables with zero frequencies were removed from the analysis. Missing university course marks were imputed, using the expectation-maximization (EM) method, in SPSS. EM estimation is based on the assumption that the sequence of missing data is associated to the observed data only (SPSS Inc, 2007). This condition is called missing at random (MAR).

The Binary Logistic Regression Model was run in SPSS. The stepwise regression approach (Forward Likelihood Ratio technique) chosen is useful in that it builds models in a sequential way and it allows examination of a collection of models which might not otherwise have been examined. The procedures can be used in cases where there is a great excess of independent variables for example in this research project.The Forward Likelihood Ratio technique, starts with no predictor variables in the model, and then enters variables one at a time, at each step adding the predictor with the largest score statistic, whose significance value is less than 0.05. At each step, SPSS checks for significance of variables already in the model to see if any should be removed. Removal is based on the Likelihood Ratio Test. The Hosmer-Lemeshow Goodness

Test and the ROC were selected to validate the models built. The Multinomial Logistic Model was run in SAS Enterprise Guide where, the outcome variable was selected as unordered and the reference level was the value *Not Completed.*

3.4.1 Model Validation Process

Cross Validation splits the sample into a number of subsamples and Tree Models are then generated, omitting the data from each subsample in turn. "The first tree is derived from predicting all of the cases, omitting those in the first sample fold, the second tree is predicted on all of the cases omitting those in the second sample fold, and so on. For each tree, misclassification risk is estimated by applying the tree to the subsample omitted in generating it" (SPSS Inc, 2004). It is very difficult to obtain information about the real predictive power of a statistical model, because some overfitting of the model may be present, leading to an apparently over optimistic error estimation rate (Stone, 1977). Cross validation can be used as an efficient general tool for evaluating the predictive ability (Van Houwelingen and Le Cessie, 1990). In Logistic Regression, the ROC Curve was utilized to check the fit of the model. In this analysis, the power of the model's predicted values, to distinguish between positive and negative cases, is quantified by the Area under the ROC curve (AUC) and the value varies from 0.5 to 1.0. To perform a ROC Curve Analysis plot, the predicted probabilities are plotted against the outcome variable in the Logistic Regression.

CHAPTER 4

ANALYSES AND RESULTS

The main purpose of this study was to examine a number of possible academic and demographic factors that impact on TTD at Wits for students enrolling in the Schools of Construction and Management; and of Architecture and Planning. The following statistical techniques; Binary Logistic Regression, Multinomial Logistic Regression and Classification Trees were used to analyse the data. This chapter presents the results of the analyses.

4.1 First Binary Logistic Model

This model was run to analyse the impact of Matric Aggregate, Gender and Race on predicting university first year success at 5% level of significance. From the 658 cases included by SPSS, 145 students (22%) failed their first year at university and 513 (78%) passed. The outcome variable *Success* had two values; Pass coded as 1 and Fail coded as 0, as shown by Table 4.1. The predictors that were tested are Matric Aggregate, Gender and Race.

Table 4.1 Dependent Variable Encoding

Original Value	Internal Value
Fail	0
Pass	1

Table 4.2 shows how well the intercept-only model predicts the outcome variable. Predicting success based on the most likely group (*Pass*), would be accurate for 78% of the time and is based on the model with no predictors.

Table 4.2 Classification Table of First Binary Model

Observed			Predicted		
			Success		Percentage Correct
			Fail	Pass	
Step 0	success	Fail	0	145	0.0
		Pass	0	513	100.0
	Overall Percentage				78.0

The overall significance of the model is tested using the Chi-Square Model (see Table 4.3), derived from the probability of observing the original data under, the assumption that the model that has been fitted is accurate. The Omnibus Tests of Model Coefficients table provides a Chi-Square significance statistics for Step, Block and Model. The Step Chi-Square tests the contribution of the specific variable(s) entered on the current step; the Block Chi-Square tests the contribution of all the variables entered with the current block, and the Chi-Square tests the fit of the overall model. There are two hypotheses to test in relation to the overall fit of the model:

H_0: The model is a good fitting model.

H_1: The model is not a good fitting model (that is the predictors have a significant effect).

Table 4.3 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12.245	3	0.007
	Block	12.245	3	0.007
	Model	12.245	3	0.007

In this case the Chi-Square Model has 3 degrees of freedom, a value of 12.245 and p= 0.07 (Table 4.3). This shows that the predictors have a significant effect and create essentially a different model. The Cox and Snell R Square and Nagelkerke R Square provide an indication of the amount of variation in the dependent variable.

The model summary (see Table 4.4) shows the Cox and Snell's R-Square statistics which attempts to imitate multiple R-Square based on likelihood. Here it is indicating that approximately 18% of the variation is explained by the logistic model.

Table 4.4 Model Summary of First Binary Model

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	681.773[a]	0.18	0.28

In this model the Nagelkerke R^2indicates an approximate relationship of 28% between the predictors and the prediction.

The Hosmer and Lemeshow Goodness of Fit Test divides subjects into deciles based on predicted probabilities, and then computes a Chi-Square value from observed and expected frequencies. The H-L Statistic value is 0.802 (see Table 4.5), we fail to reject the null hypothesis and conclude that that there is no significant difference between observed values and model-predicted values. The model's estimates fit the data at an acceptable level.

Table 4.5 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.635	4	0.802

The variables in the equation (see Table 4.6) below shows the Wald statistic and associated probabilities which provide an index of the significance of each predictor in the equation. The Wald statistic has a Chi-Square distribution and is explained by the significance value in the model. Aggregate (1) represents an average matric mark of 60% and above, Gender (1) represents Female students and Race (1) represents white students. In this model; Aggregate (1) is significant at 5% level of significance, meaning that students who obtain an average matric mark of 60% and above, have higher probability of passing university first year. Gender variable and Race are not significant at 5% level of significance. The regression coefficients in Table 4.6

represent the change in the log-odds according to a one unit change in the values of the predictor variables (Cramer, 2003).

Table 4.6 Variables in the Equation of the First Binary Model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Aggregate(1)	0.570	0.196	8.426	1	0.004	1.768
	Gender(1)	-0.011	0.201	0.003	1	0.955	0.989
	Race(1)	0.452	0.292	2.394	1	0.122	1.571
	Constant	0.961	0.148	41.920	1	0.000	2.614

The Exp (B) column in Table 4.6 shows the effect of raising the predictor variables by one unit. The odds ratio provides a more intuitive interpretation of one unit changes in the independent variables. The odds ratio is the number by which one multiplies the odds of a category occurring for a change of one unit in a predictor variable controlling for any other predictors (Cramer, 2003; Hosmer and Lemeshow, 2000). The Exp(B) value associated with Aggregate is 1.768 hence when Aggregate is raised by one unit the odds ratio is 2 times as large and therefore matric students with an average mark of 60% and above in their matric subjects are 2 more times likely to pass their first year at university.

4.2 Second Binary Logistic Model

The outcome variable of the second model had two categories; *Completed* (all students who completed in minimum time) and *Not Completed* (all students who did not complete). Table 4.7 shows that only 170 (25.8%) completed their degree programmes within the minimum time and 488 (74.2%) did not complete.

Table 4.7 Frequency Table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not Completed	488	74.2	74.2	74.2
	Completed	170	25.8	25.8	100.0
	Total	658	100.0	100.0	

The Omnibus Tests of Model Coefficients table indicates that the Second Binary Logistic Model has a Chi-Square value of 333.825 (see Table 4.8) and a probability of 0.000. Therefore, the null hypothesis that there is no difference between the model with only a constant and the model with independent variables was rejected. The overall model is significant when the predictors are entered.

Table 4.8 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 11	Step	4.155	1	0.042
	Block	333.825	8	0.000
	Model	333.825	7	0.000

The Classification Table shows that 62.9% (see Table 4.9) of those students, who completed their degree programmes, were predicted correctly. The overall rate of correctly classifying cases is 85%.

Table 4.9 Classification Table of Second Binary Model

			Predicted		
			Timetodegree		
	Observed		Not Completed	Completed	Percentage Correct
Step 11	Timetodegree	Not completed	452	36	92.6
		Completed	63	107	62.9
	Overall Percentage				85.0

Table 4.10 provides a summary of model components. The B column provides the estimated coefficients for each variable in the equation. The Wald column provides the Wald statistic, which tests the hypothesis that a coefficient is significantly different from zero. If the coefficient is significantly different from zero then we can assume that the predictor is making a significant contribution to the prediction of the outcome. The Sig. column reports the p-value for the Wald statistic.

Table 4.10 Variables in the Model of Second Binary Model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 11[a]	Gender(1)	0.538	0.265	4.114	1	0.043	1.712
	Repeat(1)	3.073	0.722	18.105	1	0.000	21.599
	HLanguage(1)	1.243	0.278	20.003	1	0.000	3.467
	BUQS110	0.073	0.036	4.219	1	0.040	1.076
	ARPL1003	0.449	0.092	23.865	1	0.000	1.566
	BUQS113	0.145	0.027	28.862	1	0.000	1.156
	BUQS1000	0.078	0.022	13.271	1	0.000	1.082
	Constant	-59.488	9.187	41.933	1	0.000	0.000

The following predictors are significant at 5% level of significance: Gender (1), Repeat (1), HLanguage (1), Theory and Practice of Quantity Surveying 1 (BUQS110), Architectural Representation 1 (ARPL1003) and Building Quantities (BUQS113). Gender (1) represents Female students and Repeat (1) represents students who did not repeat. The significance of the variable Gender (1) means that Female students (p=0.043) complete their degree programmes faster than the Male students. Not Repeats (p=0.000) in first year and students who speak English as HLanguage also have a greater chance of completing within the minimum time. The Exp (B) value associated with the predictor variable Gender (1) is 1.712. Hence, when this predictor is raised by one unit and the other predictors kept constant, the odds ratio is 2 times and therefore Female students are 2 times more likely to complete their programmes within the specified time as compared to the Male students. The odds ratio of Repeat (1) is 22 times larger

when this predictor is raised by one unit and therefore not Repeat students are 22 times more likely to complete as compared to Repeat students.

