Applying Academic Analytics

Developing a Process for Utilizing Bayesian Networks to Predict Stopping Out Among Community College Students

by

Philip Arcuria

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved December 2014 by the
Graduate Supervisory Committee:

Roy Levy, Chair
Samuel Green
Marilyn Thompson

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

Many methodological approaches have been utilized to predict student retention and persistence over the years, yet few have utilized a Bayesian framework. It is believed this is due in part to the absence of an established process for guiding educational researchers reared in a frequentist perspective into the realms of Bayesian analysis and educational data mining. The current study aimed to address this by providing a model-building process for developing a Bayesian network (BN) that leveraged educational data mining, Bayesian analysis, and traditional iterative model-building techniques in order to predict whether community college students will stop out at the completion of each of their first six terms. The study utilized exploratory and confirmatory techniques to reduce an initial pool of more than 50 potential predictor variables to a parsimonious final BN with only four predictor variables. The average in-sample classification accuracy rate for the model was 80% (Cohen's κ = 53%). The model was shown to be generalizable across samples with an average out-of-sample classification accuracy rate of 78% (Cohen's κ = 49%). The classification rates for the BN were also found to be superior to the classification rates produced by an analog frequentist discrete-time survival analysis model.

DEDICATION

To Mary & Victoria, the guiding lights of my life.

Hallelujah! Shel Silverstein was right. The sidewalk does eventually end! But the journey was not without its costs. The greatest of which was borne by the two of you. Too long the journey has required me to be sequestered either in presence or in mind from the two of you as I traversed bend after bend. I only hope the destination was worth the toll. At last, the journey has finally come to a close. I am coming home.

To Dylynn, one of the world's true superheroes who is a constant reminder that to surrender is not an option.

Although we cannot see you I know you are always there to remind us that it is not how long you live but how you live that matters most. Today is going to be a great day.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Drs. Sam Green, Marilyn Thompson, and Roy Levy who are three of the greatest people I have had the pleasure to meet. They have guided and nurtured me along every step of my journey. Their ego-less comradery, tenacity, and collective commitment to student success above all else have kept the spirit of the MSMS program alive despite its physical destruction and the stealthy attempt to scatter its memories through free agency. One could not have been blessed with a better group of faculty. I have spent a lot of time trying to formulate the best way to articulate in words my level of appreciation toward the three of them. Their wonderful qualities have already been extolled in probably every conceivable word choice in the collection of Acknowledgement pages penned by their numerous former students. Because of this, any words I write come off the pen and onto paper as unoriginal and duller than they appear in my mind. So in the end I will try to convey my feelings in a less indirect manner by simply saying that not even the seed of the greatest oak can germinate in a barren desert. Thank you for being the oasis that has allowed me to sprout and take root. I am indebted to the three of you.

A special thank you is owed to Dr. Levy for his mentorship on my dissertation. He is the rare combination of vast intellect, patience, and humility. He is a sage who is much wiser than his years. I would also like to thank Drs. Alka Arora Singh and Bryan Palmer, as well as countless colleagues, friends, and family members for their continued support throughout the process. Your support mattered more than you probably realized.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

Introduction

In 1995, the United States, along with New Zealand, had the highest first-time graduation rates among 30 Organization for Economic Co-operation and Development (OECD) countries. By 2008, the United States had slipped to 14th on the list. In 2009, President Obama sought to reverse this trend by challenging institutions of higher education to reclaim the top spot by 2020 (United States White House). This challenge has led to policy statements at both the national and local levels. For example, at the national level, in 2009 the Lumina Foundation established its Goal 2025 initiative aimed at increasing the percentage of adult Americans (ages 25-64) who hold a postsecondary degree or certificate to 60% by 2025. The College Board Advocacy and Policy Center issued a similar initiative, titled the College Completion Agenda, aimed at increasing the percentage of adults 24-35 that hold at least an associate degree to 55% by 2025. The reverberations of the President's challenge can also be observed at a state level. For example, the Arizona Community College Presidents' Council (ACCPC) set the goal of increasing the number of adults with a college degree in Arizona from 25% in 2010 to 30% in 2020 (Arizona Community Colleges, n.d.).

The Great Recession has further heightened the perceived need for postsecondary education, creating immense pressure on institutions of higher education to drive the nation's goal of increasing degree completion rates. In 2013, Georgetown University's Center on Education and the Workforce predicted that 65% of all U.S. jobs will require some college-level education by 2020 (Carnevale, Smith, & Strohl, 2013). Forty-seven percent (47%) will require an associate's degree or higher. In comparison, only 28% of

U.S. jobs required an associate's degree or higher in 1973. As of 2011, only 39% of Americans ages 25 to 64 held an associate's degree or higher (Lumina Foundation for Education, 2013). Postsecondary education results in higher incomes and lower unemployment rates, on average. In 2012, the median weekly income for full-time U.S. workers aged 25 or older who earned an associate's degree was 20% more than those who only earned a high school diploma, and their unemployment rate was 25% lower (U.S. Bureau Labor Statistics, 2013). The average economic benefits are even greater for higher levels of educational attainment.

To meet these growing needs, postsecondary institutions must become more effective at retaining and graduating the students they enroll rather than simply focusing on recruiting and enrolling more students. Exact national graduation rates are difficult to calculate due to the lack of a standardized methodology and common data set. In 2012, the National Student Clearinghouse Research Center conducted perhaps the most comprehensive study to date on graduation rates that included 1.9 million students who first enrolled at a U.S. postsecondary institution in 2006. The cohort included students at four-year and two-year public and private institutions. It also took into account students who may have started at one institution and graduated from a different institution. The authors found that only 54% of first-time, degree-seeking postsecondary students earned a degree or certification within six years (Shapiro et al., 2012). In other words, postsecondary institutions as a whole are failing to achieve their core mission with almost half of full-time degree-seeking students. The figure is even more sobering considering that not all students attend college full-time. The success rate was even lower for two-year public institutions with a six-year graduation rate of 36%. This statistic is not

surprising given the fact that the nation's community colleges tend to serve a much more diverse student body that is generally less academically prepared than students who attend four-year institutions. For example, Kopko and Cho (2013) reviewed the 2005-06 academic year transcripts of 14,617 first-time postsecondary students at eight community colleges and only 16% were considered "college ready" in math, reading, and writing.

The anemic retention and graduation rates of U.S. postsecondary institutions have directed the attention of policy makers and higher education leaders to the potential of "academic analytics", a grafting of business intelligence practices to the field of higher education (e.g., Chacon, Spicer, & Valbuena, 2012; Wagner & Ice, 2012). For years the field of business has utilized advanced statistical modeling and data mining techniques to help optimize and inform business operations. The hope is that such techniques can be applied to higher education to help identify struggling students in real time in order to provide them with additional support before stopping or dropping out. However, despite all the attention it has received, academic analytics is still a nascent field. Efforts are currently underway to better define and circumscribe the practice of academic analytics (e.g., van Barneveld, Arnold, & Campbell, 2012; Siemens, 2012) although only a few institutions have actually operationalized such techniques (e.g., Wishon & Rome, 2012; Young, 2011).

One of the earliest and perhaps best known applications of academic analytics is Purdue University's *Signals* project (Arnold, 2010). *Signals* uses a "student success algorithm" to predict which students are "at risk" based on a combination of demographic variables and behavioral data (e.g., performance within a course). Administrators, faculty, and students receive alerts based on this information with the goal of providing "at-risk"

students with the additional resources they need to succeed. The student success algorithm is based on a logistic regression model with a dichotomized outcome variable of academic success (Campbell, 2007). Logistic regression has the advantage of being a well-established analytic technique that is widely available in mainstream statistical software programs used in the social sciences (e.g., SPSS; IBM Corp., 2013). However, it does not easily lend itself to providing dynamic updates of prior estimates once new information becomes available. New approaches are needed that permit such updating in order to better predict in real-time a student's probability (i.e., risk) of stopping out of school. Traditional techniques such as logistic regression are also tethered to a frequentist paradigm. Alternative paradigms, namely Bayesian-based analysis, arguably provide a more suitable alternative to model student success; however the techniques have been rarely used toward that end. In sum, the field of educational research is in need of expanding beyond traditional techniques of modeling postsecondary student success to include the use of Bayesian analyses.

Purpose of Research

The purpose of this study is three-fold:

(1) Expand the budding corpus of academic analytics research and practice to include Bayesian approaches to modeling postsecondary students' probability of stopping out of college;

(2) Establish a model-building process for developing Bayesian networks (BNs; Pearl, 2000) that leverages educational data mining and traditional iterative model-building techniques

(3) Develop a final user-friendly model that can be used by non-methodologists to quickly and accurately calculate understandable estimates of students' probability of stopping out of college.

Toward that end, the aspiration aim of the study is to develop and evaluate a specialized form of BNs, dynamic Bayesian Networks (DBNs; Dean & Kanazawa, 1988), in order to generate probabilistic estimates of a whether a community college student will stop out of college prior to graduating or earning a degree. Stopping out will be operationally defined as any degree- or transfer-seeking student who takes at least one fall or spring semester off from school without earning an associate's degree or certificate or transferring to a 4-year college or university. A student's probability of stopping out will be estimated at the completion of each fall and spring term over a three year period.

In developing the final model, exploratory data analysis (EDA; Cleveland, 1993; Tukey, 1977) and educational data mining (EDM; Baker, 2010) techniques will be applied to a relatively large number of variables in order to identify, along with prior research, potentially meaningful predictor variables. The collection of variables will include ex post facto demographic data (e.g., age, gender, ethnicity), course placement exam data used to determine a student's content knowledge in reading, math, and English (e.g., exam scores, course placement level), developmental and college-level course taking behaviors (e.g., credits attempted, number of courses withdrawn from, pass rates), and financial aid data (e.g., types of financial aid received, amount of federal loans received), among other data. The final BN was constructed via a series of three primary steps. First, an initial prototype of the model was developed within a discrete-time

survival analysis framework using established model building procedures. Second, the model from step one was translated into a fully Bayesian framework in order to leverage the power of Bayesian inference and subjective interpretations of probability. The final step converted the fully Bayesian models into a BN that can be more readily used to support real-time decision making.

Chapter 2

Literature Review

The current study is offered as an application of academic analytics that embraces aspects of both the EDM and traditional modeling paradigms. As such, the following chapter provides an overview of academic analytics (and its relation to learning analytics) and educational data mining. It also provides a comparison of traditional and EDM analytic methodologies, a synthesis of prior attempts to model student success, an introduction to Bayesian analysis in general and Bayesian networks in particular, as well as an introduction to discrete-time survival analysis. The chapter concludes with an overview of variables found by prior research to be significant predictors of postsecondary persistence, graduation, and/or transfer within the context of a survival analyses. The overall purpose of the chapter is to provide the reader with the theoretical, contextual, and historical information underpinning this study and its justification.

An Overview of Academic Analytics

At its broadest level academic analytics is the application of a business intelligence paradigm to the field of academics. The origin of the term "academic analytics" is credited to Goldstein and Katz in 2005 (Elias, 2011). In 2007, Campbell, DeBlois, and Oblinger expanded the conception of academic analytics to emphasize the analysis and use of data: "analytics marries large data sets, statistical techniques, and predictive modeling. It could be thought of as the practice of mining institutional data to produce 'actionable intelligence'" (p. 42).

It is important to note that academic analytics is distinct from learning analytics, although the two terms are often used interchangeably. A synthesis of the research (e.g.,

Siemens, 2010, 2012; van Barneveld, Arnold, & Campbell, 2012) identifies academic and learning analytics as two distinct yet interrelated areas of thought distinguishable on two dimensions. The first dimension is the type of data that are collected and analyzed. Learning analytics is primarily interested in data produced by the learner. In contrast, academic analytics is concerned more broadly with data relevant to an academic institution. The second dimension that characterizes the two fields is the desired end product. The ultimate goal of learning analytics is to improve learning. This distinguishes itself from academic analytics which is geared more toward increasing institutional effectiveness. One dimension the two areas *do not* seem to differ on is the types of methodologies they use. Both apply a wide variety of methodologies including traditional (e.g., logistic regression) and more modern methods (e.g., naïve Bayesian classifiers) statistical techniques.

van Barneveld, et al. (2012) further conceptualized learning analytics as a subset of academic analytics. From this perspective there will always be overlap between the two areas. For instance, the current study aims to predict the probability a student will stop out within three years of first enrolling. The results could be used to inform institutional practices, such as targeted interventions, intended to produce higher retention and graduation rates. This would arguably lead to improvements in both institutional effectiveness and student learning. That said, the study is most appropriately classified as an application of academic analytics rather than learning analytics since its focus is not specifically on student learning.

Surprisingly, there is a dearth of published applications of academic analytics in peer-reviewed literature. A search for the term "academic analytics" (unconstrained by date range limitations) in eight major research databases – ABI/INFORM, Academic Search Primer, Education Full Text, ERIC, JSTOR, Primary Search, PsycInfo, Sage Premier – returned only two peer-reviewed articles on the topic; in both, academic analytics was a brief reference and not a central focus of the article. There are several possible inferences one could form from this. The absence of published articles may indicate the lack of maturity in the field and that there has not been a long enough gestation period for research and practice to find its way onto the pages of journals and periodicals. This seems unlikely given that Goldstein and Katz' initial articulation of academic analytics was almost a decade ago. A more likely explanation is that it is a field made up of practitioners, not researchers, who are more interested in applying academic analytics than publishing research related to it. This is evidenced by the large number of secondary reports in the popular and trade press on initiatives in academic analytics at work at various institutions of higher education (Grush, 2012; Kolowich, 2013; Marcus, 2012; Meyer, 2012; Parry, 2011; Parry, 2012; Young, 2011; Wagner & Ice, 2012) and an almost complete absence of peer-reviewed research in the name of academic analytics. In order to grow and mature the field it needs a body of researchers actively conducting and publishing research on its behalf. The current study attempts to take a small step in addressing this need by explicitly stating a priori that it is being conducted under the banner of academic analytics.

An Overview of Educational Data Mining

Educational data mining is a relatively new field of study that applies traditional data mining techniques to educational data. The International Educational Data Mining Society defines EDM as, "… an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in." (http://www.educationaldatamining.org). This definition conveys that EDM is fundamentally a methodological field focused on analysis tools and techniques that are more oriented toward exploratory data analysis rather than theory- or hypothesis-driven analysis. That being said, it would be overreaching to state that EDM exists in a state completely devoid of theory and hypotheses. As Romero and Ventura (2007) commented, "the application of data mining in educational systems is an iterative cycle of hypothesis formation, testing, and refinement" (p. 136). The same authors reinforced this point in a later work (Romero & Ventura, 2010) by writing, "EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that *combine data and theory* to transform practice to benefit learners" (p. 601, italics added).

At its roots EDM stems from the broader field of data mining. Data mining applies statistical techniques and modern computing in search of finding potentially meaningful patterns in large data sets (Witten, Frank, & Hall, 2011). As institutions of higher education deployed enterprise-wide software, e.g., student information systems and course/learning management systems, in the late 1990's and early 2000's, the amount

of educational data captured increased exponentially. This produced new, fertile ground for the spread of data mining and related fields into the area of education.

Although EDM has a distinct lineage and corpus of research (Baker, 2010; Baker & Yacef, 2009; Yacef, Baker, Barnes, & Beck, 2009), it is intimately related to academic and learning analytics (Siemens, 2011; Siemens and Baker, 2012) in that they are both focused on improving educational practices and developing and refining techniques for analyzing large data sets relevant to education. The current study embraces both perspectives. The overall purpose of the study is seen as falling under the overarching paradigm of academic analytics since its ultimate goal is to develop an effective model for predicting stopping out behaviors that could ultimately be used to improve institutional effectiveness. EDM is viewed as a collection of analytic techniques that will be used, along with "traditional" analytic techniques, in the name of meeting that purpose. The next section compares and contrasts non-traditional methodologies used by EDM and traditional methodologies typically used in educational research in order to lay a conceptual foundation for the techniques that will be utilized in the study. The specific analyses to be employed will be further elucidated in later sections.

Comparing Traditional and Non-Traditional Methodologies

Types of non-traditional analyses can be broadly grouped into two categories. The first category includes Bayesian-based techniques such as Bayesian networks and naïve Bayes classifiers. These techniques will be discussed later as part of a subsequent section. The second category is a group of methods stemming from the field of data mining. Generally speaking the primary purpose of data mining is to reveal previously unknown patterns among data (Nisbet, Elder, & Miner, 2009). These techniques differ from

traditional methods in multiple ways. First, and perhaps most noteworthy, they represent different statistical philosophies (Breiman, 2001). Breiman artfully illustrated the distinction between these philosophies (or "cultures") with the use of a figurative black box:

> Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables… (p. 199)

Traditional model-based approaches stem from a statistical culture that wants to know what is happening in the black box. This perspective uses a deductive process. The analyst theorizes a priori what the phenomena are that underlie the relationships between ***x*** and ***y***. Then, the analyst seeks to simulate how "nature functions" in the black box by selecting a model that is believed to serve as a simplified approximation of the phenomena dictating the relationships between ***x*** and ***y***. The data are fed into the model and analyzed in an attempt to "confirm" (i.e., provide support for) the model and test an a priori hypothesis. In other words, traditional approaches are typically theoretically-driven, deductive, and confirmatory in nature. In contrast, data mining techniques tend to be atheoretical, inductive, and exploratory. Data mining techniques are based on a statistical culture that is less concerned with what happens in the black box. Instead, it focuses its attention on utilizing mathematical algorithms to uncover previously hidden relationships between ***x*** and ***y***. It does not encumber itself with trying to confirm a pre-specified model believed to represent what is happing in the black box. At its core, data mining is exploratory and seeks to inductively "learn" the relationships between the variables directly from the data rather than presuppose the relationships in the form of a theorized model. In short, traditional approaches generally seek to explain "why" ***x*** and ***y***

are related, while non-traditional approaches are mainly interested simply in "how" they relate. It is important to note that this distinction is not a true dichotomy but more a difference in degree.

There are several other characteristics that distinguish these two methodological perspectives. Traditional techniques have a long history (Hald, 1990; Hald, 1998). In its formative years computers did not exist. This limited the field to approaches that were tractable and could be calculated by hand. Although traditional techniques have greatly expanded with the advent of early and modern computing, the field is still tethered to some degree to approaches with historical precedence. One example is its strong allegiance to models with known distributions even when this requires having to comply with assumptions that range from onerous to unattainable. Data mining is a much younger field. Its history is inextricably bound with modern computing, descending from two branches of computer science: artificial intelligence and machine learning. As a result, data mining approaches were specifically designed to leverage the power of computers. Therefore, they typically utilize automated or semi-automated algorithms able to process large quantities of data without the reliance on known distributions and associated assumptions. A second difference is that data mining is more apt to be applied to complete data sets (e.g., data on all students from an institution's learning management system) rather than sample data (e.g., data on a subset of the institution's students collected as part of a sampling design). Therefore, concepts that sit at the heart of many traditional approaches, such as statistical significance, are less relevant to data mining approaches because the values of the population parameters (e.g., the mean of the complete data set) are known and do not need to be inferred. Third, data mining

approaches are more opportunistic. They tend to be applied to data that has already been collected for another purpose (usually by enterprise-level software, e.g., a student information or learning management system). In contrast, traditional techniques are generally applied as part of larger, pre-planned experimental or quasi-experimental design that includes both the collection and analysis of data. The net effect is that data mining techniques are more flexible and less structured than traditional techniques to account for the fact that the analyst will have little to no control over how the data are collected. Lastly, data mining aims to use all available data even if the number of variables dramatically exceeds the number of cases (Brieman, 2001). This is a stark difference from traditional methods that are built on a fundamental tenet that the number of cases must be larger than the number of variables (usually many times larger). Data mining overcomes this requirement by employing automated sampling and replication techniques specifically designed to evaluate a massive number of variable combinations even with a relatively small number of cases (e.g., support vector machines). There are instances of similar procedures in traditional approaches (e.g., stepwise regression); however they are fewer in number and are usually discouraged from use in large part because they run counter to the philosophy of traditional methodologies (Cohen, Cohen, West, & Aiken, 2003).

It is important to note that these philosophical, historical, and operational differences between traditional and non-traditional methods are generalizations. Some traditional techniques may display some or all of the characteristics of data mining and vice versa. It is also worth commenting that no one technique or family of techniques is "better" than the other. In the words of Tukey (1969), "we ought to try to calculate what

will help us most understand our data, and their indications. We ought not to be bound by preconceived notions—or preconceived analyses." (p. 83).

The current study follows Tukey's words of wisdom and will use both traditional and non-traditional analytic approaches. The traditional methodologies will consist of correlational and logistic regression analyses (and related extensions). The non-traditional approaches will include Bayesian networks and data mining techniques. These approaches will be discussed in more detail in later sections.

An Overview of Prior Methodologies Used to Predict Student Success

There is an extensive corpus of research on trying to predict and model retention/persistence[1] and graduation of postsecondary students. Although there have been several thorough reviews of retention literature in general (e.g., Astin, 1975; Braxton, Hirschy, & McClendon, 2004; Pascarella & Terenzini, 2005; Tinto, 1975), there does not appear to be any published reviews of the methodological techniques used to predict and model retention and related variables. This section summarizes an extensive but by no means exhaustive review of publications predicting retention and/or graduation. The review included the following peer-reviewed journals: *American Educational Research Journal, Computers & Education, Educational Researcher, Journal of College Student Retention, Journal of Educational Data Mining, Journal of Educational Measurement, Research in Higher Education, Review of Educational Research, Review of Research in Education, The Journal of Educational Research, The Journal of Experimental Education,* and *The Journal of Higher Education*. Various keyword searches were also conducted using several search engines (e.g., Google, Library One Search). The results revealed a wide variety of methodological approaches used to predict

retention and/or graduation. Some of the methodological techniques include: Bayesian networks (Nandeshaw, Menzies, & Nelson, 2011), correlational analysis (e.g., Crawford, 1930; Potthoff, 1931), decision trees (e.g., Herzog, 2006; Pittmann, 2008), discriminant analysis (Bers & Smith, 1991; Bianchi & Bean, 1980), growth curve modeling (e.g., Hausmann, Schofield, & Woods, 2007), linear regression (e.g., Astin, 1997; Bogan Eaton & Bean, 1995), logistic regression (e.g., Andrieu & St. John, 1993; Caison, 2007), naïve Bayes (e.g., Nandeshaw, Menzies, & Nelson, 2011; Zhang, Oussena, Clark, Hyensook, 2010), neural networks (Herzog, 2006; Pittmann, 2008), support vector machines (e.g., Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009; Zhang, Oussena, Clark, Hyensook, 2010), and survival analysis (e.g., Flores & Horn, 2009; Murtaugh, Burns, & Schuster, 1999), –among others. Among these methods, the four techniques utilized the most often were correlational analysis, linear regression, discriminant analysis, and logistic regression. An overview of each method is discussed below.

Correlational analyses. One of the first published efforts to predict student success and retention was Thurstone's 1921 article, *A Cycle-Omibus Intelligence Test for College Students*. In the article Thurstone advocated correlating students' scores on "mental tests" with their academic success and retention status as a method of trying to predict the latter from the former. Over the next two decades (and beyond) an explosion of retention and student success correlational studies appeared in the published literature (e.g., Crawford, 1930; Edgerton & Troops, 1929; Freeman, 1931; Garrett, 1949; Landry, 1937; Lillis, 2012; Mort, 1932; O'Brien, 1928; Odell, 1930; Potthoff, 1931; Prediger, 1966; Schmid & Reed, 1966).

Strengths. The strength of correlational analysis is in its simplicity. It is very easy to compute and interpret. Correlational techniques also have broad applicability. They can be applied to calculate associations between variables of all types of measurement levels, including those with latent distributions (e.g., tetrachoric & polychoric correlations). Traditional bivariate correlations can also be expanded to analyze associations between more than two variables.

Limitations. Correlational analyses provide a measure of the linear association between variables. They are not true predictive methods. As a result, they lack the theoretical framework and techniques to formally predict retention and graduation. Early correlation studies are analogous to early combustible engines. They laid the framework and principles that continue to power many more sophisticated methodologies, but in and of themselves are often too simplistic to deal with today's complex predictive modeling scenarios. They are also ill-suited for measuring nonlinear associations between variables. In addition to linearity, inferential tests of correlation coefficients are based on the assumptions that the two variables have a bivariate normal distribution and homogeneity of variance. While correlation analyses may be robust to violations of these assumptions in some situations, their use is limited in instances when these assumptions are not believed to reasonably hold.

Linear regression. Linear regression (Cohen, et al., 2003; Tabachnick & Fidell, 2007) extends correlational analyses to be able to better and more formally predict a continuous outcome variable from one or more continuous or categorical predictor variables. The methodology has been widely applied in predicting retention and postsecondary academic success variables (e.g., Astin, 1964; Astin, 1997; Bogan Eaton &

Bean, 1995; Irvine, 1966; LeSure-Lester, 2004; Oseguera, 2006; Panos & Astin, 1968; Peng & Fetters, 1978; Schmid & Reed, 1966).

Strengths. Linear regression, specifically multiple linear regression, has a number of advantages. First, it allows for the ability to predict a continuous outcome variable from multiple predictors that may themselves be related (correlated). This is major improvement over bivariate correlational analysis. Second, it can accommodate nonlinear relationships between the predictors and outcome variables and/or between predictor variables. That is, not all relationships need to be linear. Third, it provides a singular, best-fitting solution (in terms of minimizing the squared differences), alleviating the concern of having feigned optimized solutions (e.g., local maxima in maximum likelihood). Fourth, the actual and predicted values of the outcome are on the same scale, making the latter very easy to interpret. Fifth, linear regression provides a very intuitive measure of model of fit (R^2) that is easy to understand. Sixth, it enables significance testing of the overall model and predictor variables. Lastly, it is a very well-established methodology based on a broad body of applied research. This can easily be taken for granted until one delves into new methodologies that have been sparsely researched and have not developed methods for addressing fundamental methodological issues (e.g., evaluating model fit, checking assumptions, etc.).

Limitations. The main drawbacks to linear regression relate to its assumptions. The technique assumes errors have a mean of zero, are normally distributed, and have homogeneity of variance. This limits the methodology to continuous outcome variables since dichotomous outcomes would violate the assumptions that the error variance is homogenous and normally distributed. This is a major limitation to its use in retention

and graduation studies where the outcome variable is typically dichotomous (e.g., student graduated/did not graduate). Linear regression also assumes the absence of a perfect relationship between predictors, that the relationship between variables has been correctly specified (e.g., using a linear model if the relationship is linear), and the variables are measured without error. Although linear regression can handle various types of relationships between variables, the relationships must follow a known parametric form. Typical linear regression is not suited to handle nonparametric relationships. Lastly, in its most practiced form, multiple linear regression is constrained to a single outcome variable, further limiting its applicability.

Predictive discriminant analysis. Discriminant analysis (DA; Cohen, et al., 2003; Tabachnick & Fidell, 2007) has two main strands: descriptive discriminant analysis (DDA) and predictive discriminant analysis (PDA). This section will focus solely on PDA. PDA has a different lineage than linear regression and stems from analysis of variance techniques. More specifically, PDA is a mirror image of multivariate analysis of variance (MANOVA). In MANOVA the goal is to evaluate whether a multivariate set of continuous variables significantly varies between categorical groups, e.g., whether the multivariate average of high school GPA, age, and SAT scores significantly vary between students who graduated from college and those who dropped out. In contrast, PDA aims to predict group membership based on the set of variables (e.g., if we know the multivariate average of high school GPA, age, and SAT scores, can we accurately predict whether a student graduated from college or dropped out). PDA became a popular technique in retention students in the mid-to-late 20th century and has remained so due to its ability to predict non-continuous outcomes (e.g., Bers & Smith, 1991; Bianchi &

Bean, 1980; Campbell & Fuqua, 2009; Finnegan, Morris, & Lee, 2008; Folly Nicpon, Huser, Hull Blanks, Sollenberger, Befort, Robinson Kurpius, 2006; Halpin, 1990; Krotseng, 1992; Pascarella & Terenzini, 1980; Robinson, 1969; Welsh, Petrosko, & Taylor, 2006).

Strengths. Perhaps the greatest benefit of PDA is that it permits the use of a set of continuous variables to classify students into two or more groups. This is a distinct advantage over linear regression which is designed to predict continuous outcome values and not group membership. Another advantage of PDA is that it is a useful byproduct of MANOVA. Similar to linear regression, analysis of variance techniques have a long, well-explored history. Those who are familiar with MANOVA can readily apply PDA with a minimized learning curve. Lastly, PDA uses priors to adjust for known or hypothesized differences in frequencies of group membership in the population. This is important in cases where a much higher proportion of students are believed to be in one group versus another.

Limitations. A major shortcoming of PDA is that, like linear regression, it is limited by some onerous assumptions. The first is that the relationship between predictor variables is multivariate normal. This excludes the use of non-continuous variables as predictors, which is a significant limitation since persistence and retention studies typically include continuous (e.g., high school GPA) and nominal variables (e.g., gender). Another assumption is homogeneity of the within-group covariance matrix across groups. This assumption is difficult to meet in practice (Cohen et al., 2003). Together these two assumptions dramatically diminish the utility of PDA compared to more flexible approaches like logistic regression, which is discussed next. A final notable drawback of

PDA is that the machinery of how the classification scores are estimated is more opaque and less intuitive than the estimation techniques employed by correlational analysis and linear regression.

Logistic regression. Logistic regression is used for modeling the relationship between a discrete outcome variable and one or more predictor variables that are either continuous or discrete. Binomial logistic regression is when the outcome variable has two categories (e.g., retained, not retained) and multinomial regression is when it has more than two categories (e.g., persisted, transferred, dropped out). Logistic regression is typically used to predict the probability a student will or will not persist/be retained based on one or more predictors. It is arguably the most popular approach for modeling retention and persistence (e.g., Andrieu & St. John, 1993; Araque, Roldan, & Salguero, 2009; Caison, 2005; Caison, 2007; Crisp & Nora, 2010; Delen, 2011; Feldman, 1993; Glynn, Sauer, & Miller, 2005; Hagedorn, Maxwell, & Hampton, 2001; Kuh, Cruce, Shoup, Kinzie, & Gonyea, 2008; Lowe & Rhodes, 2012; Perrine, 2009; Rohr, 2013; Sutton & Nora, 2008; Swenson Goguen, Hiester, & Nordstrom, 2010; Wang, 2009).

At first blush, it may seem like the task of predicting persistence or retention could be handled by standard linear regression since the outcome variable of interest, the probability of an outcome occurring, is continuous. The shortcoming of this approach is that we do not actually observe the probability of a student persisting or not. All we observe is a binary outcome of whether a student did or did not persist. The benefit of logistic regression is that it is able to estimate the probability for each value of the outcome occurring. This is accomplished by modeling the logit of an outcome occurring rather than its probability. The logit is the natural log of the predicted probability of an

outcome occurring divided by the predicted probability of it not occurring. This ratio is referred to as the odds of an outcome occurring. The logit can be also expressed as a linear function of the predictor variables:

$$ln(\widehat{odds}) = ln\left(\frac{\hat{P}(outcome)}{1-\hat{P}(outcome)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots \beta_k x_{ki} \quad (2.1)$$

This permits applying the logic and conceptual understanding of linear regression to logistic regression. For instance, in this form, β_0, is the mean logit value when the predictor variables are all zero, and the coefficient, β_k, is the expected logit change in the outcome variable for each one-unit change in x_k, holding all other predictor variables constant. Once the logit values are estimated they can be converted to a predicted probability using the following equation (one of several equivalent equations):

$$\hat{P}(outcome) = \frac{e^{\beta_0+\beta_1 x_{1i}+\cdots\beta_k x_{ki}}}{1+e^{\beta_0+\beta_1 x_{1i}+\cdots\beta_k x_{ki}}} \quad (2.2)$$

Similar to linear regression, the model fit and statistical significance of the overall model and the coefficients can be tested. However, in contrast to linear regression, there is no single accepted measure of model fit in logistic regression. Instead a variety of methods have been developed, ranging from pseudo-R^2 metrics (e.g., R_L^2, Cox & Snell index, Nagelkerke index) to likelihood ratio significance tests. The reader is directed to Hosmer and Lemeshow (2000) and Cohen et al. (2003) for a full discussion on logistic regression model goodness-of-fit techniques.

