A LONGITUDINAL ANALYSIS USING AUXILIARY INFORMATION TO MODEL RETENTION IN UNDERGRADUATE STUDENTS

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BY

LAKENDRA MIRANDA PEOPLES-MCAFEE, B.S.

DENTON, TEXAS

DECEMBER 2008

TEXAS WOMAN'S UNIVERSITY
DENTON, TEXAS

August 6, 2008

To the Dean of the Graduate School:

I am submitting herewith a thesis written by LaKendra Miranda Peoples-McAfee entitled "A Longitudinal Analysis Using Auxiliary Information to Model Retention in Undergraduate Students." I have examined this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements of the degree of Master of Science with a major in Mathematics.

Mark S. Hamner

Mark Hamner, Ph.D., Major Professor

We have read this thesis and recommend its acceptance:

David Marshall, Ph.D

Marie-Anne Demuynck

Marie-Anne Demuynck, Ph.D

Don E. Edwards

Don Edwards, Ph.D
Department Chair

Accepted:
Jennifer Martin

Dean of the Graduate School

ACKNOWLEDGMENTS

I would like to thank Dr. Hamner for agreeing to be my advisor and taking time out of his busy schedule to assist with my thesis work. I would also like to thank Dr. Demuynck and Dr. Marshall for being members of my panel. Most importantly I would like to send special thanks to my husband and son who pardoned me on completing some of my wifely and motherly duties in order to complete my thesis work.

ABSTRACT

LAKENDRA PEOPLES-MCAFEE

A LONGITUDINAL ANALYSIS USING AUXILIARY INFORMATION TO MODEL RETENTION IN UNDERGRADUATE STUDENTS

DECEMBER 2008

Attrition is an issue for colleges and universities, and attempts to retain students are becoming more and more difficult. This study focuses on predicting student attrition of first time incoming (FTIC) students over a long time period. The population of this study consists of all FTIC students from Fall 2001. The students were followed 3.5 academic years to observe whether they experienced attrition. Exploratory data analysis was conducted to examine existing independent variables and some variables that were created to determine their contribution to the model. A discrete time hazard method was used to measure the timing of event occurrence. Cumulative GPA after one semester, number of major changes, major type, and minority status were selected to be included the model. Cross-validation was performed on Fall 2002 FTIC students to assess model fit. Overall, the model did a great job of predicting attrition of students over the long term.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The best-known American college and university rankings are compiled annually by the magazine *U.S. News & World Report* (2007) and long term retention serves as one of the first four factors that account for the great majority of the ranking. *U.S. News & World Report* conducts this ranking report every year and many have found it useful when considering what draws certain individuals to a particular college or university. One of the most important aspects of the determining the rank is retention, which involves the tracking of full-time students in a degree program over time to determine whether the student has completed the program (Center for the Study of College Student Retention, 1996). University rankings are often considered by potential students and their parents when deciding what institution of higher learning to attend (U.S. News & World Report, 2007*)*. From the Online Education Database a college's retention rate reflects the student body's overall interest in what is being offered by the college. Since retention rate influences university rank and the overall selection process, several universities have began to take steps to remedy the issue of student dropout, also referred to as attrition (U.S. News & World Report). To understand attrition, a university needs to be aware of the reasons a student remains at a university after initial enrollment such as academic programs, the quality of professors, and the availability of financial assistance. In this regard, retention rates may be a significant indicator of a universities effort to create an environment that will be conducive towards students completing their degree. If a large amount of students dropout from a particular school, it is important that the school makes an effort to prevent such a deficit to help minimize costs associated with losing students. Dr. Watson Scott Swail, the president of the international organization, Education Policy Institute (EPI), identifies 3 forms of cost associated with student attrition: institutional, individual, and societal. Of major importance to colleges is the institutional loss. When a student drops out there is a loss of future revenue from tuition and fee charges, bookstore purchases, and potential room and board charges, etc. With such losses it is obvious why it is important for universities to understand and develop an agenda for student retention.

The school of interest for this paper is Texas Woman's University (TWU). Formerly known as the Girls Industrial College, the college was founded in 1901 and later became known as TWU in 1957. TWU was originally an all women's college, but men have been admitted since 1972. The main campus is located in Denton, Texas with two other campuses in Houston, Texas and Dallas, Texas. The university is accredited to offer bachelor's, master's, and doctoral degrees. TWU has approximately 92% female student body population making the university unique relative to most universities that have a fairly equal female and male population. For example, in 2004, the number of men and women who enrolled into degree granting institutions in the U.S. was 7387 and 9885, respectively (Freeman, 2004). On average, women accounted for a little over 57% of all students who enrolled into a college or university in the U.S. in that year. Women are on the rise for college enrollment, but TWU has an above average enrollment of women which makes our environment unique and essentially controls for the gender effect, especially when considering gender as a significant factor to predict retention. For example, some researchers identify male domination of classrooms and laboratories as a factor in the under-representation of women in some sciences. According to the National Center for Education Statistics online website, in the past women were underrepresented in degree obtainment and could be considered a minority in itself. Between 1970 and 2001, women went from being the minority to the majority of the U.S. undergraduate population, increasing their representation from 42 percent to 56 percent of undergraduates (Freeman, 2004). In particular, TWU does not have a problem of under-representation of women since the student population is approximately 92% women. This factor alone gives us the potential to derive a unique retention agenda or prediction model compared to a traditional university with a somewhat equal male/female population. TWU also has a very diverse student body population with approximately 15% African American, 12% Hispanic, and 58% Caucasian. Of the students who enrolled in degree granting institutions in the U.S. in 2004, 66.1% were White, non-Hispanic, 12.5% Black, non-Hispanic, 10.5% Hispanic, 6.4% Asian or Pacific Islander, 1.0% American Indian/Alaskan Native, 3.4% nonresident alien (Freeman, 2004). In America, White, non-Hispanic students are dominating in college enrollment as well as at TWU, but TWU ethnicity statistics for minority students are above average in comparison to the U. S. statistics. Furthermore, TWU was ranked third in the state and 21st in the nation among universities with the most diverse student populations by U.S.

News and World Report magazine (2008). TWU has a significant account of students in every area of ethnicity.

Before a university can develop and implement a program to prevent attrition it must first determine the indicators that may expose whether a student is at risk of attrition. A model that works for one particular university may or may not be valid for another university. Consider a 1996 study of 300 campuses which found that racially-mixed student populations have positive effects on retention, overall college satisfaction, college grade point average, and intellectual and social self-confidence (Chang, 1996). Given the positive outcomes of racially diverse campuses and retention, ethnicity could possibly prove to be a significant indicator for certain colleges with diverse student populations versus those colleges with less diverse student populations. According to Freeman (2004), approximately sixty-seven percent of all degrees conferred during the 2002–03 academic year went to White, non-Hispanic students; twenty-two percent to minority students (Black, non-Hispanics, Hispanics, Asians/Pacific Islanders, and American Indians/Alaska Natives), and the remainder to nonresident aliens or individuals whose race/ethnicity was unknown. In that same academic year, total enrollment for degree granting institutions for White, non-Hispanic students was a little over sixty-seven percent, nearly thirty percent for minority students, and about three percent for nonresident aliens (Freeman, 2004). Interestingly, while ethnic minority students account for nearly 30% of the enrollment, they only account for 22% of the graduation rate. From these statistics we see that graduation rate of 22% for minority's lags behind the enrollment rate of 30% whereas the graduation and enrollment for white students remains consistent at about 67%. This could suggest that minority students experience a higher rate of dropout than the white students and that ethnicity may be a factor in retention. Therefore, the proportion of minority students that enroll and earn a degree has been increasing, but retention is still a concern. Furthermore, the various reasons that students dropout of college and do not earn degrees could reveal other potential factors for predicting retention. According to Dr. Linda K. Lau (2003), students leave school before graduating for reasons that are many times beyond institutional control, the inability to manage normal school work or to assimilate within the student population, the lack of motivation, the lack of appropriate role models and mentors, the overwhelming stress due to the transition from high school to college, and the institution has failed to create an environment that is conducive to their needs. In view of this concern, looking at the dropout rate among college students by ethnicity can possibly help an institution focus on at risk groups. Moreover, an institution could also consider other factors such as whether or not there are freshman specific programs to aid students in the transition or if a student was provided the option of having a mentor. Nonetheless, significant understanding for the exact time a student drops out may be difficult since some colleges do not have the data which indicate whether the students leave because they are transferring to another school or dropping out entirely. All in all it is important to observe which students are more at risk of attrition when trying to develop a retention program.

The focus of this study is to perform a longitudinal analysis that will assist with the development of a model to predict student retention using auxiliary information of undergraduate students. A longitudinal analysis is important for long term retention since the characteristics that indicate the potential drop out of a student when they are a freshman may change over time as a student matures and makes progress towards their degree. From this model we will be able to identify specific variables that help predict retention as well as provide specific interpretations as to the level of importance each variable provides. To validate our model we will use cross-validation. Cross-validation uses an alternative data set, which was not used to build the model, to test its predictive accuracy. In chapter 2, a literature review of previous studies of longitudinal exploratory data analysis to model retention is discussed that serves as a comparison of the prediction model approach and analysis of other studies with the model in this study. Chapter 3 will consist of a brief overview of the data used for this study and the results from carrying out exploratory data analysis that leads to the selected statistical model used to model the retention rate. Furthermore, in chapter 3 we begin establishing some mathematical notation that will aid in developing our model. In chapter 4, we introduce and explain the concept and importance of the person period data set and the life table. Chapter 5 will include the underlying mathematical analysis that we use to help explain how we derived our model. The results and interpretation of our model are discussed in chapter 6. In chapter 7 we perform cross validation to assess the fit of our model. Finally, chapter 8 will discuss future research in student retention and ideas of what could be done to further increase the precision of the model by creating or acquiring more variables that could potentially serve as good predictors of student retention.

CHAPTER II

LITERATURE REVIEW

The literature shows that there are various covariates that determine student retention. Several of these studies identify factors for a certain population. For example, a study on student athlete retention conducted by Radcliffe, Huesman, and Kellogg (2006) to identify students at risk, use the number of particular letter grades, housing, credits, and number of remedial courses, etc to develop a model to predict retention. The study pointed out that even though the retention of student athletes of color was very promising, something needs to be done about those who do leave because they do so early in their careers. Another study by Ishitani and Snider (2006) uses a survival analysis approach to examine the longitudinal impact of high school programs on retention. This study deals with high school students as its population of interest, thus it serves as a pre-college study that addresses whether a particular group of high school students are retained, once they enroll in college, based on high school programs that were available to them prior to enrollment. The results indicated that participating in the ACT/SAT preparation program increased a student's commitment and motivation to earn a college degree. Also, receiving assistance in filing financial aid applications produced a negative effect on retention. Alternatively, Chizmar (2000) studied the target population consisting of first time freshman who initially enrolled as economics majors. His study used a discrete time hazard analysis to determine if there were gender differences in continued participation in economics courses. The results showed that the profiles of females who majored in economics were indistinguishable from males who also majored in economics. In our study, the entire TWU population is of interest, in contrast with the three examples provided. In particular, we will conduct a longitudinal analysis on all TWU first-time in college (FTIC) cohort group from a particular semester and year. No matter what particular major a student selects or whether or not they are an athlete, inclusion into the population consists of all disciplines and activities because we are developing a long-term model to predict retention for all TWU students who begin as FTIC but throughout time change their status from freshmen to sophomores, juniors, seniors, and ultimately graduates.

There are programs that currently exist at certain universities that serve to improve retention. For instance, Mangold, Bean, Adams, Schwab, and Lynch (2002) conducted a study that evaluated freshman block registration and a mentoring program as a method to improve retention. These particular programs designated that freshman students enroll in the same courses so they could attend classes as a cohort. Furthermore, the students met with their mentors on a weekly basis to help them stay on track. The results showed that the program had a positive impact on retention for freshmen. A second example of an approach to improve retention was a study by Dale (1995) to evaluate the influence of a program called HORIZONS Student Support Program. The HORIZONS program is specifically designed to increase retention of first generation, low income, or physically disabled students. This study compared all 47 freshmen who entered the program with a matched group of students who did not. The results showed that participation in the program had a significant influence on student retention and rate of graduation. The significant increase in retention and graduation rates resulted from the services provided to the students through the program. Both of the studies previously mentioned involved intervention at the freshmen level. Another approach to increase retention among African Americans and minority students is the Young Scholars Program (YSP), is intended to increase underrepresented minority youth who desire to attend college, and assists them in meeting entrance requirements and successfully earning a college degree (Newman, 1999). This program also had very positive results, in that the students that participated displayed strong motivation, aptitude and a purpose to succeed. For example, after two academic years, the YSP student's retention rate was 72%, the retention rate for the entire freshman class was 70%, and the retention rate for a comparison group, matched with the YSP students on family income, adjusted high school class rank, self-reported high school grade point average, race, and gender, was 62%. Furthermore, a greater percentage of these students, who are considered least likely to graduate from high school and go on to college, did earn a degree. The latter approach to improve retention involved a continuous process of working with the students from grade school up to college graduation. Each program proves to be an effective factor when trying to increase retention rates. Currently, there have been no formal statistical studies or sufficient data records available to indicate the statistical significance of programs such as Multi-Ethnic Biomedical Research Support (MBRS), Bridges, and the like that are being employed by TWU to help improve retention. Although there are several social groups and discipline specific clubs that students are members of that may promote degree obtainment, TWU does not have anything specifically documented, online at least that focuses on improving retention. Nonetheless, this study may provide a statistical understanding of long-term retention that TWU can use to provide interventions or programs to improve retention.

Several different methods exist by which researchers predict retention. For example, a study was conducted that utilizes survival analysis to model student retention among a sample of 8,867 undergraduate students at Oregon State University between 1991 and 1996 (Murtaugh, 1999). The results from the study concluded that attrition was found to increase with age, and decrease with increasing high school GPA and first-quarter GPA. Also, resident and international students had a lower attrition rate than did non-resident students, and students were at a decreased risk of dropping out if they took the Freshman Orientation Course. A proportional hazard regression model was developed to predict student retention based on several academic and demographic characteristics. Proportional hazard regression modeling is a technique in survival analysis to obtain models coefficients, using the hazard function (Der & Everitt, 2006). An article by Ishanti (2002) is another instance of employing a method to develop a model to predict retention. The study investigated the longitudinal effects of being a first generation student on attrition. The method used in this study was an event history model, which is another term for survival analysis. Results indicated that first-generation students were more likely to depart than their counterparts over time. After controlling for factors such as race, gender, high school GPA, and family income, the risk of attrition among first generation students was 71 percent higher than that of students with both college educated parents in the first year (Ishanti, 2002). A final example of a method used to develop a model to predict retention was conducted by Radcliffe et al. (2006) that consisted of creating a practical application to help a large doctoral research extensive public university promote student success by identifying at-risk students. A logit probability model and a longitudinal model using survival analysis were used to identify factors that impact a student's ability to persist and graduate (Radcliffe et al., 2006). In our study, we will also use survival analysis to develop a model, but the variables in our model will certainly differ from the previous studies since we are limited by the available data at TWU.

In general, statistical models may vary because of the general population of interest and the available data an institution has. In other words, different schools imply different issues and different availability of information to be able to build a model that effectively predicts student retention. Nonetheless, modeling long term retention using survival analysis methods makes sense since. Time is a factor that certainly allows the possibility that certain variables may change.

CHAPTER III

MATHEMATICAL NOTATION AND EXPLORATORY DATA ANALYSIS

Exploratory data analysis is a set of procedures aimed at understanding the data and the relationships among the variables (Refaat, 2006). The longitudinal characteristic we want to explore with our data is the time until attrition, denoted at T. The data we have available to model retention over time are snap shots of Spring and Fall TWU student census data. By the nature of our data, T is a discrete variable measured in terms of semesters. Discrete time is recorded in thicker intervals whereas continuous time is recorded in thin precise units (Singer & Willett, 2003). Before we begin our data exploration in this chapter we will specify the appropriate mathematical notation and define terminology that will facilitate explaining some of our exploratory results. Then we will explore characteristics or factors that may help explain the attrition of FTIC students over Time (i.e. $T \geq 0$).