4.2.1 Model Validation: The Receiver Operating Curve

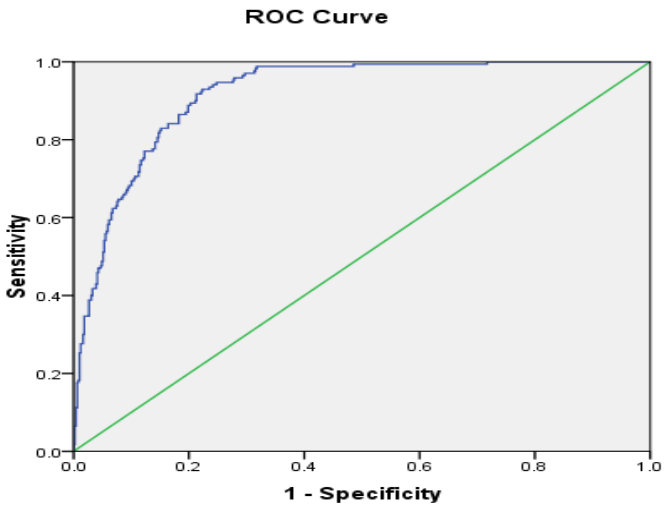
A measure of goodness of fit often used to evaluate the fit of a Logistic Regression Model is based on the simultaneous measure of sensitivity (True Positive) and specificity (True Negative) for all possible cutoff points. The sensitivity and specificity pairs for each possible cutoff point are calculated and plotted as follows: 'sensitivity' on the y-axis and '1-specificity' on the x-axis. This curve is called the ROC Curve.

Table 4.11 Area Under the Curve of Second Binary Model

			Asymptotic 95% Confidence Interval	
Area	Std. Error	Asymptotic Sig.	Lower Bound	Upper Bound
0.917	0.011	0.000	0.896	0.938

The AUC gives a quantitative indication of how good the test is. The ideal curve has an area of one. This area provides a measure of the models ability to discriminate between those subjects who experience the outcome of interest, versus those that do not. The area under the curve determined by the Mann-Whitney U Statistic is 0.917 with 95% confidence interval (0.896, 0.938) as seen in Table 4.11. This indicates that the models performance is excellent.

Figure 2 ROC Curve for Second Binary Logistic Model

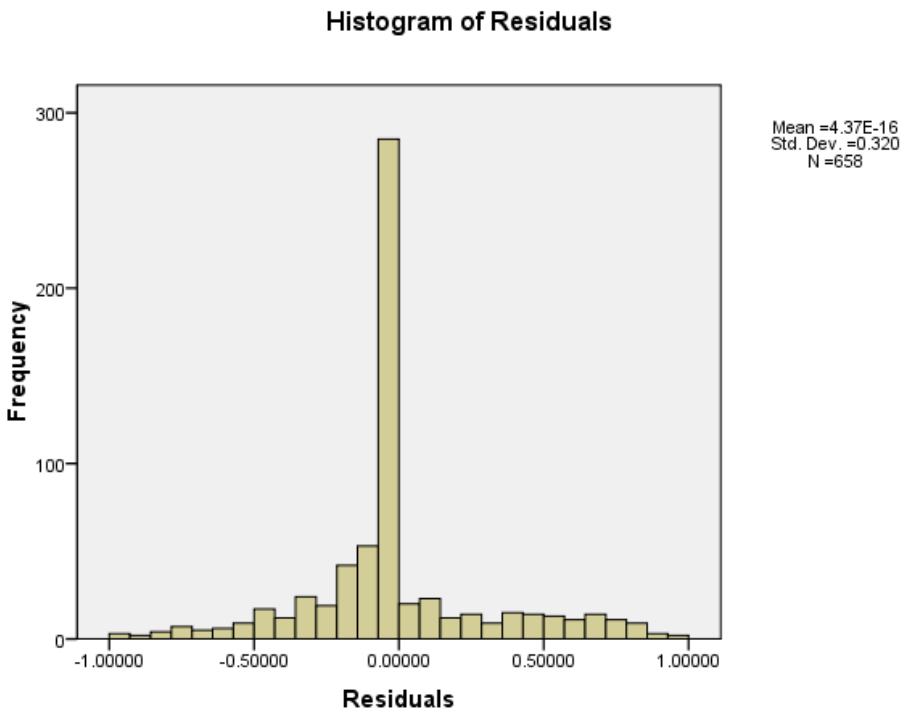


4.2.2. Residual Analysis

The residuals are examined to see how well the model fits the observed data. If the model fits the data well, then we can have a more confidence that the coefficients of the model are accurate. The main purpose of examining residuals in logistic regression is to separate points for which the model fits poorly and to separate points that exert an undue influence on the model.

The Histogram of the Residual (Figure 3) can be used to check whether the variance is normally distributed. A symmetric bell-shaped histogram which is evenly distributed around zero indicates that the normality assumption is likely to be true. If the histogram indicates that random error is not normally distributed, it suggests that the model's underlying assumptions may have been violated.

Figure 3 Histogram of Residuals for Second Binary Model



The pattern shown by the histogram indicates that the residuals are not normally distributed

The Kolmogorov-Smirnov compares the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > 0.5$), it tells that the distribution of the sample is not significantly different from a normal distribution. If the test is significant ($p < 0.5$) then the distribution is significantly different from a normal distribution. In this model the Kolmogorov-Smirnov test shown in Table 4.12 is significant ($p< 0.05$), this shows that the distribution of the sample is non normal therefore satisfying the assumption of the logistic regression model.

Table 4.12 Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residuals	.242	658	.000	.909	658	.000

a. Lilliefors Significance Correction

A normal probability plot of the standardised residuals will give an indication of whether or not the assumption of normality of the random errors is appropriate. If the normal probability plot shows a straight line, it is reasonable to assume that the observed sample comes from a normal distribution. If, on the other hand, the points deviate from a straight line, there is statistical evidence against the assumption that the random errors are an independent sample from a normal distribution. The P-P plot and the Q-Q plots (see Figure 4 and Figure 5) confirm that the distribution is not normal because the dots deviate substantially from the line.

Figure 4 Normal P-P Plot of Residuals for Second Binary Model

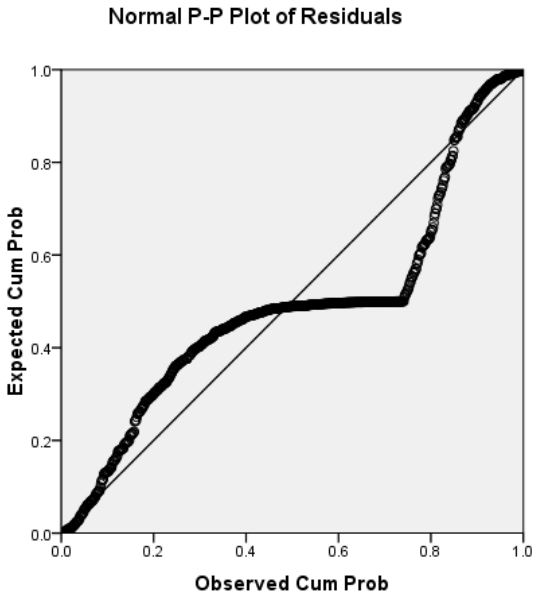
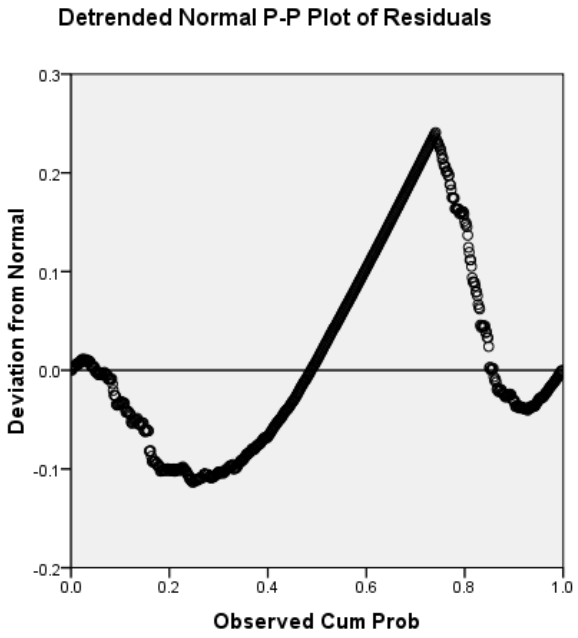


Figure 5 Detrended Normal P-P Plot of Residuals for Second Binary Model



4.3 Third Binary Logistic Model

The outcome variable of the third model had two categories: *Completed* (all students who completed) and *Not Completed* (all students who did not complete).

The Cox and Snell's R-Square indicates that 39% (see Table 4.13) of the variation is explained by the Logistic Model. Nagelkerke R-Square indicates a perfect relationship of 52% between the predictors and the prediction.

Table 4.13 Model Summary of Third Binary Model

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
9	580.090[a]	0.385	0.517

The Classification Table shows that 73% were correctly classified for the completed group (see Table 4.14). The overall rate of correctly classifying cases is 77.1%.

Table 4.14 Classification Table of Third Binary Model

Observed			Predicted		
			Timetodegree		Percentage Correct
			Not Completed	Completed	
Step 9	Timetodegree	Not Completed	299	74	80.2
		Completed	77	208	73.0
	Overall Percentage				77.1

The variables in the Equation Output Table show that Gender is Female, HLanguage is English speaking, Physics Building (PHYS1010), Architectural Representation1 (ARPL1003), Building Quantities (BUQS113) and Planning for Property Developers (ARPL1010) are significant at 5% level of significance. Female students (p=0.014) complete their degree programmes faster than Male students. Students who speak English as their HLanguage have higher chances of completing their degree programmes within the minimum time (p= 0.001).

Table 4.15 Model Variables of Third Binary Model

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 9	Gender(1)	0.539	0.220	6.005	1	0.014	1.713
	HLanguage(1)	0.743	0.230	10.447	1	0.001	2.103
	PHYS1010	0.141	0.018	59.156	1	0.000	1.152
	ARPL1003	0.246	0.055	19.730	1	0.000	1.280
	BUQS113	0.062	0.016	15.457	1	0.000	1.064
	ARPL1010	0.061	0.021	8.584	1	0.003	1.063
	Constant	-29.645	3.809	60.571	1	0.000	0.000

Table 4.15 shows the variables that were used in the model and the Exp (B) value associated with the predictor variable Female is 1.713. Hence, when this predictor is raised by one unit the odds ratio is 1.7 times as large and therefore Female students are 1.7 more times likely to complete their degree programmes within the minimum time.