Strengths. The main benefit of logistic regression is that it provides a mechanism for predicting—in probabilistic terms—a categorical outcome variable based on a set of predictors that can be either continuous or categorical. This overcomes key weaknesses of correlational analysis (limited to simple non-directional models), linear regression

(limited to continuous outcome variables) and discriminant analysis (limited to continuous predictors). Comparing all four techniques, logistic regression, in general, is best equipped to model the most common student retention and success prediction needs – those that involve a categorical outcome variable and a set of predictor variables of various measurement levels. That said, discriminant analysis has been shown to be have more power than logistic regression when its assumptions are met (Cohen et al., 2003, p. 485). This is also believed to be the case with linear regression (Tabachnick & Fidell, 2007, p. 441). Another strength of logistic regression is that it is conceptually similar to linear regression. Learning logistic regression is a natural progression from linear regression that does not require learning a new philosophical or conceptual framework (the two also fall under the broader Generalized Linear Model framework, see Hilbe, n.d., for a brief overview). Analysts and researchers use what they know. The greater the degree of separation is between what they already know and what they need to learn to use a new technique, the less apt they are to learn and use it. This is in line with learning theories that view effective learning as being moderated by the extent to which new knowledge can be linked to prior knowledge (Driscoll, 2004). Another non-technical benefit of logistic regression is that it is available in SPSS (as an add-on module). It is my belief that the single best determinant of the broad adoption and use of a statistical technique in educational research is whether or not the technique is available in SPSS. SPSS is more than a software program; it is a cultural phenomenon within the field that sets the lens through which many non-methodologist researchers view the world. This does not distinguish logistic regression from correlation analyses, discriminant analysis, and linear regression, all of which are also available in SPSS. However, this does

distinguish it from more novel methods used to predict student persistence or graduation, such as Bayesian networks.

Limitations. Perhaps the greatest limitation of logistic regression is the fact that it can be specified in so many forms, leading to confusion on various aspects and terms related to the technique (Peng & So, 2002). Aside from sports bookies, few people think in terms of odds ratios. An odds ratio represents a distinct concept but is close enough to common notions of probability to cause confusion. The concept of a logit also causes confusion to non-mathematicians due to its abstract nature. In comparison, linear regression provides an intuitive result that is on the same scale as the original outcome variable. Logistic regression also suffers from the absence of a singular method for assessing model fit. This is an instance where logistic regression's similarity to linear regression is a limitation. Those familiar with linear regression have the proclivity to misinterpret pseudo R^2 metrics as measure of variance accounted (Peng & So, 2002). Even the most robust method is rendered detrimental if its results are consistently misinterpreted.

An Overview of Bayesian Analysis

In comparison to the four techniques discussed in the prior section, Bayesian-based analyses (Gill, 2009; Kruschke, 2011; Winkler, 1972) have rarely been used to model student success. This is despite the fact that Bayesian analyses are well suited for the task. They provide an efficient and intuitive mechanism for making inferences about the complex world around us in the absence of having all the information needed to make a decision with 100% accuracy. In short, they help us reason in the face of uncertainty. For example, let us assume a college freshman arrives at college and her academic

advisor must determine if she is prepared to take a college-level algebra course. It is impossible for the advisor to know whether or not the student would successfully pass the course since the event has yet to occur. Instead, the advisor must make a probabilistic judgment about whether or not he thinks the student will be successful based on the imperfect and incomplete data available to him (e.g., the student's high school transcripts, placement test scores, results of his conversation with the student, his prior experience with similar students, etc.). In both cases, Bayesian analyses, specifically Bayesian networks (BNs; Pearl, 2000; Pearl & Russell, 2001), provide a mechanism for leveraging human expertise, prior experiences (e.g., the advisor's expertise) and the empirical data that are available (e.g., student's placement test scores) to make probabilistic inferences about an outcome (e.g., the probability the student will be successful in a college-level algebra course). Before discussing the specifics of BNs, it is important to provide an overview of two fundamental topics that underlie the proposed use of Bayesian analysis: subjective probability and Bayesian inference.

Subjective probability. There are two aspects of probability: its properties and its interpretation. The fundamental properties of probabilities are widely agreed upon and are summarized by the Kolmogorov axioms (Gill, 2009). There is far less agreement on the *interpretation* of probabilities; that is, what probabilities mean. Although there are a variety of interpretations (Hájek, 2012), the two most common perspectives in the social sciences (and many other fields) are the frequentist perspective and the subjective or personal perspective (Winkler, 1972). The frequentist perspective views probability as an almost law-like attribute that quantifies the likelihood of an outcome of an event occurring (e.g., the probability that a flipped fair coin will land on heads). It is viewed as

an unknown fixed value that is defined by the limiting relative frequency of an event occurring over repeated trials under identical conditions, and estimated as the relative frequency of the event occurring over a large number of such trials. As the number of trials increases, the difference between the estimated and true values of the probability shrinks. The two probabilities are theorized to eventually converge when the number of trials – conducted under identical conditions – reaches infinity. The true probability is, of course, an unobtainable theoretical limit that must be estimated using asymptotic assumptions. This is a relatively minor limitation given that the use of asymptotic assumptions is a standard, well-accepted practice in statistics (e.g., Weiss, 2008). A far greater limitation of the frequentist perspective of probability is its reliance on the assumption that the repeated trials must be conducted under identical conditions. This assumption is tenable in certain situations when near identical replications are possible (e.g., the probability of a defective widget being produced using a highly-calibrated, automated manufacturing process) but is not realistic in many other situations that are impossible to replicate (e.g., the probability of a given presidential candidate winning a specific election). Due to this limitation, Winkler (1972) called the frequentist interpretation of probability conceptual but not operational.

The second perspective is a subjective or personal interpretation of probability. This perspective views probability as one's degree of belief in the outcome of an event occurring. The subjective probability does not rely on the onerous assumption that all trials must be completed under identical conditions nor does it view probability as a fixed, unknown property of the world. Instead it views probability as an analyst's subjective judgment of an event occurring based on prior knowledge and the data

available. To contrast the two perspectives, the frequentist interpretation of probability views the analyst as an automaton completely removed from the process of determining the probability of an outcome; whereas subjective probability views the analyst—her beliefs, prior experiences, expertise, knowledge of prior research, knowledge of the specific context in which the event will occur, etc.—as an active agent in the process. It is important to note that the injection of subjective judgment does not mean subjective probabilities are uniformed or capricious. In reality, an analyst is never fully removed from the process, as has been well articulated in qualitative research (Bogdan & Knopp Biklen, 2007). As noted by Denzin and Lincoln (2000), "all research is interpretive; it is guided by a set of beliefs and feelings about the world and how it should be understood and studied" (p. 19). The fundamental difference then between the two perspectives is that the subjective perspective makes these beliefs explicit, while the frequentist perspective does not. Another critical difference between these two interpretations is that the subjective interpretation is not burdened by the assumption of repeated trials under identical conditions. As a result, the subjective interpretation is both conceptual and operational, to use Winkler's terms, making it more applicable for real-world use when such an assumption is unreasonable, such as when modeling student success.

The Bayesian analyses used in this study, including the BNs, are based on the latter interpretation of probability. Prior subjective beliefs related to the probability of an outcome occurring are explicitly modeled as part of the mechanics of Bayes' theorem and inference. The process for accomplishing this is discussed next.

Bayesian inference. Bayesian inference allows us to reason under uncertainty through the use of Bayes' theorem. On a conceptual level, Bayes' theorem provides a

process for inferring the probability of an unknown event based primarily on two pieces of known information: (1) prior beliefs on the probability of the event occurring, and (2) the probability of the data that are known given the unknown event. By leveraging both pieces of information, Bayes' theorem allows us to translate what we know into a probabilistic judgment about what we do not know. The theorem is expressed as follows (one of several equivalent forms, see Gill, 2009 pp. 10-11for a complete proof):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

Although it is easier to initially think about probabilities in terms of a single outcome of a specific event (e.g., probability of graduating), in actuality, Bayes' theorem deals with probability *distributions*. A probability distribution is a function that expresses the probability for each mutually exclusive outcome of an event occurring. Building on our example, let us define *A* as a discrete random variable, "freshmen performance in a college-level math course." We define the sample space (all possible, mutually exclusive outcomes) of the variable as {A_1 = passed with a grade of C or better, A_2 = did not pass with a grade C or better}. Let *B* be defined as a discrete random variable, "student's recommended course based on an incoming placement test score" with a sample space of {B_1 = placed in a college-level algebra course, B_2 = did not place in a college-level algebra course}. With that in mind, let us deconstruct Bayes' theorem.

Starting with the right side of the equation, *P*(*A*) is known as the prior distribution. It represents the analyst's belief of what the probability mass function (for discrete random variables) or probability density function (for continuous random variables) is for the sample space of *A* before a specific event occurs. In other words, it is the marginal distribution for *A.* In our example, this represents the advisor's degree of

belief in what the probabilities are of a freshman passing (A_1) or not passing (A_2) a college-level algebra with a C or better without knowing anything about the student's placement level (B). Prior distributions can take many forms (Gill, 2009). Conjugate priors follow a known distribution and, when paired with a specific likelihood function, produce a posterior distribution that is from the same family of distributions as the prior. They ensure the posterior distribution can be analytically calculated (which was essential to making Bayesian analysis tractable prior to the advent of modern computing and simulation techniques). Uninformed prior distributions are used when the analyst wants to minimize or neutralize the effect of the prior. For example, if the advisor has very little prior knowledge about what the probabilities are of a freshman passing or not passing a college-level algebra course, he may set the prior distribution to be uninform, making each outcome equally likely. Informed prior distributions are based on preexisting knowledge or research relevant to the variable(s) under study (e.g., expert opinions, prior published research, analyst's own experience, etc.). Hybrid prior distributions are a combination of informed and uniformed inputs. Our example uses informed priors based on the advisor's previous experience working with numerous similar students.

Turing back to the equation for Bayes' theorem, the prior distribution is multiplied by the conditional probability of B given A, $P(B|A)$, to form the numerator on the right side of the equation. In our example, this is the probability a student placed/did not place in a college-level algebra course given whether the student passed/did not pass a college-level algebra course. To restate this in terms of a particular outcome, it tells us the probability a student placed into a college-level algebra course given she passed a college-level algebra course. This is admittedly counterintuitive to think about the

probability of an earlier event (placement test results) given the results of a later event (success in algebra course). As a quick look ahead, this illustrates a key point that the direction of the arrows in a Bayesian Networks do not need to represent causal or chronological relationships between variables, only dependencies. Once *B* is known, $P(B|A)$ can be re-expressed as the likelihood of *A* given *B*, $L(A|B)$ (Box & Tiao, 1992; Levy, 2011). The likelihood can be broadly interpreted as the most likely value for *A* given *B*. Although the likelihood $L(A|B)$ is not a true probability because it does not sum or integrate to one, it is proportional to $P(B|A)$ (Winkler, 1972). The likelihood function plays a critical role by specifying the distributional function believed to govern the observed data.

The denominator on the right side of the theorem is the marginal probability of *B*, $P(B)$. In our example, this is the probability of a student placing (B_1) or not placing (B_2) in a college-level algebra course based on a placement test score. It serves as a normalizing constant to ensure the posterior distribution sums (or integrates) to one. The product of the prior distribution, $P(A)$, and the likelihood function, $L(A|B)$, divided by the marginal probability, $P(B)$, produces the posterior distribution, $P(A|B)$, displayed on the left side of the equation. The posterior distribution represents the conditional probability of *A* given *B*. In our example, this is the probability the student will pass or not pass a college-level algebra course given her placement test results.

A limitation of the use of Bayes' theorem is that calculating the posterior distribution can be intractable for even relatively simple models for several reasons. The first is that the marginal distribution of *B* can be difficult to calculate, especially if it relates to a continuous random variable that must be integrated. As noted above, $P(B)$ is

typically known. When this is the case, it becomes a constant and can be dropped from the equation. This increases the computational efficiency of Bayes theorem. In most cases, the normalizing constant can be reintroduced later to properly scale the posterior distribution (Gill, 2009). Removing $P(B)$ produces the following expression:

$$P(A|B) \propto L(A|B)P(A) \tag{2.4}$$

This states that the probability of A given B is proportional to the product of the likelihood function and the prior distribution. This reduces the computational burden of estimating the posterior distribution, as will be discussed later. The second limitation is number of joint probabilities that have to be specified due to the fact that the number of probability statements in a joint distribution grows exponentially as the number of variables increases. For example, a joint distribution for n binary variables requires the estimation of 2^n probabilities. A model that contains 30 discrete variables would require the estimation of 1,073,741,824 probabilities – a sizable number for even modern computing power. The third limitation relates to the difficulty in calculating the posterior distribution for models of nontrivial size. All three of these limitations are addressed with the use of BNs, which will be discussed in the next section.

Bayesian networks. Now that the conceptual and mathematical underpinnings have been laid out, it is time to formally discuss Bayesian networks (also known as Bayesian inference networks or Bayesian belief networks). A Bayesian network is a graphical representation of dependent relationships between variables. Pearl (2000, p. 13) articulated the following purposes of BNs:

1. To provide convenient means of expressing substantive assumptions;
2. To facilitate economical representation of joint probability functions; an

3. To facilitate efficient inferences from observations.

Each purpose will be discussed in turn in the subsequent sections.

Directed acyclic graphs: A convenient means of expressing substantive assumptions. BNs provide a visual representation of the relationships between variables relevant to a specific line of inquiry. This is conceptually similar to other methods of visually representing a statistical model and the relationship of its underlying variables, such as structured equation modeling and path diagrams (Kline, 2011). BNs are depicted as directed acyclic graphs (DAGs)[2]. DAGs are comprised of unidirectional edges (arrows) and nodes. The nodes represent either observable or latent variables. The edges represent dependencies (i.e., joint probabilities) between nodes. The relationships do not need be specified as casual (*A* causes *B*) or temporal (*A* precedes *B*), although it can be beneficial to do so (Pearl, 2000; Taroni, Aitken, Garbolino, & Biedermann, 2006).

If an arrow is drawn from one node, *A*, to another node, *B*, *A* is referred to as a parent of *B* and *B* its child. If an arrow is drawn from *B* to another node, *C*, *A* would be the ancestor of *C* and *C* the descendant of *A*. A node without any edges entering it is referred to as a root node. By definition, a BN must be acyclic and not include any feedback loops. That is, a node cannot be both its own ancestor and its own descendant (Taroni, Aitken, Garbolino, & Biedermann, 2006). A simple BN is presented in figure 1.

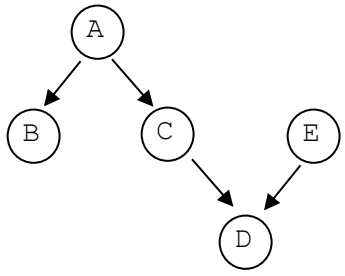


Figure 1. *A Bayesian network depicted as a directed acyclic graph (DAG).*

Similar to other modeling techniques, the decision of what variables to include in a BN and how to structure the connections between the variables in the DAG can be informed by many sources relevant to the research question(s) being investigated (e.g., theory, prior research, reasoned thought); however, there is a scarcity of literature on exploratory Bayesian network modeling building since the networks are typically used as confirmatory models. The exception is the data mining practice of "learning" the structure of a BN directly from the observed data using various algorithms (see Neapolitan, 2003, for a detailed presentation of BN learning algorithms). Once a BN is constructed it provides an effective means of visually depicting the variables under study (nodes), the nature of their relationships (edges), and a priori beliefs about the strength of those relationships (joint probability tables). Next, we turn to the idea of conditional independence and the critical role it plays in BNs.

Conditional independence: Facilitating economical representation of joint probability function. As noted above, specifying the full joint probability distribution for all variables in a model can quickly become intractable given the fact that the number of possible outcomes in a joint probability distribution increases exponentially. One of the main benefits of BNs is that it provides a method for dramatically reducing the number of probabilities that need to be specified. It does this through the use of conditional independence. Conditional independence states that the probability distributions of two random variables (or sets of variables), $\boldsymbol{X}$ and $\boldsymbol{Y}$, are independent of each other once conditioned on a third variable (or set of variables), $\boldsymbol{Z}$, such that $P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Z})$ $P(\boldsymbol{Y}|\boldsymbol{Z})$ (Mislevy & Gitomer, 1996). This holds true even if the two variables are originally dependent. Utilizing conditional independencies allows us to break the full

joint probability distribution for all the variables in a BN into smaller joint probability distributions containing only a subset of the variables. This reduces the number of probabilities that need to be specified and estimated, simplifying the process of estimating the posterior probability.

As mentioned earlier, the dependencies among the nodes are codified in a DAG. It is relatively easy to identify conditionally dependent and independent relationships among the nodes of a small BN; however, it can become very difficult to identify them as the number of nodes and connections increase. Detection of such relationships is aided by the concept of directional-separation (commonly abbreviated as *d*-separation; Pearl, 2000; Taroni et al., 2006).

Once we know the conditional independencies among the nodes, we are able to identify the subset of local joint probability distributions that need to be specified and later estimated. This is done with the aid of the chain rule and the Markov property. The chain rule states that a joint probability distribution of *n* variables can be expressed as the product of the conditional distributions of all the variables (Mislevy & Gitomer, 1996):

$$P(x_1, \dots, x_n) = \prod_{j=1}^{n} P\left(x_j \middle| x_{j-1}, \dots, x_1\right) \quad (2.5)$$

where x_j equals the j^{th} variable and *n* is the total number of variables. The Markov property specifies that the joint probability of a set of variables can be reduced to the conditional probability of a variable and the variable(s) that immediately precede it (i.e., its parents, ***PA***), if the variable is independent of all other variables in the DAG given its parents (Pearl, 2000).

If the Markov property is satisfied, the chain rule simplifies to:

$$P(x_1, \dots, x_n) = \prod_{j=1}^{n} P\left(x_j \middle| \boldsymbol{PA}(x_j)\right) \quad (2.6)$$

$\boldsymbol{PA}$ = the full set of parents for x_j

This reduces the maximum computational burden of specification from 2^n (for binary nodes) to 2^m, with *m* being the number of variables in the largest local conditional probability distribution, $P(x_j|\boldsymbol{PA}(x_j))$. The savings can be substantial especially for larger models, illustrating the efficiencies afforded by BNs.

Bayesian updating: facilitating efficient inferences from observations. Once the structure of a BN has been articulated and the necessary prior and joint probabilities have been specified, data are incorporated into the model and used to update the posterior probabilities of the nodes with unknown states. The elegance of BNs is that they facilitate bidirectional inferences (Mislevy & Gitomer, 1996; Pearl & Russell, 2001). The structure of the BN allows new information (e.g., observed data) to flow deductively down the arrows, while the use of Bayes' theorem allows new information to flow inductively up the arrows. In short, once new information about the variables is known, BNs provide the ability to make probabilistic inferences about the still unknown values in the model wherever they reside.

In Bayesian analyses in general the actual calculation of the posterior probabilities for continuous variables can be computationally difficult (i.e., "NP hard"; Charniak, 1991) due to the fact the denominator of Bayes' theorem would require integration. A major advantage of a BN is that it facilitates efficient and tractable propagation of new information throughout the network by utilizing *discrete* variables. This means estimating the posterior distribution only requires summation not integration of the denominator of Bayes' theorem. The summation of probabilities is easily handled by modern computer power and by the use of the methods covered in the prior section for breaking down a

large network into more manageable chunks for computational purposes. These techniques, along with the use of Bayes' theorem and conditional independence, make BNs a powerful technique for efficiently making inferences about what is unknown given one's prior beliefs and the observed data. For illustrative purposes, a full example of a Bayesian network is provided in Appendix A.

Dynamic Bayesian networks. Traditional Bayesian networks (as described above) assume the state of the world being modeled remains fixed over time. As new information is collected and propagated throughout the model, the posterior probabilities for the outcomes of an event change, but the structure and underlying probability distributions remain the same. For instance, in our example above, we assume that the relationship between the outcome of a student's placement test and success in a college-level algebra course remains unchanged over time even if the student attempted the placement exam several times. Dynamic Bayesian networks (DBNs) are a special case of BNs that expand Bayesian inference and model building to include changes over time (Dean & Kanazawa, 1988; Reye, 2004). Over the last decade DBNs have gained popularity in the area of educational research and measurement as a method for modeling student proficiency in intelligent tutoring systems (Mayo & Mitrovic, 2001; Millán & Pérez de la Cruz, 2002; Reye, 2004; Zapata-Rivera & Greer, 2004) and educational gaming (Manske & Conati, 2002; Shute, 2011), among other areas.

A DBN is comprised of two main components. The first component is a series of individual local BNs for given slices of time. Figure 2 illustrates a simple DBN. Let *A* be a latent random variable and *B* and *C* be observable random variables. The $B \leftarrow A \rightarrow C$ diverging clusters in the figure represent local BNs. In isolation, each one can be viewed

as a traditional BN subject to all the same conceptual (e.g., bidirectional flow of information) and operational (e.g., *d*-separation) properties previously described. Each local BN represents the hypothesized model structure for a given point in time. As such, the nodes in figure 2 are identified with a subscript, *t*, to identify an increment of time. The unit of measurement does not need to represent chronological time in a literal sense; nor does it need to increase in discrete units (Murphy, 2002). It can represent any unit of change, e.g., days, number of interactions within a tutoring system, number of exam items answered, semesters in college, etc. It should reflect a unit of change that is most relevant to the nature of the research question and real-world phenomena being modeled. Typically, the structure of each local BN is a carbon copy of each other (as depicted in figure 2); however, this is not a requirement (Reye, 2004).

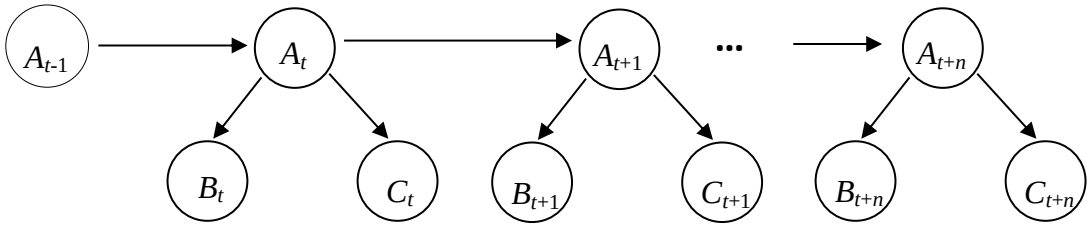


Figure 2. *A simple dynamic Bayesian network.*

The second component of a DBN is a series of nodes that form directed acyclic connections between the local networks over time. These nodes are typically specified as latent (hidden) random variables (Ghanmi, Mahjoub, & Ben Amara, 2011). Reye (2004) referred to these connections between time periods as the "belief net backbone" of a DBN (p. 75). The backbone passes information from one state in time to another. In figure 2, this is represented by the nodes A_{t-1}, A_t, A_{t+1}, A_{t+n}. The first time node in the backbone, A_{t-1}, represents the initial prior belief about the probability distribution of *A*. It can be a

single node or a full local BN. Applying the normal properties of a DAG, the arrow between nodes across time points (e.g., $A_{t-1} \rightarrow A_t$) denotes a dependent relationship between the variables. It is also possible to have cross-time dependencies between nodes in backbone and those in the local BNs. For instance, if we draw an arrow between C_t and A_{t+1}, this would indicate that the future probability distribution of *A* is directly dependent on the current states of both *A* and *C*. DBNs have the flexibility to take a variety of forms. See Ghanmi, Mahjoub, & Ben Amara (2011) and Murphy (2002) for presentations on various DBN structures of varying complexity.

DBNs are updated using a two-step process. The first step is to pass information from time point *t*-1 (*t* = 0 when the previous time point is the prior) to time point *t* for all dependent nodes connected via the backbone. In our example this is achieved by simply conditioning the probability of A_t on A_{t-1} ($P(A_t \mid P(A_{t-1}))$. The Markov property is used for more complex models. More specifically, the probability distribution of a set of backbone variables $\boldsymbol{X}$ at time point *t* conditioned on $\boldsymbol{X}$ at *t*-1, ($P(\boldsymbol{X}_t \mid P(\boldsymbol{X}_{t-1}))$, is equal to $P(\boldsymbol{X}_t \mid \boldsymbol{PA}(\boldsymbol{X}_t))$, where $\boldsymbol{PA}$ represents the parents of X_t. The second step of the process is to propagate information throughout the local BN at time point *t* using Bayes' theorem. For example, the posterior probability of A_t given B_t and C_t is expressed as:

$$P(A_t|B_t, C_t) = \frac{P(B_t,C_t|A_t)P(A_t)}{P(B_t,C_t)} \tag{2.7}$$

$$P(A_t) = P(A_t|P(A_{t-1}))$$

Once the posterior distribution is estimated for A_t, it flows to the next time slice and becomes the prior for A_{t+1}. The model continues to update in this fashion across a finite number of time points as more information is collected.

Discrete-Time Survival Analysis: An Interim Step to Building Bayesian Networks.

As previously mentioned, a major advantage of BNs over traditional techniques for modeling variables of student success is that they provide an effective means of visually depicting the variables under study and their relationships, they allow a priori beliefs about the strength of those relationships including those based on subjective interpretations of probability, and they can quickly provide updated probability estimates via Bayes' theorem once new information is available. Where BNs are less advantageous than traditional analysis is they have been primarily limited in educational research to use as confirmatory models. In contrast to traditional analytic techniques, there is a shortage of literature and established practice in educational research on how to construct BNs in non-confirmatory situations. This severely diminishes their application in situations such as the current study that require exploratory model building. Exploratory BN modeling building techniques are discussed outside the realm of education (e.g., Neapolitan, 2003; van Gerven, Taal, & Lucas; 2008), but they tend to represent approaches and underlying perspectives that are foreign or antithetical to the frequentist, null-hypothesis significance testing perspective espoused by most educational researchers. The current study seeks to provide an example of a model building approach for BNs that scaffolds existing knowledge and practices in educational research with more modern approaches. This is a modest attempt to promote branching between the two perspectives without forcing educational researchers to choose between abandoning the use of BNs in instances when exploratory modeling building is required or the modeling building analytic perspective and practices they are already familiar with.

In pursuit of that goal, discrete-time survival analysis models (Singer & Willett, 2003; also commonly referred to as a hazard model) was used as an exploratory step to inform the construction of the final BN used to predict students' stopping out behavior. Although the use of survival analysis is relatively new in the area of educational research, it has quickly become a widely used longitudinal technique for predicting discrete outcomes of student persistence (Chen & DesJardins, 2008; Gross, Torres, Zerquera, 2013; Ishitani, 2003; Murtaugh, Burns, & Schuster, 1999; Radcliffe, Huesman, Kellogg, 2006; Ronco, 1994; Willett & Singer, 1991), transfer behaviors (Bahr, 2008; Bahr, 2012; Johnson & Muse, 2012), and degree attainment (Calcagno, Crosta, Bailey, & Jenkins, 2007a; Calcagno, Crosta, Bailey, & Jenkins, 2007b; DesJardins, Ahlburg, & McCall, 2002; DesJardins, McCall, Ahlburg, & Moye, 2002; Gross, Torres, Zerquera, 2013; Zwick & Sklar, 2005) among postsecondary students.

The goal of a discrete-time survival analyses is to create a survivor function, $S(t_{ij})$, that estimates the probability a randomly selected student will not experience an event (i.e., will "survive") by a given time point. At its most basic level the estimated values of the survivor function are equal to the proportion of students in the sample at time point *j* who have not experienced the event. For example, if the event is specified as stopping out of college, $S(t_{ij})$ estimates the probability that student *i* will persist (i.e., not stop out) to time point *j*. The survivor function cannot be directly estimated if the sample includes any "censored" data. Generally speaking, censored data are missing dependent variable values for a given student that are assumed to be the result of the study's design and independent of the outcome (see Singer & Willett, 2003, for a more nuanced

definition and discussion of censored longitudinal data). In these instances the survivor function is estimated through the use of the hazard function, $h(t_{ij})$:

$$\hat{S}(t_{ij}) = \hat{S}(t_{ij-1})[1 - \hat{h}(t_{ij})] \quad (2.8)$$

$$\hat{h}(t_{ij}) = \mathrm{P}[T_i = j \mid T_i \geq j, \boldsymbol{X}_{ij} = \boldsymbol{x}] \quad (2.9)$$

where $\hat{h}(t_{ij})$ is the predicted probability that student *i* will experience the event (T_i; e.g., stop out) by time period *j*, conditioned on (1) the fact that student *i* has not already reached the event ($T_i \geq j$) and (2) a set of student-level predictor variable values ($\boldsymbol{X}_{ij} = \boldsymbol{x}$). The predictor variables may include both time-variant (e.g., term GPA) and time-invariant (e.g., gender) variables. $\hat{S}(t_{ij})$ is therefore defined as the proportion of students who survived to time point *j*-1 multiplied by the probability a student will not stop out by time period *j*. The modeling of the hazard function can be accomplished through the use of the logit function:

$$logit\ \hat{h}(t_{ij}) = [\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij}] + [\beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_P x_{Pij}] \quad (2.10)$$

The logit of the hazard function for student *i* at time *j* is equal to the sum of the intercept values, a_n, at each time point plus the sum of the weighted *p* predictor variables for student *i* at time *j*. This generalized specification of a discrete-time survival analysis can be viewed as a longitudinal extension of normal logistic regression. A key advantage of this approach is that it produces results that are easily interpretable to those familiar with standard logistic regression. For example, β_P, represents the change in log odds associated with a one-unit increase in the value of x_p for student *i* at time *j*, holding all other variables constant. Another advantage of the logistic specification of the hazard

function is that it can be easily extended to accommodate a polytomous response variable to model "competing risks" (Scott & Kennedy, 2005; Singer & Willett, 2003).

Prior Research on Potential Predictor Variables

There have been a small number of studies that specifically modeling postsecondary students' stopping out behavior (e.g., DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013); however, it has been far more prevalent for researchers to study the inverse of stopping out – persistence, transfer, and graduation – and the its sister variable, dropping out (at a conceptual level, dropping out is viewed as a student who has permanently left the institution, whereas stopping out is a student's first break in enrollment from which they may or may not return). As such, the literature review primarily centered on prior research related persistence, transfer, and graduation with the belief that knowledge gained from those studies would be directly relevant to the current study since they are inversely related to stopping out behaviors.

Persistence, transfer, and graduation have been arguably the most widely studied topics in the educational literature; so much so that entire tomes have been written on these subjects and related topics (e.g., Braxton, et al., 2014; Habley, Bloom, & Robbins, 2012; Pascarella & Terenzini, 1991; Pascarella & Terenzini, 2005; Seidman, 2012; Tinto, 1994; Tinto, 2012). It is not the purpose of this study to provide yet another voluminous summary of such research. Not only would that be impractical, it would have only a limited benefit on informing which predictor variables to potentially include in the models under study since the significance and meaningfulness of predictor variables are inextricably tied to the context in which they are collected and modeled. The more dissimilar the context is to the current study, the less relevant the results are to informing

it. Ideally, the current study would be informed by research that mirrors its context, e.g., a sample of students at a community college of similar type, with data collected and defined in the same manner, analyzed using the same modeling techniques, etc. In reality, this approach is not practical since it is unlikely to find research with such a degree of similarity. Instead, the decision was made to conduct a review of studies that (a) had postsecondary students (of any type) as the target population and (b) employed discrete-time survival analysis to predict postsecondary persistence, transfer, and/or degree attainment – all of which are inversely related to stopping out and are therefore of interest. Matching on these aspects of the context seemed to present an appropriate compromise given the focus of the study. Studies utilizing BNs were not included because, as previously mentioned, the author was not able to find any such studies except from one (Nandeshaw, Menzies, Nelson, 2011).

A review of prior research that met the criteria discussed above netted a set of variables that have been shown to be significant ($p < .05$) predictors of postsecondary persistence, transfer, and/or graduation. The magnitude, and in some cases direction, of the relationships varied by study and/or by time point within a study; because of this the below summary of research only focuses on the significance of the variables. These results will be used to inform the selection of potential predictor variables to include in the current study. The variable selection process is articulated in more detail in Chapter 3.

Persistence. The following variables were found to be significant predictors of postsecondary persistence: Academic integration (Chen & DesJardins, 2008), age (Murtaugh, Burns, & Schuster, 1999), attendance at a freshmen orientation (Murtaugh, Burns, & Schuster, 1999), college GPA (Chen & DesJardins, 2008; DesJardins, Ahlburg,

& McCall, 2002; Gross, Torres, & Zerquera, 2013; Ishitani, 2003; Johnson, 2006; Murtaugh, Burns, & Schuster, 1999), cost of attendance (Gross, Torres, & Zerquera, 2013), educational expectations (Chen & DesJardins, 2008; Ishitani, 2006), ethnicity (DesJardins et al., 1994; DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013; Ishitani, 2003; Ishitani, 2006; Johnson, 2006; Murtaugh, Burns, & Schuster, 1999; Ronco, 1994), family income (Chen & DesJardins, 2008; DesJardins, Ahlburg, & McCall, 2002; Ishitani, 2003; Ishitani, 2006), entering college within a year of graduating high school (Johnson, 2006), federal work study recipient (Chen & DesJardins, 2008; Johnson, 2006), full-time enrollment (Johnson, 2006), gender (Gross, Torres, & Zerquera, 2013; Ishitani, 2003; Ishitani, 2006), high school GPA (Ishitani, 2003; Murtaugh, Burns, & Schuster, 1999), high school academic intensity (Ishitani, 2006), high school rank (Gross, Torres, & Zerquera, 2013; Ishitani, 2006), income level (Johnson, 2006), institutional type (Ishitani, 2006), living off campus (Gross, Torres, & Zerquera, 2013), number of developmental education credits attempted (Gross, Torres, & Zerquera, 2013), number of college credits attempted (Gross, Torres, & Zerquera, 2013), number of college transfer credits (DesJardins, Ahlburg, & McCall, 2002), parent's educational expectations (Ishitani, 2006), parent's educational level (Chen & DesJardins, 2008; Ishitani, 2003; Ishitani, 2006), residency (Murtaugh, Burns, & Schuster, 1999), SAT/ACT scores (Johnson, 2006), size of hometown (Ishitani, 2003), students identified as needing help with studying (DesJardins, Ahlburg, & McCall, 2002), and type of financial aid received (Chen & DesJardins, 2008; DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013; Ishitani, 2006; Johnson, 2006).