The cohort of FTIC students we will explore and use to build our model comes from Fall 2001 census data. In general, the semester a students enters TWU as an FTIC represents $T = 0$. Using notation, we let $P = \{1, 2, \ldots, N\}$ represent the indices of units for FTIC students of size N from a particular semester of interest. For exploratory analysis, the data that we will use to build the predictive model consists of $N = 513$ students at Texas Woman's University (TWU) who entered as FTIC students in the fall of 2001. Interestingly, some FTIC students enter college for the first time as sophomores because students have the ability to complete approximately 30 semester credit hours while in high school. Given the limited amount of longitudinal data we had available, we restricted our observational period on our Fall 2001 FTIC students to 3.5 academic years worth of data: $T = 0, 1, \ldots, 6$. Now, for each $k \in P$, there may exist a time $T = t$, within our observation period, T_k such that student k drops out of TWU. To facilitate the exploration and modeling of the dichotomous state of dropping out or not dropping out before the end of the observational period of time, we formally define the random of interest as

$$y_{kt} = \begin{cases} 1 & \text{if the student drops out at time T = t} \\ 0 & \text{if the student does not drop at time T = t} \end{cases} \tag{3.1}$$

To explore and understand the retention phenomenon requires considerable data mining in order to observe patterns for the values of y_{kt}. This involves repeated measurement of enrollment over a long period of time. Furthermore, 'retention is not an instantaneous event, but rather a prolonged process' (Tinto, 1987). None the less, the prolonged process has to end at some point in time so that researchers can finish exploring data to obtain a model of the event $y_{kt} = 1$. This means that any longitudinal study involves a finite time limit to which the event $y_{kt} = 1$ will occur. To illustrate the nuisances of student attrition over our finite longitudinal period of time, six semesters, we will introduce a vector of binary indicator variables or dummy coding. For any $k \in P$, we can represent the time until attrition, $T = t$, in an alternative way

$$\boldsymbol{T}_k = [\mathrm{S}_0 \;\; \mathrm{S}_1 \;\; \mathrm{S}_2 \;\; \mathrm{S}_3 \;\; \mathrm{S}_4 \;\; \mathrm{S}_5]$$

where $\boldsymbol{T_k}$ is a 1×6 vector of indicator variables such that

$$S_j = \begin{cases} 1 \text{ , indicates a student drops at semester t} \\ 0 \text{ , Otherwise} \end{cases}$$

for $j = 0, 1, \ldots, 5$. This notation will also facilitate specification of the mathematical model used to predict attrition in chapter-4. It is important to note that an individual can only experience attrition once so that $\boldsymbol{T_k}$ will contain at most one element with a value of 1. For example, for an individual $k \in P$, who did not re-enroll the third semester, or equivalent whose attrition occurred at $T = 3$, will have $\boldsymbol{T_k} = [\ 0\ 0\ 0\ 1\ 0\ 0]$. "The only requirement for survival analysis is that, in any particular research setting, the states be both mutually exclusive (non-overlapping) and exhaustive (of all possible states) (Singer & Willett, 2003). In other words, a student cannot exhibit persistence once he or she has experienced attrition; they can be in one state or the other but not both. For instance, suppose the individual with attrition time $T = 3$ or equivalently with $\boldsymbol{T_k} = [\ 0\ 0\ 0\ 1\ 0\ 0]$ re-enters at semester number 5 and then drops out again at semester number 5, we cannot have $\boldsymbol{T_k} = [\ 0\ 0\ 0\ 1\ 0\ 1]$. For this longitudinal study, the first time a student experiences the event of attrition, they are no longer a part of those individuals who are eligible to experience the event of attrition at a later time. Thus, those individuals from P who are eligible to experience $y_{kt} = 1$ at $T = t$ are part of what is called the *risk set*.

Definition -1: The risk set at the beginning of each time period $T = t$ is: $R_t = \{k\!: k \in P$ and $y_{kt} = 0$ for each time period $T < t\}$.

The number of individuals for each risk set is N_t such that $N_t \leq N$. For example, we have $N = 513$ individuals at $T = 0$ and if 40 of those students do not re-enroll at the beginning of $T = 1$, then those $N_1 = 473$ individuals remaining are apart of the risk set for the duration of the time period $T = 1$.

In our research we are interested in predicting the total attrition we can expect from the elements in the risk set R_t at various times $T = t$. Note that the dependent variable representing attrition, equation (3.1), is a dichotomous variable where 1 represents a student drops out during the time period and a value of 0 represents a student did not drop during the time period. Using this notation, we represent total attrition at times $T = t$ as

$$n_t = \sum_{k=1}^{N_t} y_{kt} \tag{3.2}$$

where n_t can range from 0 to N_t.

In our study there is a particular circumstance, due to the finite observational period, where the response of interest, event time T, can not be observed. Specifically, since the end of the observation period is $T = 6$ or fall 2004, everyone who has not experienced the event, $y_{kt} = 1$, up to that time will eventually drop out or equivalently experience $y_{kt} = 1$, possibly due to graduation. If a student did not drop out by the end of the observation period, $\boldsymbol{T}_k = [0\ 0\ 0\ 0\ 0\ 0]$, then that student is *censored.* Censoring occurs whenever a researcher does not know an individual's event time (Singer & Willett, 2003). There are several types of censoring, but the most common types of censoring can occur as right censoring or left censoring. Left censoring occurs when all that is known about an observation on a variable is that it is less than some value (Allison, 1995). Most of the time left censoring occurs when the researcher observes a sample in which the individual has already experienced the event and the researcher can only conclude that the event occurred sometime before the actual beginning of his/her observation period but does not know specifically. It appears as if our observations of attrition are left censored because we observe attrition at the beginning of a new observation period, say $T = t + 1$, but due to the construction of our time variable and because it is discrete, then we technically know the time of attrition, which is at time $T = t$. Right censoring occurs when the observation is terminated before the event of interest occurs (Allison, 1995). From this definition, we note that everyone in our study who has not experienced the event, $y_{kt} = 1$ by the end of the observation period is considered right-censored. There is also a concept known as interval censoring that combines both right and left censoring, see Paul Allison's *Survival Analysis Using the SAS System: A Practical Guide* (1995) for further discussion on this type of censoring.

To explore data effectively requires a clear understanding of the problem being researched. Now that we have defined risk set, total attrition, dependent variable, censoring and understand the nuisances of attrition through specification of $\boldsymbol{T}_k$, we can begin to effectively explore our data. The exploratory data analysis conducted for this study helps us understand if there are other variables contained in our data sets that can help explain attrition. Additional variables that are used to explain the outcome of a response variable are called *covariates or independent variables*. In other words, for each $k \in P$, the response variable y_{kt} may be dependent upon a set of size p, $p < N$, covariates represented as

$$\mathbf{x}_{kt} = \{x_{1kt}, x_{2kt}, \ldots, x_{pkt}\} \quad (3.2)$$

Notice that that the elements of $\mathbf{x}_{kt}$ indicate that the independent variables are time-varying, however, it may be the case that some of the predictors do not vary over time. In those cases where the independent variables do not vary over time, then it is implicit that the subscript t can be dropped. What follows is our examination on a number of independent variables that may be included in equation (3.2) to help explain the dependant variable

The available covariates for this model building will focus on: grade point average, ethnicity, age, gender, ACT/SAT score, ACT/SAT provided or both, the number of major changes, minority status, major type (science vs. non-science), household income, and distance. From our fall 2001 FTIC students, Table 1 indicates that cumulative GPA after one semester may possibly help explain student attrition.

Table 1

Student Dropout Relationship with GPA

	GPA<2	2<=GPA<3	GPA>=3	Total
Not Dropout	10	76	179	265
Dropout	42	52	93	187
Total	52	128	272	452

Notice that those FTIC students with a *GPA* less than or equal to 2.0, had a large cumulative attrition rate of 80.8% over the entire observation period. In contrast, students with a *GPA* greater than 2.0 but less than 3.0, had a much smaller attrition rate of 40.6%, and for students whose *GPA* was greater than or equal to 3.0, only 34.2% experienced attrition. Those students whose GPA, after one semester, is greater that 2.0 appear to be significantly more likely to remain in school than those students whose GPA is less that 2.0. This maybe due to the fact that a GPA of 2.0 is passing in the college realm and anything less is failing. A failing GPA after only the first semester of school may discourage students and may also place them on academic probation which makes it difficult financially for some students to return. Although GPA can vary over time the GPA variable we are discussing is the GPA after one semester. From this point forward we will refer to the GPA after one semester as just GPA. Another possibly significant variable is age. For example, table A1 in appendix A suggests 47.1% of students who enroll for the first time at the age 21 or younger dropped out, while 65.7% of those students who enroll for the first time that were older than 21 dropped out. From age we see that younger FTIC students appear to be less likely to drop out and hence, exhibit persistence. In table A2 in appendix A the ethnicity covariate does show a moderate difference between the different ethnic groups. White, non-Hispanic, Black, non-Hispanic, Hispanic, Asian or Pacific Islander, American Indian/Alaskan Native, nonresident alien have 53.2%, 42.3%, 50.0%, 36.7%, 66.7%, 33.3% of students from each group that dropped out, respectively. The range of ethnicity attrition rate appear large at 33.4% with the highest percentage at 66.7% and the lowest percentage at 33.3%. Therefore, ethnicity might serve as a good covariate to include in the model because the percentage of students that dropout varies across the levels of ethnicity. Gender is another covariate that we observed from the fall 2001 FTIC students to gain insight as to whether it will have any affect on the dependant variable. Gender is one of the more interesting covariates for us because TWU is a predominately female campus and we would like to see if there is a gender difference when it comes to dropping out. In table A3 in appendix A, 47.5% of the females in the cohort group dropped out whereas 70% of the males dropped. It may appear that the females are more likely to persist because of the gender dominant campus environment in their favor. Whatever the case may be there is a 22.5% difference in the drop out rate based on gender, so gender may work as a covariate in the model. Lastly, we looked at the ACT and SAT scores to determine whether or not these two measurements would serve as good candidates for the model. In table A4 in appendix A, 48.1% of those students who scored less than a 21 for their ACT composite score dropped out, 53.3% of those students who scored greater than or equal to a 21 but less than or equal to a 25 dropped out, and 33.3% of those students who scored greater than a 25 dropped out. There is not too much variability between the categories of the ACT composite score so ACT may not be included as a covariate in the model. In table A5 in appendix A, we have students with a SAT score less than 1000, students with a SAT score was greater than or equal to 1000 but less than 1200, and students whose SAT score was greater than 1200 with drop out rate 46.7%, 40.8%, and 36.4%, respectively. The SAT covariate also does not exhibit much variability across the specified levels so it too will probably not be included as a covariate in the model.

Each of the covariates that we have discussed to this point already existed in the snap shots of data we had from students census at TWU, but we can also create covariates through programming that that may help develop an even better model for attrition. For example, we created a covariate that delineated between whether or not a student selected a science or non-science major. In A6 in appendix A, 38.4% of those students who chose a science major dropped out versus the 50.0% who dropped out that chose a non-science major. There is only an 11.6% difference in attrition between the science and non-science majors so this covariate may not be included in equation (3.2). In our original data set ethnicity was already present but we decided to create a variable that notes whether a student is a minority or non-minority student. For this variable Blacks, Hispanics, Native Americans, and Asian Americans were part of the minority group and any student not in either of those categories is part of the non-minority group. In table A7, the minority group consisted of 44% students that dropped out whereas the non-minority group had 52.7% of its students that dropped out. Although the minority group had an 8.7 % less attrition rate, the minority versus non-minority students does not appear to be a good predictor of our response variable. Another predictor that deals with major that may be more substantial is the number of times that a student changes his or her major after their initial selection of a major. From semester to semester we checked whether or not a student changed his/her major and if so counted the number of times they changed their major to create the major change covariate. The results are in table A8 in appendix A. Interestingly, for those students who did not change their major 69.4% dropped out, whereas 31.0% of those who only made 1 major change dropped out and 8.8% of those students who made more than 1 major change dropped out. It appears that the more uncertain a student is about his or her major the more likely they are to drop out. The number of major changes a student makes shows promise as a covariate to be included in the model. Colleges and universities often side more with one aptitude test over the other so we created a variable to see whether a student provided the ACT, SAT, or both the ACT/SAT scores. In table A9, only 60% of those students who did not provide their SAT scores dropped out while 44.6% of those students who did provide their SAT scores dropped out. In table A10, the number of students that provided their ACT scores 43.7% dropped out while 49.9% of those students that did not provide their ACT scores dropped out. In table A11, for those students that provided both ACT/SAT scores 37.7% dropped out while 50.5% of those that did not provide scores for both exams dropped out. Each of these measures for the different covariates created does not show a large amount difference in attrition rate. We created socioeconomic variables such as household income and distance. We did this by merging an alternative data set that contained national census information on U.S. zip codes with our existing snap shots of TWU data. Household income was categorized as household income less than or equal to $40,000, household income that is greater than $40,000 but less than or equal to $60,000, and finally household income greater than $60,000. In table A12, the percentage for each of the household categories with respect to attrition was 40.4%, 44.9%, and 46.4% respectively. There is very little variability across the household income levels and attrition rate, therefore household income will more than likely not be included in the model. The distance variable we created is the total number a miles a student's home is from the campus location. In table A13, those students who lived less than 100 miles from the campus 40.6% dropped out, greater than or equal to 100 miles but less than or equal to 500 miles 49% dropped out, and for those students who lived more than 500 miles away from the campus only about 20% of those students dropped out. There is not a noticeable difference amongst the first two groups of students based on the distance they live from the university, but for those few students that live more than 500 miles away there does appear to be slight variability between that group and the other two groups. Distance probably will not be a good predictor to be included in the model since that first two groups exhibit very little variability in rate of attrition.

Per our exploratory analysis GPA, age, ethnicity, and major change appear to be covariates that will contribute to the model while the remaining independent variables do not show signs of significant variability of attrition across their levels and thus may not serve as good predictors. However, exploratory analysis just gives us an idea beforehand of what covariates may be included in the model. The statistical procedures that we will run to build our model will use statistical significance tests to determine whether or not a particular covariate will be included in the model.

CHAPTER IV

THE LIFE TABLE AND THE PERSON PERIOD DATA SET

"The fundamental tool for summarizing the sampling distribution of event occurrence or an individual's transition from one state to another state is the *life table* (Singer & Willet, 2003). A *life table* tracks the life of a sample from the beginning of the observation period through the end of the observation period (Singer & Willet). The life table for us begins at $T = 0$ and ends at $T = 6$. Before we develop a life table for our Fall 2001 FTIC students, it is important to format our data. Originally our data is in a *person-oriented* format in which each individual's data appears on a single record or row. The format that our data needs to be in is a *person-period* data set. A person period data set is a data set in which each person has multiple records, one for each measurement occasion (Singer & Willet). Below Table 2 and Table 3 is an example of a person-oriented and a person-period data set.

Table 2

Person-Oriented Data Set

ID	*T*	Censor
20	2	0
126	3	0
129	5	1

Table 3

Person-Period Data Set

ID	Period	Attrition
20	0	0
20	1	0
20	2	1
126	0	0
126	1	0
126	2	0
126	3	1
129	0	0
129	1	0
129	2	0
129	3	0
129	4	0
129	5	0

In table 2, the person-oriented data set, there is only one record for each ID number in the example, whereas the person-period data set in table 3 contains as many records for a particular ID as is indicated by the event time, *T*, in the person-oriented data set. Once our data is in the person-period format we can now develop our life table from that particular data set. Table 5 below represents the life table from the 513 students that are in our cohort group.

Table 4

Life Table

Interval [Upper, Lower)	Number Failed	Number Censored	Risk Set	Hazard	Survival
[0, 1)	61	0	513.0	0.1189	1.0000
[1, 2)	99	0	452.0	0.2190	0.8811
[2, 3)	29	0	353.0	0.0822	0.6881
[3, 4)	57	0	324.0	0.1759	0.6316
[4, 5)	13	0	267.0	0.0487	0.5205
[5, 6)	32	0	254.0	0.1260	0.4951
[6, 7)	22	111.0	0	0.4327	0.4327

The life table provides several important bits of information regarding our data such as: upper and lower interval for time, the number that failed, the number that were censored, the effective sample size, hazard, and survival. Each of these measures plays an important role in helping us determine the probability that a student will drop out or not. The upper and lower columns in the life table represent the time intervals for each semester. The brackets represent inclusions and the parenthesis represents exclusions. For example, the first time interval [0, 1) denotes the beginning of the observation period when the students enroll at time 0 up to but not including time interval 1. In first row of the life table the number failed is 61, but we do not find that number until a student census is completed for time interval [1, 2), $T = 1$, then any students who did not reenroll is categorized as dropping out or is deemed to have experienced the event. This logical process continues for the subsequent intervals. Notice that a value for the number censored does not enter the table until time interval [6, 7) or $T = 6$. We could have witnessed censoring as soon as interval [5, 6) or $T = 5$ because we had students in our cohort group that were sophomores. Since our data did not contain information on the graduation status of the students, for FTIC students who entered at sophomore status we decided that he or she will be considered to have graduated once they completed 3 academic years ($T = 5$), and hence will be deemed censored. Yet, we did not have any students in our 2001 FTIC cohort that met that criteria, so censoring actually only occurs during our last observation period. The last interval represents both graduates and those students who did not experience the event during the observation period. Notice that the risk set is the number of students who enter the interval minus the number that failed in the previous interval. For example, in interval [0, 1) 513 students enter that time period and 61 students failed, so the effective sample size for interval [1, 2) is 452 (513-61). We next introduce the concepts of *hazard* and *survival* that are presented in Table 4.3.

"The fundamental quantity used to assess the risk of event occurrence in each discrete time period is known as hazard" (Singer & Willett, 2003). More specifically for $k \in P$ and time period $T = t$, the discrete time hazard is defined by the following probability function

$$\theta(ykt) = \Pr[ykt = 1 \text{ for } T = t \mid ykt = 0 \text{ for } T < t, Rt] \quad (4.1)$$

The discrete time hazard is the conditional probability that individual k will experience the event in time period $T = t$, given that he or she did not experience it in any early time period. For time period $T = t$, an estimate of equation (4.1) for each $k \in P$ using our 2001 fall FTIC cohort group is simply

$$\hat{\theta}(y_{kt}) = \frac{n_t}{N_t} \tag{4.2}$$

The values for hazard in our life table are calculated using equation (4.2). In our life table, Table 5, the hazard value is 0.1189 in the first row because 61 students failed out of 513. In the subsequent row notice that 99 students failed or experienced the event at time interval [1, 2), $T = 1$, out of 452 students that where at risk at that time and thus $\hat{\theta}(y_{k1})$ = 0.2190 (99/452); $\hat{\theta}(y_{k2})$ = 0.0822 (29/353) for the time interval $T = 2$; etc. The Hazard tells us a great deal of what we need to know about the time and occurrence of our event. A valuable way to examine hazard is to graph it over time. Figure 1 is a graph of hazard for student attrition over time.