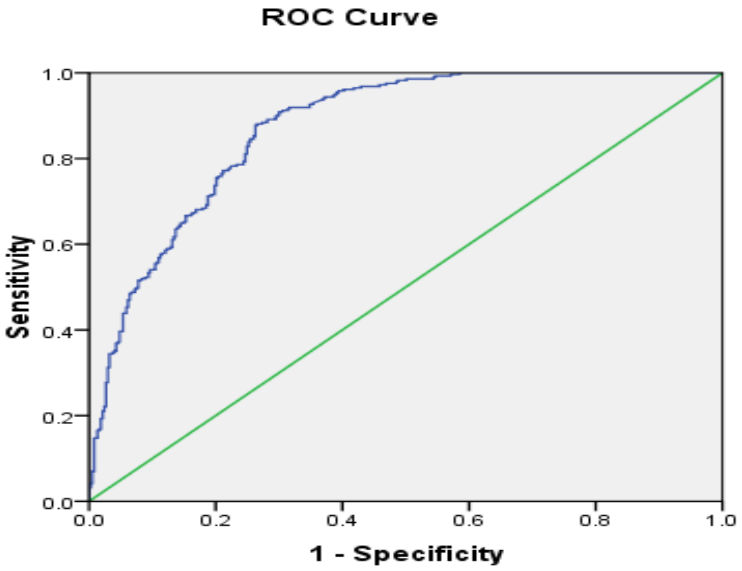
4.3.1 Model Validation: The ROC Curve

This section presents the results of the ROC Curve for Third Binary Model which tests the fit of the model.

Table 4.16 Area Under the Curve of Third Binary Model

Area	Std. Error	Asymptotic Sig.	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.872	0.013	0.000	0.846	0.898

Figure 6 ROC Curve for Third Binary Model



Diagonal segments are produced by ties.

The area under the curve is 0.872 with 95% confidence interval (0.846, 0.898) shown in Table 4.16. Also, the area under the curve is significantly different from 0.5 since p-value is 0.000 meaning that the Logistic Regression classifies the group significantly better, than by chance.

4.3.2 Residual Analysis for Third Binary Model

Figure 7 Histogram of Residuals for Third Binary Model

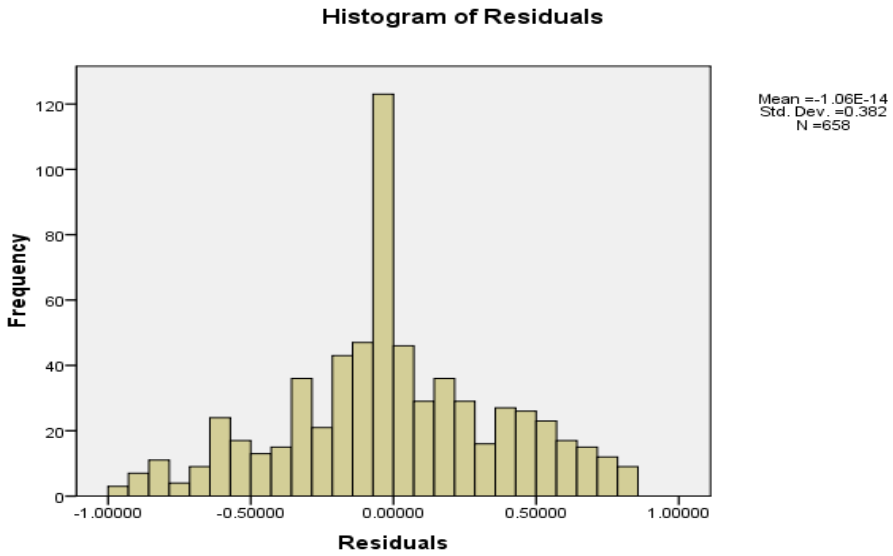


Figure 8 Normal P-P Plot of Residuals for Third Binary Model

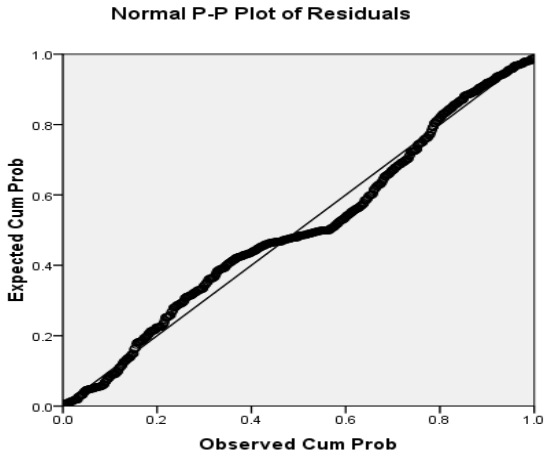
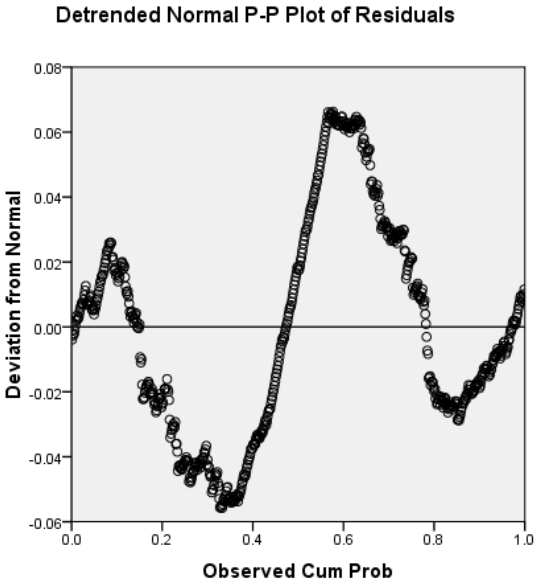


Figure 9 Detrended Normal P-P Plot of Residuals for Third Binary Model



The P-P plot and the Q-Q plots (see Figure 8 and Figure 9) also confirm that the distribution is not normal since the data points deviate substantially from the line.

4.4 Multinomial Logistic Model

The Model Information Table lists the background information about the fitting of the model. The table includes the name of the input data set, the response variable, the number of observations utilized, and the link function used (see Table 4.17). The Generalised Logit displays the output from the Generalised Logit Function. The Forward Selection Method commences with no variables in the model. For each of the independent variables, this method calculates F statistics that reflect the variable's contribution. The p-values for these F statistics are then compared to the significance level that is specified for including a variable in the model. By default, this value is 0.05. If no F statistic has a significance level greater than this value, the forward selection stops. Otherwise, the Forward Selection Method selects the variable that has the largest F statistic to the model. Thus, variables are included one by one to the model until no remaining variable produces a significant F statistic.

Table 4.17 Model Information

Data Set	WORK.SORTTEMPTABLESORTED
Response Variable	Timetodegree
Number of Response Levels	3
Model	Generalized logit
Optimization Technique	Newton-Raphson

The Response Profile Table shows the response variable values, listed according to their ordered values (see Table 4.18). The percentage of students who completed their degree programmes within the minimum time is only 26.4%, those who did not complete is 29.5% and dropouts are 44.1 %. This model predicts the odds and probabilities of completing the degree programme within the minimum time based on the explanatory variables. The reference variable was *Not Completed*.

Table 4.18 Response Profile

Ordered Value	Timetodegree	Total Frequency
1	Completed	174
2	Dropped Out	290
3	Not Completed	194

The Model Fit Statistics (see Table A2.2 in Appendix 2) contains the Akaike Information Criterion (AIC), the Schwarz Criterion (SC), and the negative of twice the log likelihood (-2 Log L) for the intercept-only model and the fitted model. The criteria, -2 Log L is used to test whether the independent variable is significant based on a Chi-Square distribution. The AIC and SC are goodness of fit measures used to compare models. Lower values of these statistics indicate a better fitting model as they reflect a trade-off between the lack of fit and the number of parameters in the model. The variable Firstyear was entered first into the model. This model has an AIC value of 1226.881, SC value of 1244.838 and the -2Log value of 1218.881(see Table A2.2 in Appendix 2) .The rest of the steps are shown in Appendix 2.

The Testing Global Null Hypotheses: Beta = 0 (Table 4.19) provides three statistics; the Likelihood Ratio, the Score Test and the Wald Statistic. The null hypothesis is that all regression coefficients of the model are 0. A significant p-value for the Likelihood Ratio Test provides evidence that at least one of the regression coefficients for an explanatory variable is nonzero. In this model, the p-value is <0.001, which is significant at the 0.05 level. The Score and Wald test are also used to test whether all the regression coefficients are 0. The Likelihood Ratio Test is the most reliable, especially for small groups.

Table 4.19 Testing Global Null Hypothesis

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	193.1013	2	<0.0001
Score	184.5290	2	<0.0001
Wald	146.5842	2	<0.0001

Overall, the model is statistically significant because the Pr>Chi-Square is less than 0.001.

The summary section is printed only for Forward Selection Method (see Table 4.20). It summarizes the order in which the explanatory variables entered the model. The output of the order is shown in Appendix 2.

Table 4.20 Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	Firstyear	2	1	184.5290	<0.0001	Firstyear
2	buqs110c	2	2	56.7311	<0.0001	buqs110c
3	HLanguage	2	3	26.7438	<0.0001	HLanguage
4	buqs113c	2	4	25.6917	<0.0001	buqs113c
5	Repeat	2	5	24.6240	<0.0001	Repeat
6	appm1014c	2	6	10.7133	0.0047	appm1014c
7	arpl1000c	2	7	11.2194	0.0037	arpl1000c
8	arpl1001c	2	8	8.4355	0.0147	arpl1001c
9	Race	2	9	8.2175	0.0164	Race
10	buqs101c	2	10	7.9575	0.0187	buqs101c
11	Gender	2	11	8.1545	0.0170	Gender
12	buqs1000c	2	12	8.0727	0.0177	buqs1000c

The Analysis of Maximum Likelihood Estimates Table lists the parameter estimates, their standard errors, and the results of the Wald test for individual parameters. The test statistics are labelled Wald Chi-Squared. They are calculated by dividing each coefficient by its standard error and squaring the result. The output shows that the predictor variable not Repeat ($p<0.0001$) impacts on the completion of degree programmes (see Table 4.21). Students who speak English as their HLanguage complete their degree programmes faster than any other languages ($p= 0.0205$). The output also shows that students who obtain a mark of 60% and above in their first year courses, have a greater chance of completing their degree programmes within the minimum time. The following courses are also significant predictors at a 5% level of significance: buqs101c ($p=0.0409$), arpl1001c ($p=0.0306$) and buqs113c ($p=0.0006$).

