EARLY ALERT ACADEMIC WARNING SYSTEM: A QUANTITATIVE STUDY OF TWO CHARACTERISTICS OF EARLY ALERT WARNINGS AND THE IMPACTS TO RETENTION AMONG FIRST-TIME FALL FRESHMEN

by

Jessica Groomer Smith

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree

Master of Science

Major Subject: Mathematics

West Texas A&M University

Canyon, Texas

August 2018

Approved:

Chairman, Thesis Committee	Date
Member, Thesis Committee	Date
Member, Thesis Committee	Date
Dean, Engineering, Computer Science, and Math	Date
Dean, Graduate School	Date

ABSTRACT

An Early Alert referral is a tool used by colleges and universities to proactively monitor student performance. Based on theories of student interaction with the institutional environment being the driving factors in student attrition, this study examines the nature of the student interaction with the Early Alert system at West Texas A&M University, a medium-sized University in rural Texas, in two ways. First, it looks at whether the timing of the initial Early Alert referral received by the student impacts retention to the next long semester. Second it examines whether being told the course performance is satisfactory, or could be satisfactory with improvement, impacts retention to the next long semester. A group of 339 full time, first-time, degree seeking freshmen who received Early Alerts during the first semester of enrollment are sampled. A logistic regression finds that the timing of the Early Alert (counted as days into the term) increases the odds of retention by about 10% each week into the term. This effect is particularly pronounced among students living on campus with lower first term GPAs.

ACKNOWLEDGEMENTS

The author would like to thank first Dr. Pam Lockwood for her patience and guidance over the course of this endeavor. I've been on and off this path, but she's been there the whole way ready any time. Dr. Dan Seth has been an invaluable source of insight, and thankfully his generous spirit and pursuit of Pippa's coffee has allowed me to partake in his wisdom. Additionally, the author would like to thank Jarvis Hampton, Pam Miller, and Janet Hanson for their support, encouragement, and sympathetic ears. These three have celebrated the victories, sympathized over the hard times, and always made me laugh. Their support has been invaluable to me. To Andrea Finchum, Robert King, Jeff Babb, Alyson Ries, Nancy Garcia, and Amber Black, Isaac Newton said, "If I have seen further than others, it is by standing on the shoulders of giants," and this is reflective of my gratitude to you all. Thanks for the interesting questions. Finally, I would like to thank my husband, Brian Smith, and my son Jonathan. They've both sat through hours of discussion of complicated math, looked at various graphs and pictures, and helped me sound it out sometimes. They may never truly know the value of their contributions, but I will never forget. And last, this thesis is dedicated to my mom. She taught me that having the courage to take the first step is the hardest, and that persistence pays. I'm proving you right every day. KOKO – LYOD.

Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

It is common in the media today to see statistics referencing the costs of obtaining a college education. Costs to individual students and to taxpayers funding education are under intense scrutiny as they rise, while the benefits of attending college are simultaneously debated. In a recent example, the Associated Press reported in early 2017 that “Americans with no more than a high school diploma have fallen so far behind college graduates in their economic lives that the earnings gap between college grads and everyone else has reached its widest point on record,” (Rugaber, 2017).

Citing dramatic differences in earnings, and impacts to employment, marriage, home ownership and retirement outcomes, the article contends few experts would propose sending more people to college. Rugaber states that as many as four out of ten students that attend college leave before earning a degree, some with debt loads they cannot afford to pay without a degree. These consequences are dire in a changing economy, both for leavers of higher education and for institutions.

Institutions and legislatures across the United States are acutely aware of statistics of the type referred to by Rugaber. Aside from economic losses to a student who does not complete college, the costs of attrition are estimated to be an average of $10 million annually to each institution of higher education (Raisman, 2013). While economic ramifications of student attrition are stark in today’s economy, for both students and schools, it makes sense to understand the nature of student attrition. Institutions of higher education have for several decades assumed some responsibility for improving persistence and retention rates.

Landmark theories in student persistence were ushered in during the late 1960s and 1970s. Prior to this time, student attrition was largely attributed to the students themselves. The early researchers took a multi-disciplinary approach to student attrition, incorporating ideas from psychology and sociology to formulate models of student attrition. In subsequent decades, researchers have applied these theories, discussed in some detail in the next chapter, to identify areas in which student support, provided by the institution, are critical. The result is a dynamic landscape of intervention programs and services available to students entering the college market today.

Many intervention programs commonly in use are focused on the first year of the college experience. Tinto, Astin, and other researchers have emphasized the importance of student integration and assimilation into the college environment as paramount to student retention particularly during the first year (Howard, 2015). The focus of this paper is one such program, namely the Early Alert Warning system as applied to entering fall freshmen enrolled in core curriculum courses at West Texas A&M University. This system allows faculty to complete a referral to Advising Services, which upon the receipt of the referral will follow up with the student to offer support services. While this system is currently available to students and faculty at all levels of undergraduate study at the University, the goal of the study is to determine if the timing of an Early Alert warning and/or the possibility of passing a course at the outset of the Early Alert impacts freshman retention to the next long semester.

STATEMENT OF THE PROBLEM AND RESEARCH QUESTIONS

Overall persistence rates remain virtually unchanged over the last several decades across the country. This is despite the fact that college and universities proactively intervene in the student experience in an effort to improve student persistence. Data harvested from the National Center for Education Statistics shows that over the last more than a decade, the fall-to-fall retention rates for two-year and four-year institutions remains virtually unchanged (see Figure 1). Though not shown in this paper, a similar exploration of college graduation rates within 150% of normal time to degree shows a similar low-variance result across the years, with just over 50% of four-year students graduating.

Much of the research outlined in the literature review focuses on the theories of student dropout, with a particular focus on the first year of the college experience. Additionally, emphasis is placed on the role of faculty in the first year experience. An Early Alert warning system can be viewed as a means of faculty outreach to the student. The focus of this study is on the characteristics of the student interactions within the Early Alert system. The primary research questions are:

1. Does the timing of the Early Alert increase the retention rate among incoming freshmen to the next long semester?
2. Does the feasibility of passing the course at the time of the first Early Alert increase retention among incoming freshmen to the next long semester?

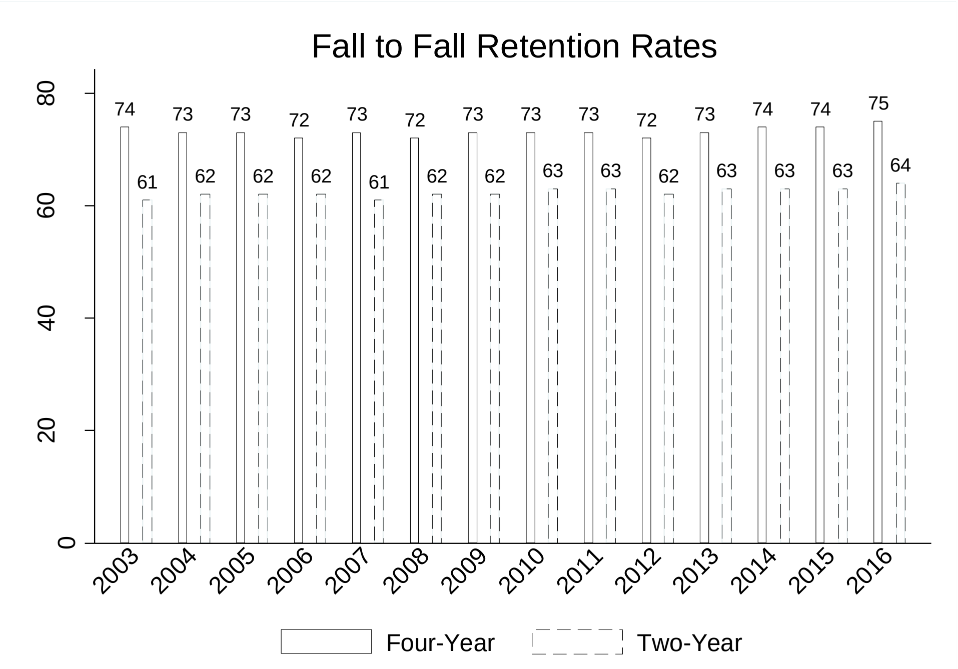


Figure 1 Median fall to fall retention rates of two-year and four-year institutions. Source: NCES, 2018

ORGANIZATION OF THIS STUDY

Subsequent to this introductory chapter, the next chapter contains a review of literature related to intervention theory and Early Alert (EA) systems. The third chapter contains an overview of the modeling strategy used along with the supporting mathematical methods. The fourth and fifth chapters detail the data and study methodology, and the details of the analysis, respectively. The final chapter contains a discussion of the results and implications of the selected model.

SUMMARY

Higher education is under intense pressure to accelerate student performance and degree completion. The call has not been unheard by institutions with the result that many programs and services have been put in place in an effort to help students succeed in college. This study will examine two characteristics, the total length of interaction with the system, and whether an opportunity to pass the course impacted retention to the next long semester.

CHAPTER II

LITERATURE REVIEW

Student attrition is widely researched with literature spanning decades under the purview of several disciplines. Modern thought on the subject conveys that in addition to the traditionally studied characteristics of students that may lead to attrition, institutions have both opportunity and obligation to provide meaningful engagement experiences toward the goal of student retention. Models aimed at understanding the process of student dropout come from several areas in the literature, but Bai and Pan (2009) classify these prominent theoretical models into four basic types: “a) student integration models (Tinto, 1975), b) student involvement models (Astin, 1975; Pascarella, 1980), c) industrial/organizational models (Bean, 1983), and d) financial impact models (St. John, 1990; Paulsen & St. John, 1993).” The first two types focus on how well the student integrates into the institutional structures, while the third type emphasizes institutional actions and the last discerns impacts to retention based on cost-benefit analysis (Bai, 2009-2010).

While several models have been proposed, all with benefits and limitations, this literature review will focus on Tinto’s 1993 comprehensive model which highlights the longitudinal process of student attrition, and frames a context under which interventions such as Early Alert are grounded. Further, arguments for early intervention are examined, as the development of such a model necessitates the discussion. Finally, a review of literature related to Early Alert use and practice is presented.

THE STUDENT INTERACTION MODEL

Early research on student attrition was largely focused on students, the pre-entry characteristics of students themselves, and flushing out differences in attrition rates among different student populations (Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition 2nd Ed., 1993) (Simons, 2011). Research on student attrition dates to the 1920s or prior (Spady W. , 1970). As colleges and universities grew, the complexities of managing institutions with growing budgets also grew. Institutions were concerned with maintaining a reputation for production of quality graduates yet efficiently managing costs and maintaining revenue (Summerskill, 1962). In his 1962 review of existing literature, John Summerskill points out "the problem has apparently never captured the active interest of any substantial segment of the social science profession, and there has been no concerted effort to pull together existing, fragmentary knowledge and deliver results of general value" (pg. 627).

Suffering from inconsistent definitions across studies, the literature of the time is scattered and lacking in theoretical cohesion. In 1970, William Spady published an extensive review of the subject literature and summarized the findings with the aim of formulating a theoretical model to describe student dropout. Citing findings that pointed to the student background (parents, potential, and past performance), the student's gender, the student's level of maturity, and the value of interpersonal relationships, Spady put forth a model to descriptively discern the characteristics of student dropout. Adapting a model from psychology, Durkheim's Theory of Suicide, Spady posited that student attrition was a terminal action, much like that of suicide, in that it is the product of student non-integration into the social and academic systems of the institution. Spady's model incorporated psychological and sociological elements into the theory of student attrition. He postulated that like one who commits suicide, a student dropout does not experience a sense of belonging or feel valued in the community. His model described a process by which a student integrates into the community. The model is premised with a personal background influenced by parents and socioeconomic status. The student then incorporates social and academic ability along with opportunities for interaction with the community into dropout decisions on the part of the student.