Degree attainment. The following variables were found to be significant predictors of postsecondary degree attainment: ACT scores (Radcliffe, Huesmann, & Kellogg, 2006), age (Calcagno, Crosta, Bailey, & Jenkins, 2007a; Calcagno, Crosta, Bailey, & Jenkins, 2007b), college GPA (DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013), cost of attendance (Gross, Torres, & Zerquera, 2013), enrollment in developmental courses (Calcagno, Crosta, Bailey, & Jenkins, 2007), ethnicity (Calcagno, Crosta, Bailey, & Jenkins, 2007; DesJardins, Ahlburg, & McCall, 2002; Radcliffe, Huesmann, & Kellogg, 2006), gender (Calcagno, Crosta, Bailey, & Jenkins, 2007), family income (DesJardins, Ahlburg, & McCall, 2002), financial aid needs unmet (Radcliffe, Huesmann, & Kellogg, 2006), full-time enrollment (Calcagno, Crosta, Bailey, & Jenkins, 2007), high school rank (Gross, Torres, & Zerquera, 2013), institution type (Gross, Torres, & Zerquera, 2013), living off campus (Radcliffe, Huesmann, & Kellogg, 2006), location of residence (Radcliffe, Huesmann, & Kellogg, 2006), math placement test scores (Calcagno, Crosta, Bailey, & Jenkins, 2007), number of C grades earned in first term (Radcliffe, Huesmann, & Kellogg, 2006), number of college transfer credits (DesJardins, Ahlburg, & McCall, 2002), number of credits attempted (Gross, Torres, & Zerquera, 2013), number of developmental educational courses taken (Gross, Torres, & Zerquera, 2013; Radcliffe, Huesmann, & Kellogg, 2006), number of courses withdrawn from in first term (Radcliffe, Huesmann, & Kellogg, 2006), ratio of completed to attempted credits for first term (Radcliffe, Huesmann, & Kellogg, 2006), students identified needing help with studying (DesJardins, Ahlburg, & McCall, 2002), type of financial aid received (Calcagno, Crosta, Bailey, & Jenkins, 2007;

DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013), and verbal placement test scores (Calcagno, Crosta, Bailey, & Jenkins, 2007).

Transfer. Fewer studies were found that used survival analysis to predict a student's transfer behavior as a stand-alone outcome. Of the applicable articles that were found, age (Scott & Kennedy, 2005), college credit hours earned (Johnson & Muse, 2012), college GPA (Johnson & Muse, 2012), English placement level (Bahr, 2008), ethnicity (Bahr, 2008; Johnson & Muse, 2012), gender (Johnson & Muse, 2012), having met with an advisor (Bahr, 2008), math placement level (Bahr, 2008), participation in a fraternity or sorority (Johnson & Muse, 2012), and residency (Johnson & Muse, 2012) were shown to be significant predictors of whether postsecondary students would transfer to a four-year institution.

Chapter 3

Methods

Participants

The models were developed using three years of previously collected longitudinal data on 1,756 degree- or transfer-seeking students who first enrolled in fall 2009 at a large two-year public institution located in the southwestern United States. A second validation sample was used of 4,859 degree- or transfer-seeking students who first enrolled in fall 2010 at the same participating institution. The difference in the size of the two samples was due to the fact that the first sample only included students who had completed the course placement exams required by the participating institution. This was needed in order to evaluate whether the placement scores and/or levels were significant predictors of stopping out. In contrast, the fall 2010 did not have the same limitation since placement test scores and levels were not needed to validate the final model. As a result, the second sample was treated as if it was drawn from a different population of students than the fall 2009 sample was drawn from. Student privacy was maintained by anonymizing all of the data prior to its use.

Dependent Variables

The current study included a single dichotomous outcome variable with two mutually exclusive events:

Stopped out: A student who did not graduate or transfer prior to a given term nor did he/she enroll in at least one course in that term.

Not stopped out: A student who re-enrolled in at least one course in that term.

If a student graduated or transferred after a given term he or she was removed from the risk set and treated as censored for all future terms.

Longitudinal Time Points

The probability of a student stopping out was calculated for each of the following time points:

- Time 1: Probability of stopping out before the start of the 2nd semester (spring 2010)
- Time 2: Probability of stopping out before the start of the 3rd semester (fall 2010)
- Time 3: Probability of stopping out before the start of the 4th semester (spring 2011)
- Time 4: Probability of stopping out before the start of the 5th semester (fall 2011)
- Time 5: Probability of stopping out before the start of the 6th semester (spring 2012)
- Time 6: Probability of stopping out before the start of the 7th semester (fall 2012)

Predictor Variables

A pool of potential predictor (independent) variables was considered for inclusion in the predictive models. A broad array of variables was selected based on (a) prior research, (b) informed judgment by the author, and (c) availability of data. The final model in this study used a parsimonious subset of the variables. The exact subset of variables that were used was determined based on the results of variable selection methods and substantive considerations. The specific techniques that were used are described below.

Variable selection. The pool of potential variables was constructed by including any variable the author had access to that was either supported by prior research (as described in Chapter 2) or that was reasonably assumed to have a potential impact on a community college student's probability of stopping out. The list was further reduced through the variable selection process described next.

The current study employed traditional statistical modeling and educational data mining variable selection methods in an attempt to produce models of acceptable fit and classification accuracy while using the most parsimonious set of predictor variables. Educational research has historically relied most heavily on prior research and theory to identify variables to serve as independent variables. This stems from the general practice in educational research of specifying the variables of interest in advance of data collection, giving the researcher more direct control to limit the study to the variables of most interest. In contrast, educational data mining relies predominately on empirical techniques to select independent variables from a pool of available data. This follows from the fact that data mining typically involves the use of pre-existing data that were collected as part of every-day practices and not as the result of a research study based on a purposeful design. Since this study exclusively used pre-existing data that were not collected as part of a research design, it utilized data mining and traditional approaches to identify a parsimonious subset of variables from the broader aforementioned list of potential independent variables. That said, the final decision of which variables to include in the models was determined by the researcher based on a synthesis of analytic results, substantive considerations, prior research, sound reasoning, and interpretability. In other words, the variable selection process was data *informed* and not solely

determined by the data. The researcher served as an active agent and final arbiter of the selection process, not a detached button-pusher driven solely by the data. As such, on a philosophical level the study adopted a Bayeisan paradigm of reasoning that readily paired observed data with subjective judgments to inform an outcome.

Variable selection methods in educational data mining can be broadly classified into two categories: variable ranking and subset selection (Kantardzic, 2011; Nisbet, Elder, & Miner, 2009). Variable ranking consists of ranking each variable on one or more metrics and using a threshold to decide which variables to include in future analyses. For example, calculating the bivariate correlation between each independent variable and the dependent variable, and including only variables that have a correlation above a pre-specified threshold. The current study pre-screened potential predictor variables using the variable ranking methods described below. Subset selection methods involve search algorithms that iterate over the entire pool of potential variables in order to identify a set that collectively best predict the outcome variable. For this study subset selection was accomplished using forward stepwise regression (Cohen, Cohen, West, Aiken, 2003). This technique is outlined in detail in subsequent sections as part of the discrete-time survival analysis modeling building process.

***Variable ranking techniques*.** First, bivariate correlations were constructed between each potential predictor variable and the outcome variable. Predictor variables that had a non-significant correlation or a correlation weaker than ± .20 were considered for exclusion from the analysis. Although ±.20 is an arbitrary threshold, it seemed reasonable to consider excluding any variable that shares less than 5% of its variance with the outcome variable. Second, bivariate correlations were calculated between each

pair of predictor variables. If a correlation was stronger than ± .90, the variable that had the highest correlation with the dependent variables was retained and the other variable was removed from future analyses to eliminate concerns related to multicollinearity.

Missing data. Although there are a variety of accepted techniques for imputing missing data (Enders, 2010), applying such techniques went beyond the scope of the current study. For the nominal variables, missing data were recoded into an "Unknown" category. This was a pragmatic decision to prevent a sizable loss of data due to the compounding effect of listwise deletion across all the variables. This also allowed the researcher to evaluate whether the Unknown category had a substantive relationship with any of the dependent variables. Any continuous predictor variable with greater than 10% missing data was removed from the analyses.

Initial pool of potential predictor variables. The variable selection process just described was applied to the below list of potential predictor variables. As previously mentioned, the initial pool of potential variables was selected based on prior research, informed judgment by the author, and availability of data. The variables are divided into two categories. The first category is *time-invariant* variables that are known at time point zero (prior to the start of Fall 2009) and remained fixed from that point on. The second category is *time-variant* variables whose values may change from one time period to the next (e.g., the grade point average for a specific term). Both categories include variables that may have high degrees of multicollinearity and/or represent the same construct only measured at different levels. For example, the ratio of credits successfully attempted and GPA are expected to have a high degree of multicollinearity. In such instances the intent was to identify the single variable that best serves as predictor of the outcome and

remove the redundant or overlapping variables. The measurement level of each variable is provided in parentheses.

Time invariant variables.

Age (ratio): The student's age derived from his/her date of birth.

Anticipated work hours (nominal): The student's self-reported number of hours he/she anticipates working while in school.

Developmental English placement (dichotomous): Indicates whether the student placed into a developmental (below college level) English course based on his/her English placement exam score.

Developmental math placement (dichotomous): Indicates whether the student placed into a developmental (below college level) math course based on his/her math placement exam score.

Developmental reading placement (dichotomous): Indicates whether the student placed into a developmental (below college level) reading course based on his/her reading placement exam score.

English placement level (ordinal): Indicates the student's English course placement rank based on his/her English placement exam score.

First generation status (dichotomous): Indicates whether the student self-reported as being a first-generation student (neither parent earned a bachelor's degree college).

Gender (dichotomous): The student's self-reported gender.

High school graduation status (dichotomous): Indicates whether the student graduated from high school.

Language now (dichotomous): Indicates whether English is the language the student spoke most often at the time the data were collected.

Math placement level (ordinal): Indicates the student's math course placement rank based on his/her math placement exam score.

Median household income (ratio): A proxy of the student's estimated median household income using U.S. Census 2011 inflation-adjusted dollars by zip code.

Needs assistance with childcare information (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with childcare information.

Needs assistance choosing a major or career (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with choosing a major or career.

Needs assistance with commuter information (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with commuter information.

Needs assistance learning English (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with learning English.

Needs assistance with financial aid (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with financial aid.

Needs assistance with finding work (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance finding work.

Needs assistance with math skills (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with math skills.

Needs assistance with mentoring (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with mentoring.

Needs assistance with reading skills (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with reading skills.

Needs assistance with study skills (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with study skills.

Needs assistance with writing skills (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with writing skills.

Needs assistance with work experience credit (dichotomous): Indicates whether the student self-reported during the time of registration as needing assistance with work experience credit.

Number of developmental subjects placed into (interval): Indicates how many developmental subjects (English, math, reading) a student placed into. The values ranged from zero to three.

Race/ethnicity (nominal): The student's self-reported race/ethnicity.

Reading placement level (ordinal): The student's reading course placement rank based on his/her reading placement exam score.

U.S. military or dependent: Indicates whether the student self-reported as being a current or former member of the U.S. Armed Forces or a dependent of a current or former member of the U.S. Armed Forces.

Time variant variables.

Attempted credit hours (ratio): The number of credit hours the student attempted in a given term.

Courses dropped (ratio): The number of courses dropped by the student in a given term.

Courses successfully completed (ratio): The number of courses the student successfully passed (earned a C or equivalent or better) in a given term.

Courses withdrawn (ratio): The number of courses withdrawn by the student in a given term.

Cumulative attempted credit hours (ratio): The cumulative number of credit hours the student attempted as of a given term.

Cumulative courses attempted (ratio): The cumulative number of courses attempted by the student as of a given term.

Cumulative courses dropped (ratio): The cumulative number of courses dropped by the student as of a given term.

Cumulative courses withdrawn (ratio): The cumulative number of courses withdrawn by the student as of a given term.

Cumulative earned credit hours (ratio): The cumulative number of credit hours earned by the student as of a given term.

Cumulative grade point average (ratio): The student's cumulative grade point average as of a given term.

Cumulative Pell grant amount (ratio): The cumulative amount of Pell grant funds the student received while attending the institution as of a given term.

Cumulative ratio of earned to attempted credits (ratio): The cumulative ratio of the number of credit hours the student has earned divided by the number of credit hours attempted as of a given term.

Cumulative ratio of earned to attempted developmental credits (ratio): The cumulative ratio of the number of developmental credit hours the student has earned divided by the number of developmental credit hours attempted as of a given term.

Cumulative student loan amount (ratio): The cumulative amount of federal student loan funds the student received while attending the institution as of a given term.

Full-time student (dichotomous)*:* Indicates whether the student was enrolled in 12 or more credit hours for a given term.

Number of days registered prior to start of first course (ratio): Indicates how many days the student registered for courses prior to the start of their first course for a given term.

Pell grant amount (ratio): The amount of Pell grant funds the student received for a given term.

Pell grant received (dichotomous): Indicates whether the student received Pell grant funds for a given term.

Primary time of attendance (nominal): Indicates whether the student primarily attended courses during the day, evening, or other for a given term.

Primary time of attendance (dichotomized): Indicates whether the student primarily attended courses during the day during a given term.

Ratio of earned to attempted credits (ratio): The ratio of the number of credit hours the student has earned divided by the number of credit hours attempted for a given term.

Received need-based aid: Indicates whether the student received a Pell grant or federal subsidized loan during the prior term.

Student loan amount (ratio): The amount of federal student loan funds the student received for a given term.

Student loan received (dichotomous): Indicates whether the student received a federal student loan for a given term.

Success rate in developmental courses (ratio): The proportion of developmental courses the student successfully passed (earned a C or equivalent or better) in a given term.

Term grade point average (ratio): The student's grade point average for a given term.

Preliminary Exploratory Data Analyses

The distribution of responses for each variable that passed the aforementioned variable ranking screening process was examined using univariate visual depictions (e.g., histograms, box plots, normal probability plots) and descriptive statistics (means, standard deviations, skewness, and kurtosis). The cell frequencies of each nominal

variable were examined. Sparsely populated response categories were collapsed into another response category if there was substantive justification to do so. If there was not sufficient justification, the responses related to the sparsely populated cells were recorded as Unknown. The frequency distributions of continuous variables were checked for potential nonsensical responses (e.g., an age > 100). If any nonsensical responses were found and they appeared to be isolated to a specific variable they were recoded as Unknown and presumed to be data entry errors in the underlying student information system the data were extracted from. If there were nonsensical results across more than one variable for a given student the case would have been removed. There were no such instances in the data set. For each independent variable, the pattern of correlations with it and each dependent variable (time point) was examined to evaluate whether a variable should potentially be analyzed as an interaction with time. For time-varying predictors, the correlation was examined between the lagged instance of the variable and the respective outcome. For example, the correlation was examined between the outcome for fall 2011 and student GPA for the prior term (spring 2011).

Model Building Phase 1: Discrete-Time Survival Analysis

Following the preliminary data analysis, a binomial discrete-time survival analysis (Singer & Willett, 2003) was conducted. As previously described, the discrete-time survival analysis provides an established model building framework for informing the development of the BN.

Life table. Prior to conducting a survival analysis a life table was constructed. The life table is a staple of survival analysis. It summarizes the number of students who were in the risk set at the start of a given time period, the estimated proportion that would

stop out (hazard function) during the time period and the estimated proportion that would not stop out (survivor function) at the end of the time period. The life table does not take into account any potential covariates/predictors. The life table is also useful for calculating the median lifetime. This is the number of time periods the model estimates would need to eclipse before 50% of students in the entire sample would experience the event (i.e., stop out).

Model building. The current study utilized an amalgamation of automated data mining subset selection techniques (stepwise regression), the model building approach articulated by Hosmer and Lemeshow (2000) for standard logistic regression, and the model building approach for discrete-time survival analysis specified by Singer and Willett (2003). To aid in the analysis, the data file was set up in "person-period" format in which there was a unique row of data for each time point for each student. Students who graduated or transferred before the start of a given term were treated as censored data and were excluded from the risk set for the following time period. The data were analyzed using SPSS. The model building approach consisted of the following steps:

Step 1: To further whittle down the number of predictor variables for each time period, a forward stepwise logistic regression (Nisbet, Elder, & Miner, 2009) was conducted for all the potential predictor variables and the outcome variable for each time point using half of the students randomly selected from the entire sample. The forward stepwise regression utilized the likelihood ratio test (p-in < .05, p-out < .10). The process was then repeated using the remaining half of the sample. The results of both analyses were then used along with the researcher's judgment to determine which variables to

include in the final exploratory model for each time point. The parameters for the final exploratory models were then estimated using the entire sample.

Step 2: The results from the individual logistic regression analyses along with informed judgment by the researcher were used to select the predictors to include in the survival analysis that simultaneously analyzed the probability of each a student stopping out at each time point. Any variable that was a significant ($p < .05$) predictor of stopping out for two or more models from Step 1 was included in the survival analysis. The researcher also considered the inclusion of any variable that was of substantive interest even if it was not shown to be a significant predictor in two or more of the individual logistic regression models. The resulting survival analysis model was considered the preliminary main effects model.

Step 3: The assumption that each continuous variable is linearly related to the logit was checked for all applicable variables in the preliminary main effects models. This was done by adding to the model an interaction term of the variable multiplied by the natural log of itself and conducting a nested χ^2 likelihood ratio test (Tabachnick & Fidell, 2007). A significant difference ($p < .05$) between the two models indicated that the assumption was not supported and that other specifications (e.g., quadratic) should be considered. The resulting model was considered the final main effects model.

Step 4: The final main effects model included a general nonparametric specification of time. For the given model, this materialized as each time point being specified as a time-dependent intercept. Other specifications of time were also evaluated. Specifically, constant, linear, quadratic, and cubic specifications of time. Nested χ^2 goodness-of-fit tests were conducted to see if a more complex specification of time

significantly increased model fit. It was known a priori that the general specification of time would be the best fitting model and the constant specification of time would be the worst fitting model. These models served as the ceiling and floor to comparatively evaluate the remaining specifications of time. The question being pursued was whether the linear, quadratic, or cubic specifications offered a material improvement above the constant-only model without having markedly worse model fit compared to the model with a general specification of time.

Step 5: Once the continuous variables and time were correctly specified, all substantively meaningful interactions with time were examined to see if they were significant ($p < .05$) using a nested χ^2 likelihood ratio test. Any significant interactions would be considered for inclusion in the full model. The result model was considered the final model.

Step 6: The fit and classification accuracy of the final model were estimated using the full sample of students as described below. The outcome with the highest predicted probability was considered the predicted outcome for each student.

Model fit. The fit of the final model was evaluated using the χ^2 goodness-of-fit test, Hosmer and Lemeshow test, and the Cox and Snell and Nagelkerke pseudo R^2 statistics. A significant χ^2 goodness-of-fit test ($p < .05$) indicates that the full model fits the data better than an intercept-only model. Although this is a commonly used model fit metric, it is a low bar to clear. The Hosmer and Lemeshow test divides the rank-ordered predicted probabilities into g groups. A g x 2 (since the outcome variable is dichotomous) contingency table is constructed. A significance test is conducted on the null hypothesis that the observed and expected frequencies are equal across the groups. In other words,

that the model-produced frequency estimates are equal to the observed frequencies. A non-significant result provides support that the model fits the data. The pseudo R^2 metrics are measures of effect size, similar to R^2 in linear regression but do not indicate proportion of variance accounted for. Cox and Snell is unbound while Nagelkerke is a rescaled version of Cox and Snell that ranges from zero to one. Larger values indicate better data-model fit but there is no broadly accepted threshold on what a "good" value is. This makes the metrics difficult to interpret.

Classification accuracy. The classification accuracy of the models was examined using the overall, marginal classification rates, and adjusted overall (i.e., Cohen's κ; Cohen, 1960) classification rates. The area under the Receiver Operating Characteristic (ROC) curve was also calculated for the survival analysis model. For survival analysis models such as those examined in this phase of the study, the classification accuracy rates represent that aggregate accuracy of the model across all time points. For each time point, it is an evaluation of how well the model predicted the outcomes for students in the risk set at the start of that time point. The adjusted classification rates, Cohen's κ (Cohen, 1960), are interpreted as the percent of cases that are correctly classified over and above what would be expected based on "chance" alone. Specifically, Cohen's κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

where p_o represents the proportion of times the observed and model-predicted outcomes were in agreement, and $p_e$is the proportion of times we would expect the observed and predicted values to be in agreement simply by chance. Cohen (1960) defined chance as the joint probabilities of the marginal distributions of each category:

$$p_e = p_{observed\ outcome} * p_{predicted\ outcome} \quad (3.2)$$

where $p_{observed\ outcome}$ is the proportion of times an outcome (e.g., stopping out) occurred in the observed data, and $p_{predicted\ outcome}$ is the proportion of times an outcome occurred in the predicted data. This provides a baseline measure of agreement that is then used to compare to the level of agreement produced by the model. A Cohen's κ value of > .50 was seen as a desirable outcome for this study since it is past the midpoint between perfect agreement ($p_o - p_e = 1 - p_e$) and chance agreement ($p_o = p_e$). The area under the Receiver Operating Characteristic (ROC) curve indicates the proportion of times a student who stopped out had a higher predicted probability of stopping out compared to the predicted probability of stopping out of a student who did not stop out. A value >=.70 is seen as being acceptable (Hosmer & Lemeshow, 2000). Unless stated otherwise, the aforementioned metrics represent "in-sample" classification accuracy rates; that is, the same sample that was used to fit the model parameter estimates was used to estimate the classification accuracy of the model. This is in contrast to "out-of-sample" classification accuracy rates in which one sample is used to fit the model parameter estimates and then a second sample is applied to the model to calculate its accuracy.

Model validation. A best practice in model building is to validate a model using a sample other than the one used to construct the model. For the current study, several forms of cross validation were employed. The first form consisted of validating the stability of the model across samples from the same population. This was achieved by randomly dividing the fall 2009 sample into two equally-sized subsamples: a model-building sample and a within-population validation sample. The model was first

estimated (as described above) using the model-building sample. The model was then refitted using the within-population validation sample. The parameter estimates, model fit, and classification accuracy of the original and validation models were compared. Any large differences between the models were used to inform further model revisions. The final model was then estimated using the full 2009 sample.

The second approached was viewed as a between-populations cross validation. This was used to evaluate the stability of the model across samples from presumably different populations. This involved re-fitting the final fall 2009 model with a complete second set of longitudinal data of new degree- or transfer-seeking students who started the same institution in fall 2010. The parameter estimates, model fit, and in-sample classification accuracy rates based the fall 2009 and fall 2010 samples were compared. Any large differences between the models were used to inform further model revisions. The final model fit and in-sample classification accuracy figures were estimated using all of the students from the original fall 2009 sample.

Model Building Phase 2: Migrating to a Fully Bayesian Approach

The final model from the discrete-time survival analysis in *Phase 1* was translated into a fully Bayesian analog and estimated using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). There were several reasons for taking this approach. The first was to provide a mechanism to more thoroughly evaluate the relationship between each predictor and the outcome variable without being constrained by the frequentist concern of whether a statistic test has sufficient power. The second reason was to pave a path for making subjective interpretations of the predicted probabilities, although this is not a requirement of using such models.

Model building. The Bayesian models were constructed using the following steps:

Step 1: The final discrete-time survival analysis model from *Phase 1* was recreated in WinBUGS as a logistic regression function using normally distributed uninformed, diffuse priors for the coefficients of each intercept and predictor ($\mu = 0$, $\sigma^2 = 1000$). The variance of the prior distributions were specified in terms of precision, $\tau = 1 / \sigma^2$, as required by WinBUGS. The convergence of the model was examined using three chains. Ten thousand (10,000) draws were made for each chain. One thousand draws from each chain were discarded as a "burn in" sample (Gill, 2009). Every third draw was also removed to account for potential autocorrelations. For each parameter, the mixing of the trace plots of the three chains and the Brooks-Gelman-Rubin diagnostic plots (Brooks & Gelman, 1998) were reviewed (Gill, 2009). Chain values that overlap and mix well were viewed as providing support for convergence. Brooks-Gelman-Rubin ratio values around one were seen as providing support of convergence by indicating there is little variance between the three chains of parameter draws. The in-sample classification accuracy of the model was evaluated as described below. The odds ratio and corresponding 95% credibility intervals were constructed and evaluated for the posterior distribution of each intercept and predictor parameter. Odds ratio credibility intervals that spanned the value of one were seen as providing evidence that the variable was not systematically related to the outcome variable and should be considered for removal.

Step 2: This step represented the ambitious goal of expanding the resulting model from *Step 1* to a dynamic fully Bayesian model. This was seen as ambitious because, to the author's knowledge, little work has been done to circumscribe the process for estimating a dynamic fully Bayesian discrete-time survival analysis using data structured

in a person-period format within WINBUGS. As with all new journeys of unforged paths, the author was optimistic as to its outcome while at the same time acknowledging the route was filled with potential challenges that made the ultimate outcome unknown. As a first step, the resulting model from *Step 1* was modified to include a new predictor variable, Prior Logit, which represented the expected logit value for student *i* at time *t*-1. This variable was included to serve as a mechanism for incorporating information into the model about a student's probability of stopping out at time *t*-1 in order to inform the same student's probability for stopping out at time *t*. For $t = 1$, Prior Logit was set as 0 (equaling a 0.5 predicted probability of stopping out) and the prior distribution of the coefficient parameter was set as diffuse, ($\mu = 0$, $\sigma^2 = 1000$). For $t > 1$, Prior Logit equaled the estimated expected value of the posterior distribution of student *i* stopping out at time *t*-1. The prior distribution for the parameter at *t* was the posterior distribution of the same parameter at time period *t*-1. The model was estimated and evaluated using the same approach outlined in *Step 1*.

Classification accuracy. The in-sample classification accuracy of the fully Bayesian models were checked using the posterior distribution of the predicted outcome variable for each student. If the mean predicted probability was >= .50 the student was classified as having stopped out; otherwise a student was classified as having not stopped out. The overall, marginal, and adjusted overall in-sample classification rates where then calculated for each model using the same methods described in *Phase 1*.

Model Building Phase 3: Confirmatory Bayesian Networks

The final phase of the model building process used the results of the fully Bayesian models to create a confirmatory BN.

Model building. The creation of the BNs was as follows:

Step 1: Based on the results from *Phase 2*, the nodes and casual structure of the local BNs at each time point were constructed in the software program Netica (Norsys Software Corp.,1992-2010). As discussed in the next chapter, the pursuit of a dynamic Bayesian network model in Phase 2 Step 2 was unsuccessful. Accordingly the resulting model from Phase 2 was that from Step 1. All continuous variables were discretized to fit within the BN framework and to increase the computational efficiency of updating new information added to the model. Prior term GPA was recoded into four categories: 0.0 – 1.0, 1.1 – 2.0, 2.1 – 3.0, 3.1 – 4.0. This approach to grouping the values was used for interpretative reasons since it is common to evaluate GPA in one-point increments from 0 to 4. Number of credit attempted the prior term was re-coded into five categories: 0.5 – 3.0, 3.1 – 6.0, 6.1 – 9.0, 9.1 – 12.0, > 12.0. These categories were chosen since they roughly represent the natural range of courses and corresponding credits students take in a given term (one course through more than four). The minimum number of credits attempted (for students who did not stop out the prior term) was 0.5. Number of days registered before first course was discretized into quartiles since there was no a prior belief or rationale for otherwise grouping the values.

Step 2: Given the person-period structure of the data set, the dynamic "backbone" (Reye, 2004) was created by simply adding an indexing node to the model. Each level of the indexing node represented one of the six discrete time points.

Step 3: The parameters (i.e., conditional probability tables; CPT) were learned with Netica using a randomly selected half of the sample.

Step 4: The remaining half of the sample was used to evaluate the in-sample classification accuracy of the model using the same methods outlined above for the discrete-time survival analysis. The in-sample classification accuracy rates were then estimated using the full sample.

Step 5: As a between-populations model stability cross-validation, the model was re-fitted using the fall 2010 sample. In other words, the model nodes and structure were based on the final fall 2009 sample but the CPTs were learned using the fall 2010 sample. The fall 2009 and 2010 in-sample classification accuracy rates were then compared. Rates that were reasonably similar across samples were seen as evidence of the generalizability of the model across populations. A limitation of this approach is that it does not provide insight on how accurately the model predicts stopping out behaviors for students not in the sample used to construct the model CPTs. To address this, the out-of-sample classification accuracy rates were calculated by first processing the fall 2010 data through the model learned with the fall 2009 sample, and then running the fall 2009 data through the model built on the CPTs learned with the fall 2010 sample. The rates were then compared and averaged.

Step 6: A second, more parsimonious model was estimated using the same procedures outlined in *Steps* 1-5. Results from *Phase 2* were used to inform the reduction of the predictor variables. The classification accuracy of the model was examined using the process outlined in *Step 5*.

Step 7: The parameter estimates and associated odds ratios were estimated for the modified reduced model using the logistic regression based survival analysis technique from *Phase 1*. This was done to produce the odds ratios for each predictor variable to aid in the interpretation of its relationship with the outcome variable across all time points.

Chapter 4

Results

Sample

The sample included 1,756 new degree or transfer-seeking students at a large community college located in the southwestern United States. Of students who reported their gender, 45% were female and 55% were male. The range of student ages, as of their first term at the college (fall 2009), was 18 to 70 with a mean of 21.32 (*SD* = 6.97). The most common self-reported race/ethnicity was White (44%), followed by Hispanic/Latino (24%) and Black/African American (8%). Approximately one-fifth of the students selected their race/ethnicity as "Other" or it was unknown.

Model Building Phase 1: Discrete-Time Survival Analysis

Life table. The full life table is presented below (Table 1). Figure 3 and 4 display the hazard and survival functions, respectively, for the model. The hazard function indicates that students were most likely to stop out prior to their third term (fall 2010) and were least likely to stop out prior to their sixth term (spring 2012) over a three year period. The survival function summarizes the percentage of students who were estimated to not stop out by a given term. Only 16% of students were estimated not to stop out by the sixth semester, according to the model. The median lifetime for the present model was 1.74. In other words, it was estimated that 50% of students had stopped out after 1.74 semesters.

Table 1

Life Table for Survival Analysis Model without Covariates

			Number of Students				Percentage of Students	
Time Period	Time Interval	Time Interval	Enrolled at Start of Time Period (Risk Set)	Stopped Out During Time Interval	Stopped Out Cumulative	Censored (Graduated or Transferred)	Stopped Out During Time Interval (Hazard Func.)	Still Enrolled at End of Time Interval (Survivor Func.)
0	[0, 1)	[0, Fall 2009)	1,756	0	0	0	0%	100%
1	[1, 2)	[Fall 2009, Spring 2010)	1,756	570	570	77	32%	68%
2	[2, 3)	[Spring 2010, Fall 2010)	1,109	387	957	22	35%	44%
3	[3, 4)	[Fall 2010, Spring 2011)	700	142	1,099	9	20%	35%
4	[4, 5)	[Spring 2011, Fall 2011)	549	159	1,258	53	29%	25%
5	[5, 6)	[Fall 2011, Spring 2012)	337	55	1,313	44	16%	21%
6	[6, 7)	[Spring 2012, Fall 2012)	238	56	1,369	43	24%	16%

Note. [= Included in the time interval.) = Not included in the time interval.