Table 4.21 Analysis of Maximum Likelihood Estimates

Parameter		Timetodegree	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		Completed	1	0.5510	0.7646	0.5194	0.4711
Intercept		Dropped Out	1	0.6179	0.8025	0.5929	0.4413
Repeat	Not Repeat	Completed	1	1.2385	0.3072	16.2519	<0.0001
Repeat	Not Repeat	Dropped Out	1	0.3744	0.1211	9.5525	0.0020
Race	Black	Completed	1	-0.5950	0.2089	8.1156	0.0044
Race	Black	Dropped Out	1	-0.2225	0.1840	1.4626	0.2265
HLanguage	English	Completed	1	0.3733	0.1611	5.3731	0.0205
HLanguage	English	Dropped Out	1	-0.2009	0.1288	2.4353	0.1186
Firstyear	60 and above	Completed	1	0.2576	0.1925	1.7911	0.1808
Firstyear	60 and above	Dropped Out	1	0.0376	0.1617	0.0541	0.8161
Gender	Female	Completed	1	0.1506	0.1370	1.2092	0.2715
Gender	Female	Dropped Out	1	-0.2617	0.1092	5.7426	0.0166
buqs110c	60 and above	Completed	1	0.2744	0.1963	1.9528	0.1623
buqs110c	60 and above	Dropped Out	1	-0.4036	0.1416	8.1272	0.0044
appm1014c	60 and above	Completed	1	0.2180	0.1999	1.1889	0.2756
appm1014c	60 and above	Dropped Out	1	-0.3459	0.1250	7.6617	0.0056
buqs101c	60 and above	Completed	1	0.3048	0.1491	4.1783	0.0409
buqs101c	60 and above	Dropped Out	1	-0.1958	0.1667	1.3805	0.2400
arpl1001c	60 and above	Completed	1	1.2901	0.5966	4.6767	0.0306
arpl1001c	60 and above	Dropped Out	1	-0.4386	0.7572	0.3355	0.5625
arpl1000c	60 and above	Completed	1	-0.6103	0.4874	1.5678	0.2105
arpl1000c	60 and above	Dropped Out	1	-1.0228	0.3175	10.3784	0.0013
buqs113c	60 and above	Completed	1	0.7079	0.2056	11.8573	0.0006
buqs113c	60 and above	Dropped Out	1	-0.1272	0.1271	1.0010	0.3171
buqs1000c	60 and above	Completed	1	0.2266	0.1591	2.0271	0.1545
buqs1000c	60 and above	Dropped Out	1	-0.2089	0.1161	3.2409	0.0718

The point estimate value associated with the predictor variable Not Repeat vs Repeat is 11.906 (see Table 4.22). Hence when Not Repeat is raised by one unit, the odds ratio is 12 times as large and therefore students who do not Repeat in first year, are 12 more times likely to complete their degree programmes.

Table 4.22 Odds Ratio Estimates

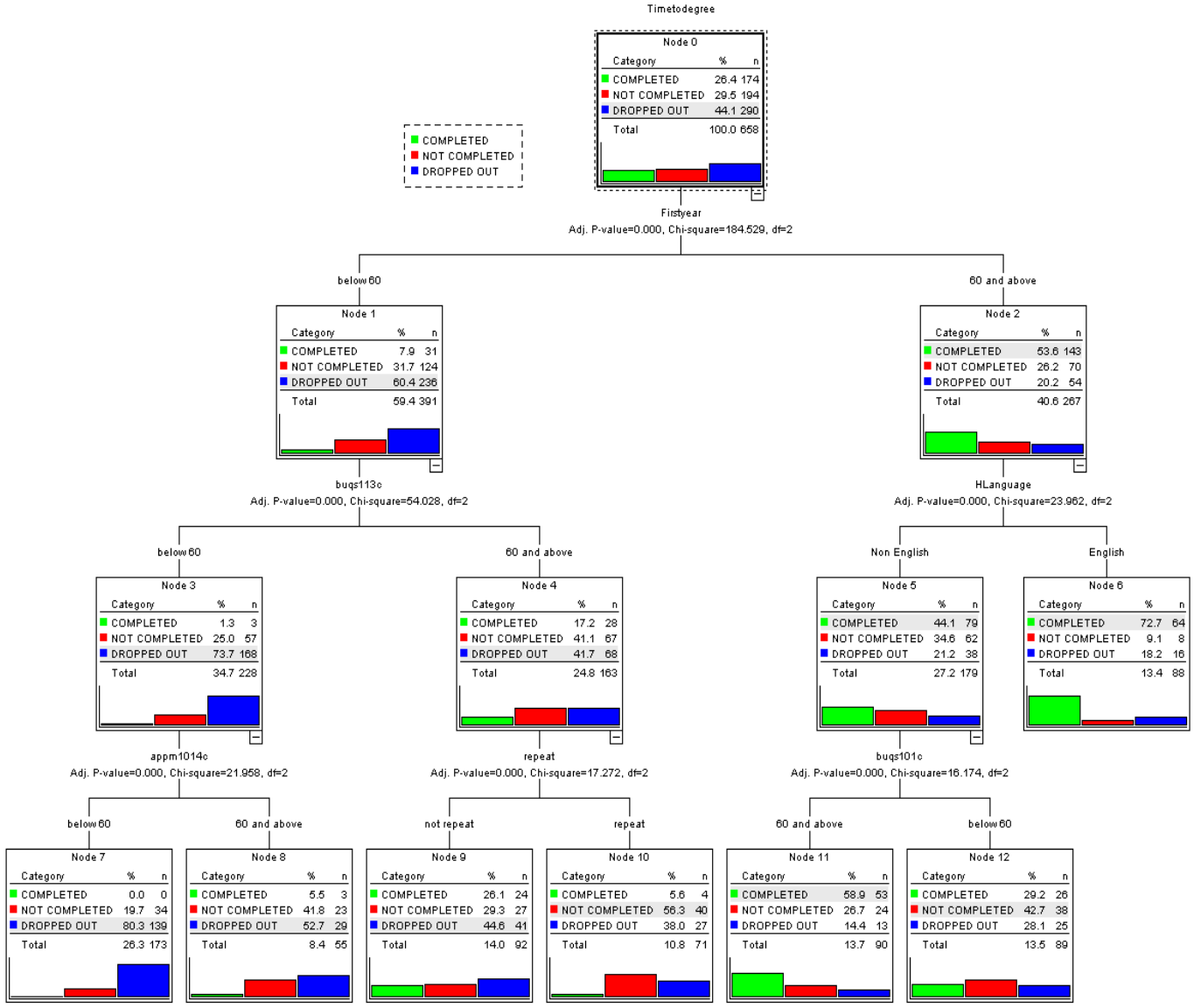
Effect	Timetodegree	Point Estimate	95% Wald Confidence Limits	
Repeat Not Repeat vs Repeat	Completed	11.906	3.571	39.699
Repeat Not Repeat vs Repeat	Dropped Out	2.114	1.315	3.399
Race Black vs White	Completed	0.304	0.134	0.690
Race Black vs White	Dropped Out	0.641	0.312	1.318
HLanguage English vs Non English	Completed	2.110	1.122	3.967
HLanguage English vs Non English	Dropped Out	0.669	0.404	1.108
Firstyear 60 and above vs below 60	Completed	1.674	0.787	3.560
Firstyear 60 and above vs below 60	Dropped Out	1.078	0.572	2.032
Gender Female vs Male	Completed	1.352	0.790	2.312
Gender Female vs Male	Dropped Out	0.593	0.386	0.909
buqs110c 60 and above vs below 60	Completed	1.731	0.802	3.737
buqs110c 60 and above vs below 60	Dropped Out	0.446	0.256	0.777
appm1014c 60 and above vs below 60	Completed	1.546	0.706	3.386
appm1014c 60 and above vs below 60	Dropped Out	0.501	0.307	0.817
buqs101c 60 and above vs below 60	Completed	1.840	1.025	3.301
buqs101c 60 and above vs below 60	Dropped Out	0.676	0.352	1.299
arpl1001c 60 and above vs below 60	Completed	13.200	1.273	136.817
arpl1001c 60 and above vs below 60	Dropped Out	0.416	0.021	8.094
arpl1000c 60 and above vs below 60	Completed	0.295	0.044	1.994
arpl1000c 60 and above vs below 60	Dropped Out	0.129	0.037	0.449
buqs113c 60 and above vs below 60	Completed	4.120	1.840	9.224
buqs113c 60 and above vs below 60	Dropped Out	0.775	0.471	1.276
buqs1000c 60 and above vs below 60	Completed	1.573	0.843	2.936
buqs1000c 60 and above vs below 60	Dropped Out	0.658	0.418	1.038

The point estimate value associated with the predictor buqs101c is 1.840 hence when buqs101c is raised by one unit, the odds ratio is 2 times as large and therefore students who pass buqs101c with a mark of 60% and above, are 2 more times likely to complete their degree programmes. Passing buqs113c with a mark of 60% and above increases the chances of completing the degree programme by 4 times. The point estimate value associated with the predictor arpl1001c is 13.200. Hence when arpl1001c is raised by one unit, the odds ratio is 13 times as large and therefore students who pass arpl1001c with a mark of 60% and above, are 13 more times likely to complete their degree programmes.

4.5 Classification Tree Model

This section presents the results of the Classification Tree Method on TTD. The CHAID growing method was selected and cross validation was used to validate the model.

Figure 10 Classification Tree for Time to Degree



The CHAID Classification Tree Method shows that Firstyear (Adj p=0.000 and Chi-Squared = 184.529) is the best predictor of TTD (see Figure 10). Students with average marks of 60 and above in their Firstyear, have a high chance of completing their degree programme within the minimum time (53.6% completed). The dropout rate for students with average first year marks of below 60%, is 60.4% and there is a 31.7% chance of not completing the programme. The next best predictor is HLanguage (Adj p=0.000 and Chi-Squared= 23.962). The Classification Tree shows that the completion rate of students who speak English as their HLanguage is 72.7% whilst for Non-English; the completion rate is 44.1%. Lastly, the other important predictor of TTD is buqs101c (Adj p=0.000 and Chi-Squared= 16.174). The completion rate of passing buqs101c with a mark of 60% and above, in first year, is 58.9%. The remaining variables not listed in the Classification Tree have no significant influence on the prediction of TTD in the model when using CHAID Growing Method.

Table 4.23 Risk Estimates

Method	Estimate	Std. Error
Resubstitution	0.386	0.019
Cross Validation	0.406	0.019

The Cross Validation Risk Estimate of 0.406 (see Table 4.23) which is the average of the risks across the 10 test samples, indicates that the category predicted by the model (Completed, Not Completed and Dropped Out), is wrong for 40.6% of the cases. The risk of misclassifying cases is approximately 41%.

The Classification Table (see Table 4.24), shows that the rate of correctly classifying students who completed within the minimum time is 67.2% and the overall classification rate is 61.4%.