In 1975, Vincent Tinto put forth a modification of Spady's model, by proposing "a theoretical model that explains the processes of interaction between the individual and the institution that lead differing individuals to drop out from institutions of higher education, and that also distinguishes between those processes that result in definably different forms of dropout behavior" (Tinto, Dropout from Higher Education: A Theoretical Synthesis of Recent Research, 1975). Tinto's original model views student dropout as a longitudinal process of interactions between the individual and the academic and social systems of the college, subject to characteristics pertaining to the individual student, including commitments and interactions with members of external community on the part of the student. Tinto's idea is that students must successfully transition from the academic and social support systems in place in high school to a new network of resources and support in the post-secondary environment.

In 1993, acknowledging criticism and growth in perspective, Tinto published an elaborated, updated model to understand student attrition. Tinto's model seeks depth of explanation of the phenomenon of student dropout. Tinto seeks to provide as much flexibility for the mechanisms of student leaving as is warranted given that he seeks to distinguish between different types of dropout behavior at the individual student level. In order to achieve this level of detail, several underlying concepts should be understood, and are described in brief.

The first concept to understand is the idea of departure. Tinto distinguishes between voluntary departure on the part of the student, which may be the result of poorly perceived integration into the college environment, and involuntary departure, which may be the result of academic dismissal. Student departure from an institution is of primary concern to the institution, though this departure may have several outcomes at the student level. A departure that results in a transfer, or even a temporary departure, is differentiated from a departure that results in the student leaving higher education altogether, called a system departure. Tinto's model is intended to be applicable to institutional departure, with a focus on the individual roots of departure at the student level. This idea is the foundation upon which the model postulates that departure decisions may be influenced by the actions of the institution, and that post-matriculation experiences are shaped by the institution and are an influencing driver in voluntary student leaving.

As much as the literature is focused on the premise that students enter college with a defined set of characteristics, Tinto’s model incorporates these into the more organic concepts of “intention” and “commitment.” It’s easy to measure pre-college ability on a standardized test, or to determine the socioeconomic status of a student, but the concepts here allow room for a student to define entering expectations of the educational experience for themselves upon arrival, and to thusly interpret the education experience. Additionally, Tinto identifies four forms of individual experience that affect departure. Students, after matriculation, may experience “adjustment,” “difficulty,” “incongruence” and/or “isolation” (Tinto, 1993 pg. 37).

To elaborate on these concepts, Tinto’s process says that students arrive at an institution with an intention for achievement, but the level of achievement may be wide in scope from one student to the next, and further is not invariant over the course of the academic pursuits of the student. For example, a student pursuing a higher degree may be more prone to complete a lower degree. A student attending for occupational reasons may be less likely to complete a degree if goals are otherwise achieved. Commitment involves whether a student is willing to spend the time to pursue educational activities, and also the level of the commitment to the institution itself on the part of the student. Family background, pre-college schooling, and skills and abilities culminate to bring a student to an institution of higher education with a certain level of intention and commitment, to both goals and the institution. This leads to the set of student experiences described by Tinto.

As much as a student leaving college is a process, a student entering college, persisting, and successfully completing college is a process. The four student experiences can disrupt the process of completion. Adjustment refers to the ability of the student to separate from the communities and associations of the past and conform to the new expectations of the academic and social communities of the institution. The term adjustment in and of itself implies that a period of time is required, and in fact, literature shows that support during the early enrollment of a new college student is predictive of success (see next section).

Difficulty refers to a student being unable to meet academic standards, which may be much more rigorous at the college level than the student has experienced before, and may result in involuntary departure from the institution. Difficulty may arise from the next experience, which is incongruence. This refers to the case when the institution is not a good fit for the student in terms of interest or needs. A student may experience incongruence at the outset of the educational endeavor, but it may also occur later in an academic career if the interests or needs of the student change. As well as academic concerns, incongruence may also be the product of social interactions on the part of the student.

The last student experience described by Tinto is isolation. This occurs when there is an absence of contact between the student and the academic and social communities of the institution. Students who fail to adjust when entering college, or those who may lack access to community interaction, are at risk for experiencing isolation. It may also arise from incongruence. While most institutions employ a large number of staff dedicated to student services, faculty interaction may be the most significant source of opportunity for a student, and has been shown to be an important factor in combatting isolation (Tinto, 1993 pg. 57-58).

The concepts described here all relate to the post-matriculation experience of the student. These are the experiences that the institution has the power to shape, but not all influences in drop out decisions are within the control of the institution. External communities and obligations also play a part in student decisions, and Tinto views institutional actions and communities as being nested within these external entities (Tinto, 1993 pg. 115).

With these ideas at play, Tinto recognizes that institutions have both academic and social systems, and that interaction in these systems may be both formal and informal. Formal interactions within the academic system occur primarily in the classroom, while within the social system they may take the form of clubs or sports. Informal interactions may occur almost anywhere. Faculty, staff, and student peers, as well as institutional action, help facilitate all of these interactions. Figure 2 is a visual rendering of the culmination of Tinto's longitudinal description of student interaction.

The figure shows that a student's pre-college characteristics lead them to the institution with a level of intention and commitment, which lead a student to experiences in both the academic and social systems. The type of experience determines the level of student integration, which then feeds into the next dropout decision. These decisions can be made continuously as college life is experienced by the student.

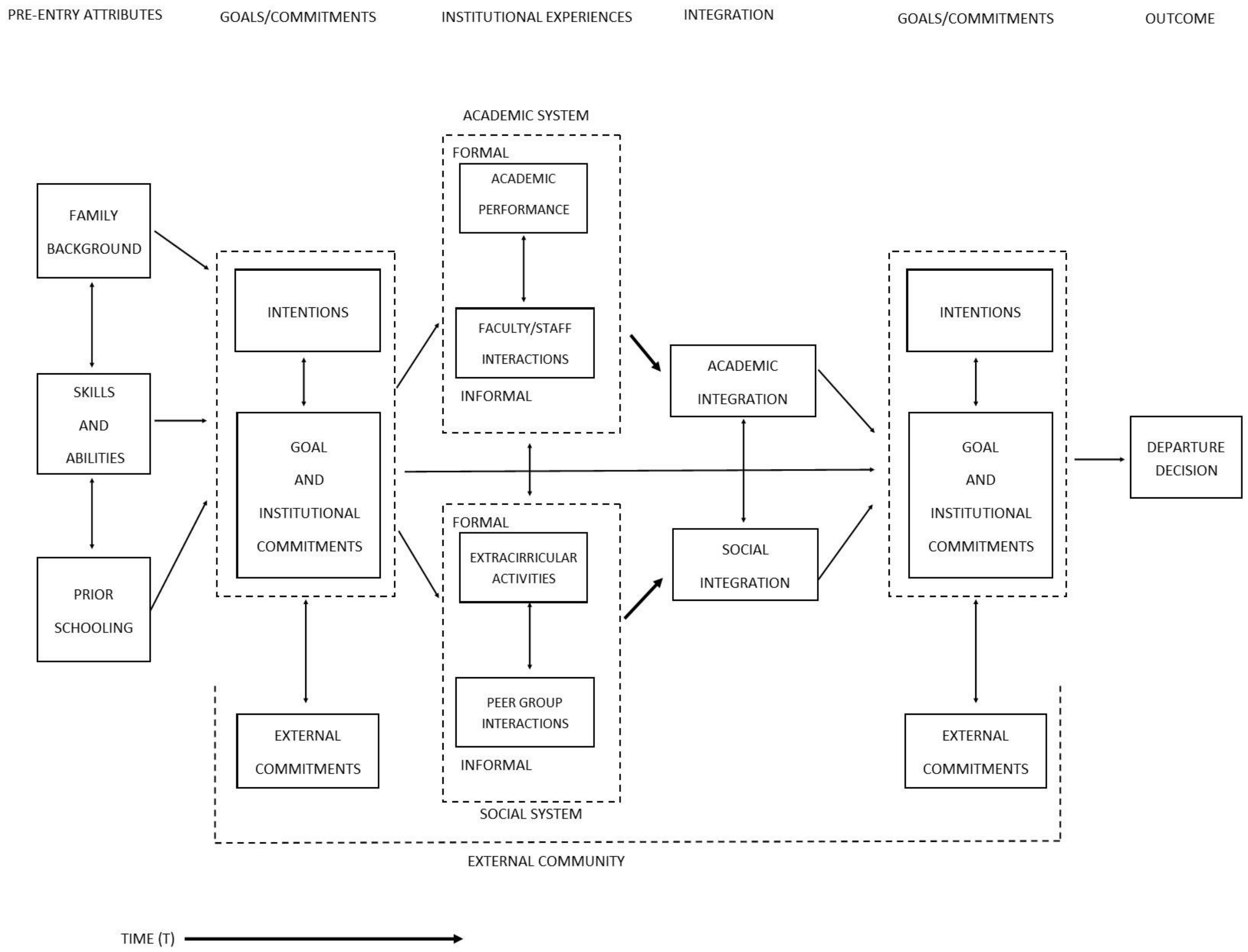


Figure 2 Tinto's longitudinal model of student interaction leading to drop out decisions.

From the figure, it can be seen how a student immediately interacts with the institutional structures that define the college experience, with academic and social integration into the environment leading into a student reassessing their levels of intention and commitment and consequently, dropout decisions. With this framework established, the question becomes about what an institution can actionably do, and when is the best time to take such action.

THE CASE FOR EARLY INTERVENTION

Early intervention is defined as initiating intervention at the earliest time possible after a problem has been identified (Seidman, 2005). Intervention should be offered early in a student's academic career, or even before the first official enrollment at the institution. Examples of interventions, designed to help incorporate students into college life before the first day of class include orientation programs and summer activities. Programs such as these are geared toward helping students understand what is expected of them, both socially and academically, as they transition into the institution (Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition 2nd Ed., 1993).

Tinto postulates that students must achieve "rites of passage" in order to successfully assimilate into the college environment. Based on Van Gennep's model, Tinto says that a student must separate from communities of the past, transition to the institutional communities, and incorporate into the social and academic systems, both formal and informal, in order to persist to degree completion. Some students have an easy adjustment period, while others may struggle to transition into the institutional framework (Tinto, 1993).

The timing of intervention is critical, and research repeatedly suggests that intervention should occur sooner rather than later at the first sign of student distress. Many researchers suggest that intervention in the first six weeks of college is critical (Simons, 2011) (Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition 2nd Ed., 1993) (Kuh, 2007). Research shows there are lasting effects for students that do poorly in the first semester of college (Nora, 2005). Nora states that how a student performs academically will impact his or her academic and social experiences, his or her commitment to attaining a degree, and ultimately his or her decision to withdraw from college (pg. 134). These students are specifically high risk for dropping out of college in a year or two.

EARLY ALERT

Most data on Early Alert systems comes from survey data and relays perceived effects of the program, rather than actual assessment. Further, many studies are longitudinal in nature, seeking to evaluate less of an immediate impact. These studies often target at-risk populations, or assess the impact of the Early Alert on utilization of another program. This study seeks to differ from those in that it looks at the immediate impacts of Early Alert on the freshman cohort.

In a 2006 article, William Hudson evaluates a pilot study of freshmen at Morehead State University in Kentucky during the spring 2003 semester. Administrators implemented a web form for instructors to send referrals related to excessive absenteeism to their Academic Support office. The forms were submitted during the 2nd, 4th, and 6th weeks of the term. Hudson reports positive gains in course success, and his recommendations are to enhance follow-through/follow-along to ensure reported students remain successful in all courses, to support interdepartmental cooperation to locate students, and implementing the early alert system on a continuous basis (Hudson, 2006).

In a 2013 article, Howard and Flora look to draw inter-institution comparisons among six liberal arts universities. They target Early Alert, but find that all six schools in the study have a system present on their campus, thus there are no institutions without an Early Alert system against which to compare.

Overall, there exists a large number of dissertations written on the subject of Early Alert. Almost all of them rely on survey data. Jill Simons (2013) did a national survey of academic officers at not-for-profit four-year institutions to determine how widespread the use of Early Alert is. She concluded that the use was more common among institutions with small campuses and at institutions with moderate to low admissions standards. She also determined that most Early Alert programs were newer initiatives on campus (less than five years old). She also finds that institutional support for Early Alert is generally limited across these campuses, and encourages institutions to demonstrate retention outcomes as related to the Early Alert program.