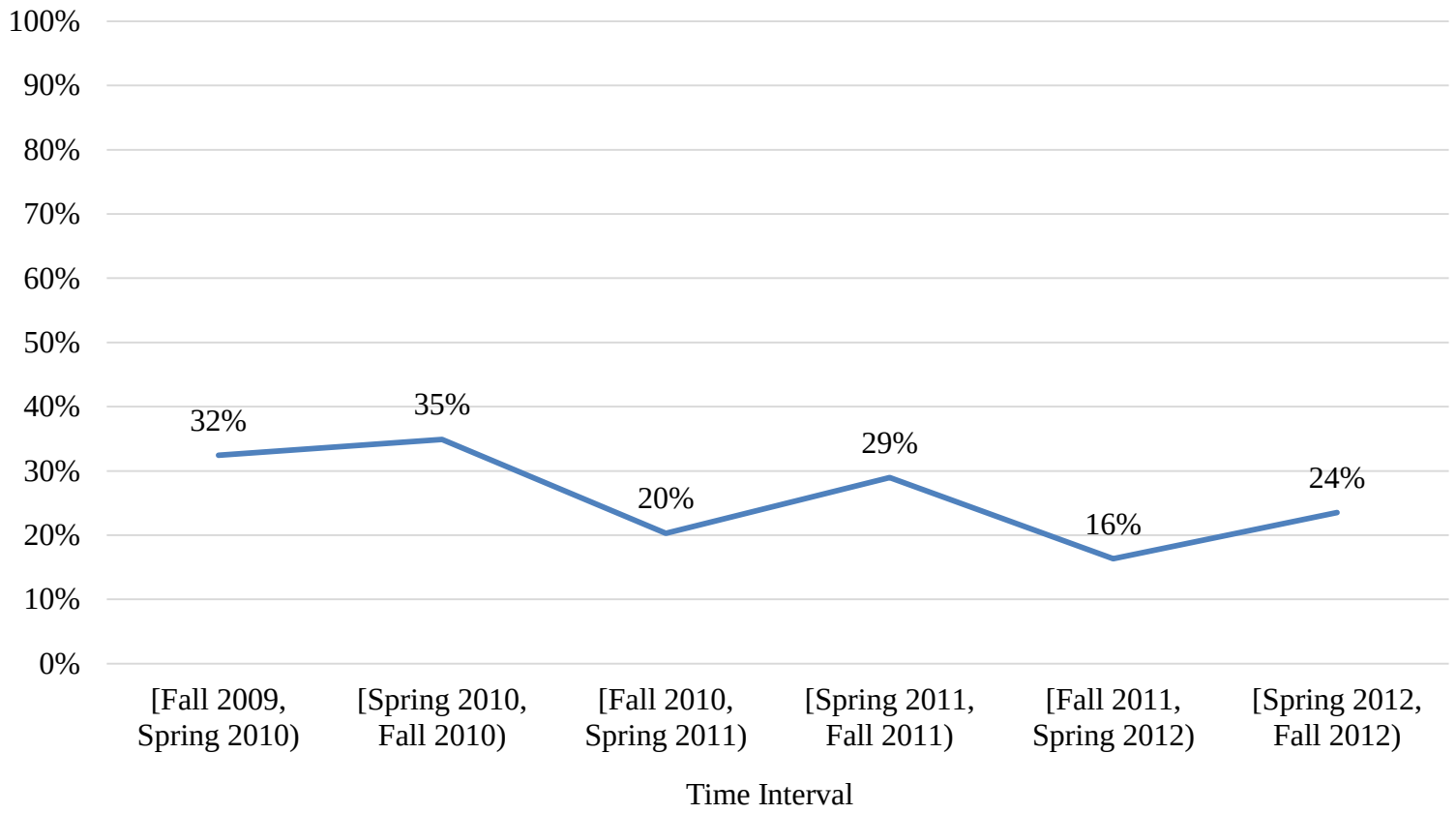


*Figure 3. P*ercentage of students who stopped out by a given time point of the survival analysis model without covariates.

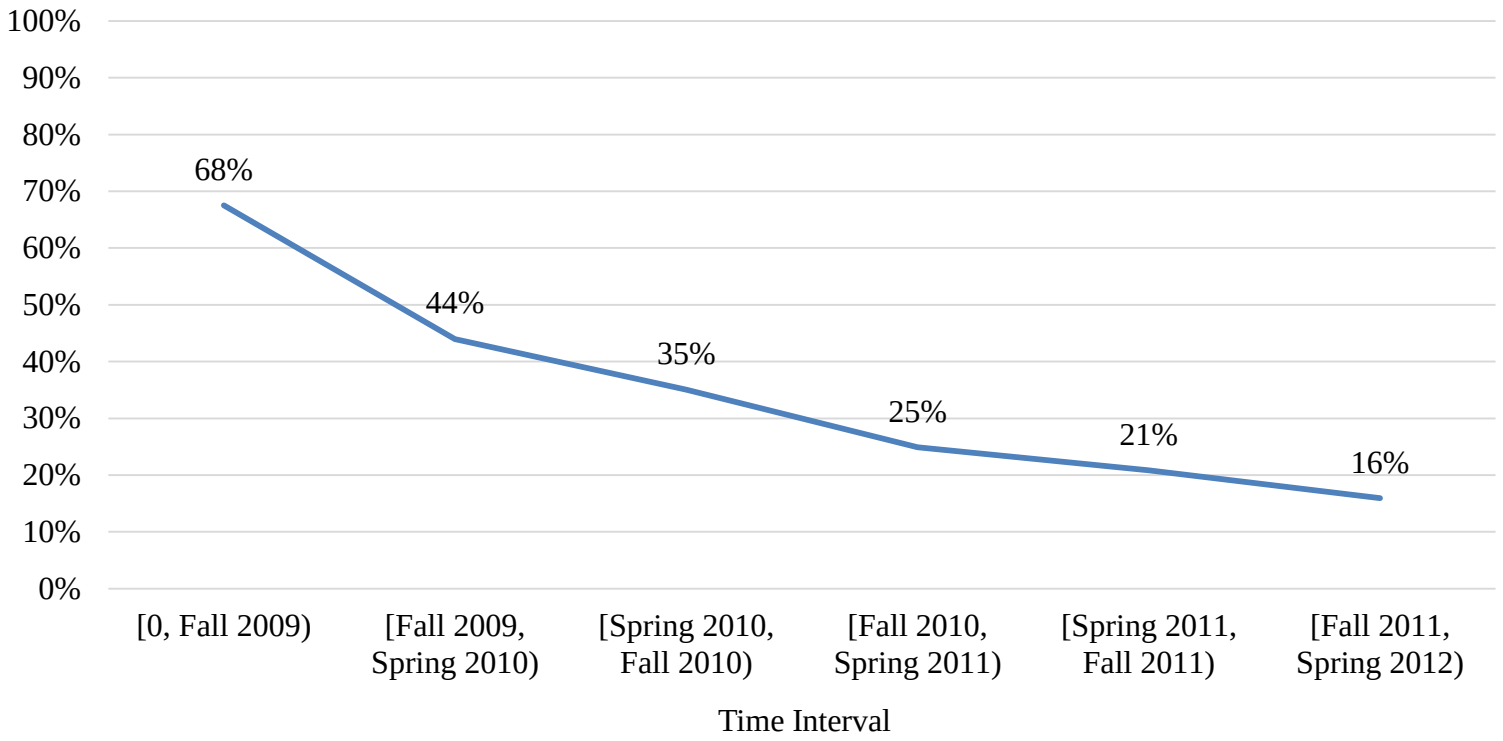


Figure 4. Cumulative percentage of students who did not stop out by a given time point.

Preliminary logistic regression models for each time point. As part of the first step of the iterative modeling building process, a forward stepwise logistic regression was conducted for all the potential predictor variables that made it through the aforementioned screening using half of the students randomly selected from the entire sample. The models were then re-fitted using the remaining half of the sample. The results of both analyses were then used along with the researcher's judgment to determine which variables to include in the final exploratory model for each outcome variable. The parameters for the final exploratory models were then estimated using the entire sample. The results of each final exploratory model based on the full sample are described below.

Time point 1: Spring 2010. The model that predicted whether a student would stop out prior to spring 2010 included seven predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), whether the student received need-based aid the prior term, prior term GPA, how far in advance the student registered for his/her first course the prior term, whether the student graduated high school, total amount of federal subsidized loans the student received the prior term, and whether the estimated median household income for the zip code the student resided in was equal to or greater than the median household income for the sample. The model fit statistics indicated the model fit the data well, R^2Cox and Snell = .37, Nagelkerke R^2 = .52, Hosmer and Lemeshow χ^2 (8) = 4.49, p = .81. The model correctly classified 82% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 2. The parameter estimates are provided in Table 3.

Table 2

Classification Accuracy for Preliminary Logistic Regression Model for Spring 2010

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	966	102	90%	43%	
	Yes	191	347	64%	9%	
Overall % Correct				82%	53%	61%

Note. Classification cut off value .50.

Table 3

Parameter Estimates for Preliminary Logistic Regression Model for Spring 2010

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Prior Term Attempted Credits	-0.24	0.02	177.82	1	< .01	0.79	0.76	0.82
Received Need-Based Aid Prior Term	-1.08	0.18	37.91	1	< .01	0.34	0.24	0.48
Prior Term GPA	-0.44	0.05	67.37	1	< .01	0.64	0.58	0.71
Days Reg. Before 1st Course Prior Term	< -0.01	< 0.01	4.60	1	.03	1.00	0.99	1.00
Cumulative Federal Sub. Loan Accepted Prior Term	< -0.01	< 0.01	9.44	1	< .01	1.00	1.00	1.00
Graduated High School	-0.24	0.20	1.35	1	.25	0.79	0.53	1.18
Est. Household Income >= Median	-0.27	0.14	3.52	1	.06	0.76	0.58	1.01
Constant	2.83	0.26	121.92	1	< .01	16.99		

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Time point 2: Fall 2010. The model that predicted whether a student would stop out prior to fall 2010 included four predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), whether the student received need-based aid the prior term, prior term GPA, and the Pell grant amount the student received the prior term. The model fit statistics indicated the model adequately fit the data, R^2Cox and Snell = .20, Nagelkerke R^2 = .27, Hosmer and Lemeshow χ^2 (8) = 4.93, p = .77. The model correctly classified 75% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 4. The parameter estimates are provided in Table 5.

Table 4

Classification Accuracy for Preliminary Logistic Regression Model for Fall 2010

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	622	78	89%	43%	
	Yes	197	190	49%	9%	
Overall % Correct				75%	52%	47%

Note. Classification cut off value .50.

Table 5

Parameter Estimates for Preliminary Logistic Regression Model for Fall 2010

							95% CI for *OR*	
Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	Lower	Upper
Prior Term Attempted Credits	-0.17	0.02	75.50	1	< .01	0.85	0.82	0.88
Received Need-Based Aid Prior Term	0.62	0.21	8.39	1	< .01	1.85	1.22	2.81
Prior Term GPA	-0.28	0.06	19.41	1	< .01	0.76	0.67	0.86
Pell Grant Amount Accepted Prior Term	< -0.01	< 0.01	2.65	1	.10	1.00	1.00	1.00
Constant	1.13	0.17	47.31	1	< .01	3.10		

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Time point 3: Spring 2011. The model for predicting whether a student would stop out by spring 2011 included seven predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), cumulative GPA for all prior terms, prior term GPA, high school graduation status, whether the student received need-based aid the prior term, whether the student requested help finding work during time of registration, and whether the student requested for assistance of any type (e.g., tutoring, career services, etc.) on the institution's incoming registration form. The model fit statistics indicated the model adequately fit the data, R^2Cox and Snell = .20, Nagelkerke R^2 = .31, Hosmer and Lemeshow χ^2 (8) = 12.28, p = .14. The model correctly classified 84% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 6. The parameter estimates are provided in Table 7.

Table 6

Classification Accuracy for Preliminary Logistic Regression Model for Spring 2011

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	515	27	95%	67%	
	Yes	87	52	37%	2%	
Overall % Correct				83%	69%	45%

Note. Classification cut off value .50.

Table 7

Parameter Estimates for Preliminary Logistic Regression Model for Spring 2011

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Prior Term Attempted Credits	-0.16	0.03	29.04	1	< .01	0.85	0.81	0.90
Received Need-Based Aid Prior Term	-1.05	0.25	17.87	1	< .01	0.35	0.21	0.57
Prior Term GPA	-0.37	0.12	9.89	1	< .01	0.69	0.55	0.87
Graduated High School	-0.56	0.35	2.54	1	.11	0.57	0.29	1.14
Requested Help	-0.38	0.26	2.09	1	.15	0.69	0.41	1.14
Cumulative GPA for Prior Terms	-0.34	0.16	4.26	1	.04	0.72	0.52	0.98
Requested Assistance Finding Work	0.55	0.29	3.60	1	.06	1.73	0.98	3.05
Constant	2.59	0.53	24.31	1	< .01	13.30		

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Time point 4: Fall 2011. The model that predicted whether a student would stop out prior to fall 2011 included five predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), cumulative GPA for all prior terms, the number of hours the student intended to work during the term as reported at the time of enrolling at the institution, total amount of federal subsidized loans the student received for all prior terms, and total Pell grant amount the student received during the prior term. The model fit statistics indicated the model adequately fit the data, R^2Cox and Snell = .14, Nagelkerke R^2 = .20, Hosmer and Lemeshow χ^2 (8) = 8.09, p = .43. The model correctly classified 72% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 8. The parameter estimates are provided in Table 9.

Table 8

Classification Accuracy for Fall 2011 Preliminary Logistic Regression Model

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	308	29	91%	52%	
	Yes	109	50	31%	5%	
Overall % Correct				72%	57%	35%

Note. Classification cut off value .50.

Table 9

Parameter Estimates for Fall 2011 Preliminary Logistic Regression Model

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Prior Term Attempted Credits	-0.11	0.03	14.06	1	< .01	0.90	0.85	0.95
Pell Grant Amount Accepted Prior Term	< -0.01	< 0.01	2.40	1	.12	1.00	1.00	1.00
Prior Term GPA	-0.27	0.09	8.11	1	< .01	0.77	0.64	0.92
Anticipated Work Hours: < 30 / Week			6.97	2	.03			
Anticipated Work Hours: >= 30 / Week	-0.04	0.38	.01	1	.92	0.96	0.46	2.01
Anticipated Work Hours: Unknown	-2.00	0.76	6.97	1	.01	0.14	0.03	0.60
Cumulative Federal Sub. Loan Accepted Prior Term	< 0.01	< 0.01	4.39	1	.03	1.00	1.00	1.00
Constant	0.85	0.25	11.62	1	< .01	2.33		

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Time point 5: Spring 2012. The model that predicted whether a student would stop out prior to spring 2012 included three predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), prior term GPA, and the cumulative number of developmental credit hours attempted for all prior terms. The model fit statistics indicated the model adequately fit the data, R^2Cox and Snell = .15, Nagelkerke R^2 = .25, Hosmer and Lemeshow χ^2 (8) = 5.48, p = .71. The model correctly classified 84% of the sample. The marginal classification rates and the associated Cohen's κ statistic are presented in Table 10. The parameter estimates are provided in Table 11.

Table 10

Classification Accuracy for Spring 2012 Preliminary Logistic Regression Model

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	230	8	97%	73%	
	Yes	40	15	27%	1%	
Overall % Correct				84%	74%	37%

Note. Classification cut off value .50.

Table 11

Parameter Estimates for Spring 2012 Preliminary Logistic Regression Model

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Prior Term Attempted Credits	-0.12	0.05	12.01	1	< .01	0.86	0.78	0.94
Cumulative Dev. Edu. Credits Attempted Prior Terms	-0.09	0.03	7.75	1	< .01	0.91	0.86	0.97
Prior Term GPA	-0.40	0.14	8.17	1	< .01	0.67	0.51	0.88
Constant	1.00	0.40	6.40	1	.01	2.72	1.00	0.40

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Time point 6: Fall 2012. The model that predicted whether a student would stop out prior to fall 2012 included three predictor variables: Number of attempted credit hours during the prior term (excluding credits withdrawn from), cumulative GPA for all prior terms, and whether the student's intent during the time of registration was to transfer to a four-year institution. The model fit statistics indicated the model adequately fit the data, R^2Cox and Snell = .11, Nagelkerke R^2 = .16, Hosmer and Lemeshow χ^2 (8) = 11.58, p = .17. The model correctly classified 75% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 12. The parameter estimates are provided in Table 13.

Table 12

Classification Accuracy for Fall 2012 Preliminary Logistic Regression Model

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	128	11	92%	56%	
	Yes	37	19	34%	4%	
Overall % Correct				75%	60%	38%

Note. Classification cut off value .50.

Table 13

Parameter Estimates for Fall 2012 Preliminary Logistic Regression Model

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Prior Term Attempted Credits	-0.09	0.05	3.67	1	.06	0.92	0.84	1.00
Intent to Transfer	-0.62	0.35	3.15	1	.08	0.54	0.27	1.07
Prior Term GPA	-0.34	0.14	6.05	1	.01	0.71	0.54	0.93
Constant	0.74	0.39	3.68	1	.06	2.10		

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Summary of individual logistic regression models. A summary of the parameter estimates from the individual logistic regression equations is presented in Table 14. The results from the individual logistic regression analyses along with informed judgment by the researcher were used to select the predictors to include in the survival analysis that simultaneously analyzed the probability of a student stopping out at each time point. The decision was made to include the following predictors in the model: a student's high school graduation status, the primary time they attended courses at the institution the preceding term, number of attempted credit hours during the prior term (excluding credits withdrawn from), whether the student received need-based aid the prior term, prior term GPA, and how far in advance the student registered for their first course the prior term. As discussed earlier, any variable that was found to be a significant predictor of stopping out for two or more time periods was included in the survival analysis model. In keeping with a Bayesian philosophical paradigm, primary time of attendance, high school graduation status, and the number of days in advance a student registered for courses in the prior term were added based on the researcher's belief they were variables of interpretive value despite the fact the first two variables were not found to be significant predictors in any of the individual logistic models and the third variable was only found to be a significant predictor for one time period (spring 2010).

Table 14

Summary of Parameter Estimates for Individual Logistic Regression Models

	Model					
Predictor	Spring 2010	Fall 2010	Spring 2011	Fall 2011	Spring 2012	Fall 2012
Anticipated Work Hours: < 30 / Week						
Anticipated Work Hours: >= 30 / Week				-0.04		
Anticipated Work Hours: Unknown				-2.00*		
Cumulative Dev. Edu. Credits Attempted Prior Terms					-0.10*	
Cumulative Federal Sub. Loan Accepted Prior Term	< 0.01			< 0.01*		
Cumulative GPA for Prior Terms			-0.34*			
Days Reg. Before 1st Course Prior Term	-0.01*					
Est. Household Income >= Median	-0.27					
Graduated High School	-0.24		-0.56			
Intent to Transfer						-0.62
Pell Grant Amount Accepted Prior Term		< 0.01		< 0.01		
Prior Term Attempted Credits	-0.24*	-0.17*	-0.16*	-0.11*	-0.16*	-0.09
Prior Term GPA	-0.44*	-0.28*	-0.37*	-0.27*	-0.40*	-0.34*
Received Need-Based Aid Prior Term	-1.10*	0.62*	-1.05*			
Requested Assistance Finding Work			0.55			
Requested Help			-0.38			

Note.$p < .05$.

Preliminary main effects survival analysis model. The six variables mentioned above were used to estimate the preliminary main effect survival analysis model. The model fit statistics indicated the survival analysis model fit the data well, R^2Cox and Snell = .36, Nagelkerke R^2 = .48, Hosmer and Lemeshow χ^2 (8) = 10.34, p = .24. The model correctly classified 79% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 15. The area under the receiver operating characteristic (ROC) curve was .81, 95% CI [.79, .82], $p < .01$. The parameter estimates are provided in Table 16. The results were considered the preliminary main effects model.

Table 15

Classification Accuracy for Preliminary Main Effects Survival Analysis Model

		Predicted Stopped Out				
		No	Yes	Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
Observed Stopped Out	No	3009	273	92%	52%	
	Yes	718	630	47%	6%	
Overall % Correct				79%	58%	49%

Note. Classification cut off value .50.

Table 16

Parameter Estimates for Preliminary Main Effects Survival Analysis Model

Variables	*α, β*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Time1: Spring 2010	1.36	0.14	96.91	1	< .01	3.91	2.98	5.13
Time2: Fall 2010	1.83	0.16	136.39	1	< .01	6.22	4.58	8.45
Time3: Spring 2011	1.17	0.18	42.64	1	< .01	3.22	2.27	4.57
Time4: Fall 2011	1.73	0.18	94.18	1	< .01	5.63	3.97	7.97
Time5: Spring 2012	0.84	0.22	14.57	1	< .01	2.31	1.50	3.55
Time6: Fall 2012	1.19	0.22	29.90	1	< .01	3.30	2.15	5.07
Graduated High School: Yes	-0.16	0.11	2.04	1	.15	0.85	0.68	1.06
Primarily Time of Attendance: Day	0.41	0.09	18.95	1	< .01	1.50	1.25	1.80
Prior Term Attempted Credits	-0.16	0.01	243.10	1	< .01	0.85	0.84	0.87
Days Reg. Before 1st Course Prior Term	< 0.01	< 0.01	8.64	1	< .01	1.00	1.00	1.00
Received Need-Based Aid Prior Term: Yes	-0.37	0.08	21.35	1	< .01	0.69	0.59	0.81
Prior Term GPA	-0.37	0.03	137.20	1	< .01	0.69	0.65	0.73

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval. *α* = Intercept parameter for time point. *β* = Parameter for predictor variable.

Final main effects survival analysis model: linearity assumption. The survival analysis as specified in the preliminary main effects survival analysis model assumed that continuous predictor variables were linearly related to the predicted logit. This assumption was tested for the three continuous variables in the model – prior term attempted hours, number of days registered before first course, and prior term GPA – by adding to the model an interaction term of the variable multiplied by the natural log of itself and conducting a nested χ^2 likelihood ratio test (Tabachnick & Fidell, 2007). None of the interactions were significant, providing support that the values of the predictor variables were linearly related to the predicted logit. The resulting model was considered the final main effects model.

Preliminary final survival analysis model: Specification of time. The final main effects model included a general nonparametric specification of time. Other specifications of time were also evaluated. Specifically, constant, linear, quadratic, and cubic specifications of time were also estimated. The results are presented in Table 17. Each successive model produced a significant increase in model fit compared to the preceding model. The largest difference in -2 Log Likelihood values was between the cubic and general specifications of time. Given the general specification had significantly and markedly better fit than the cubic specification with relatively little additional cost (three degrees of freedom), the decision was made to keep time specified in a general, nonparametric form. The resulting model was considered the preliminary final survival analysis model.

Table 17

Examination of Various Specifications of Time for Preliminary Final Survival Analysis Model

				Hosmer and Lemeshow Test			Nested Likelihood Ratio Test		
Specification of Time	-2 Log likelihood (LL)	Cox and Snell R^2	Nagelkerke R^2	X^2	*df*	*p*	Δ - 2LL	*df*	*p*
Constant Only Model	4414.32	0.35	.47	3.50	8	.90	--	--	--
Linear	4411.11	0.35	.47	4.92	8	.77	3.20	1	.05
Quadratic	4399.66	0.35	.47	2.88	8	.94	11.45	1	< .01
Cubic	4392.00	0.35	.47	7.65	8	.47	7.67	1	< .01
General (Nonparametric)	4356.50	0.36	.48	10.34	8	.24	35.50	3	< .01

Final survival analysis model: Interactions with time. A subsequent model was estimated to evaluate whether there were any significant interactions between the predictors and time in the preliminary survival analysis and time. There were no significant interactions for any of the predictors across all time points. The only significant relationships that were found were between the second time point (spring 2010) and the number of credits attempted the prior term and the second time point and the prior term GPA. The decision was made to exclude predictor-by-time interactions from the final survival analysis model given the limited and localized presence of the interactions. The parameter estimates and classification accuracy rates for the resulting model with a general specification of time and no time-by-predictor interactions are the same as the results as the preliminary main effects model (see Tables 15 and 16).

Fitted logit, hazard, and survival functions for predictors. The fitted logit, hazard, and survival functions were inspected for each of the predictor variables in the final survival analysis model. The logit function depicts the change in logits for students who stopped out compared to those who did not for each one-unit increase in predictor variable, holding all other variables constant. The hazard function provides the expected change in the probability of stopping out for each one-unit increase in predictor variable, holding all other variables constant. The survival function illustrates the change in the cumulative probability of a student not stopping out for each one-unit increase in the predictor variable, holding all other variables constant. The fitted, hazard, and survival functions for each predictor variable in the final survival analysis model are discussed below.

High school graduate. The high school graduate variable is dichotomous and indicates whether or not a student graduated from high school. Not graduating high school served as the reference category. The parameter estimate (on the logit scale) for the variable was -0.16 and was not significant, $p = .15$. The corresponding odds was 0.85, 95% CI [0.68, 1.06]. The parameter estimates for the variable at each time point along with the hazard and survival rates are presented in Table 18.

Primary time of attendance (prior term). Primary time of attendance for the prior term represents whether the student primarily took courses in the day (classes that started prior to 4:30pm) or evening during the prior term. Attending courses primarily during the day was the reference category. The parameter estimate for the variable (on the logit scale) was 0.41 and was significant, $p < .01$. The corresponding odds ratio was 1.50, 95% CI [1.25, 1.80]. This indicates that the odds of stopping out were 1.5 times greater for students who did not primarily attended courses in the day, holding all other variables constant. The parameter estimates for the variable at each time point along with the hazard and survival rates are presented in Table 19.

Term attempted hours (prior term). Term attempted hours represents the number of credit hours (excluding courses the student withdrew from) that the student attempted during the prior term. The parameter estimate for the variable (on the logit scale) was -0.16 and was significant, $p < .01$. The corresponding odds was 0.85, 95% CI [0.84, 0.87]. This indicates that the odds of stopping out decreased by a factor of 0.85 for each additional credit hour a student attempted during the prior term, holding all other variables constant. The parameter estimates for the variable at four values—3, 6, 9, and 12—along with the hazard and survival rates are presented in Table 20.

Days registered before start of first course (prior term). As the name implies, the variable denotes the number of days in the prior term a student registered for courses before the start of his/her first course for that term. For example, a value of 30 means that in the prior term a student registered for courses a month before the start of the earliest course they registered for that term. The parameter estimate for the variable (on the logit scale) was -0.003 and was significant, $p < .01$. The corresponding odds ratio was 1.00, 95% CI [1.00, 1.00]. Although the parameter estimate was significant, the odds ratio and associated 95% confidence interval provide evidence that the variable is not a meaningful predictor of whether a student will stop out. The parameter estimates, hazard and survival rates for the variable at three time points—0 days, 45 days, and 90 days—are presented in Table 21.

Received need-based aid (prior term). The variable represents whether a student received need-based financial aid during the prior term. Not receiving need-based aid served as the reference category. The parameter estimate for the variable (on the logit scale) was -0.37 and was significant, $p < .01$. The corresponding odds ratio was 0.69, 95% CI [0.59, 0.81]. This indicates that the odds of stopping out were 0.69 times lower for students who received need-based aid in the prior term compared to students who did not, holding all other variables constant. The parameter estimates for the variable at each time point along with the hazard and survival rates are presented in Table 22.

Term GPA (prior term). Term GPA provides a student's grade-point average (on a 0-4 point scale) for the prior term. The parameter estimate for the variable (on the logit scale) was -0.37 and was significant, $p < .01$. The corresponding odds was 0.69, 95% CI [0.65, 0.73]. This indicates that the odds of stopping out decreased by a factor of 0.69 for

each additional one-point increase in a student's GPA, holding all other variables constant. The parameter estimates along with the hazard and survival rates are presented for four values of the variable – 1, 2, 3, and 4 – in Table 23.

Table 18

Fitted Logit, Hazard, and Survival Estimates for Graduated High School

		Graduated from High School (Earned a High School Diploma)					
		Logit Hazard		Hazard		Survival	
Time Period	α_j	No	Yes	No	Yes	No	Yes
[Fall 2009, Spring 2010)	1.36	1.36	1.20	.80	.77	.20	.23
[Spring 2010, Fall 2010)	1.83	1.83	1.67	.86	.84	.03	.04
[Fall 2010, Spring 2011)	1.17	1.17	1.01	.76	.73	.01	.01
[Spring 2011, Fall 2011)	1.73	1.73	1.57	.85	.83	< .01	< .01
[Fall 2011, Spring 2012)	0.84	0.84	0.68	.70	.66	< .01	< .01
[Spring 2012, Fall 2012)	1.19	1.19	1.03	.77	.74	< .01	< .01

Table 19

Fitted Logit, Hazard, and Survival Estimates for Primary Time of Attendance During Prior Term

		Primary Time of Attendance During Prior Term					
		Logit Hazard		Hazard		Survival	
Time Period	α_j	Day	Not Day	Day	Not Day	Day	Not Day
[Fall 2009, Spring 2010)	1.36	1.36	1.77	.80	.85	.20	.15
[Spring 2010, Fall 2010)	1.83	1.83	2.23	.86	.90	.03	.01
[Fall 2010, Spring 2011)	1.17	1.17	1.57	.76	.83	.01	< .01
[Spring 2011, Fall 2011)	1.73	1.73	2.13	.85	.89	< .01	< .01
[Fall 2011, Spring 2012)	0.84	0.84	1.24	.70	.78	< .01	< .01
[Spring 2012, Fall 2012)	1.19	1.19	1.60	.77	.83	< .01	< .01

Table 20

Fitted Logit, Hazard, and Survival Estimates for Number of Credits Attempted During Prior Term

		Number of Credits Attempted Prior Term (Excl. Withdrawals)											
		Logit Hazard				Hazard				Survival			
Time Period	α_j	3	6	9	12	3	6	9	12	3	6	9	12
[Fall 2009, Spring 2010)	1.36	1.36	0.42	-0.06	-0.53	.80	.60	.49	.37	.20	.40	.51	.63
[Spring 2010, Fall 2010)	1.83	1.83	0.88	0.41	-0.07	.86	.71	.60	.48	.03	.12	.21	.33
[Fall 2010, Spring 2011)	1.17	1.17	0.22	-0.25	-0.73	.76	.55	.44	.33	.01	.05	.12	.22
[Spring 2011, Fall 2011)	1.73	1.73	0.78	0.31	-0.17	.85	.69	.58	.46	<.01	.02	.04	.12
[Fall 2011, Spring 2012)	0.84	0.84	-0.11	-0.59	-1.06	.70	.47	.36	.26	< .01	.01	.03	.09
[Spring 2012, Fall 2012)	1.19	1.19	0.25	-0.23	-0.70	.77	.56	.44	.33	< .01	< .01	.02	.06

Table 21

Fitted Logit, Hazard, and Survival Estimates for Number of Days Registered Before Start of First Course Prior Term

		Number of Days Registered Before Start of First Course in Prior Term								
		Logit Hazard			Hazard			Survival		
Time Period	α_j	0	45	90	0	45	90	0	45	90
[Fall 2009, Spring 2010)	1.36	1.36	1.23	1.09	.80	.77	.75	.20	.23	.25
[Spring 2010, Fall 2010)	1.83	1.83	1.69	1.56	.86	.84	.83	.03	.04	.04
[Fall 2010, Spring 2011)	1.17	1.17	1.03	0.90	.76	.74	.71	.01	.01	.01
[Spring 2011, Fall 2011)	1.73	1.73	1.59	1.46	.85	.83	.81	< .01	< .01	< .01
[Fall 2011, Spring 2012)	0.84	0.84	0.70	0.57	.70	.67	.64	< .01	< .01	< .01
[Spring 2012, Fall 2012)	1.19	1.19	1.06	0.92	.77	.74	.72	< .01	< .01	< .01

Table 22

Fitted Logit, Hazard, and Survival Estimates for Received Need-Based Aid in the Prior Term

		Received Need Based Aid in the Prior Term					
		Logit Hazard		Hazard		Survival	
Time Period	α_j	No	Yes	No	Yes	No	Yes
[Fall 2009, Spring 2010)	1.36	1.36	1.00	.80	.73	.20	.27
[Spring 2010, Fall 2010)	1.83	1.83	1.46	.86	.81	.03	.05
[Fall 2010, Spring 2011)	1.17	1.17	0.80	.76	.69	.01	.02
[Spring 2011, Fall 2011)	1.73	1.73	1.36	.85	.80	< .01	< .01
[Fall 2011, Spring 2012)	0.84	0.84	0.47	.70	.62	< .01	< .01
[Spring 2012, Fall 2012)	1.19	1.19	0.83	.77	.70	< .01	< .01

Table 23

Fitted Logit, Hazard, and Survival Estimates for Prior Term GPA

		Prior Term GPA											
		Logit Hazard				Hazard				Survival			
Time Period	α_j	1.00	2.00	3.00	4.00	1.00	2.00	3.00	4.00	1.00	2.00	3.00	4.00
[Fall 2009, Spring 2010)	1.36	1.36	0.62	0.25	-0.12	.80	.65	.56	.47	.20	.35	.44	.53
[Spring 2010, Fall 2010)	1.83	1.83	1.09	0.72	0.34	.86	.75	.67	.59	.03	.09	.14	.22
[Fall 2010, Spring 2011)	1.17	1.17	0.43	0.05	-0.32	.76	.60	.51	.42	.01	.03	.07	.13
[Spring 2011, Fall 2011)	1.73	1.73	0.99	0.61	0.24	.85	.73	.65	.56	< .01	.01	.02	.06
[Fall 2011, Spring 2012)	0.84	0.84	0.10	-0.28	-0.65	.70	.52	.43	.34	< .01	< .01	.01	.04
[Spring 2012, Fall 2012)	1.19	1.19	0.45	0.08	-0.29	.77	.61	.52	.43	< .01	< .01	.01	.02

Model validation. The between-population model stability cross validation of the final survival analysis model was evaluated by re-fitting the model using the fall 2010 sample. All other aspects of the model (e.g., predictors, specification of time, etc.) were kept the same. The fall 2010 sample model fit statistics indicated the model did not fit the data well, R^2Cox and Snell = .34, Nagelkerke R^2 = .46, Hosmer and Lemeshow χ^2 (8) = 28.72, $p < .01$. The significant Hosmer and Lemeshow χ^2 statistic indicates that there was a significant difference between the observed and model predicted values. This is not a completely unexpected occurrence whenever a new data set is used to validate a model. Although a non-significant Hosmer and Lemeshow χ^2 statistic was desirable, a significant value does not dramatically diminish the value of using the data set to evaluate the stability of the parameter estimates across samples. Table 24 provides a comparison of the parameter estimates for the fall 2009 and fall 2010 samples. For the fall 2009 sample, the terms ranged from spring 2010 (Time1) to fall 2012 (Time6). For the fall 2010 sample, the terms spanned spring 2011 (Time1) to fall 2013 (Time6).