Table 4.24 Classification Table of Multinomial Logistic Model

Observed	Predicted			
	Completed	Not Completed	Dropped Out	Percent Correct
Completed	117	30	27	67.2%
Not Completed	32	78	84	40.2%
Dropped Out	29	52	209	72.1%
Overall Percentage	27.1%	24.3%	48.6%	61.4%

4.6 Scatter Plots of Matric Subjects on University Courses

This section presents the results of the Scatter Plots of Matric Subjects against the university courses, which were found significant in this analysis:

The scatter plots of Biology vs BUQS110, Biology vs Buqs113 and Biology vs PHYS1010 shown in Appendix 3 indicates a weak relationship between the two variables. Changes in Biology marks at high school will not significantly determine whether the first year student will pass BUQS110, BUQS113, PHYS1010 and ARPL1010. There is a strong positive relationship between Physical Science, Mathematics and university first year courses. Hence, any change in Physical Science and Mathematics marks, will result in a reasonably predictable change in university first year course success rates which is consistent with the results of Eeden, Beer and Coetzee (2001), they found that school marks for Mathematics, Science and English were all related to first year performance. The plots further suggest that there is no correlation between the university first year mark and the following matric subjects: English First Language, English Second Language and Biology. Passing these subjects will not determine a pass or fail in the university first year mark.

CHAPTER 5

DISCUSSIONS AND CONCLUSION

This final chapter presents a brief overview of the study and discussions of the results of this report. This chapter will conclude with a general discussion on the models developed.

The main objective of this research is to investigate, in the cohort of 2003, 2004, 2005, 2006 and 2007 (to ensure a sizeable sample) for undergraduate students in the Schools of Construction and Management; and of Architecture and Planning at Wits, the factors that impact on TTD.
The following questions will be answered in this research:

- Do Gender, Race, Home Language, Financial Aid, Residence Status, Matric Aggregate, and University Courses affect the time taken to complete degree programmes?
- Which of these variables are the most important in predicting TTD?

The Binary Logistic Regression Model was run in SPSS and the Forward Likelihood Ratio technique selected at 5% level of significance, starts with no predictor variables in the model, and then enters variables one at a time, at each step adding the predictor with the largest score statistic, whose significance value is less than 0.05. At each step, SPSS checks for significance of variables already in the model to see if any should be removed. Removal is based on the Likelihood Ratio Test. The Hosmer-Lemeshow Goodness Test and the ROC were selected to validate the models built. The Multinomial Logistic Model was run in SAS Enterprise Guide where, the outcome variable was selected as unordered and the reference level was the value *Not Completed.*

5.1 Discussion on the Results of Binary Logistic Regression

Binary Logistic Regression is a technique used when the dependent variable is a dichotomous variable (has two values). Logistic regression uses Binomial Probability Theory in which there are only two outcome categories.

The First Binary Logistic Model was run in SPSS to analyse the impact of Matric Aggregate, Gender and Race on university first year success. The outcome variable of the first model had

the following categories: *Pass* (all students who passed their first year with an average mark of 50% and above) and *Fail* (all students who got an average first year mark of 49% and below). This model indicates that Aggregate (1) is significant at 5% which is consistent with Yathavan (2008) and Latief (2005) , meaning that students who obtain an average matric mark of 60% and above, have a higher probability of passing university first year. Gender and Race are not significant at a 5% level of significance. This result was also consistent with that of the Working Group on Retention and Throughput at Wits (2010). The model was validated using the Hosmer and Lemeshow Test. The Chi-Square (degrees of freedom = 4) value of the model is 1.635 and the corresponding H-L Statistic value is 0.802, implying the model fits the data adequately. The Second Binary Logistic Model was run to analyse the impact of selected independent variables on the TTD. The outcome variable of the second model had two categories: *Completed* (all students who completed in minimum time) and *Not Completed* (all students who did not complete). The following predictors are significant at 5% level of significance; Gender is Female, Repeat (1), HLanguage is English, BUQS110, ARPL1003 and BUQS113. Female students (p=0.043) complete their degree programmes faster than the Male students, as also indicated by Zhu (2003). Not Repeats (p=0.000) in first year and students who speak English as HLanguage also have a greater chance of completing within the minimum time. This model was validated by the ROC Curve. The area under the curve determined by Mann-Whitney U Statistic for this model which is 0.917 with 95% confidence interval (0.896, 0.938). This indicates that the model performance is excellent and it fits the data acceptably well.

The Third Binary Logistic Model also has two categories, namely: *Completed* (all students who completed their degree programmes within the minimum time and after) and *Not Completed* (all students who did not complete). In this model the following predictors were significant at 5% level of significance; Gender is Female, HLanguage is English, PHYS1010, ARPL1003, BUQS113 and ARPL1010. Therefore Female students (p=0.014) complete their degree programmes faster than Male students. Students who speak English as their HLanguage have a higher chance of completing their degree programmes within the minimum time (p= 0.001). The area under the curve in this model is 0.872 with a 95% confidence interval (0.846, 0.898). The area under the curve is significantly different from 0.5, since p-value is 0.000, meaning that the Logistic Regression classifies the group significantly better than by chance.

5.2 Discussion on the Results of Multinomial Logistic Regression

The Multinomial Logistic Regression Model was generalised from the Binary Logistic Regression (Aldrich and Nelson, 1984, Hosmer and Lemeshow, 2000). A Multinomial Logistic Model provides several equations for classifying individuals into one of many categories. This type of regression is similar to Binary Logistic Regression, but it is more general because the dependent variable is not restricted to only two categories. A Multinomial Logistic Model was also developed where the outcome variable was a polytomous outcome variable with more than two categories. The values of the outcome variable are *Completed* (all students who graduated within the minimum time of their degree programmes), *Not Completed* (all students who completed after the minimum time and those who are still registered) and *Dropped Out* (all students who were excluded from the programmes and drop outs). The following predictors were significant at 5% level of significance; HLanguage is English, Not Repeat, buqs101c, arpl1000c and buqs113c. Therefore students who do not repeat their first year and speak English as their Home Language complete their degree programmes within the minimum time and, this is consistent with Latief (2005) who also found out that Home Language was a significant predictor on Throughput. The output also indicates that students who obtain an average mark of 60% and above, in their first year courses, have a greater chance of completing their degree programmes within the minimum time. Spoerre (2010) also found out that course completion rates are significant predictors in student retention and graduation.

5.3 Discussion on the Results of the Classification Model

Classification is the process of forming groups from a large set of cases based on their similarities and dissimilarities (Fletcher *et al*, 2001). The Classification Tree procedure creates a tree-based classification model which predicts values of a dependent (target) variable based on values of independent (predictor) variables. The CHAID Classification Tree Method shows that Firstyear (Adj p=0.000 and Chi-Squared = 184.529) is the best predictor of TTD. Students with average marks of 60% and above, in their Firstyear have a high chance of completing their degree programmes within the minimum time (53.6% completed). The next best predictor is HLanguage (Adj p=0.000 and Chi-Squared= 23.962). The Classification Tree shows that the completion rate of students who speak English as their HLanguage is 72.7%. Lastly, the other

important predictor of TTD is buqs101c (Adj p=0.000 and Chi-Squared= 16.174). This model was validated by the Cross Validation Technique and has a cross validation risk estimate of 0.406 which indicates that the category predicted by the model (*Completed*, *Not Completed* and *Dropped Out*) is wrong for 40.6% of the cases. The risk of misclassifying cases is approximately 41%. The classification results indicate that the rate of correctly classifying students who completed within the minimum time is 67.2% and the overall classification rate is 61.4%.

5.4 Limitations of the Study

The ability of generalising these findings is limited since the data was collected only for the Schools of Construction and Management; and of Architecture and Planning. Caution must be taken when generalising the results to other schools unless they have students with similar characteristics. The models cannot be applied to other schools and faculties as the admission criteria requirements are different in these schools. A further limitation is that the sample size was small although it included five cohorts, and small sample sizes might lead to biased regression estimates. The actual matric subjects were not regressed in these models as in Yathavan (2008), but instead, Matric Aggregate was created by averaging the scores of the matric subjects for each student. The reason for not using the actual matric subjects was that they had two different grades (HG and SG) and some of students had missing data on the matric scores. Removing these missing cases would have further reduced the sample size.

5.5 General Discussion of the Models

The logistic regression models indicate that HLanguage-English and BUQS113 are the most important predictors of TTD. The possible reason why English speaking students do well in their studies is that most of the university courses are conducted in English and this group understand the courses better than the non-English speaking group. Matric Aggregate is an important predictor of university first year success though it has no impact on TTD. This is consistent with Mitchell et al (1997), these researchers noted that that matriculation mark is a reasonably good predictor of pass/fail at University. Robbins et al (2004) found that approximately 25 percent of the variance in the students' success can be attributed to their high school performance. The Second and Third Binary Logistic Regression Models indicate that Female students complete faster than Male students, and this finding is consistent with Zhu (2003), which determined that

the percentage of Female students completing in minimum time, was higher than their Male counterparts. The Classification Tree Method indicated that obtaining an average mark of 60% and above in first year increases the chances of completing degree programmes, as compared to a mark of below 60. The completion rates of students with a first year mark of 60% and above are 53.6%. This finding is consistent with the Shulruf et al (2010) paper, which found that high pass rates in the first year were associated with completion in the third year and high pass rates in the third year was the most significant factor for completion of the degree.

The last section of this chapter gives the conclusion of the study basing on the results obtained in the analysis section.

5.6 Conclusion

The results of the Logistic Regression Models and the Classification Tree Model shows that Gender is Female, Home Language is English, Matric Aggregate of 60% and above, BUQS110, ARPL1003, BUQS113, BUQS1000, PHYS1010, ARPL1010 and First Year Average Mark of 60% and above have, a significant impact on predicting the time taken to complete the degree programme. The results of the study on these courses which positively predict time to completion will be given to the management of the School of Construction and Management and the School of Architecture and Planning. If the schools make students aware of the findings of this research, this will assist students with passing their courses and meet the minimum requirements in order to register for the next year of study, and to avoid losing their scholarships, bursaries and financial grants. The university will in turn get more output subsidy when it graduates students' by reducing the time taken to complete degree programmes and also the objective of obtaining a higher throughput rate is achieved. Throughput is one of the factors that the government uses for funding a university (Department of Higher Education, 2001). It is suggested that students attending high school should be made aware of the importance of getting an average mark of 60% and above in their matric subjects as this has a significant impact on TTD. Race, Financial Aid, and Residence Status proved to be non significant in this study.

The Logistic Regression Models indicates that HLanguage-English and BUQS113 are the most important predictors of TTD and the Classification Tree Model indicated that passing first year at

university increased the chances of completing a degree programme in minimum time. These predictors are important in the sense that they have highly significant odds ratios in all of the models tested. Improving the TTD results in more students graduating and this potentially increases the number of graduates available and eligible for enrolling for postgraduate studies; this would be in line with the Wits' vision found in the Teaching and Learning Plan 2010- 2014 (2010), which states that by 2014, at least 40% of the students should be registered in postgraduate programmes. Students who complete within the minimum can enter the job market sooner with higher chance of obtaining the job, thereby reducing the shortage of engineering professionals in the country. Stock, Finegan and Siegfried (2009) noted that those students, who fail completely to earn a degree, are affected by costs in terms of psychological costs and delayed entry into alternative careers that better match their skills. Therefore, completing within the minimum time will significantly increase the proportion of artisans, technicians, engineers, quantity surveyors and technologists to do the hands-on, practical work required on factory floors at chemical plants.