In another dissertation, Steven Asby (2015) surveyed students that were the target of an Early Alert, and found that students perceive Early Alert as a conduit "between the student and the institution, impacting their educational satisfaction, motivation to seek resources, communication with campus officials, and to their overall sense of belonging" (pg. 2).

This study differs from other studies because it attempts to discern any significant effects of the timing of the early alert on a continuous basis, rather than at fixed points in the semester. It also seeks to gauge if indicating that a student could have a successful outcome in a course is indicative of a successful retention outcome to the next semester. Based on the interaction theory described before, the essence of these questions is whether or not Early Alert submitted at the outset of the student matriculation is a viable means of reaching a student with community, and easing the separation, transition, and incorporation of the student into college life.

SUMMARY

Retention has long been studied in the United States, but retention rates have not improved. Prominent theory in the field suggests that student engagement matters to dropout decisions, and that institutional intervention is warranted early in the students matriculation. Studies related to Early Alert systems are largely based on survey data, though some assessments of Early Alert as a driver for other student services, or as a comparison of Early Alert to other programs do exist. This study seeks to determine if the timing of the Early Alert notification, or the possibility of successfully completing the course, help a student to incorporate into the academic environment of an institution in terms of retention to the next long semester.

CHAPTER III

THE LOGISTIC MODEL

The overall design of this study includes a multivariable logistic regression on a binary outcome variable and draws heavily from the methods detailed in **Applied Logistic Regression** by David Hosmer, Stanley Lemeshow, and Rodney Sturdivant (2013). The data set contains both continuous variables and indicator variables as potential covariates. The principles of purposeful selection as outlined in Hosmer, Lemeshow, and Sturdivant are used to select the covariates to be included in the final model. The flexibility for modeling the relationship of continuous covariates with the outcome variable in the case of nonlinearity comes from the method of fractional polynomials or from linear or restricted cubic splines. The data are a sample of full-time, first-time degree seeking undergraduate freshmen for whom a faculty member completed an Early Alert, and information pertaining to the demographic and academic records of those students. Details of these concepts and the study data follow in this, and in the following, chapter.

DEVELOPMENT OF THE LOGISTIC MODEL

The purpose of undergoing a regression analysis is to select a suitable model that adequately reflects the desired outcome using the information provided by the covariates included in the model. For a given observation, the outcome response in a regression analysis can be described by $y = E(Y|x) + \varepsilon$, where $E(Y|x)$ is the conditional mean of the response variable given the covariate(s) x, and ε is an error term. Adequately modeling the outcome variable is dependent upon using the appropriate functional form that determines $E(Y|x)$, as well as knowing the distribution of ε. An examination of these two components individually provides the framework for the logistic regression model.

The Functional Form of $E(Y|x)$

When working with a set of observations where the response value is binary, either 0 or 1, the calculated mean of these observations will be strictly between 0 and 1 inclusive. Specifically, if there are n observations y_i in the data set, $0 \leq \sum_{i=1}^{n} y_i \leq n$. By definition, the arithmetic mean of $\vec{Y}$, the vector of n response observations, is $\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$. It is easily seen that this value will fall between 0 and 1, inclusive, because this calculation indicates the proportion of responses with a positive outcome in the data. Therefore, it does not make sense to choose a function that can take on any value as the value(s) of x vary, yet a function that represents a linear combination of the covariates from which we can find optimized parameter estimates is still required.

Hosmer, Lemeshow, and Sturdivant point out that many functional forms have this quality, but the logistic function is easy to work with and lends itself to meaningful interpretation of effects. To see how the logistic function exhibits properties desirable to the regression problem, consider the logistic function $h(t)$.

$$h(t) = \frac{e^t}{1+e^t} \tag{3.1}$$

A graph of this function for $t \in \mathbb{R}$ demonstrates that this function can take any real-valued input, but the output is restricted to values between 0 and 1 (see Figure 3).

Let $x' = \left(x_1, x_2, \dots, x_p\right)$ denote a set of p real-valued independent variables and suppose that t is a linear combination of those variables. Write $t(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. Then the logistic function in 3.1 becomes

$$\pi(x) = E(Y|x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \tag{3.2}$$

and describes the cumulative distribution function for the probability that $Y = 1$.

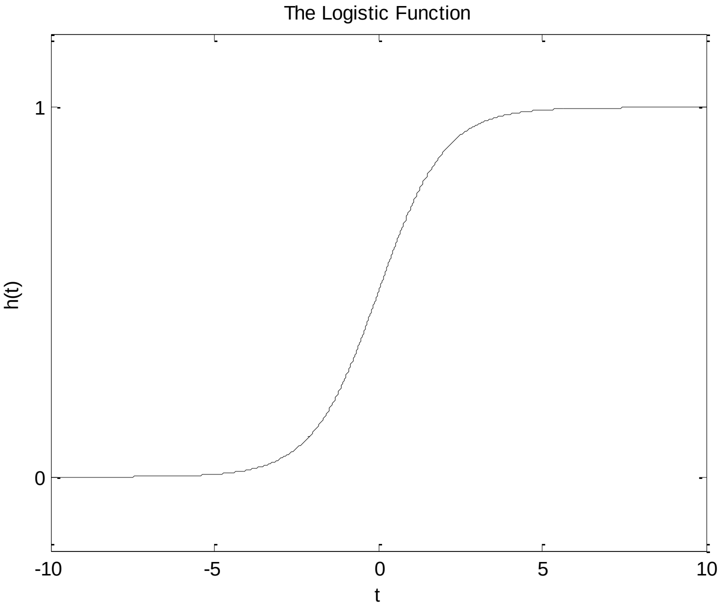


Figure 3 The logistic function in equation 3.1.

The Error Term, ε

In the case of a dichotomous outcome variable, a response, y, can only take on the values of 0 or 1. Since we seek to model y with an expression of the form $y = \pi(x) + \varepsilon$, if $y = 1$, $\varepsilon = 1 - \pi(x)$ and by equation 3.2 the probability of this is $\pi(x)$. If $y = 0$, $\varepsilon = -\pi(x)$ and the probability of this is $1 - \pi(x)$. Thus, if the sample size is sufficient, we can define a binomial distribution for the error term, with a mean of zero and variance equal to $\pi(x)[1 - \pi(x)]$ for any value(s) of x. So the appropriate model for the error term is a binomial model with the mean equal to the conditional probability found in equation 3.2, $\pi(x)$, and with variance equal to $\pi(x)[1 - \pi(x)]$.

FITTING THE LOGISTIC REGRESSION MODEL

At this point, it is not enough to establish a functional form for the desired conditional probability and a distribution for the error term. The goal of regression is to find the model with the best fit, and to do this, optimization of parameters is required. To this end, the method of maximum likelihood is used to obtain a set of parameter estimates that maximize the probability of obtaining the observed set of data. A set of likelihood functions must be determined and the system solved to find the optimum parameter values.

The logistic function in equation 3.2 gives the probability that $Y = 1$ given the value(s) of x, $Pr(Y = 1|x) = \pi(x)$. Likewise, $Pr(Y = 0|x) = 1 - \pi(x)$. For the i^{th}

observation in the data set, $Pr(Y_i = 1|x_i) = \pi(x_i)$, and $Pr(Y_i = 0|x_i) = 1 - \pi(x_i)$, where x_i may be a vector of variables, is the contribution of that observation to the likelihood. For any observation in the data set, this contribution may be expressed as $\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$. For a set of independent observations, the likelihood function obtained is

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3.3}$$

It is mathematically prudent to work with the log-likelihood, which is

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]\} \tag{3.4}$$

In the case that x_i is a vector of p variables, β will also be a vector with $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$. Differentiating equation 3.4 with respect to each of $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ gives a system of $p + 1$ equations that maximize $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ when they are set equal to zero and solved. These equations are

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

$$\text{and} \tag{3.5}$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \text{ for } j = 1, 2, \dots, p.$$

The solution to this system will be easily provided by statistical software. The values obtained from this procedure are the maximum likelihood estimates for the vector β, denoted $\hat{\beta}$, and these values can be used to obtain an estimated probability of a positive outcome conditional on the value(s) of x by evaluating expression 3.2 using $\hat{\beta}$. Denote this result as $\hat{\pi}(x)$.

THE LOGIT FUNCTION

The logit function for a logistic regression using equation 3.2 is given as

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{3.6}$$

where x and β may both be vectors. Note that the logit, $g(x)$, can take on any value as x varies, may be continuous, and is linear in the parameters. In the case of a continuous covariate, a logistic model is required to be linear in the logit for appropriate interpretability of the model. The logit function depends upon the ratio of probabilities, specifically $Pr(Y = 1|x) = \pi(x)$ and $Pr(Y = 0|x) = 1 - \pi(x)$. Evaluating the logit function using estimated parameter values provides a single estimator that allows direct comparisons among covariate values. In the instance that a continuous covariate is not linear in the logit, the method of fractional polynomials will be used to adjust the model. Further details on interpreting the value of this function and the method of fractional polynomials will be offered in later sections.

MODEL SELECTION

As is the case in any multivariable regression analysis, after fitting a model, predicted values are compared to observed values to assess the fit of the model to the data and the coefficients in the model are tested for significance. In multivariable regression, there are many possible models to choose from with varying numbers of covariates.

Methods for comparing competing models, methods for testing coefficients for significance, and the method of purposeful selection of covariates are detailed below.

Assessing Fit of the Estimated Model

In logistic regression, a likelihood ratio test is used to assess the fit of the model. The likelihood ratio test measures the deviance of the fitted model. This statistic refers to how well the proposed model fits the data as compared to the saturated model, that is, the model that has one data point for each parameter to be estimated, resulting in a deviance equal to one. For a logistic model, the deviance is defined to be

$$D = -2\ ln\left[\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model}\right]. \tag{3.7}$$

In the saturated model, there are n parameters being estimated (one corresponding to each data point), while the fitted model estimates $p + 1 < n$ parameters. D follows a $\chi^2\big(n - (p + 1)\big)$ distribution. The hypothesis for this test is

$$H_0\!: The\ smaller\ model\ holds.$$

$$H_1\!: The\ larger\ model\ holds.$$

In the event that the null hypothesis is rejected, the conclusion is that the estimated model is not a good fit for the data.

Comparing Competing Models

If two models with differing numbers of parameters are estimated on the same set of data, the partial likelihood ratio test serves to compare competing models. Because the likelihood of the saturated model is equal to one, equation 3.7 reduces to

$$D = -2\ ln[likelihood\ of\ the\ fitted\ model]. \qquad (3.8)$$

If one or more covariates are removed from the model, it can be determined if the deviance changed significantly by computing the difference

$$G = D(model\ without\ the\ covariate) - D(model\ with\ the\ covariate) \qquad (3.9)$$

Because the saturated model is common to the deviance of both models, using equations 3.7, 3.8, and 3.9, we can write a comparison of the log-likelihoods of models with differing numbers of parameters as

$$G = \ -2\ ln\left[\frac{(likelihood\ of\ model\ with\ fewer\ covariates)}{(likelihood\ of\ the\ model\ with\ more\ covariates)}\right]. \qquad (3.10)$$

If q parameters are removed from a model that had $p + 1$ parameters, G provides a test statistic that is $\chi^2(q)$ distributed under the null hypothesis that each of the q removed coefficient parameters are equal to zero based on the ratio of the log-likelihoods of the two models. If this hypothesis holds, then the covariates can be removed from the model.

Testing Coefficients

In order to know if an independent variable is significantly related to the outcome variable in a logistic regression, the individual Wald test statistic is used to test whether the individual coefficient is significantly different from zero. If a vector of parameters, $\hat{\beta}$, has been estimated, the Wald statistic for the jth element of the vector, $\hat{\beta}_j$, is

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \tag{3.11}$$

where $\widehat{SE}(\hat{\beta}_j)$ is the standard error of the parameter $\hat{\beta}_j$. This statistic follows a standard normal distribution. If the result of the test is to reject the null hypothesis, there is sufficient evidence to claim that the parameter is significantly different from zero and should be included in the model.