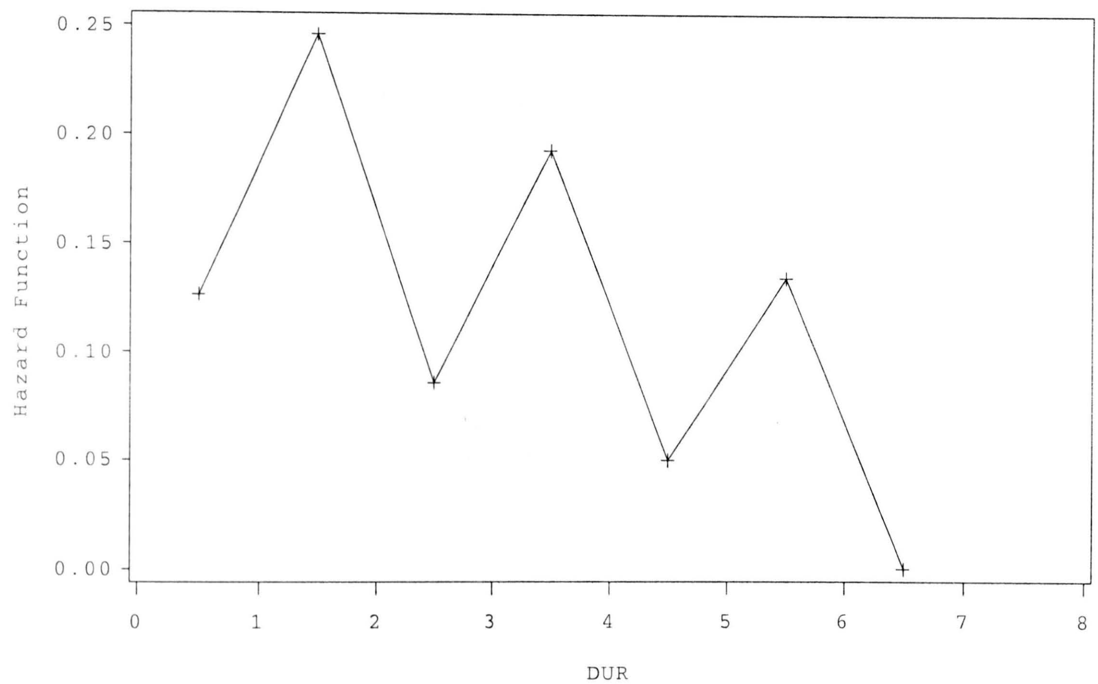

Figure 1. Graph of hazard function.

One of the main reasons that we look at the graphical representation of hazard is to identify risky period(s). In Figure 4.1, the riskiest period is between $T = 1$ and $T = 2$, which is the summer of the first academic year. Another seemingly risky period is between $T = 3$ and $T = 4$, which is also during the summer time.

An alternative method of assessing event occurrence is the *survivor* function. The survivor function cumulates the period-by-period risk of event occurrence together to assess the probability that a randomly selected individual will not experience the event (Singer & Willet, 2003). The following equation defines the population survivor function for each $k \in P$ and $T = t$,

$$S(y_{kt}) = \Pr[ykt = 0 | T \geq t] \tag{4.3}$$

Notice that equation is the probability of a student surviving past time period $T = t$. Just as with the population discrete time hazard function, for each $k \in P$ and $T = t$, we estimate the survivor values, when censoring does not occur, as

$$\hat{S}(y_{kt}) = \frac{N_t - n_t}{N} \tag{4.4}$$

Equation (4.4) is used to calculate survival in the life table all the way up to the point until censoring occurs. Every one survives in interval [0, 1) and it is not until the start of interval [1, 2) that we notice that some students in the first interval did not re-enroll in interval [1, 2) and thus they experienced the event at $T = 0$. For instance, for those 61 students that dropped out in time period [0, 1) that leaves 452 who have not experienced the event at the start of time period [1, 2) which calculates to $\hat{S}(yk1) = 0.8811$ (452/513); $\hat{S}(yk2) = 0.6881$ (353/513); etc. The preceding calculations work until we get to $T = 6$, which is when censoring first occurs. When censoring occurs, then estimate equation (4.3) with

$$\hat{S}(y_{kt}) = \left[1-\hat{\theta}(y_{kt})\right]\left[1-\hat{\theta}(y_{k(t-1)})\right]\cdots\left[1-\hat{\theta}(y_{k(t-m)})\right] \tag{4.5}$$

where $(t - m) = 0$. From equation (4.5) we notice the relationship survival is calculated by taking the cumulative product of the complement of hazard up to the time period $T = t$. We use equation (4.5) to calculate survival for the last observation period which contains censored observations. For instance, the first occurrence of censoring begins at interval [6, 7), and we have 222 students that were censored. We use equation (4.5) and calculate $\hat{S}(y_{k6})$ = (1- 0.1189) (1- 0.2190) (1- 0.0822) (1- 0.1759) (1- 0.0487) (1- 0.1260) = 0.4327. Just like with hazard, we too have a graph of the survivor function, Figure 2.

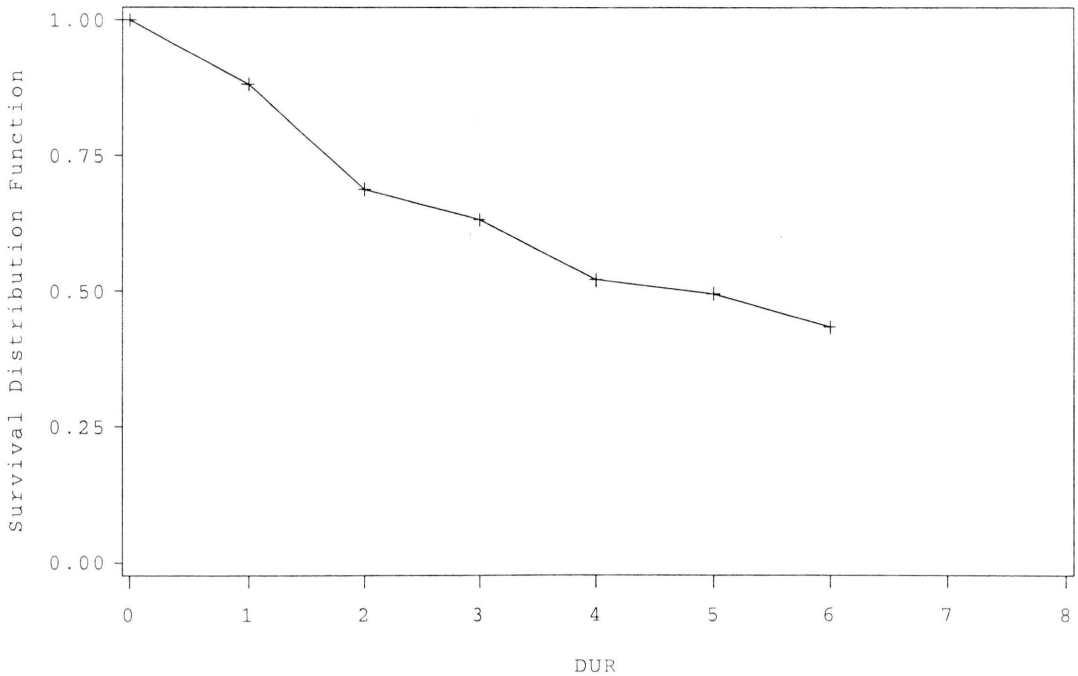

Figure 2. Graph of survivor function.

In the graph of the survivor function, the survivor function begins with a value of 1 because everyone is surviving at the beginning of time. As events occur the survivor functions drops toward a value of 0. In those time periods when hazard is low the survivor function drops slowly, and when hazard is high the survivor function drops rapidly (Singer & Willett, 2003). Unlike the hazard function, the survivor function will never increase and in time periods in which no events occur, the survivor function will remain steady at its previous level (Singer & Willett). We graph the survivor function to observe those periods where rapid drops occur, which symbolizes a significant amount of events being experienced. A rapid drop is apparent in figure 4.2 between the $T = 1$ and T

= 2. Another seemingly rapid drop is between $T = 3$ and $T = 4$. The rapid drops correspond with the risky periods that we observed from the hazard function, which again expresses the relationship between hazard and survival. Now that we have an understanding of hazard and the survivor function, next we will discuss the underlying mathematics that helps us determine whether or not an individual will experience the event of attrition.

CHAPTER V

MATHEMATICAL ANALYSIS

In our study we are concerned with the attrition of FTIC students who enter TWU at some fall semester of interest. To reiterate, the event of interest is whether or not a FTIC student drops out of school before a defined goal such as graduation. Typically the decision for a student to drop out is affected by various factors. The potential for a student to drop out could be greatly influenced by the student's GPA, the number of credit hours taken, their employment status, age, classification level, financial aid, and other important determinants. In this chapter, we will develop statistical models for predicting discrete time hazard, to help us describe a relationship between student attrition at time $T = t$, represented by a variable value of $y_{kt} = 1$ and the various factors or covariates in the set $\mathbf{x}_{kt}$ that we derived in chapter 3.

In chapter-3 for each time period $T = t$ and individual $k \in Rt$ we described the hazard, which is the corresponding probability of attrition $\theta(y_{kt})$, $0 \leq \theta(y_{kt}) \leq 1$. Accordingly, the probability a student does not drop out at time $T = t$ is $[1 - \theta(y_{kt})]$. Then, for each $k \in R_t$, the random variable y_{kt} can be modeled as a Bernoulli random variable with the following distribution:

$$f(y_{kt} \mid \theta(y_{kt}) = \theta(y_{kt})^{y_{kt}} [1 - \theta(y_{kt})]^{1 - y_{kt}} \tag{5.1}$$

with mean

$$\mu_{y_{kt}} = \mathrm{E}(y_{kt})$$

$$= \theta(y_{kt}) \tag{5.2}$$

and variance

$$\sigma^2_{y_{kt}} = E\left[\left(y_{kt} - \mu_{y_{kt}}\right)^2\right]$$

$$= \theta(y_{kt})[1-\theta(y_{kt})] \tag{5.3}$$

where E(·) denotes expectation with respect to the distribution of y_{kt}. For a more detailed discussion on expectations, E(·), see Hogg and Tanis (2006). Assuming we have independent Bernoulli trials, the joint probability distribution for all elements in the risk set R*t*, also known as a *Likelihood function*, is defined as

$$f(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{k=1}^{N_t} \theta(y_{kt})^{y_{kt}} [1-\theta(y_{kt})]^{1-y_{kt}} \tag{5.4}$$

where $\mathbf{Y} = [y_{1t}, y_{2t}, \ldots, y_{(Nt)t}]'$ is a $N_t \times 1$ vector of responses and $\boldsymbol{\theta} = [\theta(y_{1t}), \theta(y_{2t}), \ldots, \theta(y_{(N_t)t})]'$ is a $N_t \times 1$ vector of probabilities associated with the elements of $\mathbf{Y}$. Given that the likelihood function is defined by independent Bernoulli trials, we can now develop the idea behind predicting total student attrition at times $T = t$ for FTIC students at Texas Woman's University.

At the beginning of any time period $T = t$ or equivalently at the beginning of a particular semester, we will have N_t identifiable students in the risk set R_t. However, we will not know whether the N_t students will choose to drop out of TWU until the following semester, $T = t + 1$, when registration is complete and an official student census has been taken. As a result, at time $T = t$ we will not know the values of the random variable y_{kt} for any of the N_t students, which means we will not know total attrition, n_t. To predict total attrition n_t, we note that the expectation, $E(\cdot)$, is a linear operator, meaning that it can be distributed over addition or subtraction. Using the expectation of the random variable y_{kt}, equation (5.2), our estimate of the total attrition of FTIC at TWU during time $T = t$ is obtained by

$$\begin{aligned} \hat{n}_t &= E(n_t) \\ &= \sum_{k=1}^{N_t} E(y_{kt}) \\ &= \sum_{k=1}^{N_t} \theta(y_{kt}) \end{aligned} \tag{5.5}$$

Thus, our estimate is simply the sum of the individual probabilities associated with each y_{kt} for every individual in R_t.

Typically an institution will have data that will consists of useful information such as the independent variables $\mathbf{x}_{kt} = \{x1_{kt}, x2_{kt},\ldots, xp_{kt}\}$ discussed in chapter-3, which can help explain the outcome of the response variable y_{kt}. We make the association between x_{kt} and y_{kt} indirectly by focusing on the probability of a student dropping out, $\theta(y_{kt})$, and the probability the student does not drop out, $1-\theta(y_{kt})$. These probabilities indicate the chance of observing $y_{kt} = 1$ and $y_{kt} = 0$ respectively. Thus, for $k \in R_t$, we will describe how we can use independent variables or covariates x_{kt} to model the probability $\theta(y_{kt})$ in equation (5.5). To note the use of independent variables x_{kt} to help model the probability $\theta(y_{kt})$, we could simply denote $\theta(y_{kt})$ as $\theta(y_{kt} \mid x_{kt})$, which indicates the probability $y_{kt} = 1$ given the observed independent variables contained in x_{kt}. However, for convenience, we will write $\theta(y_{kt} \mid x_{kt})$ as simply $\theta(x_{kt})$.

A typical methodology used to associate a relationship between a dichotomous response variable y_{kt} such as attrition and independent factors $x_{kt} = \{x1k, x2k, \ldots, xpk\}$ is multiple logistic regression, (Peng et al., 2002). To make this indirect association we focus on the probability of a student dropping out, $\theta(x_{kt})$, and compare it to the probability the student does not drop out, $1-\theta(x_{kt})$. In multiple logistic regression, a model is formulated on the *odds* of attrition ($y_{kt} = 1$) which is defined to be

$$Odds = \frac{\theta(x_{kt})}{1-\theta(x_{kt})} \tag{5.6}$$

It is worth noting that the probability of a student dropping out, $\theta(x_{kt})$, can be rewritten in terms of odds:

$$\theta k = Odds\ [1-\theta(x_{kt})]$$

$$= \frac{Odds}{\left[\frac{1-\theta(x_{kt})+\theta(x_{kt})}{1-\theta(x_{kt})}\right]}$$

$$= \frac{Odds}{\left\{\frac{[1-\theta(x_{kt})]+\theta(x_{kt})}{1-\theta(x_{kt})}\right\}}$$

$$= \frac{Odds}{[1+Odds]} . \tag{5.7}$$

For each unit $k \in$ P, we can model the odds of dropping out ($y_{kt} = 1$) at some time $T = t$ in terms of a set of independent predictor variables using a logistic regression model of the form:

$$L_{kt}[\theta(x_{kt})] = \log_e(Odds)$$

$$= \log_e\left(\frac{\theta(x_{kt})}{1-\theta(x_{kt})}\right)$$

$$= (\alpha_0 + \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_5 S_5) +$$

$$(\beta_1 x_{1kt} + \beta_2 x_{2kt} + \ldots \beta_p x_{pkt})$$

$$= \mathbf{T}_k^* \alpha + \mathbf{x}_{kt} \boldsymbol{\beta} \tag{5.8}$$

where $\alpha = [\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_5]$ is a 6×1 vector of unknown intercept parameters, $\mathbf{T}_k^* = [1, S_1, S_2, S_3, S_4, S_5]$ is a 6×1 vector of indicator variables and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]$ is a $p \times 1$ vector of unknown coefficient parameters. It is important to note $\mathbf{T}_k^*$ does not contain S_0 because when indicators S_1-S_5 are all zero it is indicative of time period $T = 0$ with appropriate intercept α_0. The semester indicators $S_1, S_2, \ldots, S_5$ do not contain the subscript k, yet the subscript $k \in P$ is implicit since any individual who experiences attrition at $T = t$ will have the following vector of indicator variables $\boldsymbol{T}_k = [S_0\ S_1\ S_2\ S_3\ S_4\ S_5]$. It is worth noting again that the elements of $\mathbf{x}_{kt}$ indicate that the predictors or independent variables are time-varying. In those cases where the independent variables do not vary over time, then it is implicit that the subscript t can be dropped. Nonetheless, we will continue to use the subscript t, which allows our model notation the flexibility to contain either time-varying predictors or predictors that do not vary over time, or both. Notice the statistical model (5.8) defines a linear relationship between the odds and the predictor variables known as the *logit* function. A mathematically appealing aspect of equation (5.6) is that the range of the log odds is $(-\infty,\infty)$. This means that any set $\mathbf{x}_{kt}$ used in equation (5.6) will not generate a value outside the range. Also, using equation (5.8) we can rewrite the odds of attrition (i.e. $y_{kt} = 1$), equation (5.6), as

$$
\begin{aligned}
e^{L_{kt}\left[\theta(\mathbf{x}_{kt})\right]} &= e^{\left[\log_e(Odds)\right]} \\
&= e^{\left[\log_e\left(\frac{\theta(x_{kt})}{1-\theta(x_{kt})}\right)\right]} \\
&= \frac{\theta(x_{kt})}{1-\theta(x_{kt})} \\
&= \text{Odds} \qquad (5.9)
\end{aligned}
$$

The inverse transformation of the logit function, is easily understood from equations (5.7), (5.8), and (5.9) as

$$
\begin{aligned}
L_{kt}^{-1}(\mathbf{T}_k^{*}\boldsymbol{\alpha} + \mathbf{x}_{kt}\boldsymbol{\beta}) &= \frac{\mathrm{e}^{\mathbf{T}_k^{*}\boldsymbol{\alpha}+\mathbf{x}_{kt}\boldsymbol{\beta}}}{1+\mathrm{e}^{\mathbf{T}_k^{*}\boldsymbol{\alpha}+\mathbf{x}_{kt}\boldsymbol{\beta}}} \\
&= \frac{e^{(a_0S_0+a_1S_1+\ldots+a_5S_5)+(\beta_0x_{0kt}+\beta_1x_{1kt}+\ldots+\beta_px_{pkt})}}{1+e^{(a_0S_0+a_1S_1+\ldots+a_5S_5)+(\beta_0x_{0kt}+\beta_1x_{1kt}+\ldots+\beta_px_{pkt})}}
\end{aligned}
$$

$$= \frac{e^{L_{kt}\left[\theta(\mathbf{x}_{kt})\right]}}{1+e^{L_{kt}\left[\theta(\mathbf{x}_{kt})\right]}}$$

$$= \frac{Odds}{1+Odds}$$

$$= \theta(\mathbf{x}_{kt}). \tag{5.10}$$

Thus, substituting the inverses logit function into equation (5.4) we obtain the following joint distribution (i.e. likelihood function)

$$f(\mathbf{Y}|\theta) = \prod_{k=1}^{N_t} \left[L_{kt}^{-1}\left(\mathbf{T}_k^*\alpha + \mathbf{x}_{kt}\beta\right)\right]^{y_{kt}} \left[1 - L_{kt}^{-1}\left(\mathbf{T}_k^*\alpha + \mathbf{x}_{kt}\beta\right)\right]^{1-y_{kt}} . \tag{5.11}$$

Note that likelihood function above contains the unknown intercept parameters from α and unknown coefficient parameters in β. A common technique to estimate these unknown parameters is to take the log of the likelihood function and then find the values of α and β which maximize the log likelihood. Estimating the unknown coefficients by maximizing the *log-likelihood function* is referred to as maximum likelihood estimation (MLE). Using the fall 2001 FTIC data, we will use statistical software to determine the MLE's of the elements in α and β.