In most instances the parameter values were similar for the two samples. The most notable difference was between the parameter estimates for the Time5 intercepts. The odds ratio for the fall 2010 sample (4.19) was almost twice as large as the odds ratio for the fall 2009 sample (2.31). There was also a noticeable difference in the odds ratios for the predictor primary time of attendance. The odds ratio for the fall 2009 sample (1.50) was 25% greater than the odds ratio for the fall 2010 sample (1.20). Table 25 provides a comparison of the classification accuracy rates between the fall 2009 and 2010 samples, both estimated using the final survival analysis model. The overall in-sample classification accuracy rates were similar for both samples. As for the marginal rates, the

model more accurately predicted students who stopped out in the fall 2009 sample compared to the fall 2010 sample. Although the difference in parameter estimates and classification rates are not immaterial, the results were reasonably close to support the generalizability of the final survival analysis model across samples.

Table 24

Comparison of Parameter Estimates for Final Survival Analysis

	Fall 2009 Sample		Fall 2010 Sample	
Variables	α,β	*OR*	α,β	*OR*
Time1	1.36*	3.91	1.27*	3.55
Time2	1.83*	6.22	1.98*	7.26
Time3	1.17*	3.22	1.19*	3.30
Time4	1.73*	5.63	1.54*	4.66
Time5	0.84*	2.31	1.43*	4.19
Time6	1.19*	3.30	1.63*	5.10
Graduated High School	-0.16	0.85	-0.20*	0.82
Primarily Time of Attendance: Not Day	0.41*	1.50	0.18*	1.20
Prior Term Attempted Credits	-0.16*	0.85	-0.15*	0.86
Days Reg. Before 1st Course Prior Term	< -0.01*	1.00	< -0.01*	1.00
Received Need-Based Aid Prior Term	-0.37*	0.69	-0.43*	0.65
Prior Term GPA	-0.37*	0.69	-0.37*	0.69

Note. Q = Quartile. *OR* = Odds ratio.* $p < .05$. α = Intercept parameter for time point. β = Parameter for predictor variable.

Table 25

Comparison of Classification Accuracy for Final Survival Analysis

Percent Correctly Classified	Fall 2009 Sample	Fall 2010 Sample
Overall	79%	78%
Did Not Stop Out	92%	92%
Stopped Out	47%	42%
% Improvement (Cohen's κ)	49%	46%

Note. Classification cut off value .50.

Model Building Phase 2: Migrating to a Fully Bayesian Approach

The final survival analysis model was re-estimated using the fall 2009 sample and a fully Bayesian model in WinBUGS. The model estimated the final survival analysis model in order to more thoroughly evaluate the relationship between each predictor and the outcome variable. The WINBUGS code is provided in Appendix B. The model was examined using three chains, each with 10,000 draws. Every third draw was thinned to account for potential autocorrelations. The first 1,000 draws from each chain were discarded as burn in. The trace plots of the three chains overlapped and mixed well for each of the parameters (see Appendix C). The Brooks-Gelman-Rubin diagnostic plots (see Appendix C) and associated values provided support for convergence. The Brooks-Gelman-Rubin ratio values were around one for all the parameters, indicating there was little variance between the three chains of parameter draws. The remaining 9,999 iterations (after thinning and removing the burn in sample) were used to plot the posterior distributions of the parameters. The density plots of the posterior distributions for the parameters are provided in Appendix C. The mean, standard deviation, and 95% credibility intervals of the distributions are presented in Table 26. The model correctly

classified 79% of the sample. The marginal classification rates and the Cohen's κ statistic are presented in Table 27.

Table 26

Summary Statistics for Posterior Distributions of Intercept and Predictor Parameters for Bayesian Final Survival Analysis Model Using Fall 2009 Sample

Variables	β_{mean}	β_{SD}	MC Error	OR_{mean}	95% Credibility Interval for *OR*	
					Lower	Upper
Time1: Spring 2010	1.37	0.14	< 0.01	3.97	2.99	5.15
Time2: Fall 2010	1.83	0.16	< 0.01	6.33	4.57	8.46
Time3: Spring 2011	1.17	0.18	< 0.01	3.27	2.26	4.59
Time4: Fall 2011	1.73	0.18	< 0.01	5.74	3.95	8.07
Time5: Spring 2012	0.83	0.22	< 0.01	2.36	1.48	3.53
Time6: Fall 2012	1.19	0.22	< 0.01	3.38	2.15	5.06
Graduated High School	-0.16	0.11	< 0.01	0.86	0.68	1.07
Primarily Time of Attendance: Not Day	0.41	0.09	< 0.01	1.51	1.25	1.80
Prior Term Attempted Credits	-0.16	0.01	< 0.01	0.85	0.84	0.87
Days Reg. Before 1st Course Prior Term	< -0.01	< 0.01	< 0.01	1.00	0.99	1.00
Received Need-Based Aid Prior Term	-0.37	0.08	< 0.01	0.69	0.59	0.81
Prior Term GPA	-0.37	0.03	< 0.01	0.69	0.65	0.73

Note. SD = Standard deviation. *OR* = Odds ratio. *MC* = Monte Carlo.

Table 27

Classification Accuracy for Bayesian Final Survival Analysis Model with Fall 2009 Sample

		Predicted Stopped Out		Correctly Classified	Expected Agreement	% Improv. (Cohen's κ)
		No	Yes			
Observed Stopped Out	No	3,008	274	92%	52%	
	Yes	718	630	47%	6%	
Overall % Correct				79%	58%	49%

Note. Classification cut off value .50.

Final survival analysis model with dynamic predicted probability parameter. Numerous attempts were made to estimate a dynamic fully Bayesian model using a person-period format in WINBUGS as previously outlined in the Methods section. Unfortunately, the author was unable to get any of the attempts to successfully run in WINBUGS. It is unknown whether the dozens of failed attempts were due to the author misspecifying the models, a lack of expertise needed to "trick" WINBUGS into running a DBN in a person-period format, or if WINBUGS is simply not able to estimate such a model as specified. The second potential cause is believed to be the most likely source of the failed attempts. This is an area for future exploration. Regardless of the cause, this had two net effects. The first was that the results from *Phase 2,* Step 1 were used to inform *Phase 3*.The second, related effect is that it forced the author to pivot to using a BN instead of a DBN for *Phase 3*. It is arguable that a BN specified in a person-period format with longitudinal data is still a DBN. However, this deviates from the prior articulation of the DBN presented in chapters two and three (e.g., see Figure 2 on page 37). Therefore the results from *Phase 3* will be viewed as BNs to maintain continuity of thought.

Model Building Phase 3: Confirmatory Bayesian Network

Two confirmatory BNs were constructed. The first BN represented a discretized version of the final survival analysis from *Phase 2,* Step 1. The continuous variables were discretized to fit within the BN framework and allow for the efficient updating of predictor probabilities in real time. The second featured a reduced model that pared some of the predictors from the discretized final model based on the results of the *Phase 2* non-dynamic Bayesian analyses. The results of both models are detailed below.

Discretized survival analysis model. The modified (discretized) survival analysis model was replicated as a BN using Netica. A diagram of the model is provided in Figure 5. The conditional probability tables (CPTs) were learned using Netica and the fall 2009 sample. The CPTs are provided in Appendix D. The model correctly classified 85% of the sample. The marginal classification rates and the associated Cohen's κ statistic are presented in Table 28. The classification accuracy was then cross-validated by re-learning the CPTs for the model using the fall 2010 sample. The model correctly classified 82% of the fall 2010 sample. A comparison of the in-sample classification accuracy rates are provided in Table 28.

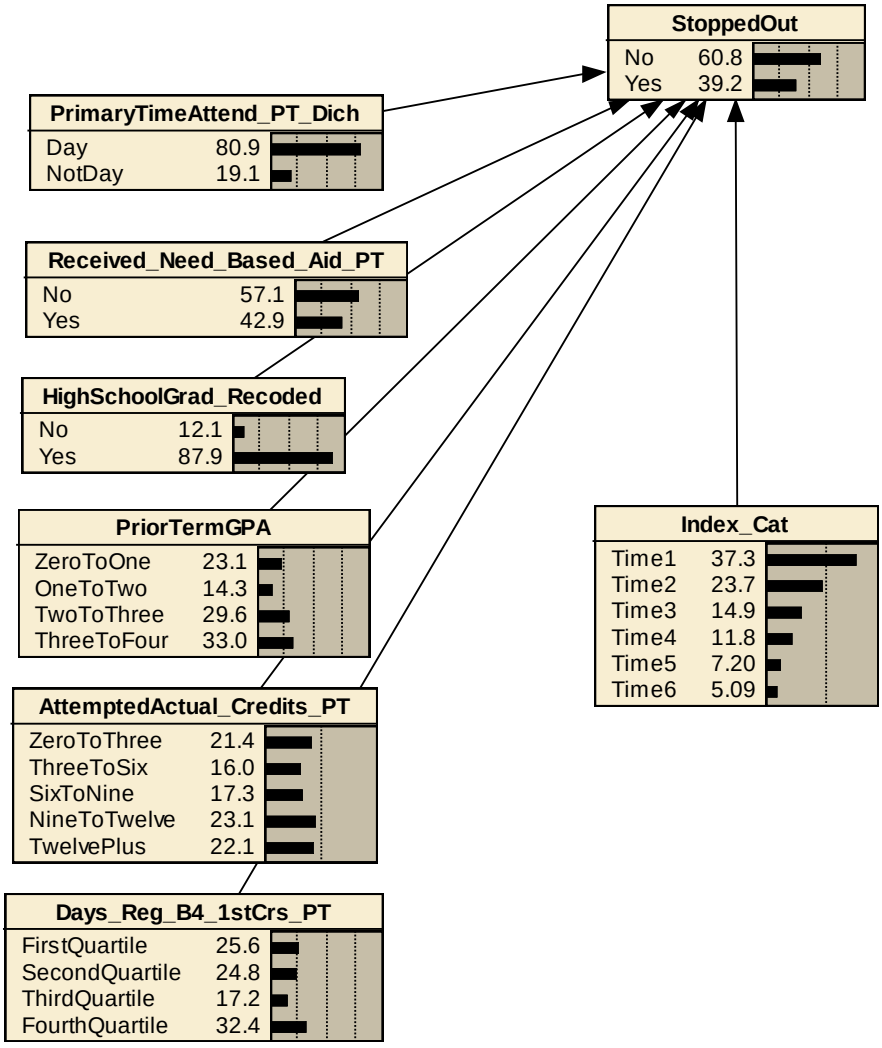


Figure 5. Modified (discretized) final survival analysis model specified as a Bayesian network.

Table 28

Comparison of Classification Accuracy for Modified Final Survival Analysis Model as a Bayesian Network

Percent Correctly Classified	Fall 2009 Sample	Fall 2010 Sample
Overall	85%	82%
Did Not Stop Out	94%	93%
Stopped Out	64%	54%
% Improvement (Cohen's κ)	65%	56%

Note. Classification cut off value .50.

Reduced discretized survival analysis model. High school graduation status and the number of days registered prior to the start of the first course the prior term were removed from the model based on the fully Bayesian analysis results. The 95% credibility interval of the odds ratios for both variables spanned both sides of one, indicating the variables were not systematically associated with the outcome variable. A diagram of the reduced model is provided in Figure 6. The CPTs were learned via Netica using the fall 2009 sample. The CPTs are provided in Appendix E. The model correctly classified 81% of the sample. The marginal classification rates and the associated Cohen's κ statistic are presented in Table 29. The classification accuracy rates were then cross-validated by re-learning the model CPTs using the fall 2010 sample. The model correctly classified 79% of the fall 2010 sample. A comparison of the in-sample classification accuracy rates are provided in Table 29.

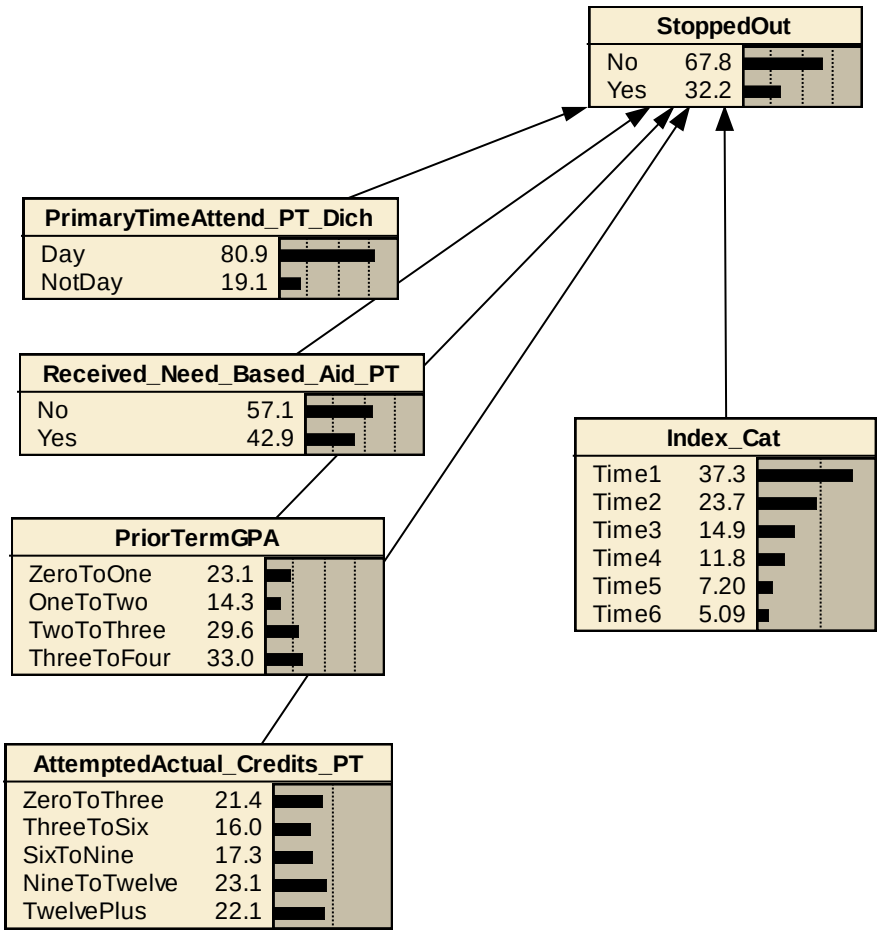


Figure 6. Reduced, modified (discretized) final survival analysis model specified as a Bayesian network.

Table 29

Comparison of Classification Accuracy for Reduced, Modified Final Survival Analysis Model as a Bayesian Network

Percent Correctly Classified	Fall 2009 Sample	Fall 2010 Sample
Overall	81%	79%
Did Not Stop Out	91%	91%
Stopped Out	54%	50%
% Improvement (Cohen's κ)	55%	51%

Note. Classification cut off value .50.

Out-of-Sample Classification Accuracy. As a further evaluation of the generalizability of the reduced model, the fall 2009 and fall 2010 out-of-sample classification accuracy rates for both the reduced BN were calculated as described in the previous chapter. This differed from the in-sample approach used to calculate all of the previously reported classification rates. For example, the fall 2010 in-sample classification accuracy rates presented in the preceding section were based on the model CPTs learned from the fall 2010 sample. In contrast, the out-of-sample fall 2010 classification accuracy rates were based on the model CPTs learned from the fall 2009 sample. A comparison of the out-of-sample and in-sample classification accuracy rates are provided in Table 30.

Table 30

Comparison of Out-of-Sample and In-Sample Classification Accuracy for Modified & Reduced Final Survival Analysis Model as a Bayesian Network

	Out-of-Sample			In-Sample		
	Learned using Fall 2009 Cohort	Learned using Fall 2010 Cohort		Learned using Fall 2009 Cohort	Learned using Fall 2010 Cohort	
Percent Correctly Classified	Fall 2010 Cohort	Fall 2009 Cohort	Avg.	Fall 2009 Cohort	Fall 2010 Cohort	Avg.
Overall	78%	78%	78%	81%	79%	80%
Did Not Stop Out	90%	89%	90%	91%	91%	91%
Stopped Out	46%	51%	49%	54%	50%	52%
% Improvement (Cohen's κ)	47%	51%	49%	55%	51%	53%

Odds ratios. The parameter estimates and associated odds ratios were estimated for the modified reduced model using the logistic regression based survival analysis technique from *Phase 1*. This was done to produce the odds ratios for each predictor

variable to aid in the interpretation of its relationship with the outcome variable across all time points. The results are presented in Table 31.

Table 31

Parameter Estimates for Modified & Reduced Final Survival Analysis Model

Variables	*B*	*SE*	*Wald*	*df*	*p*	*OR*	95% CI for *OR* Lower	95% CI for *OR* Upper
Time 1: Spring 2010	0.92	0.09	97.10	1	< .01	2.50	2.08	3.00
Time 2: Fall 2010	1.33	0.11	148.09	1	< .01	3.78	3.05	4.69
Time 3: Spring 2011	0.62	0.13	21.83	1	< .01	1.86	1.43	2.41
Time 4: Fall 2011	1.25	0.14	82.69	1	< .01	3.50	2.67	4.59
Time 5: Spring 2012	0.30	0.18	2.68	1	< .01	1.35	0.94	1.92
Time6: Fall 2012	0.72	0.19	14.44	1	< .01	2.05	1.42	2.97
Primary Time of Attendance: Not Day	0.45	0.09	23.25	1	< .01	1.56	1.30	1.87
Prior Term Attempted Credits: 0.5 – 3.0			226.48	4	< .01			
Prior Term Attempted Credits: 3.5 – 6.0	-0.66	0.11	33.88	1	< .01	0.52	0.41	0.65
Prior Term Attempted Credits: 6.5 – 9.0	-1.12	0.12	83.55	1	< .01	0.33	0.26	0.41
Prior Term Attempted Credits: 9.5 – 12.0	-1.42	0.13	126.43	1	< .01	0.24	0.19	0.31
Prior Term Attempted Credits: 12+	-2.01	0.15	187.29	1	< .01	0.13	0.10	0.18
Received Need-Based Aid Prior Term	-0.42	0.08	27.95	1	< .01	0.66	0.56	0.77
Prior Term GPA: 0.0 – 1.0			183.82	3	< .01			
Prior Term GPA: 1.1 – 2.0	-0.71	0.12	35.19	1	< .01	0.49	0.39	0.62
Prior Term GPA: 2.1 – 3.0	-1.20	0.11	114.29	1	< .01	0.30	0.24	0.38
Prior Term GPA: 3.1 – 4.0	-1.43	0.11	162.26	1	< .01	0.24	0.19	0.30

Note. SE = Standard error. *OR* = Odds ratio. CI = Confidence interval.

Chapter 5

Discussion

Methodological Conclusions

Model building process. Fully Bayesian analyses and BNs along with the broader category of data mining methodologies, are powerful methodological techniques that offer a great deal of promise to the field of education and social sciences as a whole. However, these techniques have gone relatively underutilized in the field of education for the purposes of predicting student success. This is presumably because educational researchers frequently are not aware of these techniques or are aware of them but may not be sure how to incorporate them into their methodological toolboxes. The purpose of this study was to help address this need by providing a model-building approach for developing BNs that leveraged educational data mining, Bayesian analysis, and traditional iterative model-building techniques. This was accomplished through a three-phased approach. The first phase was designed to help researchers reared in the social sciences leverage a statistical perspective (frequentist) and methodology (logistic regression) they are familiar with to build a survival analysis model using a form of logistic regression. The second phase scaffolded those efforts into the realm of Bayesian analysis. There is a growing body of work on a variety of sophisticated Bayesian-based survival analyses (e.g., Ibrahim, Chen, & Sinha, 2001). The drawback of these efforts is that they represent a significant learning curve to those raised in the frequentist perspective, making them difficult to quickly adopt and use. The approach taken in this paper sought to address this by providing a gentle path into the world of Bayesian analysis by simply translating the logistic regression survival analysis developed in *Phase*

1 into a fully Bayesian framework. The original intent was for *Phase 2* to include a dynamic fully Bayesian model to inform the development of a DBN. Unfortunately, the author was not able to successfully estimate such as model despite numerous attempts to do so. As a result, the third phase took the findings from the first two phases and converted them into a discretized BN. This was to accomplish a key goal of this study: to leverage the benefits of BNs – visual representation of variable dependencies, economical estimation of conditional probabilities, and efficient bidirectional updating – to provide a user-friendly model that can be used by non-methodologists to quickly and accurately calculate understandable estimates of a student's probability of stopping out by a given term.

To this researcher, a key requirement of all research is that it must be made accessible to those who can benefit from it or else it will remain largely unused and sit as inert words on a page. The current study fulfilled this mission by starting with a pool of more than 50 potential predictor variables and six outcome variables (time points) and reduced it to a BN with only four predictor variables that produced reasonably good levels of classification accuracy. The graphical nature of the BN, as displayed in Netica, gives any potential user, from academic counselors to faculty to administers, the ability to quickly estimate a student's probability of stopping out prior to a given term by only knowing fives bits of information – whether the student primarily attended courses in the day during the prior term, whether he/she received need-based aid the prior term, what his/her prior term GPA was, the number of credit hours he/she attempted the prior term, and what their next term is – and clicking the corresponding values in the appropriate nodes in the BN. This makes the results accessible to a broader audience of front-line

staff who are not researchers but have the ability to apply the results to make data-informed decisions aimed at increasing student success. For example, let us assume an academic counselor is scheduled to meet with two students after the completion of their first term at the institution to help them plan their schedule for the upcoming term. In preparation for the meetings, the counselor pulls up each student's information in the institution's student information system. The counselor sees that the first student attempted 15 credits during her first term, primarily attending courses during the day, earned a 3.5 GPA, and did not receive need-based aid. He opens the BN in Figure 6 in Netica and clicks the appropriate node values for the information he has just pulled on the student. The model uses the inputs to almost instantaneously estimate that the student's probably of stopping out prior to the next term is 10%. He then enters in the information for the second student. This student is more of a non-traditional student in that he only attended part time the prior term (attempting 6 credit hours), mostly during the evening. He received need-based aid and also earned a 3.5 GPA. With five quick clicks of a mouse the counselor sees that the student's predicted probability of stopping out prior to next term is 64%. The counselor is surprised to see that the predicted probability of the second student stopping out is approximately 6.5 times greater than the predicted probability of the first student stopping out even though they had the same GPA. He knows (based on training from his local institutional research office) that the results are not deterministic "truths" but are rather probabilistic estimates that simply provide one additional bit of information to help inform his meetings with both students. Based on this information he decides to spend some extra time with the second student inquiring about his experiences during the prior term in search of corroborating or disconfirming evidence that the

student might be danger of stopping out. For example, he may discover that the student is working full time to pay for tuition and thus is only able to enroll part time during the evening. Armed with this information the counselor could then try to pair the student with resources (e.g., scholarship opportunities, on-campus jobs with flexible schedules, etc.) to help him remain enrolled and successful at the institution.

Classification accuracy. There were some similarities and differences in in-sample classification accuracy rates between the differing modeling techniques that are worthy of brief discussion. Comparing the fall 2009 rates for the non-discretized final model, the overall (79%), marginal (92% did not stop out; 47% stopped out), and adjusted (49%) classification rates were equivalent for the frequentist survival analysis and its Bayesian analog. This was expected since the latter utilized uninformed priors and therefore based its estimates on the same framework (logistic regression) and data used in the frequentist survival analysis. Of more interest is the comparison of the average in-sample classification rates across the fall 2009 and 2010 samples between the modified (discretized) final model estimated via the frequentist survival analysis and the same model estimated using a BN. The overall, marginal, and adjusted classification rates for the frequentist survival analysis were 78%, 91% (did not stop), 47% (stopped out), and 49%, respectively. The corresponding average in-sample rates for the BN were 84%, 94% (stopped out), 59% (did not stop out), and 61%, respectively. This translates into the BN being approximately 7% better at classifying students overall, 3% better at predicting who would not stop out, and 26% better at predicting who would stop out. In terms of adjusted classification accuracy, the BN produced a 23% improvement over the

frequentist survival analysis model. The latter two figures represent large, material differences in the two classification rates.

For the reduced, modified survival analysis model, the magnitudes of the differences in in-sample classification accuracy rates shrank but were still large enough to be meaningful. The reduced BN was approximately 3% better at classifying students overall, showed no improvement predicting who would not stop out, was 13% better at predicting students who would stop out, and had an improvement of 10% in terms of adjusted classification rate. The most noticeable difference across both models was in terms of correctly classifying students who stopped out. This is also the most important difference since the primary goal of the model was to predict students in danger of stopping out. Increasing the accuracy of predicting these students by 26% and 13% – averaged across to separate samples – is a sizable increase worthy of note. It is unknown if this is an anomaly of the specific model and/or samples utilized or if this is indicative of a more generalized benefit of adopting a BN for modeling a discrete-time survival analysis with discrete predictors. This is an area that warrants additional research.

An attempt was made to contextualize the in-sample classification accuracy rates based on the results found in prior research. Surprisingly and unfortunately, out of all the similar studies reviewed and referenced in Chapter 3, only one published the classification accuracy rates. Radcliffe, Huesman, & Kellogg (2006) utilized a discrete-time survival analysis to predict attrition of student athletes at a U.S. university. The resulting model correctly classified 72% of students overall, 67% of students who dropped out, and 73% of student who were retained. Comparatively, the results of the current study produced higher overall and marginal classification rates for students who

did not stop out/were retained, but had lower marginal rates for students who stopped/dropped out. Of course, a single study is not sufficient to adequately contextualize the results of this study, especially considering they are applied to different populations (student athletes at a university compared to degree- or transfer-seeking students at a community college). This can only be done if and when more authors of similar studies make a concerted effort to publish classification rates.

The classification rates for the reduced BN were also shown to be resilient when predicting stopping out behaviors for students who were not in the sample used to learn the model CPTs. The out-of-sample overall, marginal, and adjusted classification rates averaged across both samples were 78%, 90% (did not stop out), 49% (stopped out), and 49%, respectively. These were only slightly lower than the in-sample overall, marginal, and adjusted classification rates averaged across both samples of 80%, 91% (did not stop out), 52% (stopped out), and 53%. In comparative terms, these represent decreases of 3%, 1% (did not stop out), 6%, and 8%, respectively. These relatively small decreases support the generalizability of the model across samples.

Substantive Conclusions

The current study aimed to produce a parsimonious model for predicting whether degree- or transfer-seeking students would stop out following each of their first six semesters at a community college. The iterative model-building process whittled down a pool of more than 50 potential predictors to a final set of four predictors through the combination of traditional and non-traditional techniques.

The first of the four final predictors was whether a student primarily attended courses in the day during the prior term. The results of the logistic regression-based

survival analysis for the fall 2009 sample indicated that the odds of stopping out were 1.56 higher for students who primarily did not take their courses during the day the prior term, holding all other variables in the model constant. This makes sense when considered in the context that students typically take courses in the evening because their time during the day is filled with other activities such as work or caring for dependents. Attending courses in the evening is not believed to be the direct cause of stopping out, but rather is a variable the institution has access to that serves as a proxy for a host of unobserved and unavailable variables (e.g., competing commitments between school and work, diminished time to focus on studies, etc.) that have an effect on whether students remain enrolled at the institution. Surprisingly, none of the prior literature reviewed as part of Chapter 2 used primary time of attendance as a predictor.

The second predictor was the number of credits a student attempted the prior term (excluding credits the student withdrew from). The more credits a student attempted the prior term, the less likely they were to stop out, holding all other variables in the model constant. Compared to the reference group of students who attempted 0.5 – 3.0 credits, the odds ratios based on the fall 2009 sample for students who took 3.5 – 6.0, 6.5 – 9.0, 9.5 – 12.0, and more than 12 credits were 0.52, 0.33, 0.24, 0.13, respectively. This means that the odds of stopping out for a student who attempted more than 12 credits the prior term was 87% less compared to the odds of stopping out for a student who only attempted 0.5 – 3.0 credits. That is a substantial difference especially considering this variable only refers to the number of credits attempted, not the number of credits passed. It is clear from the results that a student's enrollment intensity is strong indicator of whether a student will stop out prior to the next term for the samples evaluated. These

results are similar to those found by Gross, Torres, and Zerquera (2013) and Calcagno, Crosta, Bailey, and Jenkins (2007). Gross et al., found the annual number of credits attempted (modeled as a continuous variable) to be inversely related to the probability of stopping out as part of a broader competing risks model. Calcagno et al. modeled dropping out as part of a competing risks model and found student enrolled full-time had significantly lower odds of dropping out.

The third predictor was whether a student received need-based aid the prior term. Need-based aid was defined as Pell grants and/or subsidized federal student loans. Using data from the fall 2009 sample, the odds of stopping out for students who received such aid the prior term was 34% lower than students who did not receive similar aid, holding all other variables in the model constant. There are several possible explanations for this. One is that the receipt of need-based aid may have alleviated some or all of the financial burden of attending college. This may have manifested itself in requiring these students to work less, freeing up more time for them to focus on activities associated with not stopping out (e.g., studying, which leads to higher GPAs, which in turn is associated with persistence). Another possible explanation may be that the students who are able to successfully navigate the process of filing for Pell grants and/or federal subsidized student loans are also able to successfully navigate the "process" of attending college. Successfully applying for need-based aid and successfully navigating the college environment both require the capacity to deal with at times complex bureaucratic processes that require attention to detail, follow through, and an adherence to externally mandated deadlines. It is possible that the receiving need-based aid serves as a proxy for a student's ability to succeed in such environments. Alternatively, receiving need-based

aid may serve as an indirect indicator of student motivation since students who received such aid had enough motivation to apply for it. It is reasonable to assume some proportion of students who did not receive need-based aid would have qualified for such aid but were not motivated enough to apply for it. These students may also be less motivated to persist. Although there is a collection of prior research that has evaluated the relationship between various specifications of persistence, stopping out, or dropping out and various types of financial aid (Chen & DesJardins, 2008; DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013; Ishitani, 2006; Johnson, 2006), none of the prior studies reviewed specifically looked at students who received need-based aid as defined by this study. The closest comparison was the work of Chen and DesJardins (2008) who found no significant main effect between students who received a Pell grant and dropping out.

The last predictor of the four that were found to be significant predictors of stopping out behavior was a student's GPA for the prior term. The higher a student's GPA was the prior term, the less likely they were to stop out. Based on the fall 2009 sample, students who had GPAs of 1.1 – 2.0, 2.1– 3.0, and 3.1 – 4.0 the prior term had odds ratios of 0.49, 0.30, and 0.24, respectively, compared to the reference group of students who had a prior term GPA of 0.0 – 1.0. The direction of these results is intuitive and is supported by prior research (Chen & DesJardins, 2008; DesJardins, Ahlburg, & McCall, 2002; Gross, Torres, & Zerquera, 2013; Ishitani, 2003; Johnson, 2006; Murtaugh, Burns, & Schuster, 1999). We would expect students with lower GPAs to stop out at higher rates compared to students with higher GPAs since GPA is a key indicator of student success. A student who is not succeeding in the classroom is less likely to

remain enrolled at the institution for a variety of reasons. One reason is that students at most institution who have extremely low GPAs (e.g., 0.0 – 1.0) in successive semesters are barred from re-enrolling for lack of making satisfactory academic progress. Another reason is that few people continue to persist at something they are not successful at doing, especially when there are direct (e.g., tuition) and in-direct (e.g., lost wages) costs for engaging in the activity.