Purposeful Selection of Covariates

The method of purposeful selection of covariates as outlined by Hosmer, Lemeshow, and Sturdivant recommends a combination of statistical tools and analyst judgement to guide model selection. The approach is very straight-forward and at all times recommends that clinical evidence of the importance of a variable in relationship to the outcome should be considered alongside statistical evidence, and this in fact may preclude a statistical exclusion of a variable. Bearing this, the following steps will be used to select the final model for consideration:

1. Perform a thorough univariate analysis of each potential covariate with the response variable. Software accomplishes this task easily. In the case of continuous covariates, formulate a logistic model relating the single covariate to the response. Note the p-value of the likelihood ratio test. In the case that a potential covariate is categorical, a χ^2 test of independence can be used. Note that the expected cell frequency must be sufficient for this statistic to be valid. Consider any variable that is significant with a p-value of 0.25 or less upon

univariate analysis as a candidate for inclusion in the multivariable model, along with any variable that is clinically significant to the outcome.

2. Fit the multivariate model. Use the Wald statistic to assess the significance of the covariates in the model. Note the value of the estimated coefficient, the odds-ratio (based on the logit function, to be discussed in the section on interpreting the model), and confidence intervals for both parameter estimates. Use traditional levels of significance, which in this paper will be $\alpha = 0.10$, to remove covariates, one or two at a time, beginning with the variables with the largest p-values. Record all pertinent statistical output.
3. Fit a reduced model that excludes the covariates identified in the previous step. Record the pertinent statistical output. Compare the values of the estimated coefficients in the reduced model to the corresponding value in the previous model. Look for changes of 20% or more in the remaining coefficient values compared to the previous model. Any large changes may be indicative of an interactive effect. Compute a partial likelihood ratio test to compare the models. Cycle through steps two and three until all remaining variables are either clinically or statistically significant.
4. One at a time, introduce the covariates that were excluded in step one into the model remaining at the end of step three. Retain any covariates that are significant. The model produced in this step is the preliminary main effects model.
5. Examine the continuous covariates for linearity in the logit. Two methods are used to accomplish this graphically. First, a locally weighted regression, specifically a Lowess smoother is used to plot the continuous covariate against the estimated logit values. Second, the method of design variables will be used. Further details on these procedures are below. In the event that a continuous covariate is not linear in the logit, the methods of fractional polynomials or splines will be used to allow flexibility in modeling the outcome.
6. Examine the model for interaction effects. Interaction effects occur when the logit applied to a continuous covariate is linear at different rates for differing values of a categorical covariate. Interaction effects are modeled by a numerical product of the covariate values in the data set, and this term is included in addition to the two or more participating variables. Significant interactions are retained at the $\alpha = 0.05$ level. The conclusion of this step produces a preliminary final model, which then must undergo diagnostics.

ASSESSING LINEARITY OF THE MODEL

Several methods are available to assess the linearity of continuous covariates on the logit scale. Two methods will be used primarily. First, design variables based on the percentiles of the covariate are used, and then the findings are corroborated with a smoothed scatter plot.

Design Variables

To use a design variable to assess linearity, first obtain the quartiles of the continuous covariate. Code 1/0 design variables for subjects belonging to each of the upper three quartiles. Fit the model replacing the single continuous covariate with the three design variables and record the coefficients. Plot the estimated coefficients against the midpoints of each quartile. Use zero as the coefficient for the first quartile. Examine the plot to discern if a linear form is present. If this plot is linear in form, then the coefficients increase at a constant rate with increasing values of the continuous explanatory variable.

Smoothed Scatterplot

A smoothed scatterplot is a nonparametric locally weighted polynomial regression that uses a subset of data that is nearby a point being estimated to determine a fit function. The amount of data used in the calculation is specified by the bandwidth, which reaches to either side of the data point in question to determine any one estimate.

Data that is closer to the point being estimated is weighted highly compared to data that is further away. Each weighted estimate is plotted and the linearity of the covariate can be discerned. Specifically, in this paper, a lowess smoother is used. Further details on this procedure can be found in the methodology section.

MODELING NONLINEARITY

It is desirable in a regression analysis for continuous covariates to express a linear form in relation to the outcome variable, but this is not always the case. In the event that a continuous covariate is not linear on the logit scale, two methods for modeling nonlinearity are available to be applied. Fractional polynomials provides a flexible method for transforming a continuous covariate into one or two power terms, for which coefficients can be estimated. Splines allow for the data to be modeled over intervals with differing functions in different intervals.

Fractional Polynomials

If the relationship of the continuous covariate to the logit is nonlinear in form, the method of fractional polynomials provides a flexible option for comparing polynomial forms of varying degrees of the continuous covariate, while also offering a strategy to find the most parsimonious polynomial expression. Royston and Altman (1994) propose a methodical search through a limited set of powers, limited to one or two power terms for a continuous covariate. Stata software performs this analytical procedure. The premise is that for a continuous covariate, x, find the best power, p, of x that models the response and then estimate the coefficient for the x^p term in the model. This method restricts the possible set of values that p assumes to the set $\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where 0 is interpreted as the natural log of the variable, and considers at most a two term expansion for the polynomial form.

Particularly, let J be the number of power terms of the covariate to be included in the model. The logit function can be written as

$$g(x, \beta) = \beta_0 + \sum_{j=1}^{J} \beta_j F_j(x) \tag{3.12}$$

where β is a vector of estimated coefficients and $F_j(x)$ is a particular power function of the form x^{p_j} and $p_j \in \wp$ as defined above. If $J = 1$, then

$$g(x, \beta) = \beta_0 + \beta_1 x^{p_1}. \tag{3.13}$$

Methodically applying each of the $p \in \wp$ powers results in a set of eight models with power representations of the continuous covariates that can be compared to each other, and other expanded models.

If $J = 2$, then $F_j(x)$ in equation 3.12 is

$$F_j(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1} \\ F_{j-1}(x) ln(x), & p_j = p_{j-1} \end{cases}. \tag{3.14}$$

Methodically applying each of the possible $(p_1,\ p_2)$ pairs when $p_1, p_2 \in \wp$ results in 36 possible models containing two power terms of the covariate. Note that when $J = 1$ and $p_1 = 1$, the regression is just the linear form of the logit, and is the model we seek to improve by introducing fractional polynomials. Let $L(1)$ denote the log-likelihood of this model. For each of the seven remaining models in the set of eight $J = 1$ models, determine which has the largest log-likelihood, and denote this log-likelihood as $L(p_1)$. Likewise, for the set of thirty-six $J = 2$ models, select the model with the largest log-likelihood and denote this as $L(p_1, p_2)$.

Under the closed testing procedure, compare the best two-term model to the linear model using $G = -2ln[L(1) - L(p_1, p_2)]$, which is $\chi^2(3)$ distributed under the null hypothesis that the $J = 2$ model is not a better fit than the linear model. If the test in not significant, stop and use the linear model. If the test is significant, compare the best two-term model to the best one-term model using $G = -2\ ln[L(p_1) - L(p_1, p_2)]$, which is $\chi^2(2)$ distributed under the null hypothesis that the $J = 2$ model is not significantly different from the $J = 1$ model. If this test is not significant, select the one-term model. If this test is significant, select the best two-term model.

Once this procedure is complete, if a model other than the linear one is chosen, substitute the transformed covariate into the logit equation in place of the linear term and estimate coefficient parameters for the model containing the transformed covariate, along with any other applicable covariates.

One adaptation of this method applies if some instances of the continuous covariate are equal to zero, and the remaining positive values are right-skewed. In this case, a dichotomous variable, d, is coded where the variable is equal to one if there is a positive value on the covariate, and zero if the value of the covariate is zero. This variable is used in conjunction with the fractional polynomial search to estimate the best one-term and two term model. In this case, it may also make sense to examine a log or square root model.

Splines

Linear splines and restricted cubic splines both work similarly in that they divide the range of the continuous covariate into regions over which a function can be approximated based on specified cut points in the data, called knots. Linear splines estimate the slope of a linear function inside each defined region. Restricted cubic splines define a cubic function over the interior regions, and a linear function for the regions at each end of the range of the continuous variable. Many methods are available for defining these variables. Stata software will be used to generate spline variables that will be used in place of the linear term in the regression model. The exact methodology used will be outlined in the next chapter.

There are many possible choices for creating spline variables. The number and location of the knots will produce different models with differing numbers of estimated parameters as they vary. In the event a spline transformation is to be used, both the Akaike Information Criterion (AIC) and Schwartz's Bayesian Information Criterion (BIC) will be examined. These values are estimated on the computed model according to the following formulas: $AIC = -2\ell + 2p$ and $BIC = -2\ell + \ln(n)\,p$, where ℓ is the log-likelihood of the estimated model, p is the number of estimated parameters including the constant, and n is the sample size. Both values penalize the likelihood for added parameters in the model. The primary difference is the rate at which the values are penalized; the former at a constant rate per added parameter, and the latter scales the number of added parameters by a function of the sample size.

When fitting a spline model, comparing the AIC to that of the model with the linear term may well produce a different result than comparing the BIC of the same model to the BIC of the model with the linear term. There are no strict guidelines used in determining which model is best, so the decision will be made holistically, comparing changes in AIC and BIC in a linear term model to those of a model with more parameters due to spline variables. Large improvements in one or both criterion, relative to other models compared, or agreement in improvement in fit by both of these criterion will make a spline transformation a strong candidate for consideration.

MODEL ASSESSMENT

After selection of a preliminary final model, the model is assessed in two stages. The first is to compare the fitted values produced by the model to the observed values to gain some perspective on how well the model reflects the data. The second is to run diagnostics to ensure that the model performs consistently across the data. Goodness-of-fit and diagnostic methods are explained below.

Goodness-of-fit

Goodness-of-fit refers to an assessment of how well the outcome as predicted by the selected model compares to the observed distribution of the outcome variable in the data. In logistic regression, the fitted values represent the estimated probability of $y = 1$

when the logistic function is evaluated at the estimated logit obtained from the calculated coefficients. Each unique combination of covariate values will produce a unique estimated logistic probability. In the case where two or more subjects have the exact same measurements in all the covariate values, the "equivalent" subjects will all have the same estimated logistic probability. This leads to the notion of a *covariate pattern*. When two or more subjects are associated with identical measurements on all covariate values, there are fewer covariate patterns than there are subjects in the data.

Define J to be the number of covariate patterns found in the data. In the extreme case, all the subjects in the data have a unique combination of possible covariate values and then $J = n$. However, if one or more subjects have identical measurements in all the covariates, then the number of covariate patterns, J, is less than n. Let m_j denote the number of subjects with covariate pattern x_j, $j = 1, 2, 3, \ldots, J$, and let y_j denote the number of subjects in the j^{th} covariate pattern with $y = 1$. Then the number of subjects in covariate pattern j with $y = 0$ is $m_j - y_j$. Because all of the subjects within the j^{th} covariate pattern have the same values on all of the covariates, these subjects will all have the same fitted probability based on the model coefficients. Call the estimated logistic probability calculated from the logit for each covariate pattern $\hat{\pi}_j$. These definitions provide the basis for three straight-forward assessments of goodness-of-fit.

Hosmer and Lemeshow Test

The Hosmer and Lemeshow test statistic is a chi-square distribution calculated on a 2 x 10 contingency table where the rows correspond to the two possible values of the outcome variable, and the columns are created by grouping the covariate patterns based on the calculated estimated probability, $\hat{\pi}_j$, according to decile. After grouping the covariate patterns in this way, observed and expected frequencies can be calculated for each group against the outcome variable, and then a Pearson chi-square statistic computed.