5.7 Possible Future Research

Most of the independent variables which were analysed in this research are quantitative. Future studies could include both qualitative and quantitative independent variables. The qualitative variables must be collected directly from the students through administering appropriate questionnaires. Research into students' expectations about university study and students' commitment to academic success (Branxton, Bray and Berger, 2000) could also be explored in future studies. Other interesting qualitative variables to consider in future studies are: under preparedness (students not being academically strong enough to the university), students' approach to learning, their attitude and expectations, students' taking less responsibility for their learning, issues of the students' life and other pressures such as personal, social, financial or family matters. These factors were found to have a significant impact on TTD in the Faculty of Engineering and Built Environment (Working Group on Retention and Throughput, 2003). There is also a need to use large samples of students with the same characteristics and explore the original matric subjects instead of using the average matric mark. It is also suggested that future research focus on determining the financial implications of completing degree programmes within and after the prescribed time, to both the student and the university. The State of Texas

estimated that the cost to students (or parents) for a degree, completed in the prescribed four years, is $41,636, while the cost jumps to $60,264 if the degree takes six years. The cost to the state jumps from $24,948 to $31,752 per student (Texas Higher Education Coordinating Board, 1996). Research of this type could assist in the future financial planning of the university.

REFERENCES

Aldrich, J.H and Nelson, F.D. (1984) Linear probability, logit and probit models, 1st ed., Newbury Park, CA: SAGE Publications.

Astin, A., Tsui, L. and Avalos, J. (1996) *Degree attainment rates at American Colleges and Universities: Effects of Race, Gender, and Institutional Type*, University of California, Los Angeles, CA.

Berry, W. D. (1993) Understanding regression assumptions. Newbury Park, CA: SAGE Publications.

Branxton, J., Bray, N., and Berger, J. (2000) Faculty teaching skills and their influence on the college student departure process. Journal of College Student Development, vol. 4 1, pp. 215 -2 27.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) Classification and Regression Trees. Wadsworth, Belmont.

Cramer, D. (2003) Advanced Quantitative Data Analysis. Philadelphia, PA: Open University Press.

Czepiel, S.A. (2002) *Maximum likelihood estimation of logistic regression models*: Theory and Implementation.

Department of Higher Education. (2001) Funding of Public Higher Education: *A New Framework,* Ministry of Education, Pretoria, South Africa.

Dilorio, F.C. (1991) SAS Applications Programming: A gentle introduction, Belmont, California, Duxbury Press.

Eeden, V. R., Beer, D. and Coetzee, C. H. (2001) Cognitive ability, learning potential, and personality traits as predictors of academic achievement by engineering and other science and technology students. *South African Journal of Higher Education,* vol. 15, pp.171-179.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, B. A., and Shaywitz, S. E. (1993) Classification of learning disabilities: Relationships with other childhood disorders. In G. R. Lyon, D. Gray, J. Kavanagh, & N. Krasnegor (Eds.), Better understanding learning disabilities, pp. 27–55. New York: Paul H. Brookes.

Fletcher, J. M., Lyon, G. Reid, Marcia, B., Stuebing, K. K., Francis, D. J., Olson, R. K., Shaywitz, S. E., Shaywitz, B. A. (2001) Classification of Learning Disabilities: An Evidence Based Evaluation.

Frontline Systems Inc. (2010) Risk Solver Basic: Free Monte Carlo Solver In Excel, Retrieved on 05 November 2011 http://www.solver.com/xlminer/help/Ctree/ClassificationTree_intro.html

Hall, M. (1999) Why students take more than four years to graduate, *A paper presented at the Association for Institutional Research Seattle*, WA, South Eastern Louisiana University.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd ed., Toronto: John Wiley and Sons.

Knight, W.E. (2004) Time to Bachelor's Degree Attainment: An Application of Descriptive, Bivariate and Multiple Regression Techniques, *Association for Institutional Research*.

Lam, L.M.T. (1999) Assessing financial aid impacts on TTD for nontransfer undergraduate students at a large urban public university, *Paper presented at the Association for Institutional Research Forum*, Settle, WA.

Latief, A. (2005)*Throughput of UWC students who did at least one semester of third year Statistics*, Technical paper, M.Sc., University of Western Cape.

Lawless, A. (2005) Number and needs: Addressing imbalances in the civil Engineering profession, Halfway House: South African Institute for Civil Engineering.

Levesque, R. (2007) SPSS Programming and Data Management: A Guide for SPSS and SAS Users, SPSS Inc, 4 th ed., Chicago Ill.

Mitchell, G., Fridjhon, P. and Haupt, J. (1997) On the relationship between the Matriculation examination and University performance in South Africa—1980 to 1991. *South African Journal of Science*. 93, 382-387.

Morris, R., and Fletcher, J. M. (1988) Classification in neuropsychology: A theoretical framework and research paradigm. *Journal of Clinical and Experimental Neuropsychology*, vol.10, pp. 640–658.

Oracle Business Intelligence Dashboards (2011), University of Witwatersrand, Johannesburg, South Africa: http://146.141.11.136:9704/analytics/saw.dll?Dashboard.

Reeves, R.J and Haynes, A. (2008) *A descriptive analysis of bachelor's degree recipients of 2005-2006*, Clearing House Research.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstorm, A. (2004) Do psychosocial and study skill factors predict college outcomes? *Psychological Bulletin.* 130, 261-288.

Roux, N. J. Le., Bothma, A. and Botha, H. L. (2004) Statistical properties of indicators of first-year performance at University. *The Journal of ORSSA*, vol.20, pp. 161-178.

Scott, D. (2005) How long do people spend in tertiary education? Ministry of Education, Wellington, New Zealand.

Scott, J, Brown, K.J., and Yang, X. (2007) Summer school enrolment and TTD, *A paper presented at the Association for Institutional Research Forum*, Kansas, Missouri.

Shulruf, B, Tumen, S. and Hattie, J. (2010) Student pathways in a New Zealand polytechnic: Key factors for completion, *International Journal of Vocational and Technical Education*, vol.2, no.4, pp. 67-74.

Skinner, H. A. (1981) Toward the integration of classification theory and methods. *Journal of Abnormal Psychology*, vol. 90, pp. 68–87.

Spoerre, G.A. (2010) *Student Retention in two-year Construction Management Programs*, Southern Illinois University Carbondale.

SPSS Inc. (2004) SPSS Classification Tree 13.0, Copyright © 2004 by SPSS Inc, United States of America.

SPSS Inc. (2007) SPSS Missing Value Analysis 16.0, Copyright © 2007 by SPSS Inc, United States of America.

Stock, W.A., Finegan, T.A., and Siegfried, J.J. (2009) Completing an Economics PhD in five years, *American Economics Review*: *Papers and Proceedings*, vol. 99, no. 2, pp. 624-629.

Stone M. (1977) An asymptotic equivalence of choice of model by cross validation and Akaike's criterion. *Journal of the Royal Statistical Society* , Series B 39, pp. 44–47.

Sunjka, B.P. (2010) Future of Engineering in South Africa, Retrieved 30 October 2011 from http://www.saiie.co.za/index.php?option=com_content&view=article&id=222:future-of-engineering-in-south-africa&catid=95:newsletter-items&Itemid=212.

Texas Higher Education Coordinating Board. (1996) *Ten strategies and their financial implications for reducing TTD in Texas Universities*, Austin, TX.

The South African Council on Higher Education. (2009) *The State of Higher Education Report*, Higher Education Monitor no. 8, Pretoria, South Africa.

The South African Council on Higher Education. (2010) *Access and throughput in South African Higher Education Report*: Three Case Studies, HE Monitor no. 9, Pretoria, South Africa.

The Teaching and Learning Plan 2010 to 2014. (2010) Academic Planning Office, University of Witwatersrand, Johannesburg, South Africa.

Yathavan, V. (2008) Analysing first year students' performance in the commerce faculty at the University of the Witwatersrand, University of Witwatersrand, Johannesburg, South Africa.

Van Houwelingen JC, Le Cessie S. (1990) Predictive values of statistical models. Stat Med,vol. 9, pp. 1303–1325.

Wits 2013 Strategy. (2011) Towards Global Top-League Status, University of Witwatersrand, Johannesburg, South Africa.

Working Group on Retention and Throughput . (2003) Success at the University of the Witwatersrand, *Report.*

Zhu, L. (2003) Who attains a bachelor's degree in four years? *Paper presented at the Northeast Association for Institutional Research Annual Conference*, Newport, RI.

Appendix 1: Course Descriptions and Variable Coding

Course Code	Course Title
BUQS110	Theory & Practice of Qs I
PHYS1010	Physics Building
MATH1012	Mathematics BQT
APPM1014	Applied Mathematics BQ
BUQS101	Production Planning & Design I
APPM1000	Applied Mathematics 18
ARPL1005	Architectural Discourse I
ARPL1003	Architectural Representation I
ARPL1002	Introduction to Structures
ARPL1001	Theory and Practice of Construction
ARPL1000	Architectural Design ND Theory 1
BUQS113	Building Quantities 1
BUQS1000	Construction Planning and Design
ARPL1010	Planning for Property Developers
ARPL1004	Introduction to Built Environment

Table A1.1 Course Descriptions

Categorical Variables Codings

		Frequency	Parameter coding (1)
Admissionpoints	23 and above	504	1.000
	below 23	154	0.000
Repeat	Not Repeat	510	1.000
	Repeat	148	0.000
Financial Aid	No Financial	368	1.000
	Financial	290	0.000
Race	White	110	1.000
	Black	548	0.000
HLanguage	English	219	1.000
	Non English	439	0.000
Firstyear	60 and above	267	1.000
	below 60	391	0.000
Aggregate	60 and above	305	1.000
	below 60	353	0.000
Gender	Female	247	1.000
	Male	411	0.000

Table A1.2 SPSS Categorical Coding

Appendix 2: SAS Multinomial Logistic Model Output

Table A2.1 Logistic Regression Results

Model Information		
Data Set	WORK.SORTTEMPTABLESORTED	
Response Variable	Timetodegree	Timetodegree
Number of Response Levels	3	
Model	generalized logit	
Optimization Technique	Newton-Raphson	

Number of Observations Read	658
Number of Observations Used	658

Response Profile		
Ordered Value	Timetodegree	Total Frequency
1	Completed	174
2	dropped out	290
3	Not Completed	194

Logits modeled use Timetodegree='Not Completed' as the reference category.