In light of independent variable values $\mathbf{x}_{kt}$ for each individual $k \in R_t$ at time period $T = t$, we can now give an alternative representation of equation (5.5), $\hat{n}_t$. Recall that for each time period $T = t$ and individual $k \in R_t$ the variable y_{kt} is Bernoulli. Given independent variable values $\mathbf{x}_{kt}$ from k $\in R_t$ and using equation (3.2) along with the inverse logistic transformation model (5.10), the expectation of y_{kt} is

$$E(y_{kt}) = \theta(\mathbf{x}_{kt})$$

$$= \frac{e^{L_{kt}\left[\theta(\mathbf{x}_{kt})\right]}}{1+e^{L_{kt}\left[\theta(\mathbf{x}_{kt})\right]}}.$$

Thus, for the set R_t at time $T = t$ equation (5.5) becomes

$$\hat{n}_t = \sum_{k=1}^{N_t} \theta(x_{kt})$$

$$= \sum_{k=1}^{N_t} L_{kt}^{-1}\left(\mathbf{T}_k^{*}\boldsymbol{\alpha} + \mathbf{x}_{kt}\boldsymbol{\beta}\right)$$

$$= \sum_{k=1}^{N_t}\left(\frac{e^{\mathbf{T}_k^{*}\boldsymbol{\alpha}+\mathbf{x}_{kt}\boldsymbol{\beta}}}{1+e^{\mathbf{T}_k^{*}\boldsymbol{\alpha}+\mathbf{x}_{kt}\boldsymbol{\beta}}}\right). \qquad (5.12)$$

In the following chapter we will use Fall 2001 FTIC data to obtain a specification of equation (5.8).

CHAPTER VI

RESULTS AND INTERPRETATION

The model building process will be implemented using SAS software. SAS is an integrated system of software solutions that enables its users to perform the following tasks: data entry, retrieval, management, and mining, report writing and graphics design, statistical and mathematical analysis, business forecasting and decision support, operations research and project management, and applications development (SAS, 2001). In this chapter we will use Fall 2001 FTIC data to obtain the statistically significant independent variables to include as elements of $\boldsymbol{x}_{kt}$ in equation (5.8). Then, we will specify the MLE values for the unknown parameters contained in α and β.

To obtain a specification of equation 5.8 we used Fall 2001 FTIC data. The independent variables for $\mathbf{x}_{kt}$ were selected using a SAS stepwise selection procedure. The stepwise procedure starts with all potential covariates and then systematically selects variables that are statistically significant until there are no more statistically significant variables. From Table 5, we can see that there were four covariates selected for equation (3.2), $\mathbf{x}_{kt}$ = [*gpa* (x_{1t} = CUM_GPA1), *minority vs. non-minority* status (x_{2t} =MINORITY_IND), *total number of major changes* (x_{3t} = MJRCHGS6), *major type science vs. non-science* (x_{4t} = MAJOR_TYPE)]. Interestingly, the GPA variable selected is not time variant. It turns out the GPA variable selected is the GPA a FTIC student obtains after their first semester. The total number of major changes is the total number of times a student changed his/her major from semester to semester. The major type covariate denotes whether or not a student selected a non-science or science major. Lastly minority vs. non-minority categorizes students based on ethnicity as either minority students or non-minority students. The variables selected are for the most part well occupied with only GPA having 61 missing values and the remaining covariates with no missing values. From Table 5 we can also see the corresponding MLE of the coefficient parameter values for the selected independent variables are $\boldsymbol{\beta}$ = [-0.9599, 0.3441, -1.0769, 0.5177]. In addition, the MLE estimates of the intercept parameters are α = [-2.0374, 3.4975, 2.5539, 3.6246, 2.1831, 3.2168]. Thus, for each $k \in R_t$, given $\boldsymbol{x}_{kt}$ and the MLE parameter estimates, α and $\boldsymbol{\beta}$, from table 5, we can give specification of to equation (5.8)

$$
\begin{aligned}
Lkt\,[\,\theta(x_{kt})\,] &= \mathbf{T}_k^{*}\,\alpha + \mathbf{x}_{kt}\,\boldsymbol{\beta} \\
&= -2.0374 + 3.4975S_1 + 2.5539S_2 + 3.6246S_3 + 2.1831S_4 + 3.2168\,S_5 + \\
&\quad -0.9599(x_{1t}) + 0.3441(x_{2t}) + -1.0769\,(x_{3t}) + 0.5177(x_{4t}). \qquad (6.1)
\end{aligned}
$$

Table 5

MLE for Elements of α and β

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0374	0.4732	18.5379	<.0001
CUM_GPA1	1	-0.9599	0.1067	80.9177	<.0001
MINORITY_IND	1	0.3441	0.1695	4.1209	0.0424
MAJOR_TYPE	1	0.5177	0.2380	4.7319	0.0296
MJRCHGS6	1	-1.0769	0.1325	66.0419	<.0001
SEMESTER 1	1	3.4975	0.4374	63.9367	<.0001
SEMESTER 2	1	2.5539	0.4733	29.1167	<.0001
SEMESTER 3	1	3.6246	0.4582	62.5774	<.0001
SEMESTER 4	1	2.1831	0.5203	17.6016	<.0001
SEMESTER 5	1	3.2168	0.4783	45.2367	<.0001

Since we are doing a longitudinal analysis it is important to note that equation (6.1) does not represent a single model, it represents 6 different models that correspond to specific semesters in our finite observation time period:

$L_{k0}[\,\theta(x_{k0})\,] = -2.0374 - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) + 0.5177(x_{4t})$

$L_{k1}[\,\theta(x_{k1})\,] = -2.0374 + 3.4975(S_1) - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) +$

$0.5177(x_{4t})$

$L_{k2}[\,\theta(x_{k2})\,] = -2.0374 + 2.5539(S_2) - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) +$

$0.5177(x_{4t})$

$L_{k3}[\theta(x_{k3})] = -2.0374 + 3.6426(S_3) - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) + 0.5177(x_{4t})$

$L_{k4}[\theta(x_{k4})] = -2.0374 + 2.1831(S_4) - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) + 0.5177(x_{4t})$

$L_{k5}[\theta(x_{k5})] = -2.0374 + 3.2168(S_5) - 0.9599(x_{1t}) + 0.3441(x_{2t}) - 1.0769(x_{3t}) + 0.5177(x_{4t})$

For example, $L_{k1}[\theta(x_{k1})]$ is the log odds of experiencing attrition after semester 1 but before semester 2 (i.e. during $T = 1$). Each model has the intercept value for α_0 and coefficients for *GPA*, *total number of major changes*, *minority vs. non-minority status*, and *major type*. The coefficient for GPA in our model indicates that for every 1 unit increase in GPA the overall log odds for student attrition decreases by 0.9599. The number of major changes coefficient symbolizes that for every 1 unit increase in total number of major changes the overall log odds for student attrition decreases by 1.0769. The minority vs. non-minority coefficient differs from the previous two coefficients in that either a student is in one group or the other. If the student is a minority student then the overall log odds for student attrition would increase by 0.3441 otherwise if a student is a non-minority student it does not have an affect on the log odds. The major type coefficient works exactly like the minority coefficient either a student chose a non-science major or he or she did not. If the student chose a non-science major then the overall log odds for student attrition would increase by 0.5177 otherwise if a student chose a science major then it does not have an effect on the log odds. To obtain more practical interpretations of the coefficients in equation (6.1), we will use the odds ratio, which are obtained by taking the exponential of the coefficient estimates. For example, the coefficient estimate for GPA is - 0.9599 and the exponential of this value is 0.3829 . This means that for every 1 point increase in level of GPA after one semester of course work at TWU the student's has a 0.3829 times the odds (i.e. the odds decrease/shrink by a factor of 0.3829) of dropping out. To put this specific interpretation into perspective we will take the reciprocal of 0.3829 to obtain 2.61, which means that for every increase 1 point increase in GPA after one semester of course work at TWU the student's has a 2.61 times the odds (i.e. the odds increase (expand) by a factor of 2.61) of not dropping out. Lastly, each time period has a distinct coefficient for that particular semester. For example, for semester $T = 1$ the coefficient estimate is 3.4975, which has an exponential value of 33.03. In other words, for a student that did not experience the event after enrollment ($T = 0$) the odds of attrition during $T = 1$ versus time period $T = 0$ increases by a factor of 33.03 during time period $T = 1$.

Now that we have equation (6.1), it is important to assess how well the model describes the response variable, equation (3.1) over time. In particular, we want to test the model and determine how effectively the model predicts the response variable for the Fall 2001 FTIC students, from which the model was built. A means of assessing the model fit is by conducting a commonly used approach, the Hosmer and Lemeshow Goodness of Fit Test (H-L). The H-L test divides subjects into deciles based on predicted probabilities of the response variable, then computes a chi-square from observed and expected frequencies of the response variable. The chi-square statistic tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. For a more detailed explanation see Hosmer and Lemeshow (2000).

Table 6

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	F	Pr> ChiSq
14.4953	8	0.0697

In Table 6, the output from the H-L test statistic produces a Chi-Square value of 14.4953 and a p-value of 0.0697, which is not statistically significant so we fail to reject the null hypothesis at the .05 significance level. Therefore we failed to reject that there is no difference between the observed and predicted values of the response variable, which suggest that our model fits the response variable of the Fall 2001 FTIC data set well.

In the following chapter we will test the predictive accuracy of the specified logistic regression model, equation (6.1), to see how well it predicts the response variable.

CHAPTER VII

CROSS VALIDATION

In this chapter we test the predictive capabilities of our model by using Cross-validation. Cross validation will test how well the model predicts the response variable on a *validation data set*, which is an alternative data set that was not used to determine the MLE estimates of the unknown parameters. In this case, the validation data set we will use comes from Fall 2002 FTIC students.

To set up the fall 2002 FTIC data set for cross-validation we had to make sure that the independent variables found in Fall 2001 FTIC are also available for the Fall 2002 FTIC. Once that was accomplished we ran our model against the 2002 FTIC cohort data or our validation data set. Thus for each $k \in R_t$ in the validation data set, we will use their realized values to the independent variables in $\boldsymbol{x}_{kt}$ = [*gpa* (x_{1t} = CUM_GPA1), *minority vs. non-minority* status (x_{2t} =MINORITY_IND), *total number of major changes* (x_{3t} = MJRCHGS6), *major type science vs. non-science* (x_{4t} = MAJOR_TYPE)] for equation (6.1). Using the outcomes obtained from equation (6.1) into equation (5.10) we get the predictive probability for response $y_{kt} = 1$. Table (7.1) represent the results we obtained from performing cross-validation. The values in the predicted column are the predicted total number of individuals who dropped out, rounded to the nearest whole number, for the various semesters in the validation data set, using equation (5.12). The values in the actual column are the realized values of attrition using equation (5.2).

Table 7

Cross-Validation

Time Period	Risk Set	Actual (n_t)	Predicted ($\hat{n}$)	Off (n_t) - ($\hat{n}$)	% Off
T = 0	855	143	40	103	12.05%
T = 1	712	121	215	94	13.20%
T = 2	591	74	117	43	7.28%
T = 3	517	132	181	49	9.48%
T = 4	385	112	76	36	9.35%
T = 5	273	117	103	14	5.13%
From N = 855		699	732	33	3.86%

From our results in table 7, the overall prediction of attrition is excellent. We predicted that 723 will drop out over the entire time period and the actual number of students from the FTIC group that dropped out was 699. Since there were a toal of 855 FTIC students, our overall prediction was off by approximately 4% from the actual number of students that experienced attrition. However, the prediction time interval $T = 1$ is of by 103. As time goes by our model gets better at predicting attrition which is noticable from the percentages declining as we go from one semester to the next. While the model may not serve as a reliable resource to predict for time periods 0 or 1, it is reliable for long term prediction after $T > 2$ and especially for overall prediction of attrition.

CHAPTER VIII

CONCLUSION

In this study we took the 2001 FTIC students from Texas Woman's University and built a predictive model to determine the total number of students that would experience attrition over a long period of time. Our data and time frame required that we use a discrete time hazard model since we were conducting a longitudinal study over data that was collected at discrete times. We began by cleaning our data to put it in a usable form so that we could develop our model. Afterwards, we used existing variables and even created some variables to develop our model using SAS. To determine the hazard value we used a logistic regression model. Once we ran our model we obtained the model parameters which included the cumulative GPA after 1 semester, the total number of times a student changed his/her major, thr type of major a student selected, science vs. non-science, and the classification of whether on not a student was considered a minority student or not. The other covariates that were listed in chapter 3 were not selected in the stepwise selection process using SAS. After we derived our model, we assessed the model fit by performing cross-validation.

The model built in the study served its purpose in accurately estimating the total number of students that would experience attrition over long period of time. This model is more geared towards the students at Texas Woman's University and the particular characteristics that the university has. That is to say that this model may or may not work as well at another institution of higher learning. For future research, we would liked to include other variables that were not included in our existing data sets such as, total number of hours completed or attempted per semester, scholarship and financial aid information, whether on not the student would be the first in his or her family to attend college, high school grade point average, job status (full-time/part-time), and whether or not the student had any children that they provided most of the care for, or whether or not a student gave birth to a child while enrolled in school, to name just a few. Those bits of information are important when a student decides whether on not he/she will continue school after they enroll for the first time. Nonetheless, with our limited variables we demonstrated that an effective model can be obtained to predict attrition over time for FTIC students at TWU.

REFERENCES

American Association of University Women Educational Foundation. (2000). *Tech-savvy: Education girls in the new computer age.* Washington, DC: AAUW Educational Foundation Commission on Technology, Gender, and Teacher Education.

Allison, P. D. (1995). *Survival analysis using the SAS system: A practical guide.* Cary, NC: SAS Institute.

Bunderson, E. D., & Christensen, M. E. (1995). An analysis of retention problems for female students in university computer science programs. *Journal of Research on Computing in Education, 28*(1), 1-19.

Carter, D. (2001). *Research foundation for improving the representation of women in the information technology workforce.* Virtual Workshop Report. http://www.cise.nsf.gov/itwomen/itwomen_final_report.pdf

Chang, M. J. (1996). *Racial diversity in higher education: Does a racially mixed student population affect educational uutcomes?* Unpublished doctoral dissertation, University of California, Los Angeles.

Chizmar, J. F. (2000). A discrete-time hazard analysis of the role of gender in persistence in the economics major. *Journal of Economic Education, 31*(2), 107-118.

Dale, P. (1994). *A successful college retention program.* Paper published through the HORIZONS Student Support Services Program at Purdue University (Indiana). (ERIC Document Reproduction Service No. ED 380 017).

Der, G., & Everitt, B. (2006). *Statistical analysis of medical data using SAS.* Boca Raton, FL: Chapman & Hall/CRC.

Freeman, C. E. (2004). *Trends in educational equity of eirls & women: 2004 NCES 2005–016.* (U.S. Department of Education, National Center for Education Statistics). Washington, DC: U.S. Government Printing Office.

Harrington, S. M. (1990). *Barriers to women in undergraduate computer science: The effects of the computer environment on the success and continuance of female students.* Unpublished dissertation, University of Oregon, Eugene.