For each group, the observed number of subjects with $y = 1$ is the sum of the y_j across the covariate patterns found in the group, while the expected number of subjects with $y = 1$ is the sum of $\hat{y}_j = m_j\hat{\pi}_j$ across the covariate patterns found in the group. Likewise, the observed number of subjects with $y = 0$ is the sum of the $(m_j - y_j)$ across the covariate patterns in the group, and the expected number of subjects with $y = 0$ is the sum of $m_j(1 - \hat{\pi}_j)$ across the covariate patterns. The Hosmer and Lemeshow test statistic, with c_k covariate patterns in group k is

$$\hat{C} = \sum_{k=1}^{10}\left[\frac{\left(\sum_{j=1}^{c_k} y_j - \sum_{j=1}^{c_k} m_j\hat{y}_j\right)^2}{\sum_{j=1}^{c_k} m_j\hat{y}_j} + \frac{\left(\sum_{j=1}^{c_k}(m_j - y_j) - \sum_{j=1}^{c_k} m_j(1-\hat{\pi}_j)\right)^2}{\sum_{j=1}^{c_k} m_j(1-\hat{\pi}_j)}\right]. \tag{3.15}$$

This statistic is $\chi^2(8)$ when $J = n$ or $J \approx n$ under the null hypothesis that the model conforms well to the observed distribution. If the test rejects, conclude that the model does not model the outcome data well.

Some controversy over this test exists. It was derived as a method for overcoming the difficulties of the Pearson chi-square when the expected number of positive outcomes is low in one or more covariate groups. The choice of ten groups is arbitrary, and Hosmer and Lemeshow showed that this number of groups is ideal for adhering to the $\chi^2(8)$ distribution. However, others have shown that the test statistic may be heavily influenced by the choice made for the number of groups. For this reason, two other methods for testing goodness-of-fit are also used.

Osius and Rojek Test

In addition to the Hosmer and Lemeshow test, the Osius and Rojek Test evaluates the hypothesis that the model fits the observed data well based on a test statistic that is a normal approximation to the Pearson chi-square. The test statistic is derived from the data aggregated by covariate pattern. The method for deriving this statistic is as follows:

1. Obtain the estimated $\hat{\pi}_j, j = 1, 2, 3, \dots, J$.
2. Create the variance for each covariate pattern, $v_j = m_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right), j = 1, 2, 3, \dots, J$.
3. Create $c_j = \frac{\left(1-2\hat{\pi}_j\right)}{v_j}, j = 1, 2, 3, \dots, J$.
4. Compute the Pearson chi-square statistic $\mathrm{X}^2 = \sum_{j=1}^{J} \frac{\left(y_j - m_j \hat{\pi}_j\right)^2}{v_j}$.
5. Do a linear regression of c on the covariates using v as a weighting variable. Let RSS denote the residual sum-of-squares from this regression.
6. Compute the correction factor for the variance, $A = 2\left(J - \sum_{j=1}^{J} \frac{1}{m_j}\right)$.
7. Compute a two-tailed p-value from the standard normal distribution using the test statistic $z_{\mathrm{X}^2} = \frac{\left[\mathrm{X}^2 - \left(J - (p+1)\right)\right]}{\sqrt{A + RSS}}$.

If this hypothesis holds, conclude that the model is a good fit for the observed data.

Stukel's Test

Both of the goodness-of-fit assessments have been shown to not perform well under certain circumstances. The Hosmer and Lemeshow test is sensitive to the grouping of the data, while the Osius and Rojek test is sensitive to very large or very small probabilities. Stukel devised a test of the hypothesis that the parameters of a generalized logistic model are equal to zero. This test is devised as follows:

1. Obtain the estimated $\hat{\pi}_j, j = 1, 2, 3, \dots, J$.
2. Compute the estimated logits $\hat{g}_j = ln\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right) = x_j'\hat{\beta}, j = 1, 2, 3, \dots, J$.
3. Compute two new covariates: $z_{1j} = 0.5 \times \hat{g}_j^2 \times I(\hat{\pi}_j \geq 0.5)$ and $z_{2j} = -0.5 \times \hat{g}_j^2 \times I(\hat{\pi}_j < 0.5), j = 1, 2, 3, \dots, J$ and $I(arg) = 1$ if arg is true and zero otherwise.
4. Add z_1 and/or z_2 to the model and perform the likelihood ratio test. If this test rejects the hypothesis, conclude that the reduced model does not fit the data better, and does not display conformity to the logistic function.

Stukel's test is recommended in addition to the first two methods for assessing goodness-of-fit. The model is tested with all three of the methods, and agreement between them will be considered as an adequate fit.

Sensitivity and Specificity

With any model selected, the desire is that the model will assign high probability of the outcome to those that experience the outcome and low probability to those who do not experience the outcome. A two by two classification table is constructed, the columns of which classify subjects according to whether or not they experienced the outcome. The subjects into the two rows is based on a dichotomous classification variable, the division of which is determined by a cut point, c, in the estimated probabilities. If the estimated probability is higher than the cut point, this variable is coded equal to one, and zero otherwise.

The result of this coding is that subjects are divided into one of four possible classifications based on the new variable, c, and the outcome variable, y: positive outcome, positive high probability; positive outcome, zero high probability; zero outcome, positive high probability; zero outcome, zero high probability. The sensitivity of the model refers to the model correctly assigning higher probabilities to the subjects that have the outcome. This can be calculated by dividing the number of subjects with positive outcome, positive high probability by the number of those that have a positive outcome, which equates to a column percentage in the two by two table.

The specificity of the model refers to correctly classifying the subjects who did not experience the outcome. The number of subjects with zero outcome, zero high probability is divided by the number of subjects with zero outcome to give the proportion of those who did not experience the outcome and were correctly classified. The complement of this proportion is the percentage of subjects that did not experience the outcome and were incorrectly classified, and is found by subtracting the specificity from one. The underlying component that determines how subjects are classified is the cut point chosen when assigning the classification variable. A good starting point might be to choose $c = 0.50$, but the sensitivity and specificity are calculated over a the entire range of cut points. A plot of these against the range of cut points will show an intersection that maximizes both of these measures.

In addition to this plot, the estimated probabilities are plotted in histograms according the value of the outcome variable. The distribution of these histograms are compared relative to the cut point that maximizes both sensitivity and specificity. A high degree of overlap indicates that the model does not classify the subjects well. The further apart these distributions are, the better the model does at separating those that experience the outcome from those that do not experience the outcome in terms of the estimated probability.

Sensitivity can be plotted against the complement of specificity based on the cut point that maximizes both measures for each subject to obtain the Receiver Operating Characteristic (ROC) curve. The area under this curve measures the *discrimination* of the model; that is, the estimated probability that under the fitted model a subject with $y = 1$ will have a higher estimated probability, $\hat{\pi}$, than a subject with $y = 0$. At a minimum, the area under the ROC curve is 0.50, and the following guidelines are used to assess model discrimination:

ROC = 0.50	No Discrimination
0.50 < ROC < 0.7	Poor Discrimination
0.70 ≤ ROC < 0.80	Acceptable Discrimination
0.80 ≤ ROC < 0.90	Good Discrimination
ROC ≥ 0.90	Excellent Discrimination

Pseudo-R^2

While not a measure of goodness-of-fit, the pseudo-R^2 provides a method for assessing the improvement in fit by adding covariates to the model as compared to the overall mean in the outcome variable. Rather than explaining the amount of variability of the response that is attributable to the covariates, this statistic indicates the amount of improvement achieved over a model without covariates. While several R^2 measures have been proposed for logistic regression, the $Pseudo\text{-}R^2$ calculated by Stata during the logistic regression will be used.

RESIDUAL ANALYSIS

In addition to overall goodness-of-fit, the model is assessed for fit by individual observations or covariate patterns. As with any regression study, the deviation from the observed data and the model predictions is of interest in that the relative errors should be small and unsystematic. It is desirable to identify observations where there are large disagreements between the observed and predicted values. Two types of residual are used in this analysis, the Pearson residual and the deviance residual. An overall summary statistic for each is computed by summing the squared residuals.

For a fitted model containing p covariates with J covariate patterns, the Pearson residual is $r_j\left(y_j, \hat{\pi}_j\right) = \frac{\left(y_j - m_j\hat{\pi}_j\right)}{\sqrt{m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}}$, where y_j is the number of subjects in the jth covariate pattern with $y = 1$ and m_j is the number of subjects in the jth covariate pattern. The related summary statistic is $\mathrm{X}^2 = \sum_{j=1}^{J}\left[r_j\left(y_j, \hat{\pi}_j\right)\right]^2$. The deviance residual is defined as $d_j\left(y_j, \hat{\pi}_j\right) = \pm\left\{2\left[y_j ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + \left(m_j - y_j\right) ln\left(\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right)\right]\right\}^{1/2}$ where the sign is the same as the sign of $\left(y_j - m_j\hat{\pi}_j\right)$. The corresponding summary statistic is $D = \sum_{j=1}^{J} d_j\left(y_j, \hat{\pi}_j\right)^2$.

If a covariate pattern is omitted from the model fit, then the X^2 and D summaries of residuals will change from a model fit including the covariate pattern. The magnitude of this change depends on the magnitude of the underlying residual of the covariate pattern. This makes it convenient to assess the impacts of each individual covariate pattern on the overall fit of the model and to identify patterns that deviate from the model predictions. By plotting the change in X^2, denoted $\Delta\mathrm{X}^2$ and called the Hosmer-Lemeshow $\Delta\mathrm{X}^2$, against the estimated probabilities, any covariate pattern that produces a large change in the summary relative to the rest of the data will be visible. Similarly, a plot of the Hosmer-Lemeshow $\Delta\mathrm{D}$ against the estimated probabilities will reveal cases where the observation showed a large deviance residual. Further examination of cases identified in this way is warranted to determine if removal is justified.

INTERPRETING THE MODEL

Once the model is selected and any covariate patterns of interest are examined for impact, the model can be interpreted. Suppose that a generic covariate x is to be investigated assuming all other covariates remain constant. Using the fact that equation 3.2, $\hat{\pi}(x)$,gives the estimated conditional probability, $Y = 1|x$ the odds that $Y = 1$ are computed as $\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}$. If two values of the covariate are being compared, say $x = a$ and $x = b$, the odds of $Y = 1|x = b$ can be compared to the odds of $Y = 1|x = a$ as a ratio. Express this as

$$\widehat{OR} = \frac{\frac{\hat{\pi}(x=b)}{1-\hat{\pi}(x=b)}}{\frac{\hat{\pi}(x=a)}{1-\hat{\pi}(x=a)}} \qquad (3.16).$$

Applying equation 3.2 and simplifying derives a simple expression for the odds ratio:

$$\widehat{OR} = e^{\beta_x(b-a)} \tag{3.17}$$

Evaluating this expression with the obtained model coefficient and the calculated difference between the values of the covariate gives a scalar value that describes how the odds of the outcome compares at the designated values b and a. This enables a determination about whether an outcome is more or less likely to occur among subjects with $x = b$ than subjects with $x = a$, and also how much more or less likely that outcome is for those in the former group than for those in the latter.

If the covariate is dichotomous, this simply compares the odds of the outcome for one group to the odds of the outcome for the other. Interpretation for a polychotomous variable that is coded as a series of dichotomous design variables will compare subjects in the target level with everyone not in that target level, as opposed to those with no level. In the case of a continuous covariate, the calculated odds ratio will describe a difference in odds that is dependent upon the scale of the covariate. A difference in units that is meaningful should be used. This method is applicable to covariates of any type, and also generalizes to more complex models where interactions are involved. If this is the case, then the values of one variable should be varied while the other is held constant and the simplified expression for the estimated odds ratio will contain more than one of the estimated coefficients.

For covariates not involved in transformations resulting in a combination of variable representation or involved in interactions, calculation of a confidence interval is given by $\hat{\beta}_x \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_x)$ and is output by software. However, when a covariate is represented by a combination of transformed variables, or is involved in an interaction, the algebraic simplification that derives the odds ratio will contain more than one estimated coefficient. Because of this, the calculation of confidence intervals for the odds ratios requires obtaining the estimated variance associated with each coefficient and the estimated covariance associated with each pair among the included coefficients.