The four predictors did a reasonably good job at accurately predicting whether or not a student would stop out following each of their first six semesters at the participating community college. The final BN utilizing the four predictors was found to correctly classify 80% of students averaged across the fall 2009 and 2010 samples. The classification rate is respectable but needs to be evaluated by taking into account the expected agreement by "chance" alone. The average adjusted classification rate, as measured by Cohen's κ, was 53%, meaning the model was 53% better at correctly classifying students than is estimated could be achieved using no model at all. This meets the previously stated goal of the study to develop a parsimonious model that had an adjusted classification accuracy rate of greater than 50%.

The model-building process used to ultimately produce the final BN provided some other substantive discoveries along the way. First, no significant interactions were found between any of the predictors and outcome variable across all time points. The only significant interactions were between the number of credits attempted the prior term and prior term GPA with stopping out prior to the second semester. This provides evidence that the relationship between the predictors and outcome variable are relatively fixed over time with the caveat that prior term attempted credits and GPA may have a

greater effect predicting students who are likely to stop out during the transition period between their first to second semesters. Additionally, the results speak to the benefits of marrying both traditional and non-traditional perspectives and techniques when modeling student stopping out behaviors. Taking a purely traditional approach would have made it unwieldy to narrow down the more than 50 potential predictor variables into a parsimonious subset to use in the final model. While taking a purely data mining approach would have resulting in the exclusion of the variable primary time of attendance even though it was found to be a significant and meaningful predictor since none of the preliminary stepwise regression models identified it as such. It was only added, despite the stepwise regression results, based on the researcher's belief in its value as a predictor. This follows the traditional philosophy that variable selection is primarily dictated by the researcher based on prior research and informed judgment. Neither approach taken on its own would have successfully guided us to the same final model that resulted from this study. It was only through the blending of these two paradigms that the final, meaningful model was derived.

Limitations

The study had a number of key limitations. First, the size of the fall 2009 sample was limited due to the initial inclusion of 50+ predictor variables. This resulted in a large loss of students due to missing data, especially since the researcher only included students who had completed the required placement tests. By contrast the fall 2010 sample was more than two-and-half times larger due to the fact it only required pulling data on six predictor variables, none of which required students to have completed the placement tests. There are not believed to be any meaningful qualitative differences

between the two samples, although it is possible there were some differences and that had a material effect on the early exploratory analyses and resulting models.

Second, as mentioned earlier, the intent of the study was to provide an easy to follow path for those currently unfamiliar with Bayesian analyses. A drawback of this approach is that it did not utilize more sophisticated Bayesian-based survival analyses (Ibrahim, Chen, & Sinha, 2001). Of particular disappoint was the inability to estimate a dynamic model as original intended. It is possible that one of those techniques would have resulted in a more accurate and meaningful model than the one produced by this study.

Third, the study did not engage in any advanced data mining techniques for determining the cut off points for discretizing continuous data (e.g., ChiMerge; Kantardzic, 2011). The cut offs were instead selected based on interpretive considerations or for ease of use (e.g., quartiles). Utilizing data mining techniques may have resulted in enhancing the predictive power of a variable or aided in the identification of a predictor that was otherwise determined to have a non-significant relationship with the outcome variable. For example, it is possible that specifying the number of days a student registered for their first course the prior term in quartiles, as was done, did not maximize its predictive potential. There may be have been a better, non-intuitive cut off, such as dichotomizing the variable around the value of a specific day threshold (e.g., 8 days before the start of a course), that could have been identified through the aid of a data mining algorithm.

Fourth, the study only included data that were readily available to the researcher. Other variables have been shown to be significant predictors of college persistence, such

as high school GPA (Ishitani, 2003; Murtaugh, Burns, & Schuster, 1999) and engagement in an institution's learning management system (Campbell, 2007), but were not obtainable for this study.

Fifth, the current study did not evaluate any potential interactions between predictor variables. This was driven by a desire to keep the study somewhat constrained in scope so as not to distract from the primary focus of developing a model building process that utilizes traditional and non-traditional techniques to model student progression. It is possible that there were significant predictor-by-predictor interactions in the sample data that could have enriched the meaningfulness and/or classification accuracy of the final model. For similar reasons, the model treated students who graduated or transferred prior to a given term as censored. Strictly speaking, students should only be treated as censored if the missing dependent variable values for those students are assumed to be the result of the study's design and independent of the outcome (Singer & Willett, 2003). That was not the case in this study. In such instances a competing risk model is recommended (Scott & Kennedy, 2005). The desire to keep the study focused in scope also lead to the decisions to not impute missing data and to only use uninformed priors for the fully Bayesian analysis. Both of these represent limitations of the study.

Lastly, although the final BN produced respectable levels of classification accuracy, it still misclassified 48% of students who stopped out, averaging across the fall 2009 and 2010 samples. In other words, the model failed to correctly identify almost half of the students of most import. This level of marginal accuracy needs to be improved before the model can be more broadly and accurately used as a tool to benefit students by

identifying those likely to stop out and providing them with the additional resources needed in an attempt to prevent them from actually do so. More importantly, the accuracy rates need to be compared to the accuracy rates produced by the status quo method of reasoning in the face of uncertainty. In other words, the utility of the model is not how well it does compared to an ideal (100% agreement) but rather whether it provides meaningful improvement to an institution's current ability to accurately and efficiently predict students' stopping out behaviors. For example, how much more accurately and efficiently, if at all, can academic advisors predict whether a student will stop out in a given term with the aid of the BN compared to making the judgment without it? This is an area that deserves future exploration.

Recommendations for Future Methodological Research

This study laid the groundwork for several lines of future inquiry. One area for future research is to compare the classification accuracy rates of logistic regression-based discrete-time survival analysis with discretized predictors and those from an equivalent BN for a wide range of models and samples to examine whether the latter consistency produces increased level of classification accuracy even when the models and data are fundamentally the same, as was found in this study. There are several lines of inquiry this might follow. For example, one line of inquiry might be to investigate whether the differences are the result of one or more of the predictors in the model (indicating the difference is more a function of specific predictors than the modeling approaches); whereas another line of inquiry might examine whether the differences manifest from distinctions in machinery of the different modeling techniques. Additionally, it would be of merit to explore the effect different cut off values for

classifying student's predicted outcome has on classification accuracy. This study used a predicted probability of >= .50 to classify student's predicted outcome as stopping out. This assumes an equal cost of misclassifying students as false positives (predicting students will stop out when in fact they do not) and false negatives (predicting students will not stop out when they do). It is arguable that there is a greater cost associated with false negatives since it could result in students at risk of stopping out being missed by an early alert system meant to identify students in need of additional resources. The effect of various cut scores should be explored in this context. Another channel for additional research is to extend the current research to multinomial (i.e., competing risks) discrete-time survival analysis and associated BNs for outcomes with more than two categories (e.g., Scott & Kennedy, 2005). For example, separately modeling the probably of a student stopping out, persisting, graduating, or transferring prior to a given term. Future research should also explore more advanced methods for accounting for missing data (e.g., Enders, 2010), as well as more thoroughly explore the relationships between predictor-by-time interactions and predictor-by-predictor interactions and stopping out behaviors. Although the current study did not find the widespread presence of significant predictor-by-time interactions, it did find evidence suggesting prior term GPA and number of credits attempted may have a greater effect predicting students who are likely to stop out during the transition period between their first to second semesters. Further investigation is called for to more definitively determine the nature and magnitude of the effect, if any. Lastly, there is a need to create a similar model-building process for more advanced Bayesian-based survival analyses (e.g., Ibrahim, Chen, & Sinha, 2001),

including the use of informed priors and DBNs, in a way that is readily digestible for researchers most familiar with frequentist methodologies.

References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*(4), 341-359.

Andrieu, S. C., St. John, E. P. (1993). The influence of prices on graduate student persistence. *Research in Higher Education*, *34*(4), 399-425.

Araque, F., Roldan, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*, *53*, 563-574.

Arizona Community Colleges. *Arizona community colleges: Long-term strategic vision.* Retrieved from http://www.arizonacommunitycolleges.org/strategicPlan/strategicPlanAzCC.pdf

Arnold, K. (2010, March 3). Signals: Applying academic analytics. *EDUCAUSE Review Online*. Retrieved from http://www.educause.edu/ero/article/signals-applying-academic-analytics

Astin, A. W. (1964). Personal and environmental factors associated with college dropouts among high aptitude students. *Journal of Educational Psychology*, *55*(4), 219-227.

Astin, A. W. (1975). *Preventing students from dropping out*. San Francisco, CA: Jossey-Bass.

Astin, A. W. (1997). How "good" is your institution's retention rate? *Research in Higher Education*, *38*(6), 647-658.

Bahr, P. R. (2008). Cooling out in the community college: What is the effect of academic advising on students' chances of success? *Research in Higher Education*, *49*(8), 704-732.

Bahr, P. R. (2012). Student flow between community colleges: Investigating lateral transfer. *Research in Higher Education*, *53*, 94-121.

Baker, R. S. J. D. (2010). Data mining. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education* (3rd ed.; pp. 112-118). Oxford, UK: Elsevier.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*, 3-17.

Bers, T., H., & Smith, K. E. (1991). Persistence of community college students: The influence of student intent and academic and social integration. *Research in Higher Education*, *32*(5), 539-556.

Bianchi, J. R., Bean, A. G. (1980). The prediction of voluntary withdrawals from college: An unsolved problem. *The Journal of Experimental Education, 49*(1), 29-33.

Bogdan, R. C., & Knopp Biklen, S. (2007). *Qualitative research for education: A introduction to theory and methods* (5th ed.). Boston: Pearson Education.

Bogdan Eaton, S., & Bean, J. P. (1995). An approach/avoidance behavioral model of college student attrition. *Research in Higher Education, 36*(6), 617-645.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York, NY: John Wiley & Sons.

Braxton , J. M., Doyle, W. R., Hartley III, H. V., Hirschy, A. S., Jones, W. A., & McLendon, M. K. (2014). *Rethinking college student retention*. San Francisco, CA: Jossey-Bass.

Braxton, J. M., Hirschy, A. S., McClendon, S. A. (2004). *Understanding and reducing college student departure*. ASHE-ERIC Higher Education Report: Volume 30, Number 3. San Francisco, CA: Jossey-Bass.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199-231.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434.

Caison, A. L. (2005). Determinants of systemic retention: Implications for improving retention practice in higher education. *Journal of College Student Retention, 6*(4), 425-441.

Caison, A. L. (2007). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education, 48*(4), 435-451.

Calcagno, J. C., Crosta, P., Bailey, T., & Jenkins, D. (2007a). Does age of entrances affect community college completion probabilities? Evidence from a discrete-time hazard model. *Educational Evaluation and Policy Analysis, 29*(3), 218-235.

Calcagno, J. C., Crosta, P., Bailey, T., & Jenkins, D. (2007b). Stepping stones to a degree: The impact of enrollment pathways and milestones on community college student outcomes. *Research in Higher Education, 48*(7), 775-801.

Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study.* Purdue University). *ProQuest Dissertations and Theses*, , 219. Retrieved from http://login.ezproxy1.lib.asu.edu/login?url=http://search.proquest.com/docview/304837810?accountid=4485. (304837810).

Campbell, J. P., DeBlois, P. B., Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE Review*, *42*(4), 40-57.

Campbell, C. K., & Fuqua, D. R., (2009). Factors predictive of student completion in a collegiate honors program. *Journal of College Student Retention*, *10*(2), 129-153.

Carnevale, A.P., Smith, N., & Strohl, J. (2013). *Recovery: Job growth and education requirements through 2020.* Washington, DC: Georgetown University Center on Education and the Workforce.

Chacon, F., Spicer, D., Valbuena, A. (2012). *Analytics in support of student retention and success. Educause Center for Applied Research* (Research Bulletin 3). Louisville, CO: EDUCAUSE Center for Applied Research. Retrieved from http://www.educause.edu/library/resources/analytics-support-student-retention-and-success

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, *12*(4), 50-63.

Chen, R., & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education*, *49*(1), 1-18.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen J., Cohen, P., West, S. G., Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

College Board Advocacy & Policy Center. *The College Completion Agenda*. Retrieved from http://completionagenda.collegeboard.org/sites/default/files/reports_pdf/

Crawford, A. B. (1930). Forecasting freshman achievement. *School and Society*, *31*, 125-32.

Crisp, G., & Nora, A. (2010). Hispanic student success: Factors influencing the persistence and transfer decisions of Latino community college students enrolled in developmental education. *Research in Higher Education*, *51*, 175-194.

Dean, T., & Kanazawa, K. (1988). A model for reasoning about persistence and causation. *Computational Intelligence*, *5*, 142-150.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of Student Retention*, *13*(1), 17-35.

Denzin, N. K. & Lincoln, Y. S. (2000). *Introduction: The discipline and practice of qualitative research*. In N. K. Denzin, Y. S. Lincoln (Eds.), Handbook of qualitative research (2nd ed.; pp. 1-28). Thousand Oaks, ca: Sage.

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

DesJardins, S. L., McCall, B. P., Ahlburg, D. A., & Moye, M. J. (2002). Adding a timing light to the "tool box". *Research in Higher Education*, *43*(1), 83-114.

Driscoll, M. (2004). *Psychology of learning for instruction* (3rd ed.). Boston, MA: Allyn & Bacon.

Edgerton, H., A., & Troops, H. A. (1929). Academic progress. *Contributions in Administration*, *1*, 150.

Elias, T. (2011). *Learning analytics: Definitions, processes, and potential*. Retrieved from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Feldman, M. J. (1993). Factors Associated with one-year retention in a community college. *Research in Higher Education*, *34*(4), 503-512.

Finnegan, C., Morris, L. V., & Lee, K. (2008). Differences by course discipline on student behavior, persistence, and achievement in online courses of undergraduate general education. *Journal of College Student Retention*, *10*(1), 39-54.

Flores, S. M., & Horn, C. L. (2009). College persistence among undocumented students at a selective public university: A quantitative case study analysis. *Journal of College Student Retention*, *11*(1), 57-76.

Folly Nicpon, M., Huser, L., Hull Blanks, E., Sollenberger, S., Befort, C., Robinson Kurpius, S. E. (2006). The relationship of loneliness and social support with college freshmen's academic performance and persistence. *Journal of College Student Retention*, *8*(3), 345-358.

Freeman, F. S. (1931). Predicting academic survival. *The Journal of Educational Research*, *23*(2), 113-123.

Garrett, H. F. (1949). A Review and interpretation of investigations of factors related to scholastic success in college of arts and sciences and teachers colleges. *The Journal of Experimental Education*, *18*(2), 91-138.

Ghanmi, N., Mahjoub, M. A., Ben Amara, N. E. (2011). Characterization of dynamic Bayesian network. *International Journal of Advanced Computer Science and Applications*, *2*(7), 53-60.

Gill, J. (2009). *Bayesian methods: A social and behavioral sciences approach* (2nd Ed.). Boca Raton, FL: Chapman & Hall/CRC.

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2005). Configural invariance of a model of student attrition. *Journal of College Student Retention*, *7*(3-4), 263-281.

Goldstein, P. J., & Katz, R. N. (2005). *Academic analytics: The uses of management Information and technology in higher education*. Retrieved from the EDUCAUSE Center for Applied Research (ECAR) website: http://www.educause.edu/library/resources/academic-analytics-uses-management-information-and-technology-higher-education

Gross, J. P. K., Torres, V., Zerquera, D. (2013). Financial aid and attainment among students in a state of changing demographics. *Research in Higher Education*, *54*, 383-406.

Grush, M. (2012, December 5). Opening up learning analytics for the community college: A Q&A with Josh Baron and JoAnna Schilling. *Campus Technology*. Retrieved from www.campustechnology.com.

Habley, W. R., Bloom, J. L., Robbins, S. (2012). *Increasing persistence: Research-based strategies for college student success*. San Francisco, CA: Jossey-Bass.

Hagedorn, L. S., Maxwell, W., & Hampton, P. (2001). Correlates of retention for African-American males in community college. *Journal of College Student Retention*, *3*(3), 243-263.

Hájek, A. (2012). Interpretations of probability. *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/win2012/entries/probability-interpret

Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York, NY: John Wiley & Sons.

Hald, A. (1998). *A history of mathematical statistics 1750 to 1930*. New York, NY: John Wiley & Sons.

Halpin, R. L. (1990). An application of the tinto model to the analysis of freshman persistence in a community college. *Community College Review, 17*, 22-32.

Hausmann, L. R. M., Schofield, J. W., & Woods, R. L. (2007). Sense of belonging as a predictor of intentions to persist among African American and white first-year college students. *Research in Higher Education, 48*(7), 803-839.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research, 131*, 17-33.

Hilbe, J.M. (n.d.). Generalized linear models (version 17). *StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies*. Retrieved from http://statprob.com/encyclopedia/GeneralizedLinearModels.html

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Ibrahim, J. G., Chen, M-H., & Sinha, D. (2001). *Bayesian survival analysis*. New York, NY: Springer.

Ishitani, T. T. (2003). A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college students. *Research in Higher Education, 44*(4), 433-449.

Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. *The Journal of Higher Education, 77*(5), 861-885.

Irvine, D. W. (1966). Multiple prediction of college graduation from pre-admission data. *The Journal of Experimental Education, 35*(1), 84-89.

Johnson, I. Y. (2006). Analysis of stopout behavior at a public research university: The multi-spell discrete-time approach. *Research in Higher Education, 47* (8), 905-934.

Johnson, I. Y., & Muse, W. B. (2012). Student swirl at a single institution: The role of timing and student characteristics. *Research in Higher Education, 53*, 152-181.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*. Hoboken, NJ: John Wiley & Sons, Inc.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Kolowich, S. (2013, January 24). Arizona St. and Knewton's grand experiment with adaptive learning. *Inside Higher Education*. Retrieved from www.insidehighereducation.com

Kopko, E., & Cho, Sung-Woo. (2013). Timing of concentration, completion, and exit in community colleges. *CCRC Analytics*. New York, NY: Community College Research Center. Retrieved from http://ccrc.tc.columbia.edu/publications/timing-of-concentration-completion-exit.html

Krotseng, M. V. (1992). Predicting persistence from the Student Adaption to College Questionnaire: Early warning or siren song? *Research in Higher Education, 33*(1), 99-111.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education, 79*(5), 540-563.

Landry, H. A. (1937). The relative predictive value of certain college entrance criteria. *The Journal of Experimental Education, 5*(3), 256-260.

LeSure-Lester, G. E. (2004). Effects of coping styles on college persistence decisions among Latino students in two year colleges. *Journal of College Student Retention, 5*(1), 11-22.

Levy, R. (2011). *Bayesian inference* [PowerPoint slides]. CDE 591: Bayesian Analyses in the Social Sciences course materials. Arizona State University.

Lillis, M. P. (2012). Faculty emotional intelligence and student-faculty interactions: Implication for student retention. *Journal of College Student Retention, 13*(2), 155-178.

Lowe, S. R., & Rhodes, J. E. (2012). Community college re-enrollment after hurricane Katrina. *Journal of College Student Retention, 14*(2), 227-249.

Lumina Foundation for Education. (2013). *A stronger nation through higher education: Visualizing data to help us achieve a big goal of college attainment*. Retrieved from

www.luminafoundation.org/publications/A_stronger_nation_through_higher_education-2013.pdf

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*, 950-965.

Manske, M., Conati, C. (2002). *Modelling learning in an educational game*. In Proceedings of AIED 2005: World conference on artificial intelligence & education, 12, 24-153.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education*, *40*(3), 355-371.

Marcus, J. (2012, November 29). Student advising plays key role in college success – just as it's being cut. *The Herchinger Report*. Retrieved from http://usnews.nbcnews.com

Mayo, M., & Mitrovic, A. (2001). Optimising ITS behavior with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, *12*, 124-153.

Meyer, L. (2012, October 30). Community college of Rhode Island to implement predictive analytics to identify at-risk students. *Campus Technology*. Retrieved from http://campustechnology.com/articles/2012/10/30/community-college-of-rhode-island-to-implement-predictive-analytics-to-identify-at-risk-students.aspx?=CTCC.

Millán, E., & Pérez de la Cruz, J.-L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*, 281-330.

Mislevy, R. J. & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, *5*, 253-282.

Mort, P. R. (1932). The general uses of psychological tests. *Review of Educational Research*, *2*(4), 300-307.

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, inference and learning.* Retrieved from ProQuest Dissertations & Theses. (UMI 3082340).

Murtaugh, P. A., Burns, L. D., Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(3), 355-371.

Nandeshaw, A., Menzies,T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38,* 14984-14996.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis & data mining applications*. Burlington, MA: Academic Press.

Norsys Software Corporation. (2013). Netica [software] (version 5.12). Retrieved from http://www.norsys.com/

O'Brien, F. P. (1928). Mental ability with reference to selection and retention of college students. *The Journal of Educational Research, 18*(2), 136-143.

Odell, C. W. (1930). Predicting the scholastic success of college students. *University of Illinois Bulletin, 28*(5), 43.

Organization for Economic Co-operation and Development (OECD). *Education at a Glance 2010: OECD Indicators*. Retrieved from http://www.oecd.org/edu/highereducationandadultlearning/educationataglance2010oecdindicators.htm#navigator

Oseguera, L. (2006). Four and six-year baccalaureate degree completion by institutional characteristics and racial/ethnic groups. *Journal of College Student Retention, 7*(1-2), 19-59.

Panos, R. J., & Astin, A. W. (1698). Attrition among college students. *American Educational Research Journal, 5*(1), 57-72.

Parry, M. (2011, December 11). Colleges mine data to tailor students' experience. *Chronicle of Higher Education*. Retrieved from http://chronicle.com

Parry, M. (2012, July 18). A conversation with 2 developers of personalized-learning software. *Chronicle of Higher Education*. Retrieved from http://chronicle.com

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from the theoretical model. *The Journal of Higher Education, 51*(1), 60-75.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco, CA: Jossey-Bass.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco, CA: Jossey-Bass.

Pearl, J. (2000). *Causality*. Cambridge, England: Cambridge University Press.

Pearl, J., & Russell, S. (2001). *Bayesian networks*. In M. Arbib (Ed.). Handbook of brain theory and neural networks (pp. 157-159). Cambridge, MA: MIT Press.

Peng, S. S., & Fetters, W. B. (1978). Variables involved in withdrawal during the first two years of college: Preliminary findings from the national longitudinal student of high school class of 1972. *American Educational Research Journal, 15*(3), 361-372.

Peng, C. J., & So, T. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics, 1*(1), 31-70.

Perrine, R. M. (2009). Impact of a pre-semester college orientation program: Hidden benefits? *Journal of College Student Retention, 10*(2), 155-169.

Pittman, K. (2008). *Comparison of data mining technique's used to predict student retention* (dissertation). Retrieved from ProQuest Dissertations & Theses. (UMI 3297573).

Potthoff, E. F. (1931). Predicting the ultimate failure of college students on the basis of their fist quarter's records. *School and Society, 33*, 203-4.

Prediger, D. J. (1966). Application of moderated scoring keys to prediction of academic success. *American Educational Research Journal, 3*(2), 105-111.

Radcliffe, P., Huesman, R., & Kellogg, J. (2006). *Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis*. Paper presented at the annual meeting of the Association for Institutional Research in the Upper Midwest (AIRUM).

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education, 14*, 63-96.

Robinson, L. F. (1969). Relation of student persistence in college to satisfaction with "environmental" factor. *The Journal of Educational Research, 63*(1), 6-10.

Rohr, S. L. (2013). How well does the SAT and GPA predict the retention of science, technology, engineering, mathematics, and business students. *Journal of College Student Retention, 14*(2), 195-208.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, *40*(6), 601-618.

Ronco, S. L. (1994). *Meandering ways: Studying student stop out with survival analysis*. Paper presented at the annual Forum of the Association for Institutional Research, New Orleans, LA

Schmid J., & Reed, S. R. (1966). Factors in retention of residence hall freshmen. *The Journal of Experimental Education*, *35*(1), 28-35.

Scott, M. A., & Kennedy, B. B. (2005). Pitfalls and pathways: Some perspectives on competing risks event history analysis in higher education. *Journal of Educational and Behavioral Statistics*, *30*(4), 413-442.

Seidman, A. (2012). *College student retention* (2nd ed.). Lanham, MD: Rowman & Littlefield.

Siemens, G., (2012, May). *Learning analytics: Envisioning a research discipline and a domain of practice*. Paper presented at the Second International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia. Retrieved from http://learninganalytics.net/LAK_12_keynote_Siemens.pdf

Siemens, G. (2011, August 5). *Learning and academic analytics*. [blog post]. Message posted to http://www.learninganalytics.net/?p=131

Siemens, G. (2010, August 25). *What are learning analytics?* [blog post]. Message posted to http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

Siemans, G., & Baker, R. S. J. d. (2012). *Learning analytics and educational data mining: Towards communication and collaboration*. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.

Shapiro, D., Dundar, A., Chen, J., Ziskin, M., Eunkyoung, P, Torres, V., Chiang, Y. (2012). *Completing college: A national view of student attainment rates.* National Student Clearinghouse Research Center. Retrieved from http://www.studentclearinghouse.info/signature/4/

Shute, V. J. (2011). *Stealth assessment in computer-based games to support learning*. In S. Tobias & J. D. Fletcher (Eds.), Computer games and instruction (pp. 503-524). Charlotte, NC: Information Age Publishing.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Sutton, S. C., & Nora, A. (2008). An exploration of college persistence for students enrolled in web-enhanced courses: A multivariate analytic approach. *Journal of College Student Retention*, *10*(1), 21-37.

Swenson Goguen, L. M., Hiester, M. A., & Nordstrom, A. H. (2010). Associations among peer relationships, academic achievement, and persistence in college. *Journal of College Student Retention*, *12*(3), 319-337.

Tabachnick, B. G., Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Taroni, F., Aitken, C., Garbolino, P, & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. West Sussex, England: John Wiley & Sons.

Thurstone, L. L. (1921). A cycle-omnibus intelligence test for college students. *The Journal of Educational Research*, *4*(4), 265-278.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89-125.

Tinto, V. (1994). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago, IL: The University of Chicago Press.

Tinto, V. (2012). *Completing college: Rethinking institutional action*. Chicago, IL: The University of Chicago Press.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*(2), 83-91

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

United States White House. *Education: Knowledge and skills for the jobs of the future.* Retrieved from http://www.whitehouse.gov/issues/education/higher-education

United States Bureau of Labor Statistics (2013, May 22). *Earnings and unemployment rates by educational attainment*. Retrieved from http://www.bls.gov/emp/ep_chart_001.htm

van Barneveld, A., Arnold, K. E., Campbell, J. P. (2012). *Analytics in higher education: Establishing a common language*. EDUCAUSE Learning Initiative. Retrieved from http://www.educause.edu/library/resources/analytics-higher-education-establishing-common-language

van Gerven, M. A. J., Taal, B. G., & Lucas, P. J. F. (2008). Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of Biomedical Informatics*, *41*, 515–529.

Wagner, E., & Ice, P. (2012, July-August). Data changes everything: Delivering on the promise of learning analytics in higher education. *EDUCAUSE Review*, 33-42. Retrieved from http://www.educause.edu/ero/article/data-changes-everything-delivering-promise-learning-analytics-higher-education

Wang, X. (2009). Baccalaureate attainment and college persistence of community college transfer students at four-year institutions. *Research in Higher Education*, *50*(6), 570-588.

Weiss, N. A. (2008). *Elementary Statistics (*7th ed.). San Francisco, CA: Pearson Addison-Wesley.

Welsh, J. F., Petrosko, J., & Taylor, H. (2006). The school-to-college transition in the context of educational reform: student retention and the state of policy process. *Journal of College Student Retention*, *8*(3), 307-324.

Willett, J. B., & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research*, *61*(4), 407-450.

Winkler, R. L. (1972). *Introduction to Bayesian inference and decision*. New York, NY: Holt, Rinehart & Winston.

Wishon, G. D., & Rome, J. (2012, August 13). Enabling a data-driven university. *EDUCAUSE Review Online*. Retrieved from http://www.educause.edu/ero/article/enabling-data-driven-university

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Elsevier.

Yacef, K., Baker, R. S. J. D., Barnes, T., & Beck, J. E. (2009). Editorial welcome. *Journal of Educational Data Mining*, *1*(1) 1-2.

Young, J. R. (2011, April 10). The Netflix effect: When software suggests students' courses. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/The-Netflix-Effect-When/127059/

Zapata-Rivera, J-D., Greer, J. (2004). Inspectable Bayesian student modelling servers in multi-agent tutoring systems. *International Journal of Human-Computer Studies*, *61*, 535-563.

Zhang, Y., Oussena, S., Clark, T., Hyensook, K. (2010, June). *Using data mining to improve student retention in HE: a case study*. Paper presented at the 12th International Conference on Enterprise Information Systems, Portugal, Spain.

Zwick, R., & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, *42*(3), 439-464.

Footnotes

[1]There are differences of opinion on the definitions of retention and persistence. For the sake of simplicity, the two terms will be used interchangeably in this paper to mean a continuing student who re-enrolls at the same institution the following semester.

[2]Technically speaking, directed acyclic graphs (DAGs) are more accurately expressed as acyclic directed graphs (ADGs) according to graphing theory (Almond, DiBello, Moulder, & Zapata-Rivera, 2007); however, DAG is the more common expression and therefore will be used throughout this paper.

APPENDIX A

A STEP-BY-STEP ILLUSTRATION OF A BAYESIAN NETWORK

Figure A.1 provides a hypothetical Bayesian network. The BN was created using the software program Netica. Node *A* is the probability a student suffers from math anxiety {A_1 = No, A_2 = Yes}. Node *B* is the probability a student passed an algebra course in high school {B_1 = No, B_2 = Yes}, which is hypothesized to be dependent on whether she has a math anxiety. Node *C* represents whether a student placed into college-level algebra course based on her placement test score {C_1 = No, C_2= Yes}. It is also believed to be dependent on the presence or absence of a math anxiety. Node *E* indicates the probability a student in a college algebra course will get assistance from the tutors in the college's math center {E_1 = No, E_2 = Yes}. Finally, Node *D* is the probability a student will pass or fail a college algebra course with a grade of C or better {D_1 = No, D_2 = Yes}. The structure makes explicit that success in a college algebra course is dependent on a student's performance on the placement exam and whether he/she gets help from a tutor.

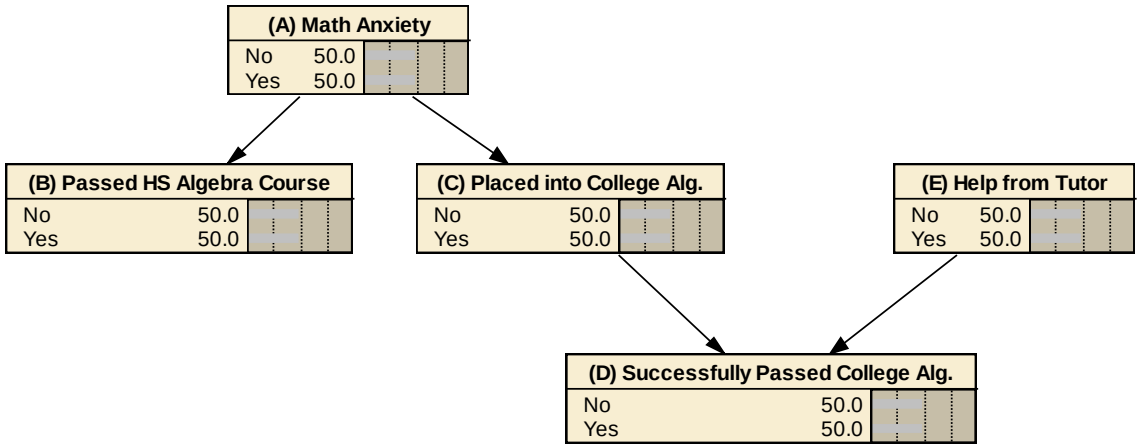


Figure A.1. *A hypothetical Bayesian network of success in a college-level algebra course.*

Before the collection of any information, marginal (prior) distributions need to be specified for the root nodes (*A*, *E*) and joint probabilities for the non-root nodes (*B, C*, *D*). To illustrate this point, the marginal and joint probabilities for the fictitious BN in figure A.1 are presented in Table A.1.

Priors can come from a variety of sources. In this instance the priors were "informed" based on the advisor's 15-year experience advising college students. To highlight a few of the probabilities, the advisor estimates that one out of every five students has math anxiety. He estimates that only 10% of students with math anxiety have passed a high school algebra course compared to 70% of students without math anxiety. In his experience, math anxiety also adversely influences a student's performance on the placement exam. Additionally, he has discovered that tutors can have a significant positive impact on a student's ability to successfully pass a college algebra course. He believes that students who do not place into a college-level algebra course but utilize a tutor have a 50% chance of successfully passing the course compared to only a 10% chance for those who do not place into the course and do not use a tutor. Unfortunately, based on his experience, only a small percentage of students (2%) seek help from a tutor.

Table A.1

Marginal and joint probabilities for the Bayesian network displayed in figure A.1.