Ishitani, T. T., & Snider, K. G. (2006). Longitudinal effects of college preparation programs on college retention. *IR Applications, 9.* http://www.airweb.org/page.asp?page=808

Lau, L. (2003). Institutional factors affecting student retention. *Education*, *124*(1), 126-136.

Mangold, W., L. Bean, D. Adams, W. Schwab & S. Lynch. (2002). Who goes, who stays: The effect of a freshman mentoring and block registration program on college student retention. *The Journal of College Student Retention*, *4*(1), 95-122.

Morgan, C.S. (1992). College students' perception of barriers to women. *Youth and Society, 24*(2), 228-236.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40,* 355-371.

Newman, P. R., & Newman, B. M. (1999). What does it take to have a positive impact on minority students' college retention? *Adolescence, 34*(135), 483-492.

Radcliffe, P., Huesman, R., & Kellogg, J. (2006). *Identifying students at risk: Utilizing survival analysis to study student athlete attrition.* Paper presented at the National Symposium on Student Retention, Albuquerque, NM.

Refaat, M. (2006). *Data preparation for data mining using SAS.* The Morgan Kaufmann series in data management systems. Amsterdam: Morgan Kaufmann.

SAS Institute Inc. (2001). *Step-by-step programming with base SAS software.* Cary, NC: Author.

Seidman, A. (1996). *Retention definitions.* Center for the Study of College Student Retention. Retrieved May 6, 2008, from http://www.cscsr.org/retention_issues_definitions.htm

Swail, W. S. (2004). *The art of student retention.* Education Policy Institute. http://www.studentretention.org/pdf/ART_OF_STUDENT_RETENTION.pdf

Seymour, E. (1995). The loss of women from science, mathematics, and engineering undergraduate majors: An explanatory account. *Science Education, 79*(4), 437-473.

U. S. News and World Report. (2007). *America's best colleges 2008.* http://colleges.usnews.rankingsandreviews.com/usnews/edu/college/rankings/about/cofaq_brief.php

APPENDIX A

FREQUENCY CHARTS

Table A1 Student Dropout Relationship with Age

Table of Student Dropout by Age			
	AGE<=21	AGE>21	
NOT DROPOUT	253 52.93%	12 34.29%	265
DROPOUT	225 47.07%	23 65.71%	248
Total	478	35	513

Table A2 Student Dropout Relationship with Ethnicity

Table of Student Dropout by Ethnicity							
	WHITE, NON-HISPANIC	BLACK, NON-HISPANIC	HISPANIC	ASIAN AMER/PAC. ISL.	AMER. INDIAN-ALASKAN	INTER-NATIONAL	
NOT DROPOUT	117 46.80%	90 57.69%	34 50.00%	19 63.33%	1 33.33%	4 66.67%	265
DROPOUT	133 53.20%	66 42.31%	34 50.00%	11 36.67%	2 66.67%	2 33.33%	248
Total	250	156	68	30	3	6	513

Table A3 Student Dropout Relationship with Gender

Table of Student Dropout by Gender			
	F	M	
NOT DROPOUT	259 52.54%	6 30.00%	265
DROPOUT	234 47.46%	14 70.00%	248
Total	493	20	513

Table A4 Student Dropout Relationship with ACT Composite

Table of Student Dropout by ACT Composite				
	ACT<21	21<=ACT<=25	ACT>25	
NOT DROPOUT	243 51.92%	15 46.88%	6 66.67%	264
DROPOUT	225 48.08%	17 53.13%	3 33.33%	245
Total	468	32	9	509
Frequency Missing = 4				

Table A5 Student Dropout Relationship with SAT Score

Table of Student Dropout by SAT Score				
	0<=SAT<1000	1000<=SAT<=1200	SAT>1200	
NOT DROPOUT	139 53.26%	61 59.22%	14 63.64%	214
DROPOUT	122 46.74%	42 40.78%	8 36.36%	172
Total	261	103	22	386
Frequency Missing = 127				

Table A6 Student Dropout Relationship with Major Type (Science vs. Non-Science)

Table of Student Dropout by Major Type			
	SCIENCE	NON-SCIENCE	
NOT DROPOUT	45 61.64%	220 50.00%	265
DROPOUT	28 38.36%	220 50.00%	248
Total	73	440	513

Table A7 Student Dropout Relationship with Minority vs. Non-Minority

Table of EVENT by MINORITY_IND			
	MINORITY	NON-MINORITY	
NOT DROPOUT	144 56.03	121 47.27	265
DROPOUT	113 43.97	135 52.73	248
Total	257	256	513

Table A8 Student Dropout Relationship with No. of Major Changes

Table of Student Dropout by No. of Major Changes				
	NO MAJOR CHANGES	1 MAJOR CHANGE	MORE THAN 1 MAJOR CHANGE	
NOT DROPOUT	83 30.63%	120 68.97%	62 91.18%	265
DROPOUT	188 69.37%	54 31.03%	6 8.82%	248
Total	271	174	68	513

Table A9 Student Dropout Relationship with SAT Provided

Table of Student Dropout by SAT Provided			
	DID NOT TAKE	DID TAKE	
NOT DROPOUT	50 40.00%	215 55.41%	265
DROPOUT	75 60.00%	173 44.59%	248
Total	125	388	513

Table A10 Student Dropout Relationship with ACT Provided

Table of Student Dropout by ACT Provided			
	DID NOT TAKE	DID TAKE	
NOT DROPOUT	194 50.13%	71 56.35%	265
DROPOUT	193 49.87%	55 43.65%	248
Total	387	126	513

Table A11 Student Dropout Relationship with SAT & ACT Provided

Table of Student Dropout by SAT & ACT Provided			
	BOTH	NOT BOTH	
NOT DROPOUT	53 62.35%	212 49.53%	265
DROPOUT	32 37.65%	216 50.47%	248
Total	85	428	513

Table A12 Student Dropout Relationship with Household Income

Table of Student Dropout by Household Income				
	INCOME<=40K	40k<INCOME<=60k	INCOME>60K	
NOT DROPOUT	81 59.56%	98 55.06%	52 53.61%	231
DROPOUT	55 40.44%	80 44.94%	45 46.39%	180
Total	136	178	97	411

Table A13 Student Dropout Relationship with Distance

Table of Student Dropout by Distance				
	DISTANCE<100	100<=DISTANCE< =500	DISTANCE>500	
NOT DROPOUT	208 59.43%	52 50.98%	4 80.00%	264
DROPOUT	142 40.57%	50 49.02%	1 20.00%	193
Total	350	102	5	457
Frequency Missing = 56				