The odds ratio estimator will involve a sum of coefficients, for example, β_{x1} and β_{x2}. Variances always add, and the standard error of the combination of coefficients is estimated by

$$\widehat{SE}(\hat{\beta}_{x1} + \hat{\beta}_{x2}) = \sqrt{\widehat{Var}(\hat{\beta}_{x1}) + \widehat{Var}(\hat{\beta}_{x2}) + \widehat{Cov}(\hat{\beta}_{x1}, \hat{\beta}_{x2}\ )} \tag{3.18}$$

and the confidence interval is estimated as above.

SUMMARY

The logistic regression model is a powerful tool used for modeling a binomial outcome. The conditional probability of the binary outcome is modeled using a logistic function, $\pi(x) = E(Y|x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$ with a binomial error term. Optimization of the log-likelihood function is used to estimate the model parameters.

The method of purposeful selection, which examines the relationship between the potential covariates and the outcome on a univariate basis first, and then on a multivariate basis, relies on both statistical significance and analyst discretion to select a parsimonious model that adequately reflects the experience of the data. Fractional polynomials and splines are presented as methods for modeling nonlinearity in continuous covariates, and for selecting from competing models when considering transformations. Additionally, summary measures for goodness-of-fit, as well as methods for assessing the fit of the model across the covariate patterns are discussed. Finally, a discussion of the interpretation of the fitted model shows that logistic regression provides a succinct way to describe the likelihood that a group of subjects presenting with a covariate characteristic will experience the outcome compared to the likelihood that subjects without the characteristic will experience the outcome in terms of odds.

CHAPTER IV

DATA AND METHODOLOGY

The data for this observational study are a subset of data from a much larger context study. The original data frame consists of all core course enrollments from the fall and spring long semesters of 2014-2015 and 2015-2016. All course enrollments were identified by the course prefix and the course number, but section numbers were removed and no identifying faculty information was included because of considerations related to the larger context study for which this data were harvested. Students were identified with unique anonymized identifiers that allowed for aggregation at the student level. Information about the student associated with the enrollment was also present. For course enrollments associated with an Early Alert submission, Advising Services provided the date and the reasons for the warning. The following sections contain a detailed description of the sample selection, the resulting data, and a statement of the intended methods of model development. A detailed overview of the coding scheme of all variables can be found in Appendix A.

Student Data

The data frame contains several points of data on each student, with both indicators and continuous measures as potential predictors. Because of the larger study context from which these data are drawn, there are several variables in the file that are not relevant for this analysis because they represent the student status at the time the data were harvested in the spring of 2017. Examples of this are any transfer hours credited, the cumulative GPA, and the current (at the time of collection) standing. Variables excluded from analysis for this reason are also excluded here.

In terms of student demographics, a fairly extensive set of descriptors is available. Veteran status, the student's gender, and first-generation status are dichotomous variables, while the student's race/ethnicity was coded as one of Hispanic, African American, Asian, Native Hawaiian/Pacific Islander, Native American/Alaska Native, White, multiple races/ethnicities, or unknown race/ethnicity. Additionally, this field identifies International students, as no race/ethnicity data is collected on them. Due to small numbers in several of the minority categories, this variable is collapsed into a variable with four categories; White, Hispanic, African American/Black, and Other. The student's age at the time of enrollment is included as a continuous covariate.

In addition to the demographic descriptors, each student's ACT or SAT score is a potential covariate as a measure of pre-college ability. All students admitted were tested prior to the term of entry. For each student with an ACT score present, the ACT score was retained for analysis regardless of whether the student had an SAT score. For students presenting with only an SAT score, the scores were concorded to an equivalent

ACT score in accordance with the table released in a joint statement by the ACT/College Board. The concordance table can be found in appendix B.

Several variables to describe the residency of the students are included in the data set. Residency can refer to the tuition rate at which the student is billed, or it can refer to the actual living situation of the student. Data for the student's tuition rate, either in-state, out-of-state, or foreign country, which is based on the related THECB county code, the student's zip code which may in fact represent the parent's zip code, and whether or not the student lives on campus is available. However, Tinto's model is focused on student interaction so this study only considers whether a student lives on campus or not. It is WTAMU policy that entering freshmen live in the dorms, but exceptions to this policy are made.

Information was obtained about the financial aid status of each student. For those that completed a FAFSA as an entering student, the family income, the expected family contribution, and the calculated financial need were all present. These data elements were summarized in polychotomous indicator variable with categories as follows: no information, FAFSA with no need, and FAFSA with need, as a summary of the student's financial security. Those with no information did not submit an application for need-based financial aid, while those with no need did apply for this type of aid, but were found to not have a need. This variable, in combination with indicators for whether or not the student received a Pell grant and whether or not the student received a student loan (parent loans are excluded), will be used to compare financial aid statuses among students as it relates to retention. The total amount of tuition and fees charged to the student is available in the file. Because WTAMU only charged out-of-state students $30 per credit hour more than in-state students as a matter of institutional policy during the time of these enrollments, the student bills are very similar with a few exceptions.

For students that participated in developmental education, three variables are summarized into a single variable that reflects the overall preparation of the student. Participation in an NCBO only, participation in one developmental subject, and participation in both developmental English and Math are compared to students who did not require any developmental preparation.

The remaining potential covariates are related to student performance during the term of entry. The total credit hours attempted is used on a continuous scale. Additionally, the total number of online hours attempted by the student is included. While midterm and final grades for all core courses were included in the data where available, these were not considered because not all students carried their entire course load in the core curriculum, and the term GPA is considered as a continuous measure of academic performance in all courses. Last, the year of entry is included as a dichotomous variable to control for any influence that may be due to the differing points of entry for the students in the sample.

Selecting the Sample

From the data frame with core course enrollments for all undergraduates, course enrollments associated with early alert submissions were isolated. The enrollments were further narrowed to those associate with first-time freshmen using the enrollment codes corresponding to admission as a recent high school graduate or entering with a GED, though none of the latter were present in the data. Enrollments by full-time students were selected based upon the student carrying twelve or more hours during the first fall term. Finally, enrollments by non-degree seeking students were excluded based upon the student having a non-degree major code. A list of remaining student identifiers was made and duplicates were removed to obtain a list of 341 students. Further examination of the potential covariates resulted in the exclusion of one further student for lack of a viable ACT/SAT score. Otherwise, all students were complete in all data points available. The final result is 340 students selected for analysis.

Early Alert Data and Research Variables

Advising Services compiled a data set for the Early Alerts related to core course enrollments during the semesters in question. Some students received Early Alerts in multiple courses, while some received multiple Early Alerts in a single course. All students in the sample were the subject of at least one Early Alert. For each Early Alert issued, the date the Early Alert was submitted, along with pertinent information regarding the instructor's recommendations and the outcome by Advising Services was provided. For each student, the first occurrence of Early Alert was isolated and the number of days elapsed in the term was calculated. Since early intervention is a focus of these research questions, the timing of the alert into the term could indicate when the institution has the first indication that a student is in peril.

Also included by Advising Services were the “reasons” the instructor has for issuing the EA. A dichotomous variable was set equal to one if the faculty member indicated that the student was either “passing” or “not passing, but with improvement could pass” and was set to zero if a drop was recommended. If faculty encouraged a student to continue by marking either “passing” or “not passing, but with improvement could pass” and the student re-enrolls in the next long semester, this could be evidence that the institutional community successfully reached the student with the necessary support for persistence.

The Outcome Variable

Retention to the following spring semester is coded as a dichotomous variable, with a value equal to one representing a student that persisted and a value equal to zero indicating the student did not return for the second semester of the freshman year.

SAMPLE SUMMARIES

The sample data consists of 340 students that were full-time first time freshmen in either the Fall 2014 or Fall 2015 entering cohorts. Further, students belonging to this population were enrolled in core academic courses in which a faculty member submitted an Early Alert warning to Advising Services. Because the original data file was limited to enrollments in core courses, the numbers of full-time first time degree-seeking freshmen and available demographics were compared to publicly available data sources

(NCES Data Center, THECB Accountability System, Office of Institutional Research website). The number of freshmen found in the original data file, as well as the number in the sample were compared to assess the viability of the sample compared to the entire entering class to ensure all freshmen were captured for potential sampling and highlight any subpopulations that may be more often the targets of Early Alert warnings.

The original data file captured all but five full-time first time degree-seeking freshmen that entered the university in the two falls under consideration. The combined total of entering freshmen is 2,705 of which 2,700 had at least one core course in their course load, and thus were present in the data set. Since the sample was selected based on the presence of Early Alert data for the student, comparisons are made between the sample data and the available cohort demographics.

Remarkably, the sample differs from the fall cohorts in the distributions of gender, developmental education status, and the outcome variable. The entering cohorts are 47.4% male students, while the sample is 60.6% male. In terms of developmental education status, 73% of the entering combined cohort entered prepared and did not participate in developmental courses, while only 57% of the sampled students entered prepared. In the sample, about 10% of students enrolled only in an NCBO, 22% took developmental courses in one subject, and 11% took developmental courses in both English and Math, compared to 5%, 16%, and 6% respectively among all the entering students. The retention rates to the next long semester differ by over 20% from the combined entering cohort and the sample. 86.7% of the overall entering cohort retained to the next spring semester, but only 65.9% of the sample students retained to the next spring.

Also notably, the percentage of sample students that achieved a 2.0 GPA or better the first term of enrollment is considerably higher among the freshmen in general than in the sampled freshmen, at 80.2% and 36.2% respectively. The average term GPA for entering freshmen is 2.67, while it is 1.37 for the sampled students.

These comparisons indicate that differences between the general freshmen and those that are the subject of an Early Alert may exist and the results of this analysis should not be generalized to an entire freshmen class. Appendix C contains tables with detailed comparisons between the combined cohort of entering freshmen and the sampled freshmen on all the potential covariates. The aim of this study is to assess the impact of the timing of the Early Alert, and the impact of the possibility of passing, on the retention outcome, while controlling for these variables. There is no outside of sample data available for further model validation.

Sample Retention Rates

Table 1 shows how the sample is divided between levels of each of the categorical covariates with counts and the percentage of the sample these subjects comprise. Additionally, the count of retained students in that level is shown, along with the percentage of each level that is retained and not retained. The percentage of students retained is markedly different for the levels of most potential covariates, except year of entry and gender. Additionally, the research variable indicating the Early Alert reasons does not show a large difference in the retention rates between the two levels.

Table 1 A summary of categorical covariate levels.

Retention by Categorical Covariate Levels

Variable	Level	Count	Retained	% Retained	% Not Retained
Year	2014	155	99	64	36
	2015	185	125	68	32
Dorm	On-Campus	267	191	72	28
Resident	Off-Campus	73	33	45	55
Gender	Male	206	134	65	35
	Female	134	90	67	33
Race/	White	167	102	61	39
Ethnicity	Hispanic	106	75	71	29
	African-American/Black	40	28	70	30
	Other	27	19	70	30
First	First Generation	167	103	62	38
Generation	Not First Generation	173	121	70	30
Financial	No Information	53	40	56	44
Status	Has Need	244	160	66	34
	No Financial Need	43	24	75	25
Pell	Pell Received	147	120	71	29
	No Pell Received	193	104	62	38
Loan	Loan Received	182	140	77	23
	No Loan Received	158	84	53	47
Preparation	Prepared	195	131	67	33
Status	NCBO Only	32	20	63	37
	One Subject Required	76	55	72	28
	Two Subjects Required	37	18	49	51
Possible	Passing or Could Pass	275	179	65	35
Pass*	Drop Recommended	65	45	69	31

The continuous covariates are summarized in appendix C for the sample. In terms of retention, some interesting patterns emerge in these covariates. First, what is not surprising is that 97% of the sampled freshmen are 17, 18 or 19 years old, and that retention among each of those ages is about 66%, which is similar to the overall retention rate. About half of the older students were retained to the next long semester. Also not surprising is that 70% of the sampled students attempted between 12 and 14 credit hours, inclusive. Those attempting 15 or more credit hours retained at a slightly higher rate, 74%, compared to 62% among those that took less than 15 hours.