Math Anxiety	
No	Yes
0.8	0.2

	Passed HS Algebra Course	
Math Anxiety	No	Yes
No	0.3	0.7
Yes	0.9	0.1

	Placed into College Alg. Course	
Math Anxiety	No	Yes
No	0.4	0.6
Yes	0.95	0.05

Placed into College Algebra Course = No		
	Success in College Alg. Course	
Help from Tutor	No	Yes
No	0.9	0.1
Yes	0.5	0.5

Placed into College Algebra Course = Yes		
	Success in College Alg. Course	
Help from Tutor	No	Yes
No	0.25	0.75
Yes	0.1	0.9

Help from Tutor	
No	Yes
0.98	0.02

Figure A.2 represents the conditional probabilities for each node based on the prior and joint probability distributions in Table A.1. According to the initial probabilities in Node *C*, a student has a 42.4% chance of successfully passing a college-level algebra

course based solely on prior beliefs about the variables in the model. The advisor then receives a copy of the student's high school transcript and sees that she did not pass a high school college algebra course. Knowing this, the outcome of *B* becomes known and is fixed to "No." This information propagates throughout the network and decreases the probability of passing the college-level algebra course to 34.3% (see figure A.3.a). Next, the advisor finds out in talking to the student that she does not have math anxiety. This information is entered into the model and results in increasing the probability of success to 49.5% (figure A.3.b). This increase in probability illustrates that a student's passing grade in a high school algebra course is no longer relevant (at least in terms of this model) to the student's probability of successfully passing a college-level algebra course once it is known that she does not have a math anxiety because knowing *A* makes *B* and *D* conditionally independent. The advisor reviews the student's placement test scores and determines that she did not place into the algebra course, decreasing the probability of success in a college-level algebra course all the way down to10.8% (figure A.3.c). As a result, he recommends that she enroll in a pre-algebra course. However, he is impressed by her stated level of motivation and determination to succeed (two variables not included in the model), so he tells her that she may enroll in a college algebra course if she feels she can handle the challenge as long as she agrees to see a tutor three times a week at the college's math center. Based on the BN model, this would increase her chance of success to 50% (figure A.4.d).

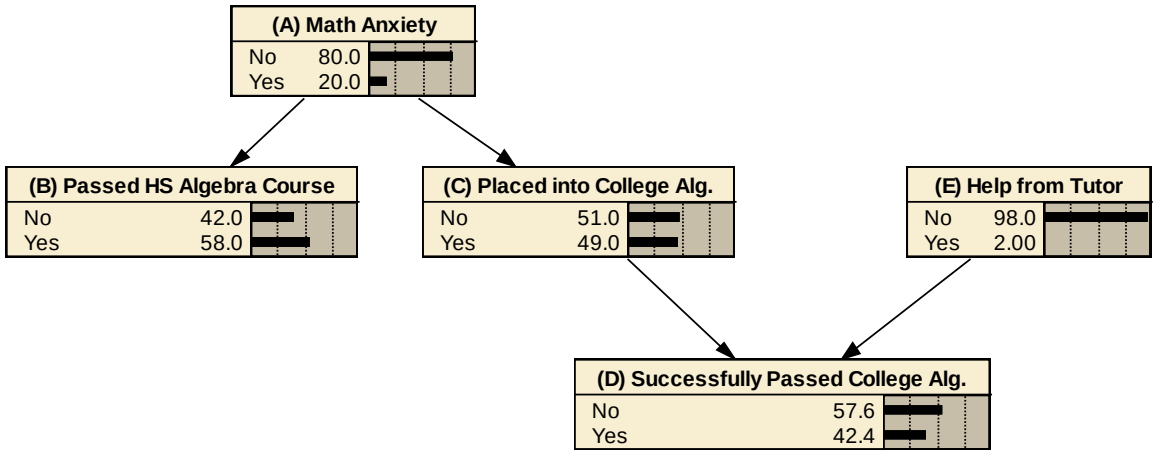
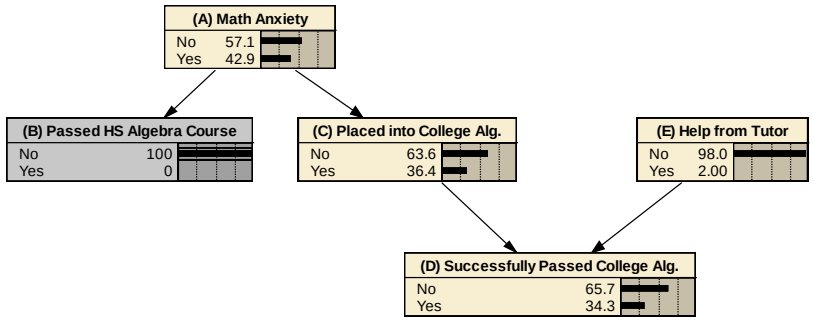
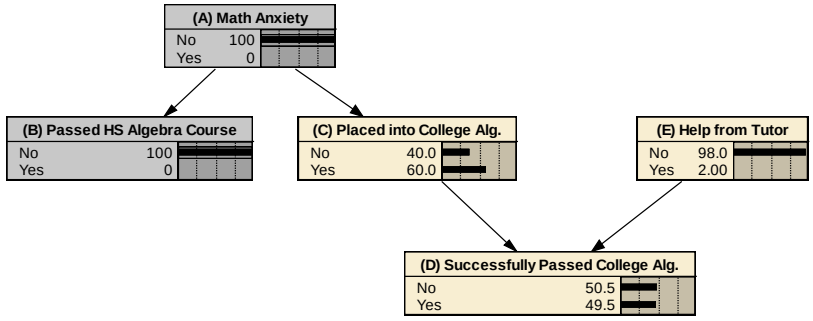


Figure A.2. *A hypothetical Bayesian network of success in a college-level algebra course after incorporating in prior and joint probability distributions.*

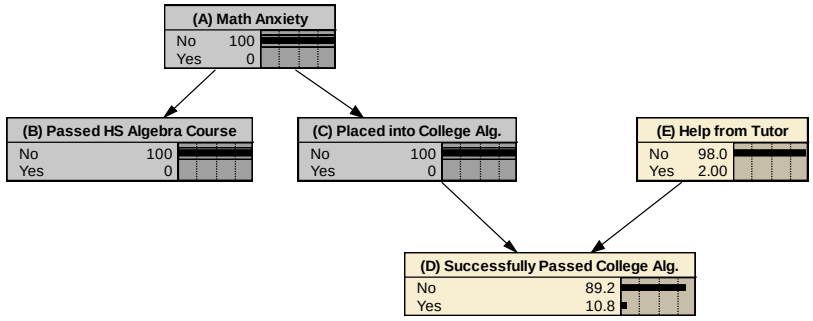
A.3.a



A.3.b



A.3.c



A.3.d

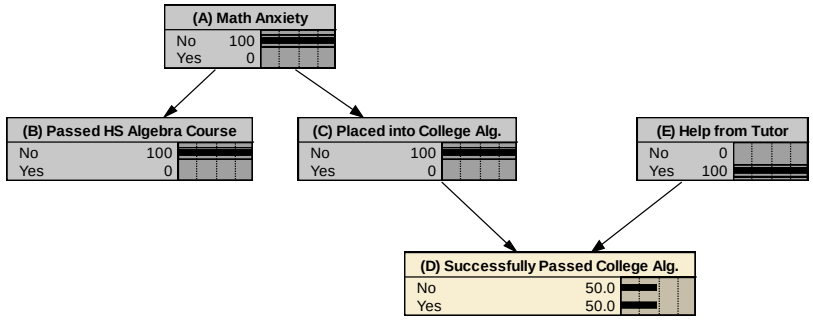


Figure A.3 *Propagation of evidence through a hypothetical Bayesian network of success in a college-level algebra course.*

APPENDIX B

WINBUGS CODE

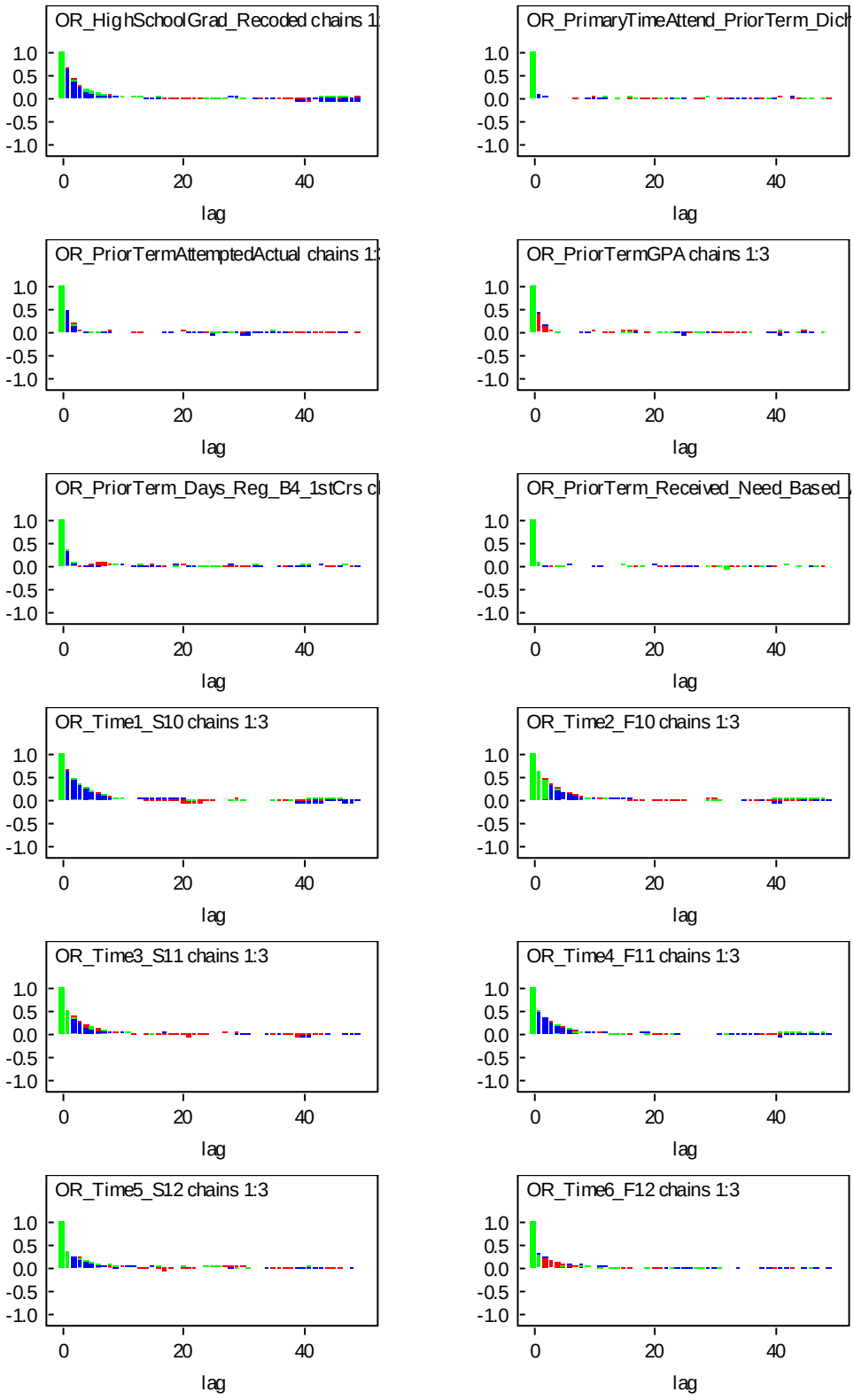
```
#Final Survival Analysis Model

model
{
for (i in 1:n) {

  # Linear regression on logit
  logit(p[i]) <- b.Time1_S10*Time1_S10[i] +
            b.Time2_F10*Time2_F10[i]+
            b.Time3_S11*Time3_S11[i] +
            b.Time4_F11*Time4_F11[i] +
            b.Time5_S12*Time5_S12[i] +
            b.Time6_F12*Time6_F12[i] +
     b.HighSchoolGrad_Recoded*HighSchoolGrad_Recoded[i] +
     b.PrimaryTimeAttend_PriorTerm_Dich*PrimaryTimeAttend_PriorTerm_Dich[i]+
     b.PriorTermAttemptedActual*PriorTermAttemptedActual[i]+
            b.PriorTerm_Days_Reg_B4_1stCrs*PriorTerm_Days_Reg_B4_1stCrs[i]+
     b.PriorTermGPA*PriorTermGPA[i]+
     b.PriorTerm_Received_Need_Based_Aid*PriorTerm_Received_Need_Based_Aid[i]

# Likelihood function for each data point
  StoppedOut[i] ~ dbern(p[i])
}

#Set diffuse priors for distribution of beta coefficients
b.Time1_S10 ~ dnorm(0.0,0.001)
b.Time2_F10 ~ dnorm(0.0,0.001)
b.Time3_S11 ~ dnorm(0.0,0.001)
b.Time4_F11 ~ dnorm(0.0,0.001)
b.Time5_S12 ~ dnorm(0.0,0.001)
b.Time6_F12 ~ dnorm(0.0,0.001)
b.HighSchoolGrad_Recoded ~ dnorm(0.0,0.001)
b.PrimaryTimeAttend_PriorTerm_Dich ~ dnorm(0.0,0.001)
b.PriorTermAttemptedActual ~ dnorm(0.0,0.0001)
b.PriorTerm_Days_Reg_B4_1stCrs ~ dnorm(0.0,0.001)
b.PriorTermGPA ~ dnorm(0.0,0.001)
b.PriorTerm_Received_Need_Based_Aid ~ dnorm(0.0,0.001)

#Odds Ratios
OR_Time1_S10 <- exp(b.Time1_S10)
OR_Time2_F10 <- exp(b.Time2_F10)
OR_Time3_S11 <- exp(b.Time3_S11)
OR_Time4_F11 <- exp(b.Time4_F11)
OR_Time5_S12 <- exp(b.Time5_S12)
OR_Time6_F12 <- exp(b.Time6_F12)
OR_HighSchoolGrad_Recoded <- exp(b.HighSchoolGrad_Recoded)
OR_PrimaryTimeAttend_PriorTerm_Dich <- exp(b.PrimaryTimeAttend_PriorTerm_Dich)
OR_PriorTermAttemptedActual <- exp(b.PriorTermAttemptedActual)
OR_PriorTerm_Days_Reg_B4_1stCrs <- exp(b.PriorTerm_Days_Reg_B4_1stCrs)
OR_PriorTermGPA <- exp(b.PriorTermGPA)
OR_PriorTerm_Received_Need_Based_Aid <- exp(b.PriorTerm_Received_Need_Based_Aid)
}

list( n = 4630)
```

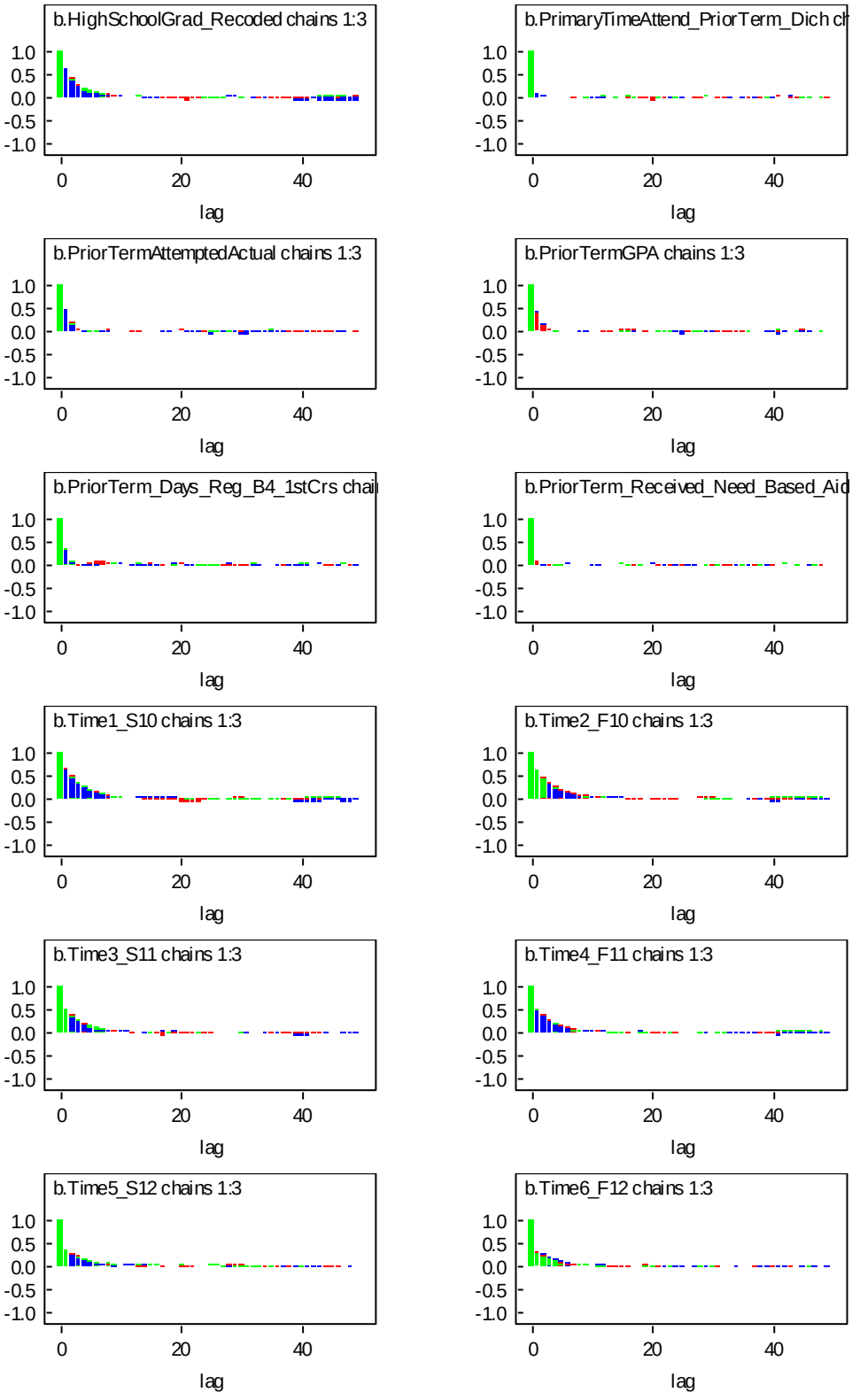
```
#Set Initial Values
list(
        b.Time1_S10 = 0,
        b.Time2_F10 = 0,
        b.Time3_S11 = 0,
        b.Time4_F11 = 0,
        b.Time5_S12 = 0,
        b.Time6_F12 = 0,
        b.HighSchoolGrad_Recoded=0,
        b.PrimaryTimeAttend_PriorTerm_Dich=0,
        b.PriorTermAttemptedActual=0,
        b.PriorTerm_Days_Reg_B4_1stCrs=0,
        b.PriorTermGPA=0,
        b.PriorTerm_Received_Need_Based_Aid=0)

list(
        b.Time1_S10 = -0.1,
        b.Time2_F10 = -0.1,
        b.Time3_S11 = -0.1,
        b.Time4_F11 = -0.1,
        b.Time5_S12 = -0.1,
        b.Time6_F12 = -0.1,
        b.HighSchoolGrad_Recoded=-0.1,
        b.PrimaryTimeAttend_PriorTerm_Dich=-0.1,
        b.PriorTermAttemptedActual=-0.01,
        b.PriorTerm_Days_Reg_B4_1stCrs=-0.1,
        b.PriorTermGPA=-0.1)

list(
   b.Time1_S10 = 0.1,
        b.Time2_F10 = 0.1,
        b.Time3_S11 = 0.1,
        b.Time4_F11 = 0.1,
        b.Time5_S12 = 0.1,
        b.Time6_F12 = 0.1,
        b.HighSchoolGrad_Recoded=0.1,
        b.PrimaryTimeAttend_PriorTerm_Dich=0.1,
        b.PriorTermAttemptedActual=0.1,
        b.PriorTerm_Days_Reg_B4_1stCrs=0.1,
        b.PriorTermGPA=0.1)
```

APPENDIX C

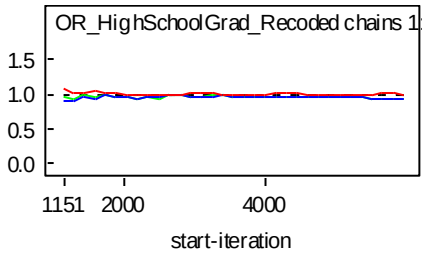
WINBUGS OUTPUT FOR FULLY BAYESIAN SURVIVAL ANALYSIS MODEL

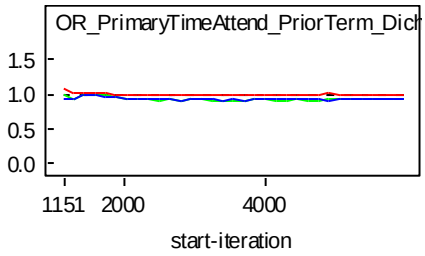
Autocorrelation Plots

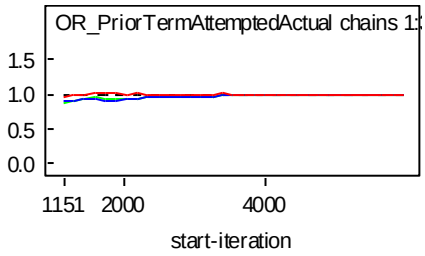


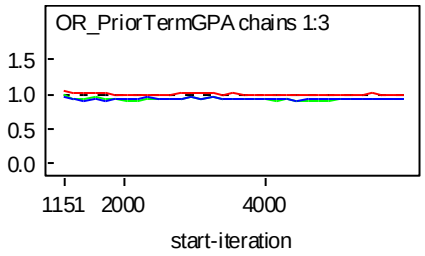

b.HighSchoolGrad_Recoded chains 1:3
b.PrimaryTimeAttend_PriorTerm_Dich ch
b.PriorTermAttemptedActual chains 1:3
b.PriorTermGPA chains 1:3
b.PriorTerm_Days_Reg_B4_1stCrs chai
b.PriorTerm_Received_Need_Based_Aid
b.Time1_S10 chains 1:3
b.Time2_F10 chains 1:3
b.Time3_S11 chains 1:3
b.Time4_F11 chains 1:3
b.Time5_S12 chains 1:3
b.Time6_F12 chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

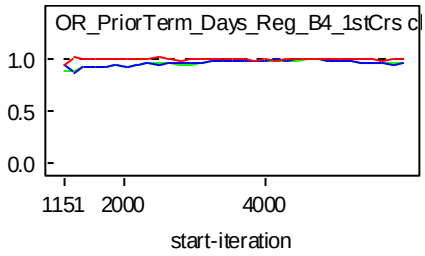
Brooks-Gelman-Rubin (BRG) Plots

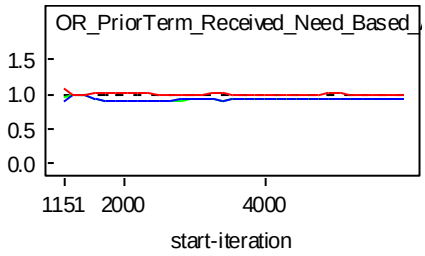


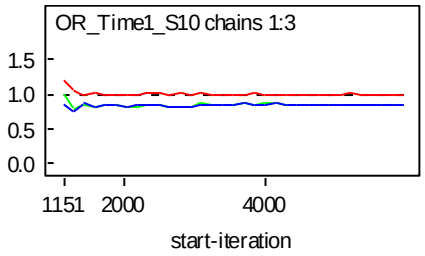


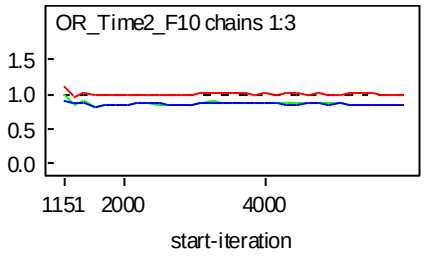


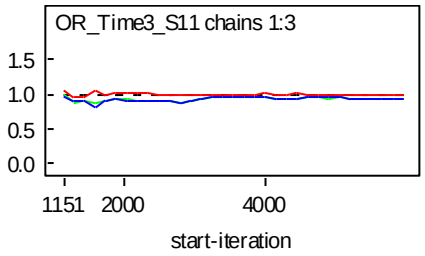


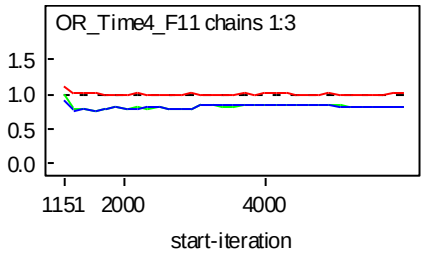


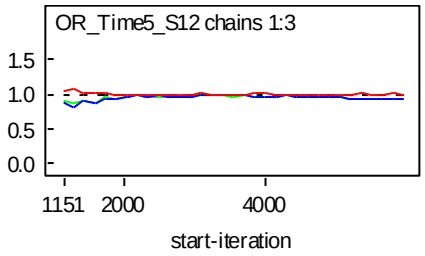


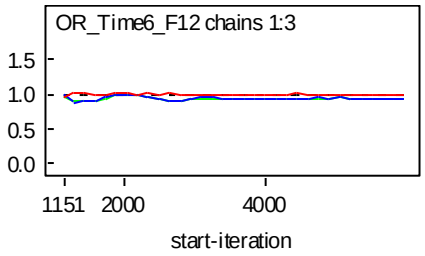


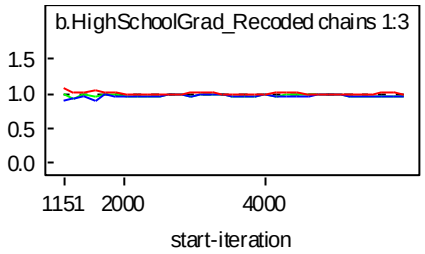


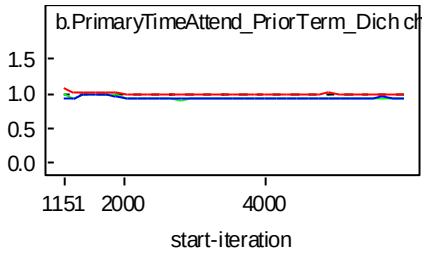


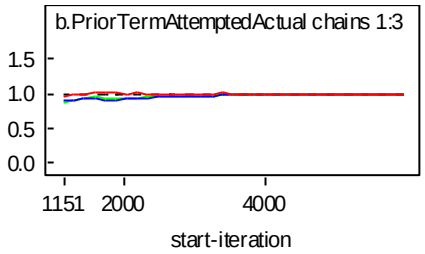


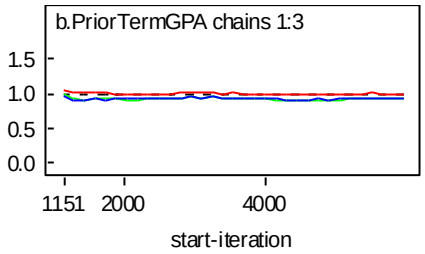


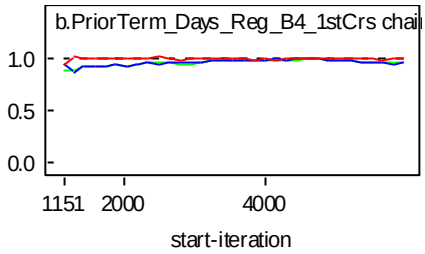


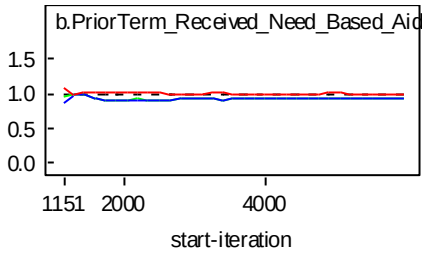

b.HighSchoolGrad_Recoded chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

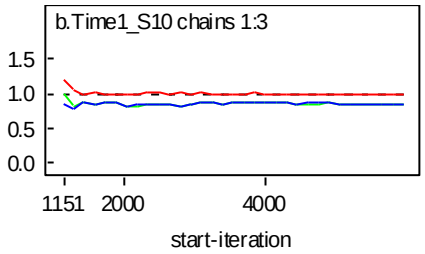

b.PrimaryTimeAttend_PriorTerm_Dich ch
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

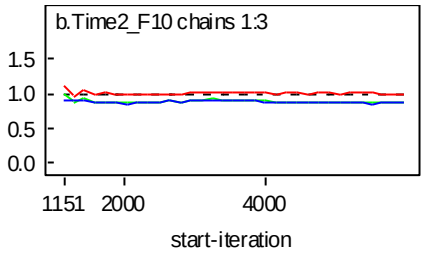

b.PriorTermAttemptedActual chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

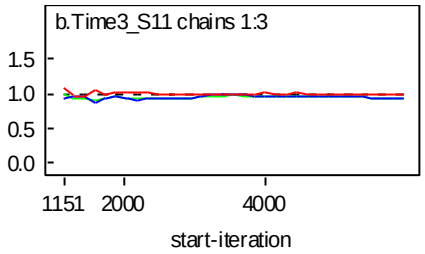

b.PriorTermGPA chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

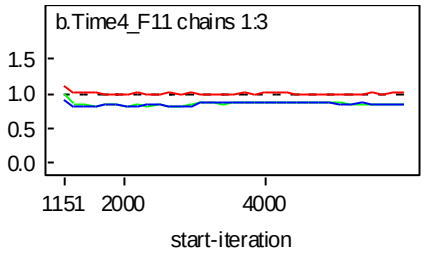

b.PriorTerm_Days_Reg_B4_1stCrs chai
1.0
0.5
0.0
1151
2000
4000
start-iteration

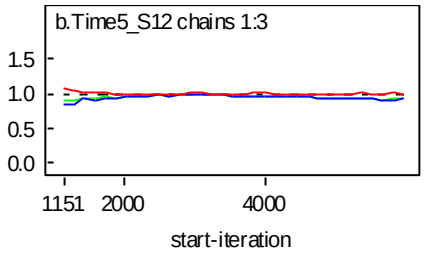

b.PriorTerm_Received_Need_Based_Aid
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