APPENDIX B

SAS CODE

```
/*-----------------------------------------------------------------------*/
/*LAKENDRA PEOPLES-MCAFEE
/*-----------------------------------------------------------------------*/
/*THESIS:  LOGITUDINAL ANALYSIS USING AUXILARRY DATA TO MODEL RETENTION
IN UNDERGRAUDATE STUDENTS  */
/*MAJOR PROFESSOR:  DR. MARK HAMNER*/
/*-----------------------------------------------------------------------*/
/*CREATE A LIBRARY TO STORE PERMANENT SAS DATA SETS
/*-----------------------------------------------------------------------*/

LIBNAME ret 'e:\Retention';
OPTIONS NODATE NONUMBER FMTSEARCH=(ret.FORMAT_LIBRARY);
/*-----------------------------------------------------------------------*/
/*CREATE DATA SET WITH FALL AND SPRING DATA FROM FALL 01' - SPRING 04'
AND                          */
/*TAG FIRST TIME INCOMING FRESHAMN
*/
/*-----------------------------------------------------------------------*/
DATA FTIC_START (KEEP=GENDER ETHNIC DOB APPLY_TERM STU_LEVEL APPLY_ST
ADMIT_ST APPLY_DATE TERM
                        YEAR NEW STATUS_LV SEMESTER APPLY CLASS MAJOR1
FIRST_TERM TOTAL_HOURS TERM_12TH
                        NEW ID_num SSN_num DOB GENDER CUM_GPA GMAT
GRE_A GRE_Q GRE_V SAT_V SAT_M
                        SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP);
    SET ret.student_data_fall ret.student_data_spring;
      IF ETHNIC='XX' THEN ETHNIC=08;
      IF ETHNIC='ZZ' THEN ETHNIC=09;
RUN;
/*-----------------------------------------------------------------------*/
/*DATA SETS CREATED AS A BREAKDOWN OF SEMESTER AND YEAR
*/
/*-----------------------------------------------------------------------*/
DATA FTIC_FA (KEEP=GENDER ETHNIC DOB STU_LEVEL_0 TERM_0 YEAR_0
SEMESTER_0 CLASS_0 MAJOR1_0 TOTAL_HOURS_0 FTIC_CH T_0
                   ID_num SEMCNT GMAT GRE_A GRE_Q GRE_V SAT_V SAT_M
SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     SP_1    (KEEP=GENDER ETHNIC DOB STU_LEVEL_1 TERM_1 YEAR_1
SEMESTER_1 CLASS_1 MAJOR1_1 TOTAL_HOURS_1 CUM_GPA1 T_1
                         ID_num SEMCNT_1 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     FA_2    (KEEP=GENDER ETHNIC DOB STU_LEVEL_2 TERM_2 YEAR_2
SEMESTER_2 CLASS_2 MAJOR1_2 TOTAL_HOURS_2 CUM_GPA2 T_2
                         ID_num SEMCNT_2 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     SP_3    (KEEP=GENDER ETHNIC DOB STU_LEVEL_3 TERM_3 YEAR_3
SEMESTER_3 CLASS_3 MAJOR1_3 TOTAL_HOURS_3 CUM_GPA3 T_3
                         ID_num SEMCNT_3 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     FA_4    (KEEP=GENDER ETHNIC DOB STU_LEVEL_4 TERM_4 YEAR_4
SEMESTER_4 CLASS_4 MAJOR1_4 TOTAL_HOURS_4 CUM_GPA4 T_4
```

```
                              ID_num SEMCNT_4 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     SP_5    (KEEP=GENDER ETHNIC DOB STU_LEVEL_5 TERM_5 YEAR_5
SEMESTER_5 CLASS_5 MAJOR1_5 TOTAL_HOURS_5 CUM_GPA5 T_5
                              ID_num SEMCNT_5 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP)
     FA_6    (KEEP=GENDER ETHNIC DOB STU_LEVEL_6 TERM_6 YEAR_6
SEMESTER_6 CLASS_6 MAJOR1_6 TOTAL_HOURS_6 CUM_GPA6 T_6
                              ID_num SEMCNT_6 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP);
       SET FTIC_START;

/*-------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 00', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.          */
/*-------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='01' THEN
       DO;
               /* AND CLASS IN ('FR', 'SO')*/
               IF STATUS_LV='01' AND NEW=1 THEN FTIC_CH=1; /*used to
mark the cohort group*/
               SEMCNT=1;
                 T_0='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                 STU_LEVEL_0=STU_LEVEL;
           TERM_0=TERM;
           YEAR_0=YEAR;
           SEMESTER_0=SEMESTER;
           CLASS_0=CLASS;
           MAJOR1_0=MAJOR1;
           TOTAL_HOURS_0=TOTAL_HOURS;
           TERM_0=TERM_12TH;
          OUTPUT FTIC_FA;
       END;
/*-------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 01', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.          */
/*-------------------------------------------------------------------*/
      IF SEMESTER='SP' AND YEAR='02' THEN
       DO;
           SEMCNT_1=1;
                 T_1='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                 STU_LEVEL_1=STU_LEVEL;
           TERM_1=TERM;
           YEAR_1=YEAR;
           SEMESTER_1=SEMESTER;
           CLASS_1=CLASS;
```

```
            MAJOR1_1=MAJOR1;
            TOTAL_HOURS_1=TOTAL_HOURS;
            TERM_1=TERM_12TH;
            CUM_GPA1=CUM_GPA;
            OUTPUT SP_1;
        END;
/*-------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 01', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.        */
/*-------------------------------------------------------------------*/
      IF SEMESTER='FA' AND YEAR='02' THEN
        DO;
            SEMCNT_2=1;
                  T_2='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_2=STU_LEVEL;
            TERM_2=TERM;
            YEAR_2=YEAR;
            SEMESTER_2=SEMESTER;
            CLASS_2=CLASS;
            MAJOR1_2=MAJOR1;
            TOTAL_HOURS_2=TOTAL_HOURS/100000;
            TERM_2=TERM_12TH;
            CUM_GPA2=CUM_GPA;
            OUTPUT FA_2;
        END;
/*-------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 02', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.        */
/*-------------------------------------------------------------------*/
    IF SEMESTER='SP' AND YEAR='03' THEN
        DO;
            SEMCNT_3=1;
                  T_3='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_3=STU_LEVEL;
            TERM_3=TERM;
            YEAR_3=YEAR;
            SEMESTER_3=SEMESTER;
            CLASS_3=CLASS;
            MAJOR1_3=MAJOR1;
            TOTAL_HOURS_3=TOTAL_HOURS;
            TERM_3=TERM_12TH;
            CUM_GPA3=CUM_GPA;
            OUTPUT SP_3;
        END;

/*-------------------------------------------------------------------*/
```

```
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 02', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*-----------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='03' THEN
        DO;
            SEMCNT_4=1;
                  T_4='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
            STU_LEVEL_4=STU_LEVEL;
            TERM_4=TERM;
            YEAR_4=YEAR;
            SEMESTER_4=SEMESTER;
            CLASS_4=CLASS;
            MAJOR1_4=MAJOR1;
            TOTAL_HOURS_4=TOTAL_HOURS;
            TERM_4=TERM_12TH;
            CUM_GPA4=CUM_GPA;
            OUTPUT FA_4;
        END;
/*-----------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 03', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*-----------------------------------------------------------------------*/
    IF SEMESTER='SP' AND YEAR='04' THEN
        DO;
            SEMCNT_5=1;
                  T_5='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
            STU_LEVEL_5=STU_LEVEL;
            TERM_5=TERM;
            YEAR_5=YEAR;
            SEMESTER_5=SEMESTER;
            CLASS_5=CLASS;
            MAJOR1_5=MAJOR1;
            TOTAL_HOURS_5=TOTAL_HOURS;
            TERM_5=TERM_12TH;
            CUM_GPA5=CUM_GPA;
            OUTPUT SP_5;
        END;
/*-----------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 03', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*-----------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='04' THEN
        DO;
            SEMCNT_6=1;
```

```
                    T_6='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                STU_LEVEL_6=STU_LEVEL;
                TERM_6=TERM;
                YEAR_6=YEAR;
                SEMESTER_6=SEMESTER;
                CLASS_6=CLASS;
                MAJOR1_6=MAJOR1;
                TOTAL_HOURS_6=TOTAL_HOURS;
                TERM_6=TERM_12TH;
                CUM_GPA6=CUM_GPA;
                OUTPUT FA_6;
            END;
RUN;
/*-----------------------------------------------------------------------*/
/*SORTING THE PREVIOUSLY CREATED DATA SETS                                */
/*-----------------------------------------------------------------------*/
PROC SORT DATA=FTIC_FA;
    BY ID_num;
RUN;

PROC SORT DATA=SP_1;
    BY ID_num;
RUN;

PROC SORT DATA=FA_2;
    BY ID_num;
RUN;

PROC SORT DATA=SP_3;
    BY ID_num;
RUN;

PROC SORT DATA=FA_4;
    BY ID_num;
RUN;

PROC SORT DATA=SP_5;
    BY ID_num;
RUN;

PROC SORT DATA=FA_6;
    BY ID_num;
RUN;
/*-----------------------------------------------------------------------*/
/*MERGES EACH OF THE DATA SETS PREVIOUSLY SORTED ON THE COMMON VARIABLE
ID_num                           */
/*-----------------------------------------------------------------------*/
DATA FTICMERG;
    MERGE FTIC_FA SP_1 FA_2 SP_3 FA_4 SP_5 FA_6;
      *INFORMAT T $CHAR6.;
 BY ID_num;
```

```
 IF FTIC_CH=. THEN DELETE;
 /*IF CLASS_0='SR' THEN DELETE;
 IF CLASS_0='JR' THEN DELETE;
 IF CLASS_0='MM' THEN DELETE;
 IF CLASS_0='PB' THEN DELETE;*/
 /*CREATE RANDOM VARIABLE T*/
 ARRAY VARIABLE{7} T_0 T_1 T_2 T_3 T_4 T_5 T_6;
 DO I=1 TO 7;
   IF VARIABLE{I}=' '  THEN VARIABLE{I}='1';/*ONE MEANS THAT THEY DID
EXPERIENCE DROP OUT*/
 END;
 /* THIS IS THE RAW RANDOM VECTOR WHICH WILL CONTIAN VALUES SUCH AS
VECTOR=0010010, WHICH WILL NEED TO
    REFORMATTED TO SAY VECTOR=0010000<--NOTICE THAT ALL VALUES AFTER
THE INTITAL '1' WERE ZEROED OUT*/
     VECTOR =T_1||T_2||T_3||T_4||T_5||T_6; /*ONLY NEED 1-6 B/C AT TIME
0 ARE DETERMINED TO DROP OUT AT T=1*/
 /* DETERMINE ERRONEOUS RESPONSES CONSIDERING THE SURVIVAL ANANLYSIS
VECTOR*/
    DUR=(INDEX(VECTOR,'1'))-1;/*RETURNS THE POSITION THAT THE FIRST
OBSERVED VALUE OF '1' OCCUREED*/
      VECTOR_ERROR=0;
    IF INDEX(VECTOR,'10')>0 THEN VECTOR_ERROR=1;/* A '10' VALUE
INDICATES THAT AN INDIVIDUAL RETURNED AFTER DROPPING OUT*/
      IF DUR=-1 THEN DUR=6;    /*THE STUDENTS DID NOT EXPERIENCE EVENT
SO CENSORED AT END OF DATA COLLECTION*/
      IF DUR<6 THEN EVENT=1; ELSE EVENT=0; /*1 MEANS THEY DROPPED OUT*/
      IF (CLASS_0 IN ('SO','JR', 'SR', 'MM', 'PB') AND DUR>5) THEN
EVENT=0;  /*FOR PEOPLE WHO ENTER ABOVE FRESHMAN IF COMPLETE 3 YEARS WE
ASSUME GRADUATE*/

      /*RETAIN EVENT_SUM 0;
      IF EVENT^='.' THEN EVENT_SUM=EVENT+EVENT_SUM;*/
RUN;
/*ODS RTF FILE="E:\FREQ_SM.RTF";
proc gchart data=FTICMERG;

   block DUR /type=FREQ;
   block DUR /type=pct;
run;
ODS RTF CLOSE;*/ /*END RTF-OUTPUT*/
/*proc gchart data=FTICMERG;
   *format sales dollar8.;
   VBAR VECTOR_ERROR /type=FREQ;
   *block VECTOR_ERROR /type=PCT;
 run;*/
/*-----------------------------------------------------------------*/
/*MERGED DATA SET INCLUDING ONLY THOSE STUDENTS IN THE FTIC COHORT
GORUP                         */
/*-----------------------------------------------------------------*/
DATA FINAL (DROP=AGE1_1 AGE2_2 AGE3_3 AGE4_4 AGE5_5 AGE6_6 AGE7_7
AGE8_8 AGE9_9 AGE10_10 AGE11_11 AGE12_12);
```

```
    SET FTICMERG;
      **FORMAT APPLY_DATE MMDDYY8. GENDER $SEXFMT. DOB DATE7. SUCCESS
SUCFMT. MAJOR SCIFMT.;
/*-----------------------------------------------------------------------*/
/*DETERMINES STUDENT'S AGE BASED ON THE CORRESPONDING SCHOOL YEAR
*/
/*-----------------------------------------------------------------------*/
      AGE1_1=(('15AUG01'D-DOB)/365.25);
      AGE1=ROUND(AGE1_1,.01);
      AGE2_2=(('15OCT01'D-DOB)/365.25);
      AGE2=ROUND(AGE2_2,.01);
      AGE3_3=(('15JAN02'D-DOB)/365.25);
      AGE3=ROUND(AGE3_3,.01);
      AGE4_4=(('15MAR02'D-DOB)/365.25);
      AGE4=ROUND(AGE4_4,.01);
      AGE5_5=(('15AUG02'D-DOB)/365.25);
      AGE5=ROUND(AGE5_5,.01);
      AGE6_6=(('15OCT02'D-DOB)/365.25);
      AGE6=ROUND(AGE6_6,.01);
      AGE7_7=(('15JAN03'D-DOB)/365.25);
      AGE7=ROUND(AGE7_7,.01);
      AGE8_8=(('15MAR03'D-DOB)/365.25);
      AGE8=ROUND(AGE8_8,.01);
      AGE9_9=(('15AUG03'D-DOB)/365.25);
      AGE9=ROUND(AGE9_9,.01);
      AGE10_10=(('15OCT03'D-DOB)/365.25);
      AGE10=ROUND(AGE10_10,.01);
      AGE11_11=(('15JAN04'D-DOB)/365.25);
      AGE11=ROUND(AGE11_11,.01);
      AGE12_12=(('15MAR04'D-DOB)/365.25);
      AGE12=ROUND(AGE12_12,.01);

/*-----------------------------------------------------------------------*/
 /*CREATE AN INDICATOR VARIABLE FOR STUDENT AGE GROUPS     */
 /*-----------------------------------------------------------------------
*/
      IF AGE1<=21 THEN AGE_IND=0;
            IF AGE1>21 THEN AGE_IND=1;
/*-----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR GPA AFTER
ONE SCHOOL YEAR                    */
/*-----------------------------------------------------------------------*/
    IF CUM_GPA1 = . THEN GOODSTART=0;
    ELSE IF (CUM_GPA1<2) THEN GOODSTART=1;
      ELSE IF (2<=CUM_GPA1<3) THEN GOODSTART=2;
      ELSE IF (CUM_GPA1>=3) THEN GOODSTART=3;
/*-----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR GOODSTART*/
/*-----------------------------------------------------------------------*/
    IF GOODSTART=0 THEN DO GS1=0; GS2=0; GS3=0; END;
    IF GOODSTART=1 THEN DO GS1=1; GS2=0; GS3=0; END;
    IF GOODSTART=2 THEN DO GS1=0; GS2=1; GS3=0; END;
```

```
    IF GOODSTART=3 THEN DO GS1=0; GS2=0; GS3=1; END;

/*----------------------------------------------------------------------*/
/*CREATE VARIABLES FOR STUDENTS WHO ENROLLED WITH 0 HOURS*/
/*----------------------------------------------------------------------*/
      IF TOTAL_HOURS_0=0 THEN FRESHSTART=1;
      ELSE FRESHSTART=0;
/*----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR ETHNICITY WHITE, HISPANIC, BLACK, AND
NATIVE AMERICAN                          */
/*----------------------------------------------------------------------*/
    IF ETHNIC=01 THEN DO; E1=0; E2=0; E3=0; E4=0; E5=0; E6=0; E7=0;
END; /*WHITE, NON-HISPANIC*/
    IF ETHNIC=02 THEN DO; E1=1; E2=0; E3=0; E4=0; E5=0; E6=0; E7=0;
END; /*BLACKS*/
    IF ETHNIC=03 THEN DO; E1=0; E2=1; E3=0; E4=0; E5=0; E6=0; E7=0;
END; /*HISPANICS*/
    IF ETHNIC=04 THEN DO; E1=0; E2=0; E3=1; E4=0; E5=0; E6=0; E7=0;
END; /*ASIAN AMERICAN*/
    IF ETHNIC=05 THEN DO; E1=0; E2=0; E3=0; E4=1; E5=0; E6=0; E7=0;
END; /*NATIVE-AMERICAN*/
    IF ETHNIC=06 THEN DO; E1=0; E2=0; E3=0; E4=0; E5=1; E6=0; E7=0;
END; /*INTERNATIONAL*/
    IF ETHNIC=07 THEN DO; E1=0; E2=0; E3=0; E4=0; E5=0; E6=1; E7=0;
END; /*OTHER*/
    IF ETHNIC=08 THEN DO; E1=0; E2=0; E3=0; E4=0; E5=0; E6=0; E7=1;
END; /*MISSING*/
/*----------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR MINORITY STUDENTS                   */
/*----------------------------------------------------------------------*/
   IF (ETHNIC=02 OR ETHNIC=03 OR ETHNIC=04 OR ETHNIC=05) THEN
MINORITY_IND=0; /*MINORITY*/
   ELSE MINORITY_IND=1;
                                /*NON-MINORITY*/
/*----------------------------------------------------------------------*/
/*CREATE A SAT COMPOSITE SCORE VARIABLE*/
/*----------------------------------------------------------------------*/
    SAT_COMP=SAT_V+SAT_M;
/*----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR SAT MATH AND
VERBAL SCORES                            */
/*----------------------------------------------------------------------*/
    IF SAT_COMP=. THEN SAT_IND=0;
    ELSE IF (0<=SAT_COMP<1000) THEN SAT_IND=1;
    ELSE IF (1000<=SAT_COMP<=1200) THEN SAT_IND=2;
    ELSE IF (SAT_COMP>1200) THEN SAT_IND=3;
/*----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR SAT VERBAL AND MATH SCORE
                                                       */
/*----------------------------------------------------------------------*/
    IF SAT_IND=0 THEN DO S1=0; S2=0; S3=0; END;
    IF SAT_IND=1 THEN DO S1=1; S2=0; S3=0; END;
```

```
    IF SAT_IND=2 THEN DO S1=0; S2=1; S3=0; END;
    IF SAT_IND=3 THEN DO S1=0; S2=0; S3=1; END;
/*--------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR ACT
COMPOSITE SCORE*/
/*--------------------------------------------------------------------*/
    IF ACT_COMPOSIT =. THEN ACT_IND_COMP=0;
    ELSE IF (ACT_COMPOSIT<21) THEN ACT_IND_COMP=1;
    ELSE IF (21<=ACT_COMPOSIT<=25) THEN ACT_IND_COMP=2;
    ELSE IF (ACT_COMPOSIT>25) THEN ACT_IND_COMP=3;
/*--------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR ACT COMPOSITE SCORE*/
/*--------------------------------------------------------------------*/
    IF ACT_IND_COMP=0 THEN DO A1=0; A2=0; A3=0; END;
    IF ACT_IND_COMP=1 THEN DO A1=1; A2=0; A3=0; END;
    IF ACT_IND_COMP=2 THEN DO A1=0; A2=1; A3=0; END;
    IF ACT_IND_COMP=3 THEN DO A1=0; A2=0; A3=1; END;
/*--------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR SAT */
/*--------------------------------------------------------------------*/
      IF (SAT_COMP=.) THEN STAK_IND=0;     /*DID NOT PROVIDE SAT SCORE*/
      ELSE STAK_IND=1;                                      /*PROVIDE
SAT SCORE*/
/*--------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR ACT*/
/*--------------------------------------------------------------------*/
      IF (ACT_M=. AND ACT_E=.) THEN ATAK_IND=0;        /*DID NOT PROVDIE
ACT SCORE*/
      ELSE ATAK_IND=1;                                      /*DID
PROVIDE ACT SCORE*/