Relatedly, 89% of the sampled students did not attempt any hours in the credit load online. Another 10% attempted three hours online. Two students took two online courses, and one took a full-time schedule online. Retention rates are comparable among those that attempted online coursework and those that did not, and it is comparable to the overall retention rate.

The variable for the total bill charged to the student is highly right-skewed, with a single case having a bill that is almost double the next highest bill. The side-by-side boxplots show that the distribution of the bills is largely concentrated in the neighborhood of $3,500 to $5,000. The distribution of the bills is similar for students that retained and those that did not retain, but for the one extreme outlying value that did not retain (see Figure 4).

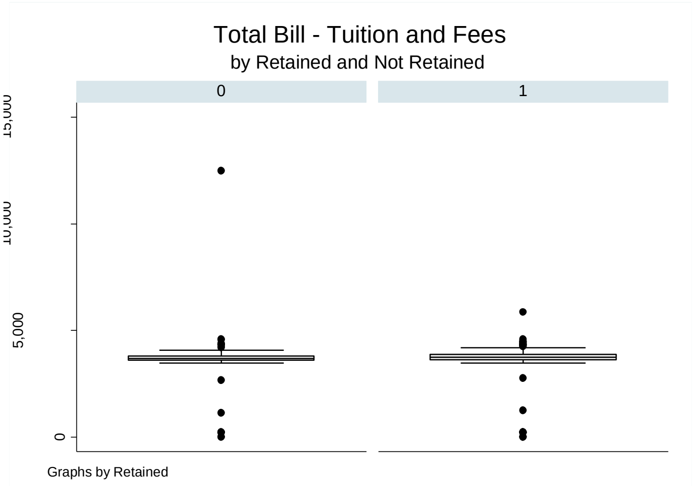


Figure 4 The distribution of the total bill charged to the student according to whether or not the student retained to the next long semester.

The ACT scores are fairly normally distributed among the sample. The histogram in Figure 5 shows a unimodal distribution, with most students scoring 14 and 19. The darker shaded portion of the bars shows the proportion of the students in that value bin that retained. The proportions are fairly similar in the bins where scores most commonly occur.

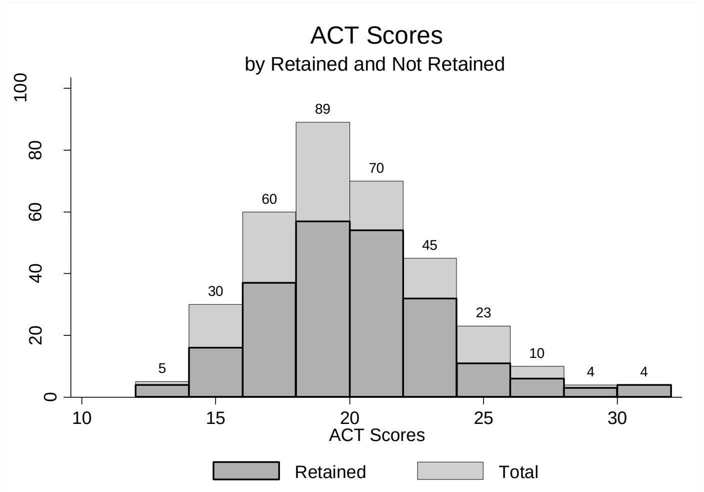


Figure 5 The distribution of ACT scores among the sampled freshmen. The darker shading shows the number of students in the bin that retained to the next long semester.

The average term GPA among the sampled students is well below 2.0, at around 1.37. Figure 6 shows that almost one third of these students achieved a GPA between 0.00 and 0.50. The shaded portion of the bar shows that the majority of these students do not retain to the following semester. Also unremarkably, at the higher GPA values, the retention rates are higher.

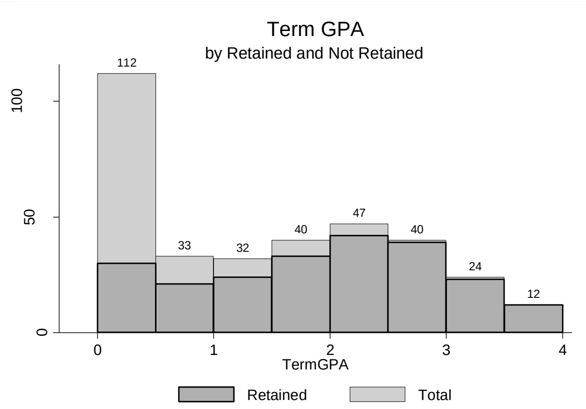


Figure 6 The first term GPA of the sampled freshmen. The darker shading shows the number of students retained to the next long semester.

The histogram for the timing of the Early Alert notifications in Figure 7 shows a bimodal distribution with modes around 40 days into the term, and again after 60 days into the term, which is approaching the drop date. Interestingly, retention rates seem to be higher the later into the term the first Early Alert is triggered. This is counter to common wisdom that says earlier intervention is better.

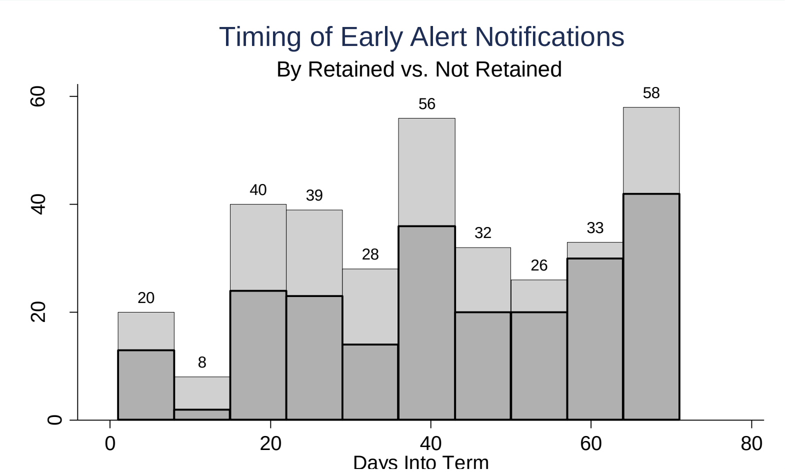


Figure 7 The distribution of the number of day into the term at the time the first Early Alert was submitted. The darker shading shows the number of students that were retained to the next long semester.

METHOD OF ANALYSIS

The methods outlined in Hosmer, Lemeshow, and Sturdivant Applied Logistic Regression are applied to the dataset via Stata 14.2. Automated commands include those used to obtain logistic coefficients and odds ratios in certain cases. Additionally, Stata automatically produces all statistical tests except Osius-Rojek and Stukel's test. Post-estimation and graphics commands produces all figures contained in this study.

SUMMARY

From a much larger data frame containing core course enrollments, 340 full-time, first-time degree-seeking freshmen in core courses in the Fall 2014 and Fall 2015 semesters that were the subjects of Early Alert notifications are sampled. The sample compared to the overall entering cohorts in those two falls has a higher proportion of male students, a higher proportion of minority students, and a higher proportion of underprepared students. The sampled students also carried a much lower term GPA and exhibited a much lower retention rate than the entering freshmen overall.

An extensive set of potential covariates is available. There are interesting differences in retention among students in almost all levels of the potential covariates. The total bill charged to the student in tuition and fees shows an extreme outlying value. The model will be formulated according to the method of purposeful selection in Hosmer, Lemeshow, and Sturdivant using Stata 14.2.

CHAPTER V

MODEL SELECTION AND ANALYSIS

Univariate Analysis

The first step in purposeful selection of the covariates is to perform an individual analysis of all covariates against the outcome variable. For continuous covariates, a univariate logistic regression is run. For each covariate, and univariable logistic regression is conducted. Variables that are significant at the $\alpha = 0.25$ significance level are carried to the initial multivariable model. The results of univariate analysis for all covariates are presented in Table 2.

The univariate analysis statistically excludes almost half of the available variables. The retained variables are used to formulate the preliminary main effects model. Notably, the research variable pertaining to the reasons for the Early Alert notification is not significant with a p-value of 0.53. Since this model seeks to determine if this variable is significantly related to the outcome of retention, the p-value excludes this variable as a candidate for the initial modeling step.

Table 2 Results of Univariate Analysis

Categorical Covariates				Continuous Covariates			
Variable	Pearson chi-squared	p-value	Carry forward	Variable	LRT chi2(1)	p-value	Carry forward
Year	0.513	0.474	No	Age	0.010	0.913	No
Dorm Status	17.681	0.000	Yes	ACT Score	0.920	0.338	No
Gender	0.162	0.688	No	Total Bill (hundreds)	0.000	0.95	No
Race/Ethnicity	3.378	0.337	No	Total Hours	3.220	0.073	Yes
First Generation	2.583	0.108	Yes	Total Online	0.610	0.433	No
Financial Status	4.118	0.128	Yes	Term GPA	136.30	0.000	Yes
Pell Status	2.728	0.099	Yes	Days Into Term	10.04	0.002	Yes
Loan Status	21.239	0.000	Yes				
Preparation Status	6.620	0.085	Yes				
Poss. Pass	0.401	0.527	no				

The Preliminary Main Effects Model

The results of fitting the initial multivariable model are shown in Table 3. Based on the Wald statistics for each coefficient found in the table, the variables are ordered for removal according to the p-value. In the case of a categorical variable with more than two levels, the smallest p-value among all the levels is considered when ranking for removal. The results of this fit indicate that the variable representing the total hour load is the first candidate for removal.

Table 3 The results of fitting the first multivariable model.

	Initial Multivariable Model				
	LL = -132.78137		Pseudo R2 = 0.3915		
	LR chi2(12) = 170.87		Prob > chi2 = 0.0000		
Variable	Coefficient	Std. Err.	z	p-value	Remove
Dorm student	0.589	0.372	1.58	0.114	3
First Generation	-0.262	0.333	-0.79	0.431	2
Financial Status					
Has Financial Need	-1.049	0.563	-1.86	0.062	5
No Financial Need	0.016	0.631	-0.03	0.980	5
Pell	1.447	0.399	3.63	0.000	
Loan	0.752	0.348	2.16	0.031	
Total Hours	0.018	0.121	0.15	0.884	1
Preparation Status					
NCBO Only	0.239	0.545	0.44	0.661	4
One Dev. Subject	0.430	0.393	1.09	0.275	4
Two Dev. Subjects	-0.898	0.531	-1.69	0.091	4
Term GPA	1.522	0.187	8.12	0.000	
EA Days Into Term	0.020	0.009	2.26	0.024	
constant	-2.542	1.814	-1.40	0.161	

Removing the variable for the total hour load from this model shows no large impacts to the coefficients remaining in the model. Additionally, a likelihood ratio test does not support keeping the variable in the model (chi2(1) = 0.02, p = 0.88). Removal of the other variables, one at a time, produces large changes in the values of the coefficients. Table 4 shows the impacts on the coefficients as compared to the initial model excluding the total hours attempted (as a percent change), as well as the results of the partial likelihood ratio test. The test rejects for the financial status of the student, but all of these variables are retained in the model due to the impacts on the remaining coefficients.

Table 4 The impacts of removing variables from the preliminary main effects model.

Variable Removed	Coefficient Impacted	% Change	LR Test (p-value)
First Generation	Financial Status – No Need	453	Chi2(1) = 0.63 (0.43)
Dorm Status	First Generation	274	Chi2(1) = 2.57 (0.11)
	Financial Status – Has Need	105	
	Financial Status – No Need	7,592	
	Pell	100	
	Loan	-98	
	Preparation – NCBO Only	169	
	Preparation – One Subject	-37	
	Preparation – Two Subjects	107	
	Term GPA	-100	
	Days Before Drop	170	
Preparation	First Generation	-28	Chi2(3) = 5.67 (0.13)
	Financial Status – No Need	-438	
Financial Status	First Generation	-35	Chi2(2) = 6.24 (0.04)
	Pell	-36	
	Loan	-20	
	Preparation – NCBO Only	-47	

Adding the variables excluded at the conclusion of univariate analysis one at a time to the model does not produce statistically significant results except in the case of the variable for the total bill. The results of this step are shown in Table 5. The likelihood ratio test of this model including the total bill supports the larger model (chi2(1) = 5.04, p = 0.025). This variable is included in the analysis going forward, though it should be noted that there exists an extremely high outlying bill in the data.