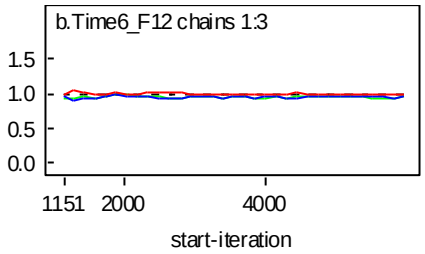

b.Time1_S10 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration


b.Time2_F10 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

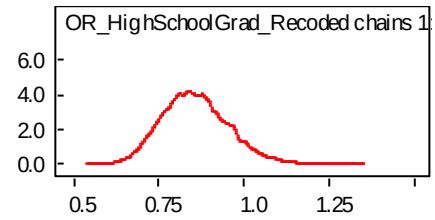

b.Time3_S11 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

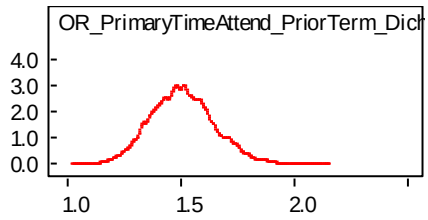

b.Time4_F11 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

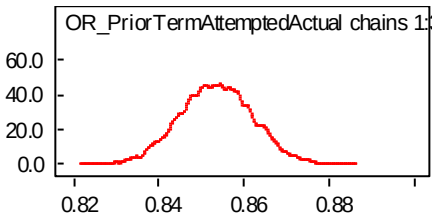

b.Time5_S12 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

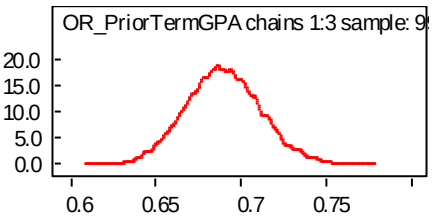

b.Time6_F12 chains 1:3
1.5
1.0
0.5
0.0
1151
2000
4000
start-iteration

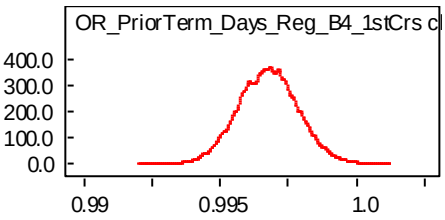
Density Plots

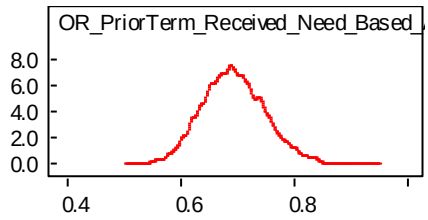


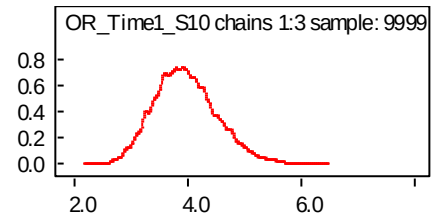


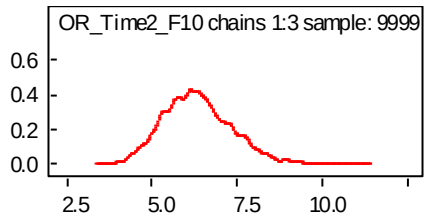


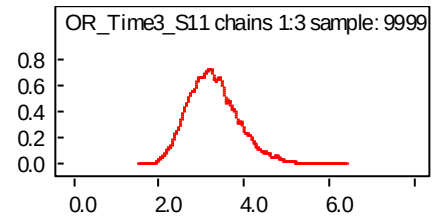


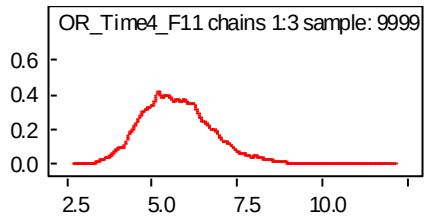


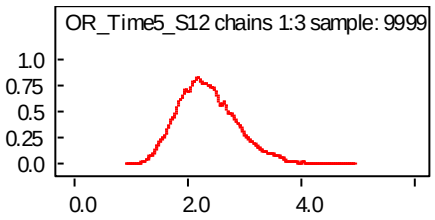


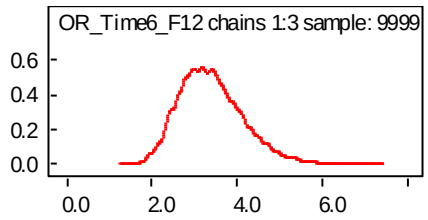


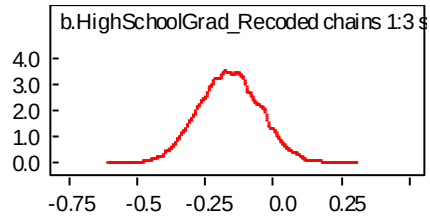


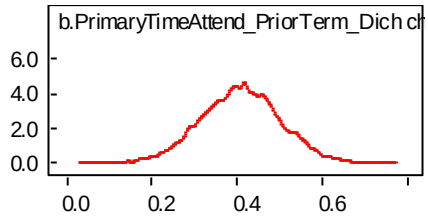


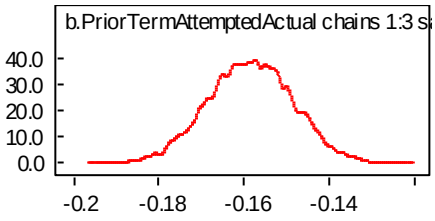


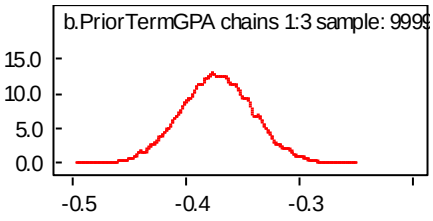


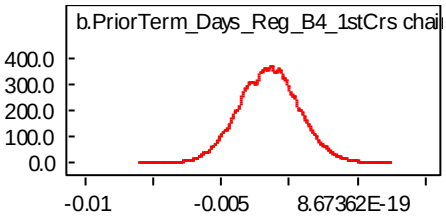


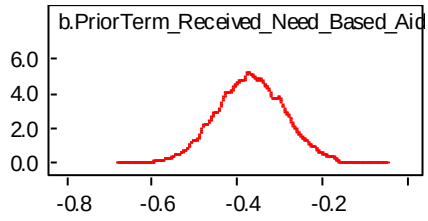

b.HighSchoolGrad_Recoded chains 1:3 s
4.0
3.0
2.0
1.0
0.0
-0.75
-0.5
-0.25
0.0
0.25

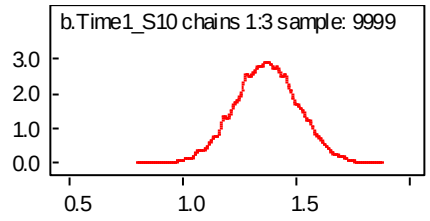

b.PrimaryTimeAttend_PriorTerm_Dich ch
6.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6

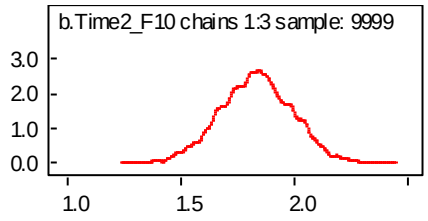

b.PriorTermAttemptedActual chains 1:3 s
40.0
30.0
20.0
10.0
0.0
-0.2
-0.18
-0.16
-0.14

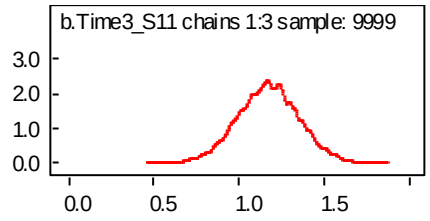

b.PriorTermGPA chains 1:3 sample: 999
15.0
10.0
5.0
0.0
-0.5
-0.4
-0.3

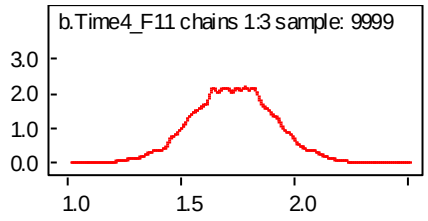

b.PriorTerm_Days_Reg_B4_1stCrs chai
400.0
300.0
200.0
100.0
0.0
-0.01
-0.005
8.67362E-19

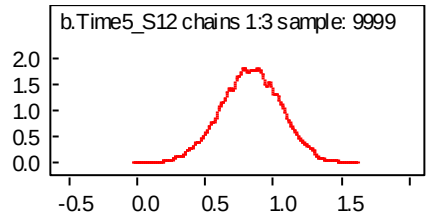

b.PriorTerm_Received_Need_Based_Aid
6.0
4.0
2.0
0.0
-0.8
-0.6
-0.4
-0.2

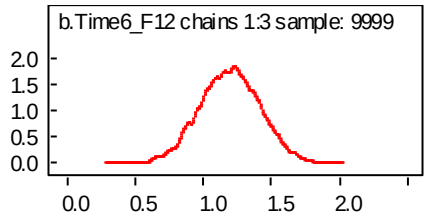

b.Time1_S10 chains 1:3 sample: 9999
3.0
2.0
1.0
0.0
0.5
1.0
1.5


b.Time2_F10 chains 1:3 sample: 9999
3.0
2.0
1.0
0.0
1.0
1.5
2.0

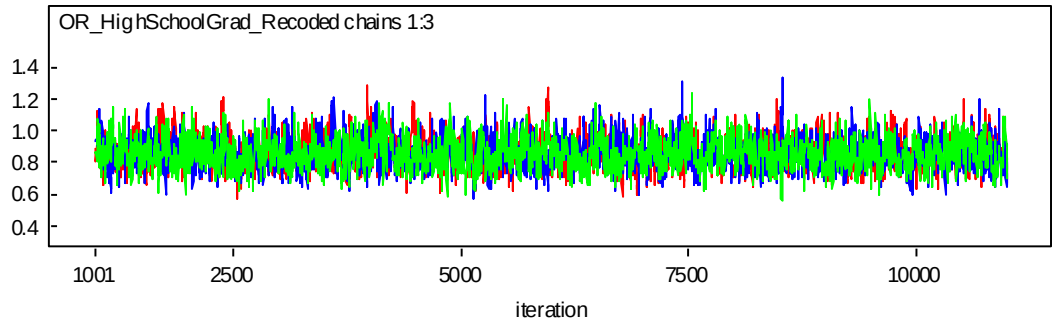

b.Time3_S11 chains 1:3 sample: 9999
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5

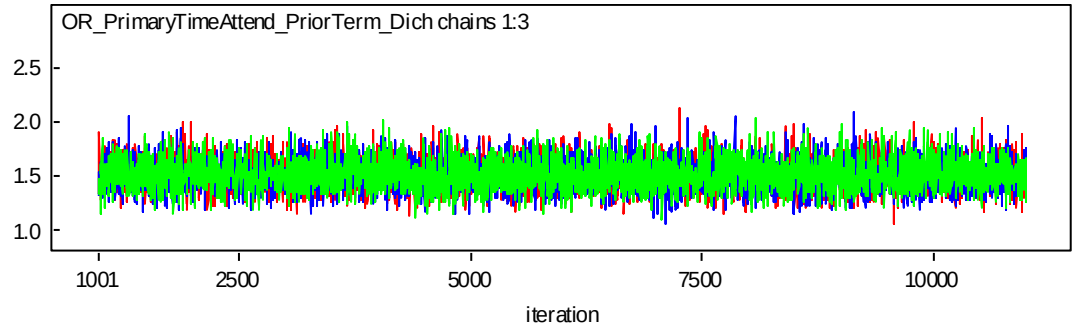

b.Time4_F11 chains 1:3 sample: 9999
3.0
2.0
1.0
0.0
1.0
1.5
2.0

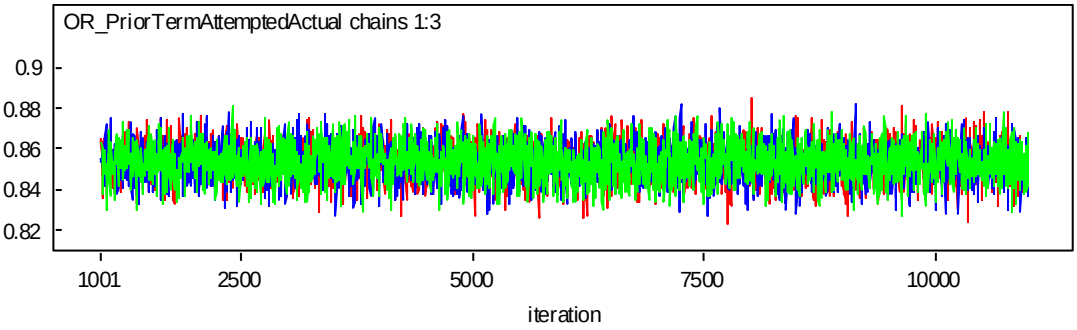

b.Time5_S12 chains 1:3 sample: 9999
2.0
1.5
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5

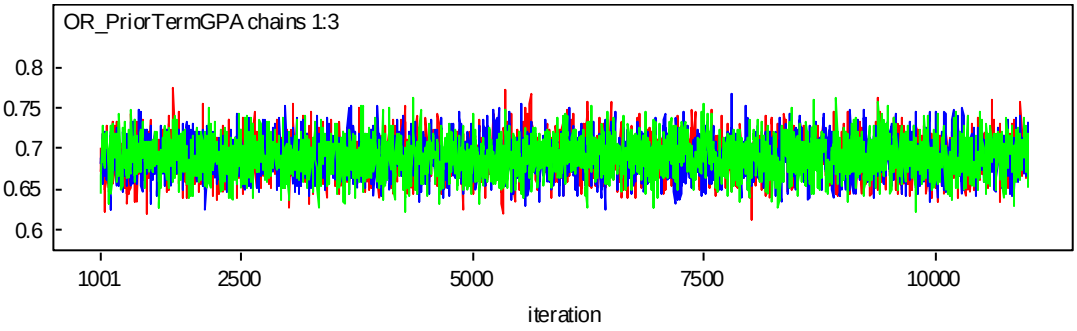

b.Time6_F12 chains 1:3 sample: 9999
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0

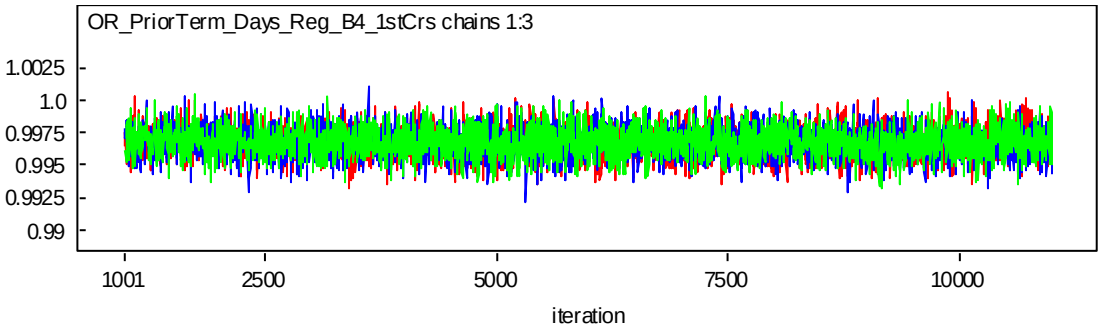
Trace Plots

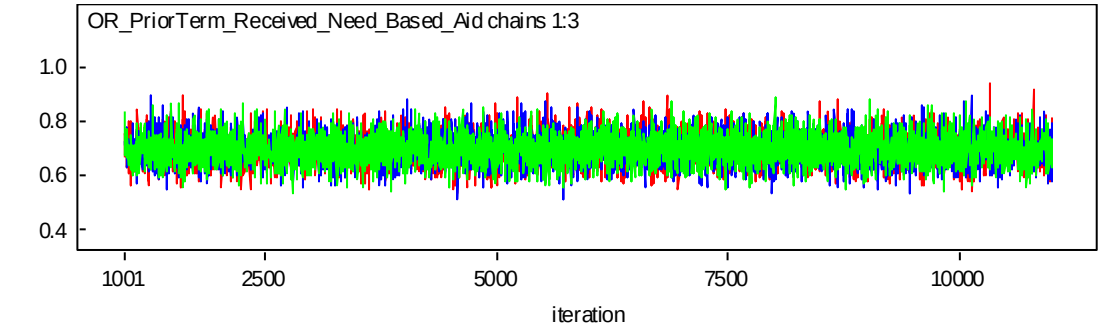

OR_HighSchoolGrad_Recoded chains 1:3
1.4
1.2
1.0
0.8
0.6
0.4
1001
2500
5000
7500
10000
iteration

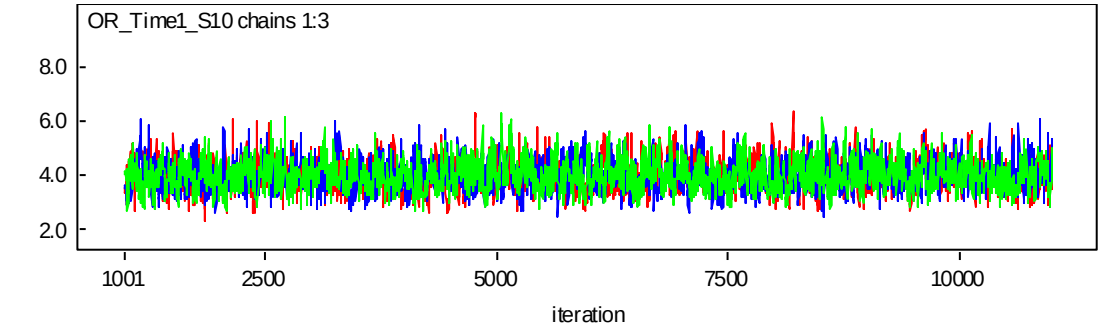

OR_PrimaryTimeAttend_PriorTerm_Dich chains 1:3
2.5
2.0
1.5
1.0
1001
2500
5000
7500
10000
iteration

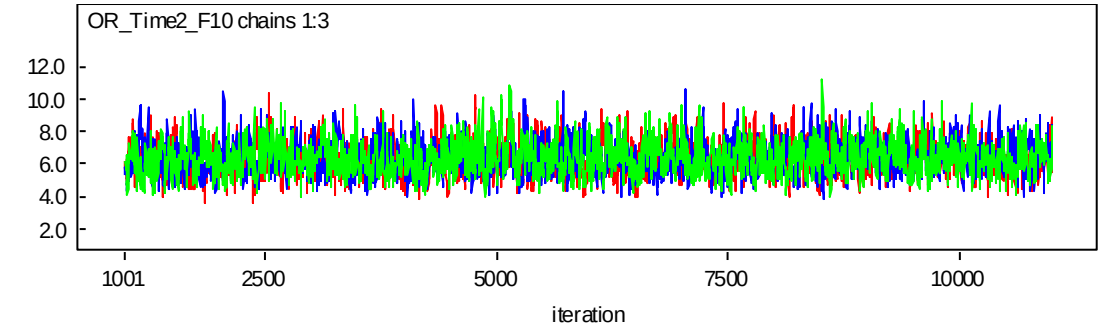

OR_PriorTermAttemptedActual chains 1:3
0.9
0.88
0.86
0.84
0.82
1001
2500
5000
7500
10000
iteration

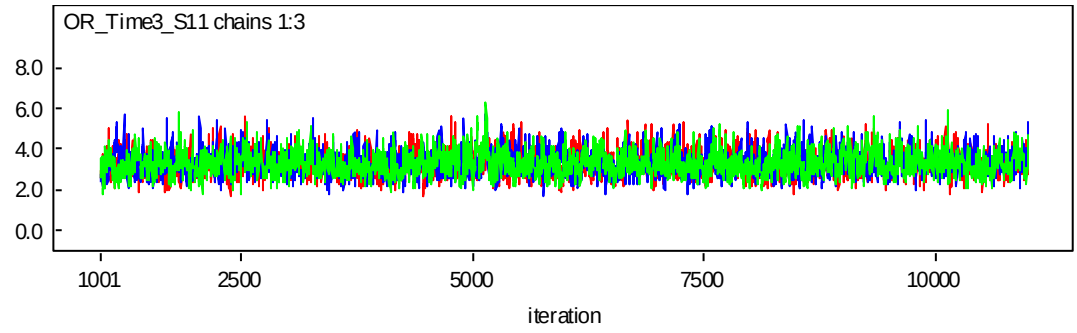

OR_PriorTermGPA chains 1:3
0.8
0.75
0.7
0.65
0.6
1001
2500
5000
7500
10000
iteration

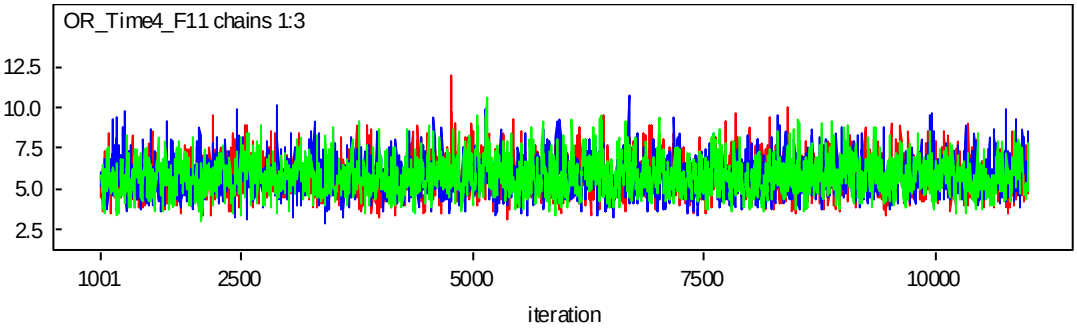

OR_PriorTerm_Days_Reg_B4_1stCrs chains 1:3
1.0025
1.0
0.9975
0.995
0.9925
0.99
1001
2500
5000
7500
10000
iteration

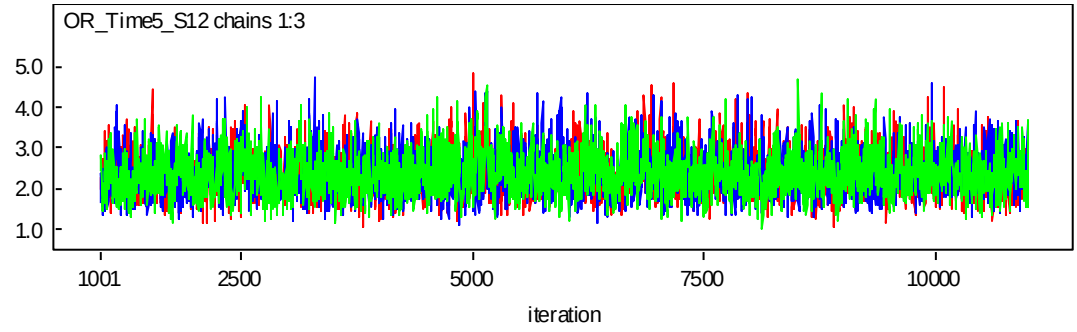

OR_PriorTerm_Received_Need_Based_Aid chains 1:3
1.0
0.8
0.6
0.4
1001
2500
5000
7500
10000
iteration

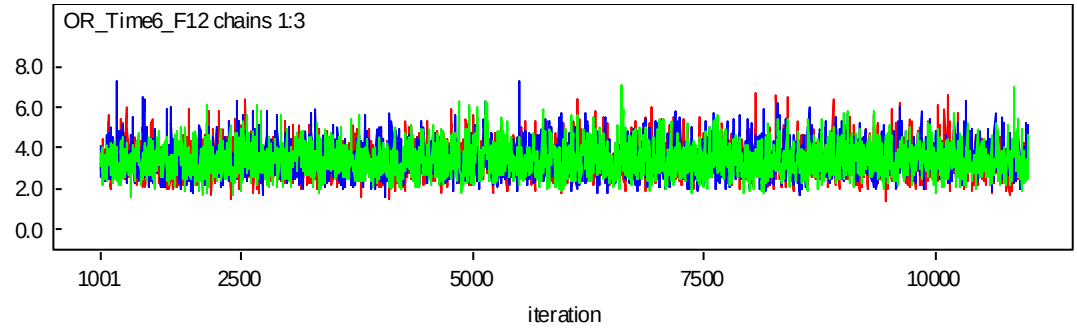

OR_Time1_S10 chains 1:3
8.0
6.0
4.0
2.0
1001
2500
5000
7500
10000
iteration

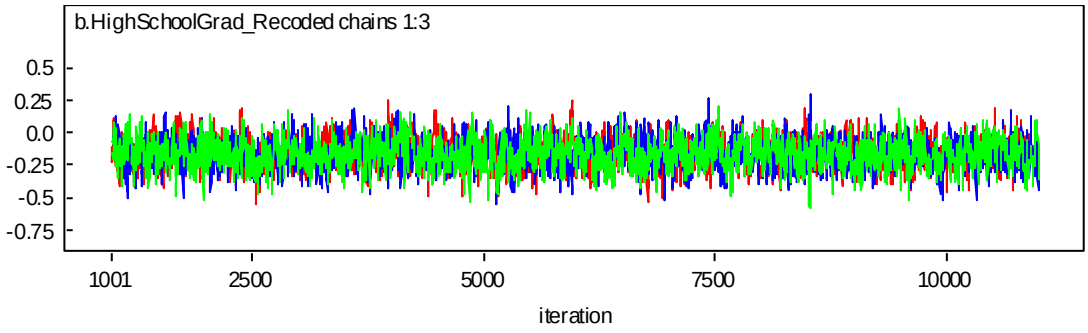

OR_Time2_F10 chains 1:3
12.0
10.0
8.0
6.0
4.0
2.0
1001
2500
5000
7500
10000
iteration

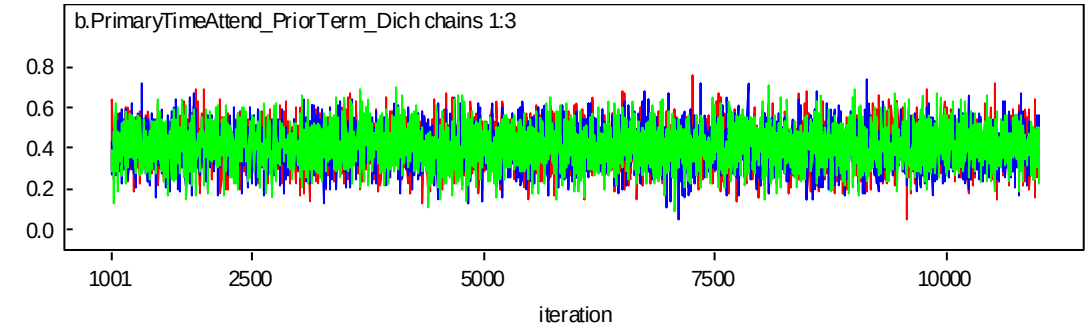

OR_Time3_S11 chains 1:3
8.0
6.0
4.0
2.0
0.0
1001
2500
5000
7500
10000
iteration

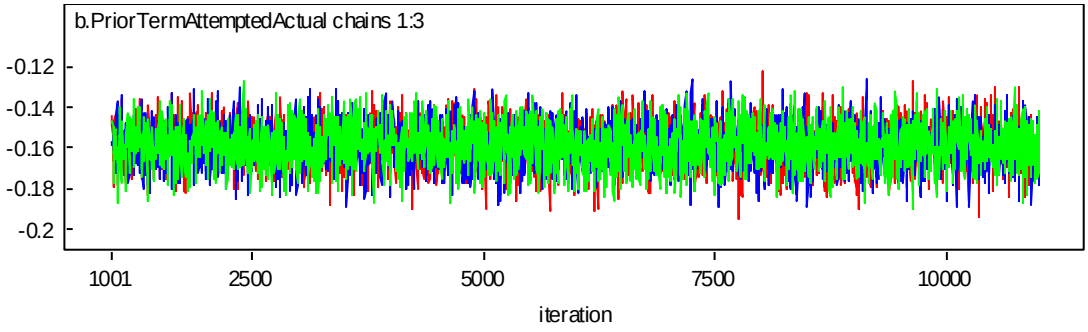

OR_Time4_F11 chains 1:3
12.5
10.0
7.5
5.0
2.5
1001
2500
5000
7500
10000
iteration

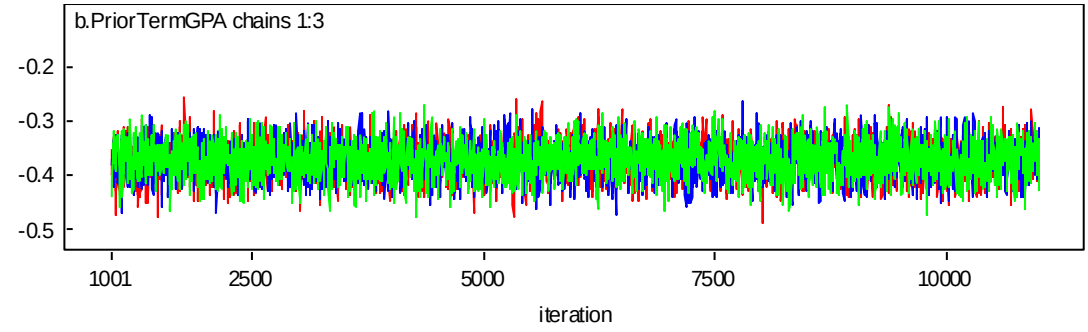

OR_Time5_S12 chains 1:3
5.0
4.0
3.0
2.0
1.0
1001
2500
5000
7500
10000
iteration

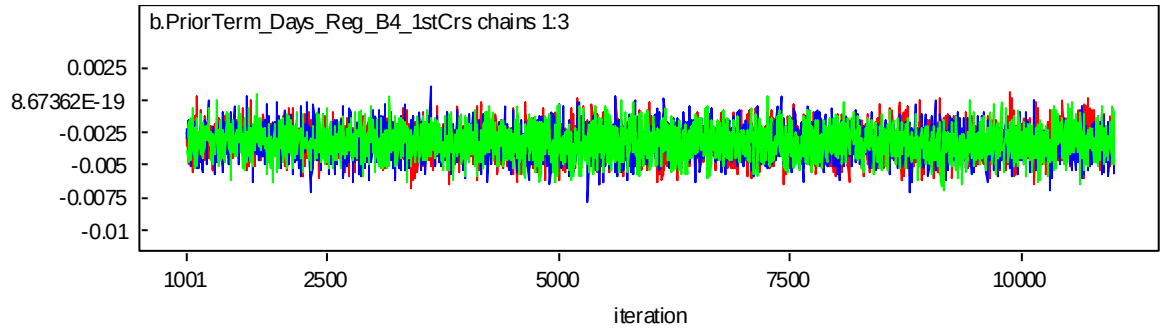

OR_Time6_F12 chains 1:3
8.0
6.0
4.0
2.0
0.0
1001
2500
5000
7500
10000
iteration

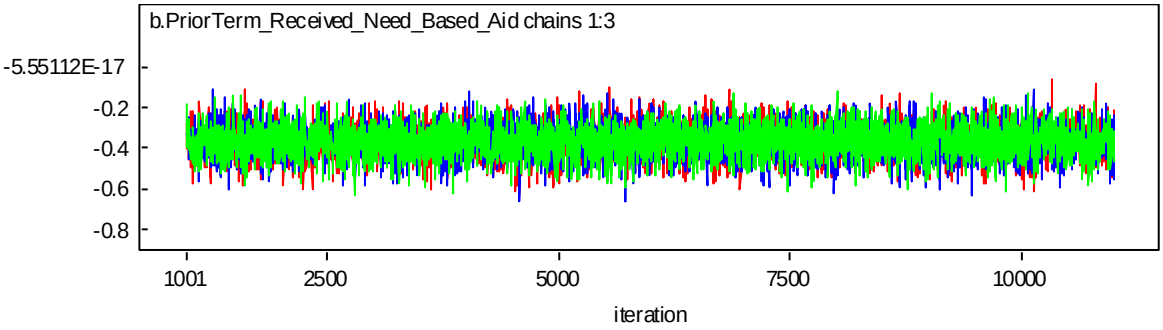

b.HighSchoolGrad_Recoded chains 1:3
0.5
0.25
0.0
-0.25
-0.5
-0.75
1001
2500
5000
7500
10000
iteration


b.PrimaryTimeAttend_PriorTerm_Dich chains 1:3
0.8
0.6
0.4
0.2
0.0
1001
2500
5000
7500
10000
iteration


b.PriorTermAttemptedActual chains 1:3
-0.12
-0.14
-0.16
-0.18
-0.2
1001
2500
5000
7500
10000
iteration


b.PriorTermGPA chains 1:3
-0.2
-0.3
-0.4
-0.5
1001
2500
5000
7500
10000
iteration


b.PriorTerm_Days_Reg_B4_1stCrs chains 1:3
0.0025
8.67362E-19
-0.0025
-0.005
-0.0075
-0.01
1001
2500
5000
7500
10000
iteration


b.PriorTerm_Received_Need_Based_Aid chains 1:3
-5.55112E-17
-0.2
-0.4
-0.6
-0.8
1001
2500
5000
7500
10000
iteration

APPENDIX D

CONDITIONAL PROBABILITY TABLES FOR BAYESIAN NETWORKS

Conditional Probability Tables for Modified (Discretized) Final Survival Analysis Model as a Bayesian Network

The conditional probability table for the Stopping Out parameter is not presented below due to its size.

Time Period (Index_Cat)

Time1	Time2	Time3	Time4	Time5	Time6
0.37	0.24	0.15	0.12	0.07	0.05

Primary Time of Attendance Prior Term (PrimaryTimeAttend_PT_Dich)

Day	NotDay
0.81	0.19

Received Need-Based Aid Prior Term (Received_Need_Based_Aid_PT)

No	Yes
0.57	0.43

Graduated High School (HighSchoolGrad_Recoded)

No	Yes
0.12	0.88

Prior Term GPA (PriorTermGPA)

ZeroToOne	OneToTwo	TwoToThree	ThreeToFour
0.23	0.14	0.30	0.33

Number of Credit Hours Attempted Prior Term (AttemptedActual_Credits_PT)

ZeroToThree	ThreeToSix	SixToNine	NineToTwelve	TwelvePlus
0.21	0.16	0.17	0.23	0.22

Number of Days Registered Before First Course Prior Term (Days_Reg_B4_1stCrs_PT)

FirstQuartile	SecondQuartile	ThirdQuartile	FourthQuartile
0.26	0.25	0.17	0.32

Conditional Probability Tables for the Reduced, Modified (Discretized) Final Survival Analysis Model as a Bayesian Network

The conditional probability table for the Stopping Out parameter is not presented below due to its size.

Time Period (Index_Cat)

Time1	Time2	Time3	Time4	Time5	Time6
0.37	0.24	0.15	0.12	0.07	0.05

Primary Time of Attendance Prior Term (PrimaryTimeAttend_PT_Dich)

Day	NotDay
0.81	0.19

Received Need-Based Aid Prior Term (Received_Need_Based_Aid_PT)

No	Yes
0.57	0.43

Prior Term GPA (PriorTermGPA)

ZeroToOne	OneToTwo	TwoToThree	ThreeToFour
0.23	0.14	0.30	0.33

Number of Credit Hours Attempted Prior Term (AttemptedActual_Credits_PT)

ZeroToThree	ThreeToSix	SixToNine	NineToTwelve	TwelvePlus
0.21	0.16	0.17	0.23	0.22