/*--------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR GENDER      */
/*--------------------------------------------------------------------*/
      IF GENDER='F' THEN GEN_IND=0;         /*GENDER IS FEMALE*/
      ELSE IF GENDER='M' THEN GEN_IND=1;                    /*GENDER IS
MALE*/
/*----------------------------------------------------------------------
-----------------------------------*/
/*CREATE  A VARIABLE FOR THOSE STUDENTS WHO PROVIDED BOTH ACT AND SAT
SCORE*/
/*--------------------------------------------------------------------*/
      IF (ACT_COMPOSIT^=0 AND SAT_COMP^=.) THEN BOTH_TEST=1;
      ELSE BOTH_TEST=0;
/*--------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS WHO PROVIDE A MARITAL
STATUS                                            */
/*--------------------------------------------------------------------*/
    IF (MARITAL_ST ^=' ' AND MARITAL_ST='M') THEN MARITAL_IND=0;  /*DID
NOT PROVIDE MARITAL STATUS*/
      ELSE MARITAL_IND=1;                                   /*DID
PROVIDE MARITAL STATUS*/
/*--------------------------------------------------------------------*/
```

```
/*CATEGORIZE THE VARIOUS MAJORS AS SCIENCE OR NON-SCIENCE MAJORS*/
/*---------------------------------------------------------------------*/
    /*SCIENCE MAJORS INCLUDE CHEMISTRY, BIOLOGY, MATHEMATICS AND
COMPUTER SCIENCE*/
    IF MAJOR1_1 IN (11604, 11608, 11612, 10204, 10208, 10206, 10212,
10216, 10604, 10606, 10616)
      THEN
      DO
                SCI_COUNT = 1; /*COUNT SCIENCE STUDENTS*/
           MAJOR_TYPE = 0;
      END;

    IF MAJOR1_1 NOT IN (11604, 11608, 11612, 10204, 10208, 10206,
10212, 10216, 10604, 10606, 10616)
      THEN
      DO
           NON_SCI_COUNT = 1; /*COUNT NON-SCIENCE MAJORS*/
           MAJOR_TYPE = 1;
      END;
/*---------------------------------------------------------------------*/
/*FOLLOWING CODE DETERMINES WHETER A STUDENT CHANGED HIS OR HER MAJOR
FROM EACH SEMESTER TO THE NEXT*/
/*---------------------------------------------------------------------*/
      MJRCHG0=0;
      IF MAJOR1_0^=MAJOR1_1 AND MAJOR1_0^=' ' AND MAJOR1_1^=' ' THEN
MJRCHG1=1;
    ELSE MJRCHG1=0;
    IF MAJOR1_1^=MAJOR1_2 AND MAJOR1_1^=' ' AND MAJOR1_2^=' ' THEN
MJRCHG2=1;
    ELSE MJRCHG2=0;
    IF MAJOR1_2^=MAJOR1_3 AND MAJOR1_2^=' ' AND MAJOR1_3^=' ' THEN
MJRCHG3=1;
    ELSE MJRCHG3=0;
    IF MAJOR1_3^=MAJOR1_4 AND MAJOR1_3^=' ' AND MAJOR1_4^=' ' THEN
MJRCHG4=1;
    ELSE MJRCHG4=0;
    IF MAJOR1_4^=MAJOR1_5 AND MAJOR1_4^=' ' AND MAJOR1_5^=' ' THEN
MJRCHG5=1;
    ELSE MJRCHG5=0;
    IF MAJOR1_5^=MAJOR1_6 AND MAJOR1_5^=' ' AND MAJOR1_6^=' ' THEN
MJRCHG6=1;
    ELSE MJRCHG6=0;
      MJRCHGS1=MJRCHG1;
      MJRCHGS2=MJRCHG1+MJRCHG2;
      MJRCHGS3=MJRCHG1+MJRCHG2+MJRCHG3;
      MJRCHGS4=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4;
      MJRCHGS5=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4+MJRCHG5;
      MJRCHGS6=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4+MJRCHG5+MJRCHG6;
/*---------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR MAJOR CHANGES
                */
/*---------------------------------------------------------------------*/
```

```
      IF MJRCHGS6=0 THEN MJR=1;
      ELSE IF MJRCHGS6=1 THEN MJR=2;
      ELSE IF MJRCHGS6>=2 THEN MJR=3;
/*----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR MAJOR CHANGES*/
/*----------------------------------------------------------------------*/
    IF MJR=1 THEN DO MJ1=0; MJ2=0; END;
    IF MJR=2 THEN DO MJ1=1; MJ2=0; END;
    IF MJR=3 THEN DO MJ1=0; MJ2=1; END;
RUN;
PROC SORT DATA=FINAL;
    BY ZIP;
RUN;
PROC SORT DATA=RET.ZIP_DISTANCE;
    BY ZIP;
RUN;
DATA FINAL_ZIP;
   MERGE FINAL RET.ZIP_DISTANCE;
   BY ZIP;
RUN;
DATA FULL (KEEP=AGE_IND ETHNIC ACT_COMPOSIT CUM_GPA1 GOODSTART DUR
EVENT AGE1 MINORITY_IND D1 D2 D3
                        SAT_COMP GEN_IND MAJOR_TYPE MJRCHGS6 MJR DIST
HHLD_IND HS_IND SAT_IND BOTH_TEST
                        ACT_IND_COMP MARITAL_IND FRESHSTART STAK_IND
ATAK_IND DISTANCE HOUSE_VALUE HHLD_INCOME);
      SET FINAL_ZIP;
      IF ID_num=' ' THEN DELETE;

/*----------------------------------------------------------------------*/
/*CATEGORIZE DISTANCE*/
/*----------------------------------------------------------------------*/
      DIS=ROUND(DISTANCE,1);
      IF DIS=. THEN DIST=0;
      ELSE IF (DIS<100) THEN DIST=1;
      ELSE IF (100<=DIS<=500) THEN DIST=2;
      ELSE IF (DIS>500) THEN DIST=3;
/*----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR DISTANCE*/
/*----------------------------------------------------------------------*/
      IF DIST=0 THEN DO D1=0; D2=0; D3=0; END;
    IF DIST=1 THEN DO D1=1; D2=0; D3=0; END;
    IF DIST=2 THEN DO D1=0; D2=1; D3=0; END;
      IF DIST=3 THEN DO D1=0; D2=0; D3=1; END;
/*----------------------------------------------------------------------*/
/*CATEGORIZE HOUSEHOLD INCOME*/
/*----------------------------------------------------------------------*/
      IF HHLD_INCOME=. THEN HHLD_IND=0;
      ELSE IF (HHLD_INCOME<=40000) THEN HHLD_IND=1;
      ELSE IF (40000<HHLD_INCOME<=60000) THEN HHLD_IND=2;
      ELSE IF (HHLD_INCOME>60000) THEN HHLD_IND=3;
/*----------------------------------------------------------------------*/
```

```
/*CREATE DUMMY VARIABLES FOR HOUSEHOLD INCOME*/
/*----------------------------------------------------------------------*/
      IF HHLD_IND=0 THEN DO H1=0; H2=0; H3=0; END;
    IF HHLD_IND=1 THEN DO H1=1; H2=0; H3=0; END;
    IF HHLD_IND=2 THEN DO H1=0; H2=1; H3=0; END;
      IF HHLD_IND=3 THEN DO H1=0; H2=0; H3=1; END;
/*----------------------------------------------------------------------*/
/*CATEGORIZE HOUSE VALUE*/
/*----------------------------------------------------------------------*/
      IF HOUSE_VALUE=. THEN HHLD_IND=0;
      ELSE IF (HOUSE_VALUE<=100000) THEN HS_IND=1;
      ELSE IF (100000<HOUSE_VALUE<=300000) THEN HS_IND=2;
      ELSE IF (HOUSE_VALUE>300000) THEN HS_IND=3;
/*----------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR HOUSE VALUE*/
/*----------------------------------------------------------------------*/
      IF HS_IND=0 THEN DO HS1=0; HS2=0; HS3=0; END;
    IF HS_IND=1 THEN DO HS1=1; HS2=0; HS3=0; END;
    IF HS_IND=2 THEN DO HS1=0; HS2=1; HS3=0; END;
      IF HS_IND=3 THEN DO HS1=0; HS2=0; HS3=1; END;

      IF CUM_GPA1=. THEN CUM_GPA1=0;
      IF GOODSTART=. THEN GOODSTART=0;
      IF SAT_COMP=. THEN SAT_COMP=0;
      IF SAT_IND=. THEN SAT_IND=0;
      IF DIST=. THEN DIST=0;
      IF HHLD_IND=. THEN HHLD_IND=0;
      IF HS_IND=. THEN HS_IND=0;
RUN;
ODS RTF FILE="E:\Lifetest.RTF";
PROC LIFETEST DATA=FULL METHOD=LIFE INTERVALS=1 TO 7 BY 1 PLOTS=(S,H);
    TIME DUR*EVENT(0);
RUN;
ODS RTF CLOSE;
/*----------------------------------------------------------------------*/
/* PERSON PERIOD DATA SET CREATED TO RUN THE LOGISITIC PROCEDURE
                                                                     */
/*----------------------------------------------------------------------*/
DATA ENROLL;
      SET FULL;
      DO SEMESTER=0 TO MIN(DUR,5);
            IF SEMESTER=DUR AND EVENT=1 THEN ATTRITION=1;
            ELSE ATTRITION=0;
            OUTPUT;
      END;
RUN;
ODS RTF FILE="E:\LOGISTIC.RTF";
PROC LOGISTIC DESC DATA=ENROLL OUTEST=MODEL_DATA;
CLASS SEMESTER(REF='0')/PARAM=REF;
MODEL ATTRITION=AGE_IND ACT_COMPOSIT CUM_GPA1 GOODSTART AGE1
MINORITY_IND
```

```
                        SAT_COMP GEN_IND MAJOR_TYPE MJRCHGS6 DIST
HHLD_IND HS_IND SAT_IND
                        BOTH_TEST ACT_IND_COMP MARITAL_IND FRESHSTART
ATAK_IND DISTANCE
                        HOUSE_VALUE HHLD_INCOME SEMESTER/
                  SELECTION = STEPWISE LACKFIT;
RUN;
ODS RTF CLOSE;
/********************************************************/
/*              2002 CROSS VALIDATATION DATA SET         */
/********************************************************/
DATA TEST_START (KEEP=GENDER ETHNIC DOB APPLY_TERM STU_LEVEL APPLY_ST
ADMIT_ST APPLY_DATE TERM
                        YEAR NEW STATUS_LV SEMESTER APPLY CLASS MAJOR1
FIRST_TERM TOTAL_HOURS TERM_12TH
                        NEW ID_num SSN_num DOB GENDER CUM_GPA GMAT
GRE_A GRE_Q GRE_V SAT_V SAT_M
                        SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP);
    SET ret.student_data_fall ret.student_data_spring;
      IF ETHNIC='XX' THEN ETHNIC=08;
      IF ETHNIC='ZZ' THEN ETHNIC=09;
    **FORMAT APPLY_DATE MMDDYY8. ETHNIC ETHFMT. MAJOR1 $MAJ_FMT. GENDER
$SEXFMT. DOB MMDDYY8.;
RUN;
/*----------------------------------------------------------------------*/
/*DATA SETS CREATED AS A BREAKDOWN OF SEMESTER AND YEAR                  */
/*----------------------------------------------------------------------*/
DATA FTIC_FAT (KEEP=GENDER ETHNIC DOB STU_LEVEL_0 TERM_0 YEAR_0
SEMESTER_0 CLASS_0 MAJOR1_0 TOTAL_HOURS_0 FTIC_CH T_0
                   ID_num SEMCNT GMAT GRE_A GRE_Q GRE_V SAT_V SAT_M
SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
     SP_1T    (KEEP=GENDER ETHNIC DOB STU_LEVEL_1 TERM_1 YEAR_1
SEMESTER_1 CLASS_1 MAJOR1_1 TOTAL_HOURS_1 CUM_GPA1 T_1
                          ID_num SEMCNT_1 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
     FA_2T    (KEEP=GENDER ETHNIC DOB STU_LEVEL_2 TERM_2 YEAR_2
SEMESTER_2 CLASS_2 MAJOR1_2 TOTAL_HOURS_2 CUM_GPA2 T_2
                          ID_num SEMCNT_2 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
     SP_3T    (KEEP=GENDER ETHNIC DOB STU_LEVEL_3 TERM_3 YEAR_3
SEMESTER_3 CLASS_3 MAJOR1_3 TOTAL_HOURS_3 CUM_GPA3 T_3
                          ID_num SEMCNT_3 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
     FA_4T    (KEEP=GENDER ETHNIC DOB STU_LEVEL_4 TERM_4 YEAR_4
SEMESTER_4 CLASS_4 MAJOR1_4 TOTAL_HOURS_4 CUM_GPA4 T_4
                        ID_num SEMCNT_4 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
     SP_5T    (KEEP=GENDER ETHNIC DOB STU_LEVEL_5 TERM_5 YEAR_5
SEMESTER_5 CLASS_5 MAJOR1_5 TOTAL_HOURS_5 CUM_GPA5 T_5
                          ID_num SEMCNT_5 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV)
```

```
     FA_6T     (KEEP=GENDER ETHNIC DOB STU_LEVEL_6 TERM_6 YEAR_6
SEMESTER_6 CLASS_6 MAJOR1_6 TOTAL_HOURS_6 CUM_GPA6 T_6
                              ID_num SEMCNT_6 GMAT GRE_A GRE_Q GRE_V SAT_V
SAT_M SAT_V ACT_M ACT_E ACT_COMPOSIT MARITAL_ST ZIP STATUS_LV);
      SET TEST_START;

/*-------------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 00', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*-------------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='02' THEN
        DO;
                /* AND CLASS IN ('FR', 'SO')*/
                IF STATUS_LV='02' AND NEW=1 THEN FTIC_CH=1; /*used to
mark the cohort group*/
                SEMCNT=1;
                  T_0='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_0=STU_LEVEL;
            TERM_0=TERM;
            YEAR_0=YEAR;
            SEMESTER_0=SEMESTER;
            CLASS_0=CLASS;
            MAJOR1_0=MAJOR1;
            TOTAL_HOURS_0=TOTAL_HOURS;
            TERM_0=TERM_12TH;
           OUTPUT FTIC_FAT;
        END;
/*-------------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 01', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*-------------------------------------------------------------------------*/
      IF SEMESTER='SP' AND YEAR='03' THEN
        DO;
            SEMCNT_1=1;
                  T_1='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_1=STU_LEVEL;
            TERM_1=TERM;
            YEAR_1=YEAR;
            SEMESTER_1=SEMESTER;
            CLASS_1=CLASS;
            MAJOR1_1=MAJOR1;
            TOTAL_HOURS_1=TOTAL_HOURS;
            TERM_1=TERM_12TH;
            CUM_GPA1=CUM_GPA;
            OUTPUT SP_1T;
        END;
```

```
/*--------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 01', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*--------------------------------------------------------------------*/
      IF SEMESTER='FA' AND YEAR='03' THEN
        DO;
            SEMCNT_2=1;
                  T_2='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_2=STU_LEVEL;
            TERM_2=TERM;
            YEAR_2=YEAR;
            SEMESTER_2=SEMESTER;
            CLASS_2=CLASS;
            MAJOR1_2=MAJOR1;
            TOTAL_HOURS_2=TOTAL_HOURS/100000;
            TERM_2=TERM_12TH;
            CUM_GPA2=CUM_GPA;
            OUTPUT FA_2T;
        END;
/*--------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 02', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*--------------------------------------------------------------------*/
    IF SEMESTER='SP' AND YEAR='04' THEN
        DO;
            SEMCNT_3=1;
                  T_3='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
                  STU_LEVEL_3=STU_LEVEL;
            TERM_3=TERM;
            YEAR_3=YEAR;
            SEMESTER_3=SEMESTER;
            CLASS_3=CLASS;
            MAJOR1_3=MAJOR1;
            TOTAL_HOURS_3=TOTAL_HOURS;
            TERM_3=TERM_12TH;
            CUM_GPA3=CUM_GPA;
            OUTPUT SP_3T;
        END;

/*--------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 02', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,  */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.         */
/*--------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='04' THEN
```

```
        DO;
            SEMCNT_4=1;
                  T_4='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
            STU_LEVEL_4=STU_LEVEL;
            TERM_4=TERM;
            YEAR_4=YEAR;
            SEMESTER_4=SEMESTER;
            CLASS_4=CLASS;
            MAJOR1_4=MAJOR1;
            TOTAL_HOURS_4=TOTAL_HOURS;
            TERM_4=TERM_12TH;
            CUM_GPA4=CUM_GPA;
            OUTPUT FA_4T;
        END;
/*----------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN SPRING 03', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,*/
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.          */
/*----------------------------------------------------------------------*/
    IF SEMESTER='SP' AND YEAR='05' THEN
        DO;
            SEMCNT_5=1;
                  T_5='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
            STU_LEVEL_5=STU_LEVEL;
            TERM_5=TERM;
            YEAR_5=YEAR;
            SEMESTER_5=SEMESTER;
            CLASS_5=CLASS;
            MAJOR1_5=MAJOR1;
            TOTAL_HOURS_5=TOTAL_HOURS;
            TERM_5=TERM_12TH;
            CUM_GPA5=CUM_GPA;
            OUTPUT SP_5T;
        END;
/*----------------------------------------------------------------------*/
/*DATA SET OF ALL STUDENTS ENROLLED IN FALL 03', INCREMENTS SEMESTER
COUNT IF TWELTH DAY STUDENT,   */
/*AND RENAMES CERTAIN VARIABLES SO THEY WILL NOT BE REPLACED WHEN THE
DATA SETS ARE MERGED.          */
/*----------------------------------------------------------------------*/
    IF SEMESTER='FA' AND YEAR='05' THEN
        DO;
            SEMCNT_6=1;
                  T_6='0';/*ZERO MEANS THAT THEY DID NOT EXPERIENCE
DROP OUT*/
            STU_LEVEL_6=STU_LEVEL;
            TERM_6=TERM;
            YEAR_6=YEAR;
            SEMESTER_6=SEMESTER;
```

```
            CLASS_6=CLASS;
            MAJOR1_6=MAJOR1;
            TOTAL_HOURS_6=TOTAL_HOURS;
            TERM_6=TERM_12TH;
            CUM_GPA6=CUM_GPA;
            OUTPUT FA_6T;
        END;
RUN;
/*------------------------------------------------------------------*/
/*SORTING THE PREVIOUSLY CREATED DATA SETS                        */
/*------------------------------------------------------------------*/
PROC SORT DATA=FTIC_FAT;
    BY ID_num;
RUN;

PROC SORT DATA=SP_1T;
    BY ID_num;
RUN;

PROC SORT DATA=FA_2T;
    BY ID_num;
RUN;

PROC SORT DATA=SP_3T;
    BY ID_num;
RUN;

PROC SORT DATA=FA_4T;
    BY ID_num;
RUN;

PROC SORT DATA=SP_5T;
    BY ID_num;
RUN;

PROC SORT DATA=FA_6T;
    BY ID_num;
RUN;
/*------------------------------------------------------------------*/
/*MERGES EACH OF THE DATA SETS PREVIOUSLY SORTED ON THE COMMON VARIABLE
ID_num                              */
/*------------------------------------------------------------------*/
DATA FTICMERGT;
    MERGE FTIC_FAT SP_1T FA_2T SP_3T FA_4T SP_5T FA_6T;
      *INFORMAT T $CHAR6.;
 BY ID_num;
 IF FTIC_CH=. THEN DELETE;
 /*CREATE RANDOM VARIABLE T*/
 ARRAY VARIABLE{7} T_0 T_1 T_2 T_3 T_4 T_5 T_6;
 DO I=1 TO 7;
   IF VARIABLE{I}=' '  THEN VARIABLE{I}='1';/*ONE MEANS THAT THEY DID
EXPERIENCE DROP OUT*/
```

```
 END;
 /* THIS IS THE RAW RANDOM VECTOR WHICH WILL CONTIAN VALUES SUCH AS
VECTOR=0010010, WHICH WILL NEED TO
    REFORMATTED TO SAY VECTOR=0010000<--NOTICE THAT ALL VALUES AFTER
THE INTITAL '1' WERE ZEROED OUT*/
      VECTOR =T_1||T_2||T_3||T_4||T_5||T_6; /*ONLY NEED 1-6 B/C AT TIME
0 ARE DETERMINED TO DROP OUT AT T=1*/
 /* DETERMINE ERRONEOUS RESPONSES CONSIDERING THE SURVIVAL ANANLYSIS
VECTOR*/
    DUR=(INDEX(VECTOR,'1'))-1;/*RETURNS THE POSITION THAT THE FIRST
OBSERVED VALUE OF '1' OCCUREED*/
      VECTOR_ERROR=0;
    IF INDEX(VECTOR,'10')>0 THEN VECTOR_ERROR=1;/* A '10' VALUE
INDICATES THAT AN INDIVIDUAL RETURNED AFTER DROPPING OUT*/
      IF DUR=-1 THEN DUR=6;   /*THE STUDENTS DID NOT EXPERIENCE EVENT
SO CENSORED AT END OF DATA COLLECTION*/