Forward Selection Procedure

Class Level Information		
Class	Value	Design Variables
Repeat	Not Repeat	1
	Repeat	-1
Financialaid	Financial	1
	No Financial	-1
Race	Black	1
	White	-1
HLanguage	English	1
	Non English	-1
Aggregate	60 and above	1
	below 60	-1
Firstyear	60 and above	1
	below 60	-1
Admissionpoints	23 and above	1
	below 23	-1
Gender	Female	1
	Male	-1

buqs110c	60 and above	1
	below 60	-1
phys1010c	60 and above	1
	below 60	-1
math1012c	60 and above	1
	below 60	-1
appm1014c	60 and above	1
	below 60	-1
buqs101c	60 and above	1
	below 60	-1
Laws1000c	60 and above	1
	below 60	-1
appm1000c	60 and above	1
	below 60	-1
arpl1005c	60 and above	1
	below 60	-1
arpl1003c	60 and above	1
	below 60	-1
arpl1002c	60 and above	1
	below 60	-1
arpl1001c	60 and above	1
	below 60	-1
arpl1000c	60 and above	1
	below 60	-1
buqs113c	60 and above	1
	below 60	-1
arch118c	60 and above	1
	below 60	-1
buqs1000c	60 and above	1
	below 60	-1
arpl1010c	60 and above	1
	below 60	-1
arpl1004c	60 and above	1
	below 60	-1

Table A2.2. Step 1: Effect Firstyear entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1226.881
SC	1424.961	1244.838
-2 Log L	1411.983	1218.881

R-Square	0.2543	Max-rescaled R-Square	0.2880

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	193.1013	2	<0.0001
Score	184.5290	2	<0.0001
Wald	146.5842	2	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
201.5258	48	<0.0001

Table A2.3. Step 2: Effect buqs110c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1178.075
SC	1424.961	1205.011
-2 Log L	1411.983	1166.075

R-Square	0.3118	Max-rescaled R-Square	0.3531

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	245.9071	4	<0.0001
Score	224.7124	4	<0.0001
Wald	163.1891	4	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
152.6448	46	<0.0001

Table A2.4. Step 3: Effect HLanguage entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1155.182
SC	1424.961	1191.096
-2 Log L	1411.983	1139.182

R-Square	0.3394	**Max-rescaled R-Square**	0.3843

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	272.8005	6	<0.0001
Score	242.7821	6	<0.0001
Wald	166.3167	6	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
129.8308	44	<0.0001

Table A2.5. Step 4: Effect buqs113c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1133.532
SC	1424.961	1178.424
-2 Log L	1411.983	1113.532

R-Square	0.3646	**Max-rescaled R-Square**	0.4129

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	298.4501	8	<0.0001
Score	256.3970	8	<0.0001
Wald	164.8690	8	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
107.6487	42	<0.0001

Table A2.6. Step 5: Effect Repeat entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1112.024
SC	1424.961	1165.895
-2 Log L	1411.983	1088.024

R-Square	0.3888	**Max-rescaled R-Square**	0.4403

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	323.9581	10	<0.0001
Score	278.2898	10	<0.0001
Wald	180.3173	10	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
88.9062	40	<0.0001

Table A2.7. Step 6: Effect appm1014c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1105.513
SC	1424.961	1168.362
-2 Log L	1411.983	1077.513

R-Square	0.3985	Max-rescaled R-Square	0.4513

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	334.4699	12	<0.0001
Score	286.4054	12	<0.0001
Wald	183.1486	12	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
78.6801	38	0.0001

Table A2.8. Step 7: Effect arpl1000c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1097.738
SC	1424.961	1169.565
-2 Log L	1411.983	1065.738

R-Square	0.4092	Max-rescaled R-Square	0.4634

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	346.2449	14	<0.0001
Score	298.6921	14	<0.0001
Wald	191.9354	14	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
68.2159	36	0.0009

Table A2.9. Step 8: Effect arpl1001c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1096.083
SC	1424.961	1176.889
-2 Log L	1411.983	1060.083

R-Square	0.4142	**Max-rescaled R-Square**	0.4691

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	351.8995	16	<0.0001
Score	299.9801	16	<0.0001
Wald	188.5650	16	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
60.4787	34	0.0034

Table A2.10. Step 9: Effect Race entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1091.694
SC	1424.961	1181.478
-2 Log L	1411.983	1051.694

R-Square	0.4216	**Max-rescaled R-Square**	0.4775

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	360.2888	18	<0.0001
Score	308.8084	18	<0.0001
Wald	190.8826	18	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
52.8882	32	0.0115

Table A2.11. Step 10: Effect buqs101c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1087.734
SC	1424.961	1186.497
-2 Log L	1411.983	1043.734

R-Square	0.4286	**Max-rescaled R-Square**	0.4854

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	368.2484	20	<0.0001
Score	319.5028	20	<0.0001
Wald	194.9578	20	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
44.9866	30	0.0387

Table A2.12. Step 11: Effect Gender entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1083.475
SC	1424.961	1191.216
-2 Log L	1411.983	1035.475

R-Square	0.4357	**Max-rescaled R-Square**	0.4934

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	376.5077	22	<0.0001
Score	325.6438	22	<0.0001
Wald	196.9887	22	<0.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
37.3432	28	0.1115

Table A2.13. Step 12: Effect buqs1000c entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

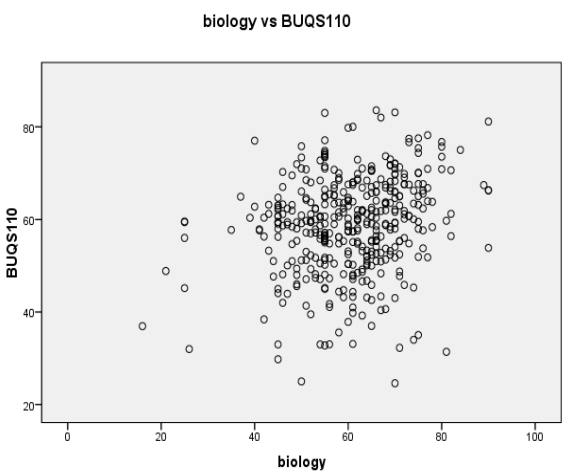
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1415.983	1079.523
SC	1424.961	1196.242
-2 Log L	1411.983	1027.523

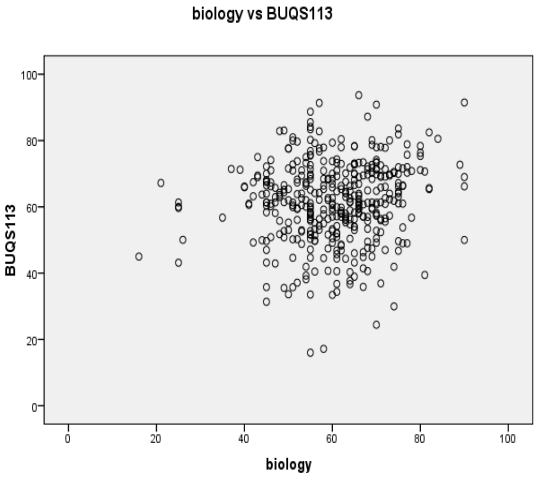
R-Square	0.4425	**Max-rescaled R-Square**	0.5011

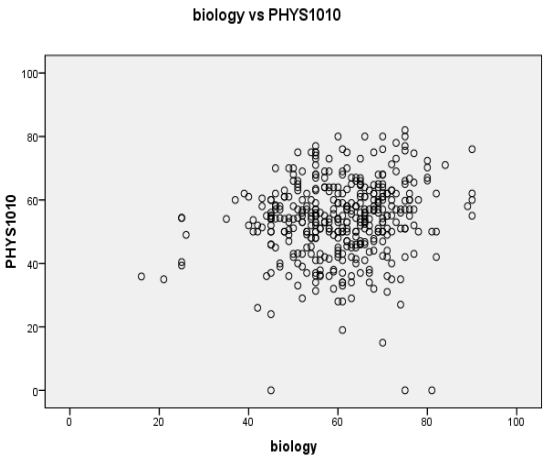
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	384.4598	24	<0.0001
Score	329.3983	24	<0.0001
Wald	195.9516	24	<0.0001

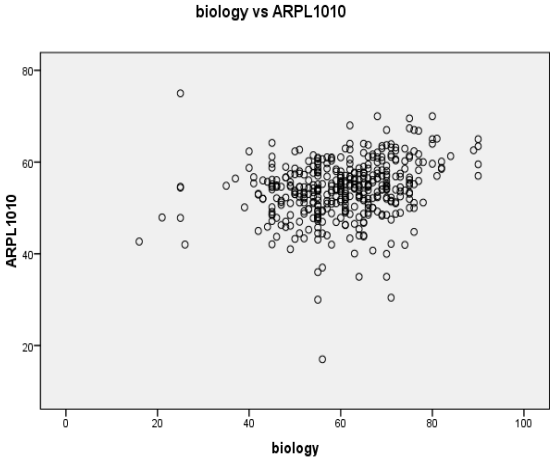
Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
29.5850	26	0.2852