      IF DUR<6 THEN EVENT=1; ELSE EVENT=0; /*1 MEANS THEY DROPPED OUT*/
      IF (CLASS_0 IN ('SO','JR', 'SR', 'MM', 'PB') AND DUR>5) THEN
EVENT=0;  /*FOR PEOPLE WHO ENTER ABOVE FRESHMAN IF COMPLETE 3 YEARS WE
ASSUME GRADUATE*/
RUN;
/*------------------------------------------------------------------*/
/*MERGED DATA SET INCLUDING ONLY THOSE STUDENTS IN THE FTIC COHORT
GROUP                                   */
/*------------------------------------------------------------------*/
DATA FINALT (DROP=AGE1_1 AGE2_2 AGE3_3 AGE4_4 AGE5_5 AGE6_6 AGE7_7
AGE8_8 AGE9_9 AGE10_10 AGE11_11 AGE12_12);
    SET FTICMERGT;
      **FORMAT APPLY_DATE MMDDYY8. GENDER $SEXFMT. DOB DATE7. SUCCESS
SUCFMT. MAJOR SCIFMT.;
/*------------------------------------------------------------------*/
/*DETERMINES STUDENT'S AGE BASED ON THE CORRESPONDING SCHOOL YEAR*/
/*------------------------------------------------------------------*/
      AGE1_1=(('15AUG02'D-DOB)/365.25);
      AGE1=ROUND(AGE1_1,.01);
      AGE2_2=(('15OCT02'D-DOB)/365.25);
      AGE2=ROUND(AGE2_2,.01);
      AGE3_3=(('15JAN03'D-DOB)/365.25);
      AGE3=ROUND(AGE3_3,.01);
      AGE4_4=(('15MAR03'D-DOB)/365.25);
      AGE4=ROUND(AGE4_4,.01);
      AGE5_5=(('15AUG03'D-DOB)/365.25);
      AGE5=ROUND(AGE5_5,.01);
      AGE6_6=(('15OCT03'D-DOB)/365.25);
      AGE6=ROUND(AGE6_6,.01);
      AGE7_7=(('15JAN04'D-DOB)/365.25);
      AGE7=ROUND(AGE7_7,.01);
      AGE8_8=(('15MAR04'D-DOB)/365.25);
      AGE8=ROUND(AGE8_8,.01);
      AGE9_9=(('15AUG04'D-DOB)/365.25);
      AGE9=ROUND(AGE9_9,.01);
      AGE10_10=(('15OCT04'D-DOB)/365.25);
```

```
      AGE10=ROUND(AGE10_10,.01);
      AGE11_11=(('15JAN05'D-DOB)/365.25);
      AGE11=ROUND(AGE11_11,.01);
      AGE12_12=(('15MAR05'D-DOB)/365.25);
      AGE12=ROUND(AGE12_12,.01);
/*----------------------------------------------------------------------*/
 /*CREATE AN INDICATOR VARIABLE FOR STUDENT AGE GROUPS         */
 /*----------------------------------------------------------------------
*/
      IF AGE1<=21 THEN AGE_IND=0;
            IF AGE1>21 THEN AGE_IND=1;
/*----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR GPA AFTER
ONE SCHOOL YEAR                          */
/*----------------------------------------------------------------------*/
    IF CUM_GPA1 = . THEN GOODSTART=.;
    ELSE IF (CUM_GPA1<2) THEN GOODSTART=1;
      ELSE IF (2<=CUM_GPA1<3) THEN GOODSTART=2;
      ELSE IF (CUM_GPA1>=3) THEN GOODSTART=3;
/*----------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR MINORITY STUDENTS      */
/*----------------------------------------------------------------------*/
   IF (ETHNIC=02 OR ETHNIC=03 OR ETHNIC=04 OR ETHNIC=05) THEN
MINORITY_IND=0; /*MINORITY*/   ELSE MINORITY_IND=1;/*NON-MINORITY*/
/*----------------------------------------------------------------------*/
/*CREATE A SAT COMPOSITE SCORE VARIABLE
                                                     */

/*----------------------------------------------------------------------*/
    SAT_COMP=SAT_V+SAT_M;
/*----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR SAT MATH AND
VERBAL SCORES                         */
/*----------------------------------------------------------------------*/
    IF SAT_COMP=. THEN SAT_IND=.;
    ELSE IF (0<=SAT_COMP<1000) THEN SAT_IND=1;
    ELSE IF (1000<=SAT_COMP<1200) THEN SAT_IND=2;
    ELSE IF (SAT_COMP>1200) THEN SAT_IND=3;
/*----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS BASED ON THEIR ACT
COMPOSITE SCORE*/
/*----------------------------------------------------------------------*/
    IF ACT_COMPOSIT =. THEN ACT_IND_COMP=.;
    ELSE IF (ACT_COMPOSIT<21) THEN ACT_IND_COMP=1;
    ELSE IF (21<=ACT_COMPOSIT<=25) THEN ACT_IND_COMP=2;
    ELSE IF (ACT_COMPOSIT>25) THEN ACT_IND_COMP=3;
/*----------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR SAT*/
/*----------------------------------------------------------------------*/
      IF (SAT_COMP=.) THEN STAK_IND=0;     /*DID NOT PROVIDE SAT SCORE*/
      ELSE STAK_IND=1;                                      /*PROVIDE
SAT SCORE*/
/*----------------------------------------------------------------------*/
```

```
/*CREATE INDICATOR VARIABLES FOR ACT*/
/*-----------------------------------------------------------------------*/
      IF (ACT_M=. AND ACT_E=.) THEN ATAK_IND=0;        /*DID NOT PROVDIE
ACT SCORE*/
      ELSE ATAK_IND=1;                                        /*DID
PROVIDE ACT SCORE*/
/*-----------------------------------------------------------------------*/
/*CREATE INDICATOR VARIABLES FOR GENDER                           */
/*-----------------------------------------------------------------------*/
      IF GENDER='F' THEN GEN_IND=0;          /*GENDER IS FEMALE*/
      ELSE IF GENDER='M' THEN GEN_IND=1;                      /*GENDER IS
MALE*/
/*-----------------------------------------------------------------------
---------------------------------*/
/*CREATE  A VARIABLE FOR THOSE STUDENTS WHO PROVIDED BOTH ACT AND SAT
SCORE*/
/*-----------------------------------------------------------------------*/
      IF (ACT_COMPOSIT^=0 AND SAT_COMP^=.) THEN BOTH_TEST=1;
      ELSE BOTH_TEST=0;
/*-----------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR STUDENTS WHO PROVIDE A MARITAL
STATUS*/
/*-----------------------------------------------------------------------*/
    IF (MARITAL_ST ^=' ' AND MARITAL_ST='M') THEN MARITAL_IND=0;  /*DID
NOT PROVIDE MARITAL STATUS*/
      ELSE MARITAL_IND=1;                                     /*DID
PROVIDE MARITAL STATUS*/
/*-----------------------------------------------------------------------*/
/*CATEGORIZE THE VARIOUS MAJORS AS SCIENCE OR NON-SCIENCE MAJORS*/
/*-----------------------------------------------------------------------*/
    /*SCIENCE MAJORS INCLUDE CHEMISTRY, BIOLOGY, MATHEMATICS AND
COMPUTER SCIENCE*/
    IF MAJOR1_1 IN (11604, 11608, 11612, 10204, 10208, 10206, 10212,
10216, 10604, 10606, 10616)
      THEN
      DO
                SCI_COUNT = 1; /*COUNT SCIENCE STUDENTS*/
            MAJOR_TYPE = 0;
      END;

    IF MAJOR1_1 NOT IN (11604, 11608, 11612, 10204, 10208, 10206,
10212, 10216, 10604, 10606, 10616)
      THEN
      DO
            NON_SCI_COUNT = 1; /*COUNT NON-SCIENCE MAJORS*/
            MAJOR_TYPE = 1;
      END;
/*-----------------------------------------------------------------------*/
/*FOLLOWING CODE DETERMINES WHETER A STUDENT CHANGED HIS OR HER MAJOR
FROM EACH SEMESTER TO THE NEXT*/
/*-----------------------------------------------------------------------*/
    MJRCHG0=0;
```

```
      IF MAJOR1_0^=MAJOR1_1 AND MAJOR1_0^=' ' AND MAJOR1_1^=' ' THEN
MJRCHG1=1;
      ELSE MJRCHG1=0;
    IF MAJOR1_1^=MAJOR1_2 AND MAJOR1_1^=' ' AND MAJOR1_2^=' ' THEN
MJRCHG2=1;
    ELSE MJRCHG2=0;
    IF MAJOR1_2^=MAJOR1_3 AND MAJOR1_2^=' ' AND MAJOR1_3^=' ' THEN
MJRCHG3=1;
    ELSE MJRCHG3=0;
    IF MAJOR1_3^=MAJOR1_4 AND MAJOR1_3^=' ' AND MAJOR1_4^=' ' THEN
MJRCHG4=1;
    ELSE MJRCHG4=0;
    IF MAJOR1_4^=MAJOR1_5 AND MAJOR1_4^=' ' AND MAJOR1_5^=' ' THEN
MJRCHG5=1;
    ELSE MJRCHG5=0;
    IF MAJOR1_5^=MAJOR1_6 AND MAJOR1_5^=' ' AND MAJOR1_6^=' ' THEN
MJRCHG6=1;
    ELSE MJRCHG6=0;
      MJRCHGS1=MJRCHG1;
      MJRCHGS2=MJRCHG1+MJRCHG2;
      MJRCHGS3=MJRCHG1+MJRCHG2+MJRCHG3;
      MJRCHGS4=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4;
      MJRCHGS5=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4+MJRCHG5;
      MJRCHGS6=MJRCHG1+MJRCHG2+MJRCHG3+MJRCHG4+MJRCHG5+MJRCHG6;
/*-------------------------------------------------------------------*/
/*CREATE AN INDICATOR VARIABLE FOR MAJOR CHANGES*/
/*-------------------------------------------------------------------*/
      IF MJRCHGS6=0 THEN MJR=1;
      ELSE IF MJRCHGS6=1 THEN MJR=2;
      ELSE IF MJRCHGS6>=2 THEN MJR=3;
/*-------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR MAJOR CHANGES*/
/*-------------------------------------------------------------------*/
    IF MJR=1 THEN DO MJ1=0; MJ2=0; END;
    IF MJR=2 THEN DO MJ1=1; MJ2=0; END;
    IF MJR=3 THEN DO MJ1=0; MJ2=1; END;
RUN;
PROC SORT DATA=FINALT;
    BY ZIP;
RUN;
PROC SORT DATA=RET.ZIP_DISTANCE;
    BY ZIP;
RUN;
DATA FINAL_ZIPT;
   MERGE FINALT RET.ZIP_DISTANCE;
   BY ZIP;
RUN;
DATA FULLT (KEEP=ID_num AGE_IND ETHNIC ACT_COMPOSIT CUM_GPA1 GOODSTART
DUR EVENT AGE1 MINORITY_IND MARITAL_IND
                        SAT_COMP GEN_IND MAJOR_TYPE MJRCHGS6 MJR DIST
HHLD_IND HS_IND SAT_IND
                        ACT_IND_COMP T_0 T_1 T_2 T_3 T_4 T_5 MJ1 MJ2);
```

```
      SET FINAL_ZIPT;
      IF ID_num=' ' THEN DELETE;
/*------------------------------------------------------------------------*/
/*CATEGORIZE DISTANCE*/
/*------------------------------------------------------------------------*/
      DIS=ROUND(DISTANCE,1);
      IF DIS=. THEN DIST=.;
      ELSE IF (DIS<100) THEN DIST=1;
      ELSE IF (100<=DIS<=500) THEN DIST=2;
      ELSE IF (DIS>500) THEN DIST=3;
/*------------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR DISTANCE*/
/*------------------------------------------------------------------------*/
      IF DIST=. THEN DO D1=0; D2=0; D3=0; END;
    IF DIST=1 THEN DO D1=1; D2=0; D3=0; END;
    IF DIST=2 THEN DO D1=0; D2=1; D3=0; END;
      IF DIST=3 THEN DO D1=0; D2=0; D3=0; END;
/*------------------------------------------------------------------------*/
/*CATEGORIZE HOUSEHOLD INCOME
                                                                   */
/*------------------------------------------------------------------------*/
      IF HHLD_INCOME=. THEN HHLD_IND=.;
      ELSE IF (HHLD_INCOME<=40000) THEN HHLD_IND=1;
      ELSE IF (40000<HHLD_INCOME<=60000) THEN HHLD_IND=2;
      ELSE IF (HHLD_INCOME>60000) THEN HHLD_IND=3;
/*------------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR HOUSEHOLD INCOME*/
/*------------------------------------------------------------------------*/
      IF HHLD_IND=. THEN DO H1=0; H2=0; H3=0; END;
    IF HHLD_IND=1 THEN DO H1=1; H2=0; H3=0; END;
    IF HHLD_IND=2 THEN DO H1=0; H2=1; H3=0; END;
      IF HHLD_IND=3 THEN DO H1=0; H2=0; H3=0; END;
/*------------------------------------------------------------------------*/
/*CATEGORIZE HOUSE VALUE
                                                                   */
/*------------------------------------------------------------------------*/
      IF HOUSE_VALUE=. THEN HHLD_IND=.;
      ELSE IF (HOUSE_VALUE<=100000) THEN HS_IND=1;
      ELSE IF (100000<HOUSE_VALUE<=300000) THEN HS_IND=2;
      ELSE IF (HOUSE_VALUE>300000) THEN HS_IND=3;
/*------------------------------------------------------------------------*/
/*CREATE DUMMY VARIABLES FOR HOUSE VALUE*/
/*------------------------------------------------------------------------*/
      IF HS_IND=. THEN DO HS1=0; HS2=0; HS3=0; END;
    IF HS_IND=1 THEN DO HS1=1; HS2=0; HS3=0; END;
    IF HS_IND=2 THEN DO HS1=0; HS2=1; HS3=0; END;
      IF HS_IND=3 THEN DO HS1=0; HS2=0; HS3=0; END;

      IF CUM_GPA1=. THEN CUM_GPA1=0;
      IF GOODSTART=. THEN GOODSTART=0;
      IF SAT_COMP=. THEN SAT_COMP=0;
      IF SAT_IND=. THEN SAT_IND=0;
```

```
        IF DIST=. THEN DIST=0;
        IF HHLD_IND=. THEN HHLD_IND=0;
        IF HS_IND=. THEN HS_IND=0;
RUN;

DATA MODEL_FIT (KEEP=ID_num T_1 T_2 T_3 T_4 T_5 EVENT LOGIT LOGITA
LOGITB LOGITC LOGITD LOGITE
                                    PROB_SUM PROB_SUMA PROB_SUMB PROB_SUMC
PROB_SUMD PROB_SUME);
        SET FULLT;
        B0=-2.0374;
        CGPA=-.9599;
        CMIN=.3441;
        CMTP=.5177;
        CMJR=-1.0769;
        CT1=3.4975;
        CT2=2.5539;
        CT3=3.6246;
        CT4=2.1831;
        CT5=3.2168;
        IF T_1='0' THEN T_1='1';
        ELSE T_1='0';
        IF T_2='0' THEN T_2='1';
        ELSE T_2='0';
        IF T_3='0' THEN T_3='1';
        ELSE T_3='0';
        IF T_4='0' THEN T_4='1';
        ELSE T_4='0';
        IF T_5='0' THEN T_5='1';
        ELSE T_5='0';
LOGIT=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJO
R_TYPE);
LOGITA=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJ
OR_TYPE)+(CT1*T_1);
LOGITB=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJ
OR_TYPE)+(CT2*T_2);
LOGITC=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJ
OR_TYPE)+(CT3*T_3);
LOGITD=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJ
OR_TYPE)+(CT4*T_4);
LOGITE=B0+(CGPA*CUM_GPA1)+(CMJR*MJRCHGS6)+(CMIN*MINORITY_IND)+(CMTP*MAJ
OR_TYPE)+(CT5*T_5);

PROB = EXP(LOGIT)/(1+EXP(LOGIT));
RETAIN PROB_SUM 0;
IF PROB^=. THEN PROB_SUM=PROB+PROB_SUM;
PROBA = EXP(LOGITA)/(1+EXP(LOGITA));
RETAIN PROB_SUMA 0;
IF PROBA^=. THEN PROB_SUMA=PROBA+PROB_SUMA;
PROBB = EXP(LOGITB)/(1+EXP(LOGITB));
RETAIN PROB_SUMB 0;
IF PROBB^=. THEN PROB_SUMB=PROBB+PROB_SUMB;
```

```
PROBC = EXP(LOGITC)/(1+EXP(LOGITC));
RETAIN PROB_SUMC 0;
IF PROBC^=. THEN PROB_SUMC=PROBC+PROB_SUMC;
PROBD = EXP(LOGITD)/(1+EXP(LOGITD));
RETAIN PROB_SUMD 0;
IF PROBD^=. THEN PROB_SUMD=PROBD+PROB_SUMD;
PROBE = EXP(LOGITE)/(1+EXP(LOGITE));
RETAIN PROB_SUME 0;
IF PROBE^=. THEN PROB_SUME=PROBE+PROB_SUME;
RUN;
PROC LIFETEST DATA=FULLT METHOD=LIFE /*PLOTS=(S,H)*/;
    TIME DUR*EVENT(0);
RUN;
/*-----------------------------------------------------------------------------*/
/*FREQUENCY TABLES FOR DEPENDENT VARIABLE AND INDEPENDENT VARIABLES  */
/*-----------------------------------------------------------------------------*/
ODS RTF FILE="E:\ETHNIC.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*ETHNIC/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. ETHNIC $ETHFMT.;
TITLE 'DROP BY ETHNICITY';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\MINOR.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*MINORITY_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. MINORITY_IND MINRFMT.;
TITLE 'DROP BY MINORITY VS. NON-MINORITY';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\GENDER.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*GENDER/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. GENDER $SEXFMT.;
TITLE 'DROP BY GENDER';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\AGE.RTF";
PROC FREQ DATA=FINAL;           /*FREQUNCY OF ATTRITION BASED ON INITIAL
AGE AT ENROLLMENT*/
TABLE EVENT*AGE_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT  EVENT EVENTFMT. AGE_IND AGE_FMT.;
TITLE 'DROP BY AGE';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\SAT.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*SAT_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. SAT_IND SAT_FMT.;
TITLE 'DROP BY SAT COMPOSITE SCORE';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
```

```
ODS RTF FILE="E:\STAK.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*STAK_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. STAK_IND EXMFMT.;
TITLE 'DROP BY SAT TAKEN';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\ACT.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*ACT_IND_COMP/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. ACT_IND_COMP ACT_FMT.;
TITLE 'DROP BY ACT COMPOSITE SCORE';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\ATAK.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*ATAK_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. ATAK_IND EXMFMT.;
TITLE 'DROP BY ACT TAKEN';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\MAJOR.RTF";
PROC FREQ DATA=FINAL;         /*FREQUNCY OF ATTRITION BASED WHETHER
AJOR IS SCIENCE OR NON-SCIENCE*/
TABLE EVENT*MAJOR_TYPE/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. MAJOR_TYPE SCIFMT.;
TITLE 'DROP BY MAJOR (SCIENCE VS. NON-SCIENCE)';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\MARITAL.RTF";
PROC FREQ DATA=FINAL;         /*FREQUNCY OF ATTRITION BASED ON GPA
AFTER 1 YEAR OF ENROLLMENT*/
TABLE EVENT*MARITAL_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. MARITAL_IND MARFMT.;
TITLE 'DROP BY MARITAL STATUS';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\MJRCHG.RTF";
PROC FREQ DATA=FINAL;
TABLE EVENT*MJR/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. MJR CHGFMT.;
TITLE 'DROP BY NUMBER OF MAJOR CHANGES';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\HHLD.RTF";
PROC FREQ DATA=FULL;
TABLE EVENT*HHLD_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. HHLD_IND HHLD_FMT.;
TITLE 'DROP BY NUMBER OF MAJOR CHANGES';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\HOUSE.RTF";
```

```
PROC FREQ DATA=FULL;
TABLE EVENT*HS_IND/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. HS_IND HS_FMT.;
TITLE 'DROP BY NUMBER OF MAJOR CHANGES';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\DISTANCE.RTF";
PROC FREQ DATA=FULL;
TABLE EVENT*DIST/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. DIST DIST_FMT.;
TITLE 'DROP BY DISTANCE';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\BOTH.RTF";
PROC FREQ DATA=FULL;
TABLE EVENT*BOTH_TEST/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. BOTH_TEST BOTH_FMT.;
TITLE 'DROP BY BOTH ACT/SAT PROVIDED';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
ODS RTF FILE="E:\GOODSTUDENT.RTF";
PROC FREQ DATA=FULL;
TABLE EVENT*GOODSTART/ NOPERCENT NOROW /*NOCOL*/;
FORMAT EVENT EVENTFMT. GOODSTART GOOD_FMT.;
TITLE 'DROP BY GOODSTART';
RUN;
ODS RTF CLOSE; /*END RTF-OUTPUT*/
```