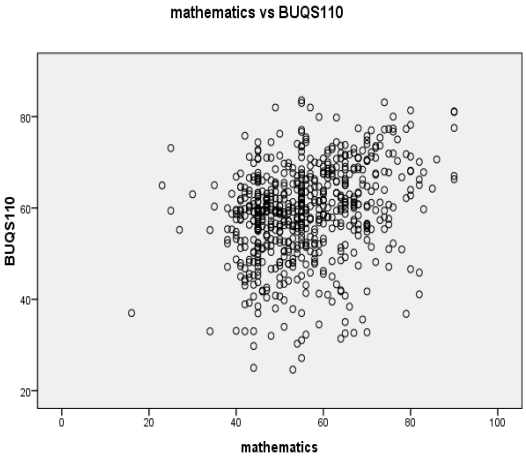
Appendix 3: Scatter Plots

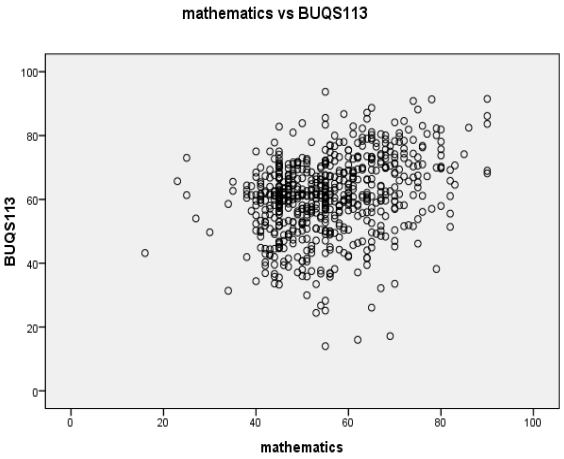


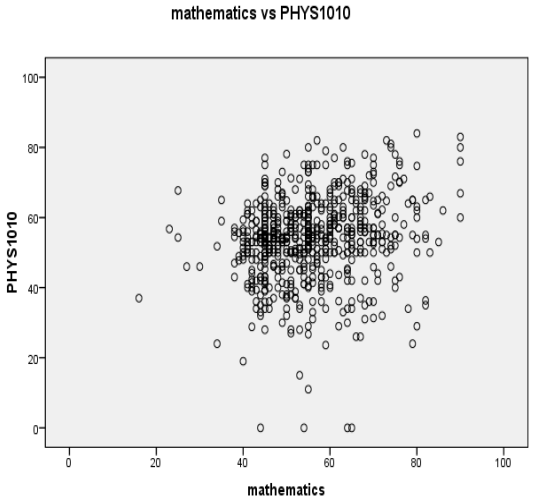


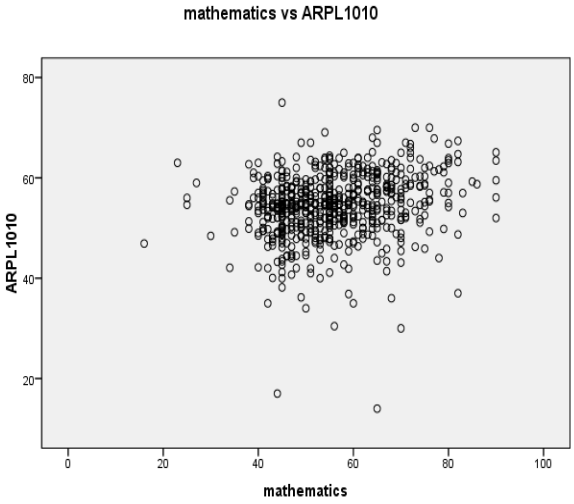


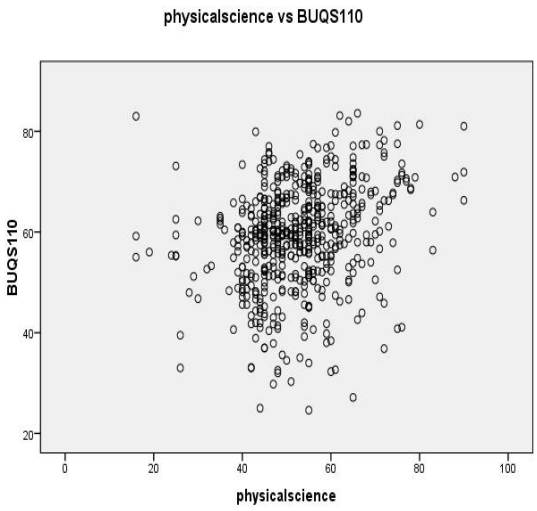


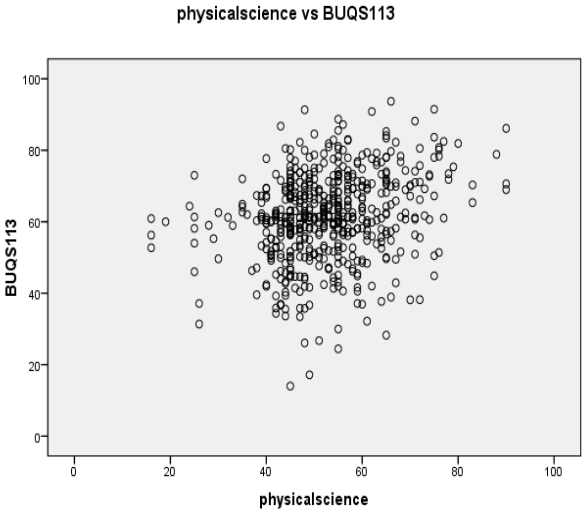


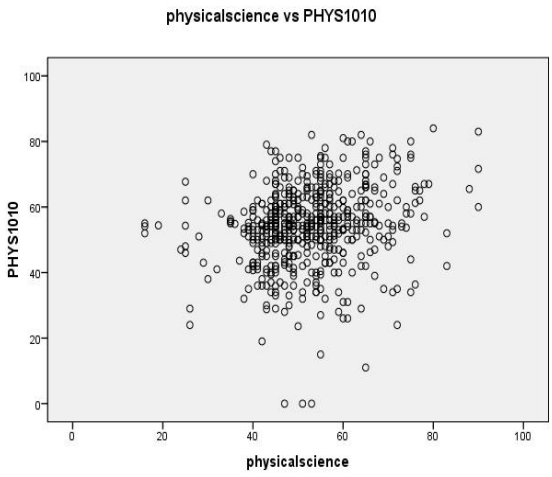


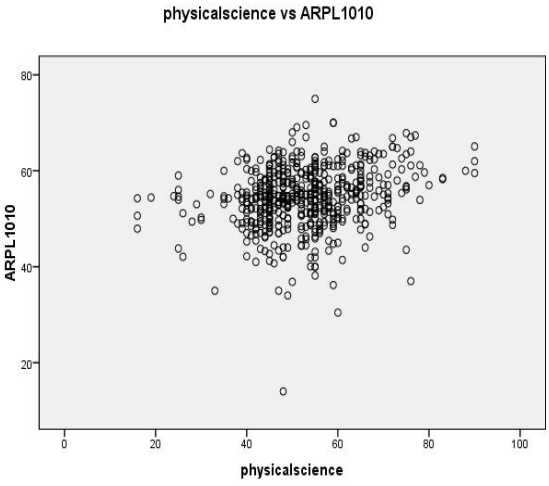

mathematics vs PHYS1010
PHYS1010
mathematics

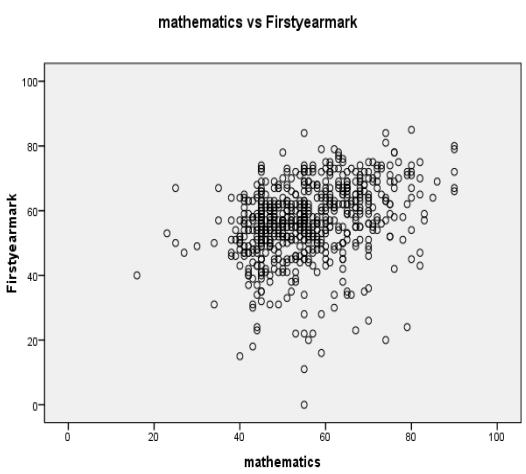

mathematics vs ARPL1010
ARPL1010
mathematics

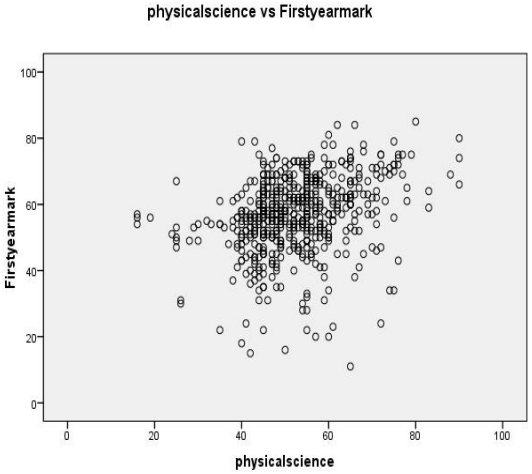

physicalscience vs BUQS110
BUQS110
physicalscience

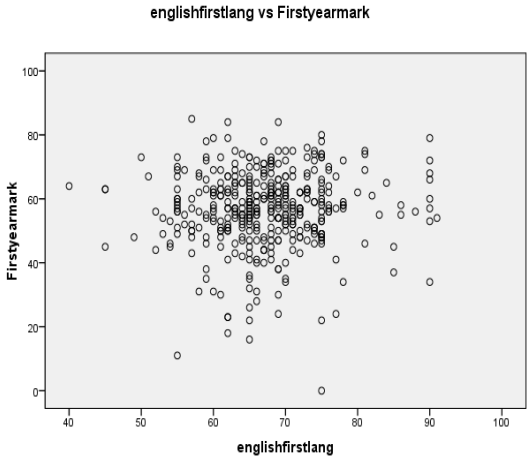

physicalscience vs BUQS113
BUQS113
physicalscience

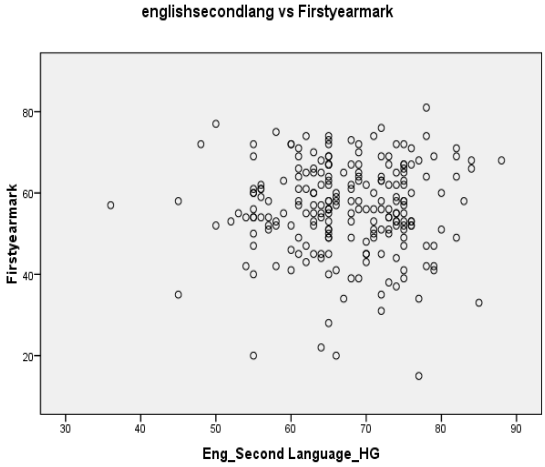

physicalscience vs PHYS1010
PHYS1010
physicalscience

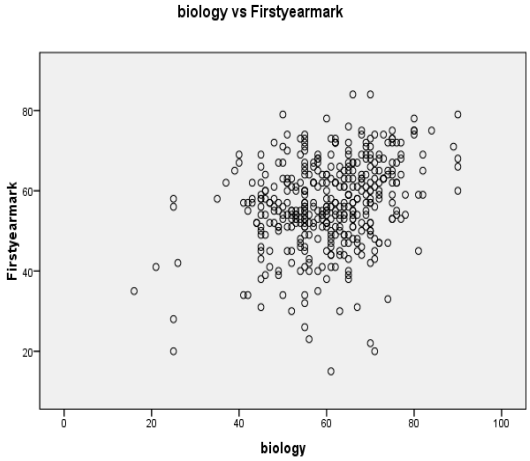

physicalscience vs ARPL1010
ARPL1010
physicalscience


mathematics vs Firstyearmark
Firstyearmark
mathematics
0
20
40
60
80
100


physicalscience vs Firstyearmark
Firstyearmark
physicalscience
0
20
40
60
80
100


englishfirstlang vs Firstyearmark
Firstyearmark
englishfirstlang
0
20
40
60
80
100
40
50
60
70
80
90


englishsecondlang vs Firstyearmark
Firstyearmark
Eng_Second Language_HG
20
40
60
80
30
40
50
60
70
80
90


biology vs Firstyearmark
Firstyearmark
biology
20
40
60
80
0
20
40
60
80
100