Table 5 The results of adding excluded variables. * indicates a significant p-value.

Variable	Coefficient	Std. Error	z	p-value
Year	0.205	0.327	0.63	0.53
Gender	-0.488	0.340	-1.44	0.15
Race/Ethnicity – Hispanic	0.429	0.370	1.16	0.25
Race/Ethinicity – African American	0.444	0.554	0.80	0.42
Race/Ethnicity – Other	0.393	0.635	-1.74	0.08
Possible Pass	-0.234	0.452	-0.52	0.60
Age	-0.036	0.126	-0.29	0.77
ACT Score	0.028	0.051	0.55	0.58
Total Bill (in hundreds)	-0.0004	0.0002	-2.30	0.02*
Online Hours	-0.077	0.129	-0.60	0.55

The results of fitting the model without the total hours and with the total bill are shown in Table 6. The likelihood ratio test supports the addition of the variable and rejects the model without the total bill (chi2(1) = 5.04, p = 0.025). This model is adopted as the final preliminary main effects model. The continuous covariates in this model, the Term GPA, the Total Bill, and the number of days before the drop date of the EA, will all be assessed for linearity in the logit.

Table 6 The results of fitting the preliminary main effects model.

	Preliminary Main Effects Model			
	LL = -130.2734		LR chi2(12) = 175.89	
	Pseudo R2 = 0.4030		Prob > chi2 = 0.0000	
Variable	Coefficient	Std. Err.	z	p-value
Dorm student	0.714	0.380	1.88	0.06
First Generation	-0.250	0.337	-0.74	0.46
Financial Status				
Has Financial Need	-1.306	0.584	-2.24	0.03
No Financial Need	-0.251	0.643	-0.39	0.70
Pell	1.543	0.410	3.77	0.00
Loan	0.780	0.351	2.22	0.03
Developmental Status				
NCBO Only	0.249	0.568	0.44	0.66
One Dev. Subject	0.431	0.391	1.10	0.27
Two Dev. Subjects	-0.861	0.535	-1.61	0.11
Term GPA	1.618	0.199	8.12	0.00
Total Bill in hundreds	-0.036	0.016	-2.30	0.02
EA Days Into Term	0.018	0.009	2.04	0.04
constant	-0.929	0.840	-1.11	0.27

Assessing Linearity

Visual assessment of linearity of each of the continuous covariates on the logit scale is shown using design variable and a lowess smoothed scatterplot. In the case of apparent nonlinearity, the variable is transformed with fractional polynomials. If no suitable fractional polynomial representation of the variable is found, splines may be applied.

Term GPA

The smoothed scatter plot of the term GPA on the logit scale found in Figure 8 shows a positive linear form. The plot of the design variable coefficients against the midpoints of the quartiles supports this. Neither of these plots suggest a departure from linearity in the logit for this variable. The covariate will be retained as a linear term in the model.

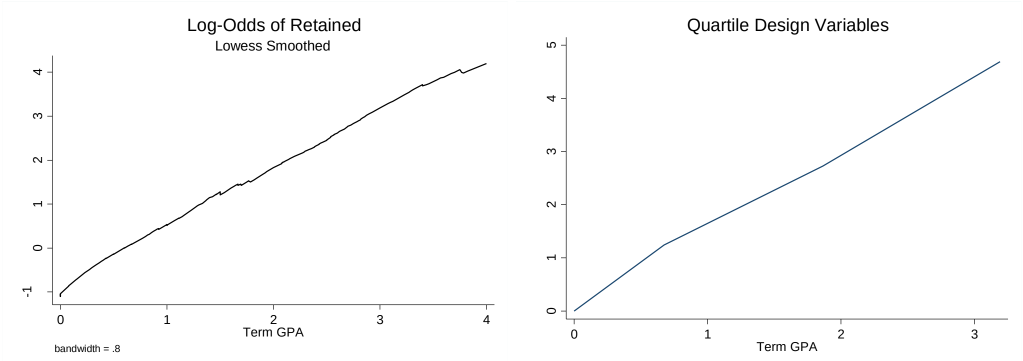


Figure 8 Left: The lowess smoother on Term GPA on the logit scale. Right: The quartile design variable coefficients plotted against the midpoint of each quartile.

Total Bill

The total bill charged to the student appears to have a nonlinear form in the lower covariate values, and then a sharp change to a relatively linear form in the lowess plot in Figure 9. This shape is not surprising given the distribution of this variable. It is worth noting that a single student carries a very large bill relative to the rest of the students and is the likely source of this decline. The IQR of this variable is only $236.25, and in fact 74% of the students have a bill between $3,500 and $3,900. The plot for the design variables suggests there is a large jump in the rate of change of the logit for this variable for students that have lower bills, compared to higher bills. Both plots suggest that methods for compensating for the nonlinearity are required.

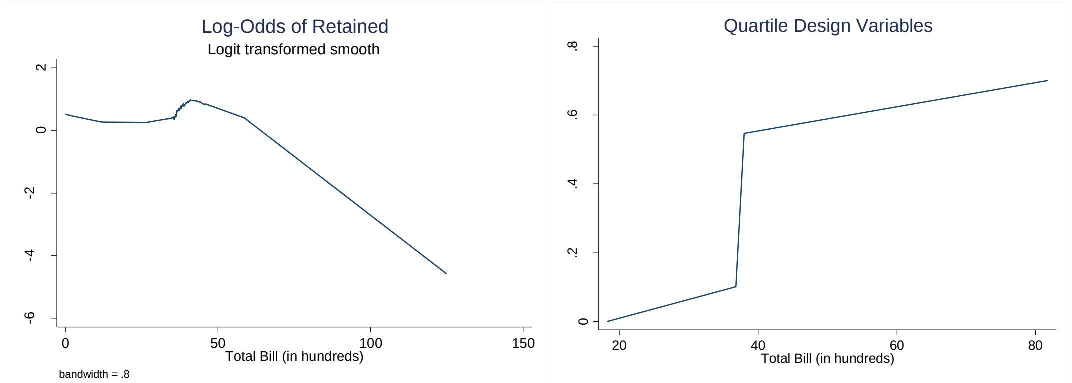


Figure 9 Left: The lowess smoother on the total tuition and fees charged to the student. Right: The coefficients of the quartile design variables plotted against the midpoints of the quartiles.

The command **fp** is used in Stata to generate variables containing fractional polynomial terms for both the FP1 and FP2 models. In both models, the five students with a total bill of $0 are coded as zero in the generated fractional polynomial variables. The output of this procedure is shown in Table 7. The closed procedure fails to reject the linear model.

Table 7 The results of the **fp** procedure on the variable for the total bill.

	Df	Deviance	Dev. Diff	p	powers
Linear	1	436.429	3.414	0.332	1
m = 1	2	435.157	2.142	0.343	3
m = 2	4	433.015	0.000	---	3 3

The **mkspline** command was used to generate variables corresponding to three-knot and five-knot linear splines, as well as three-knot and five-knot restricted cubic splines. The knots for the linear splines are placed at percentiles with equal numbers of subjects in each interval, while the knots for the restricted cubic splines are as recommended in the table by Harrell. The values at which the knots are placed are listed in Table 8.

Table 8 The knot values used in fitting spline models for the total bill variable.

Three Knot	Linear	36.336	37.978			
	Cubic	35.631	37.248	40.962		
Five Knot	Linear	35.841	36.581	37.449	38.878	
	Cubic	11.481	36.336	37.248	38.518	43.309

The linear model and the four spline models were all fit. For each model, Table 9 shows the log-likelihood, the deviance, the AIC, the BIC, and the change in each of these criterion compared to the linear term model. The three-knot linear spline model offers modest improvement in the AIC but the BIC does not agree. There is no strong evidence that any of these models fit the data better than the linear term. The plot in Figure 10 shows the log-odds for lowess smoothed values and the three-knot linear spline model. While the spline model does well at following the trends seen in the lowess plot, the large outlying value in this variable still exerts a noticeable effect. For the lack of a noticeable improvement in the model, and for simplicity, the term will be retained as linear.

Table 9 The results of fitting the spline models for the total bill variable.

Model	LL	-2*LL	AIC	AIC Diff.	BIC	BIC Diff.
Linear	-218.21	436.42	440.43	----	448.09	----
K = 3 Linear Spline	-214.52	429.04	437.05	-3.38	452.36	4.27
K = 3 Cubic Spline	-217.57	435.14	441.13	0.70	452.62	4.53
K = 5 Linear Spline	-213.91	427.82	439.81	-0.62	462.78	14.69
K = 5 Cubic Spline	-214.85	429.70	439.70	-0.73	458.84	10.75

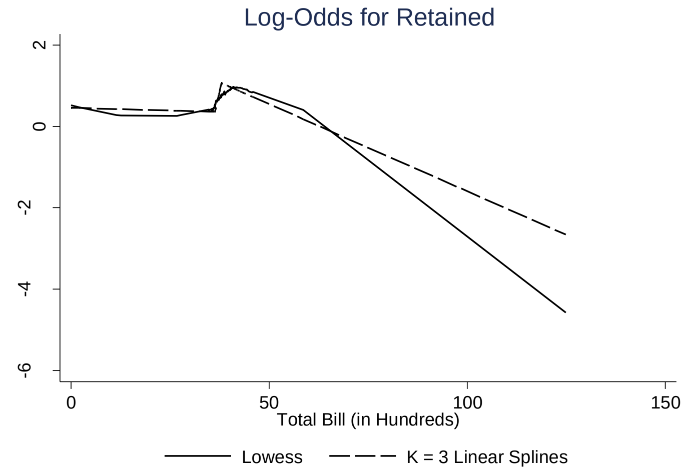


Figure 10 The log-odds of the lowess smoother and three-knot linear splines.

EA Days Into Term

The lowess smoother shows a somewhat nonlinear form for this investigative variable due to the tails, but the quartile design variables show a linear form (see Figure 11). The lowess fit is reminiscent of a cubic polynomial. Results of the **fp** fit support the linear model as the best model. The output from this procedure is in Table 10. It recommends the linear term as the best FP1 model, and fails to reject the linear model compared to the FP2 model.

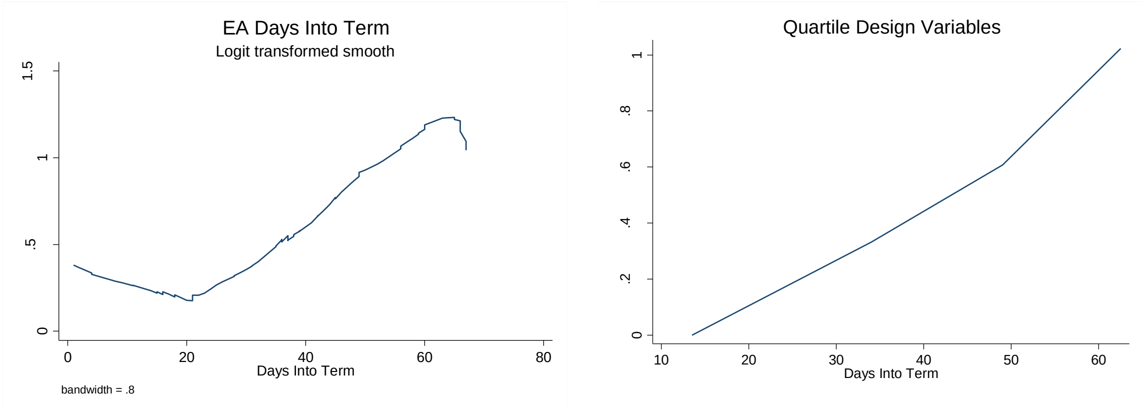


Figure 11 Left: The lowess smoother on the EA days into the term. Right: The coefficients of the quartile design variables plotted against the midpoints of the quartiles.

Table 10 The results of the fp procedure on the EA days into term variable.

	Df	Deviance	Dev. Diff	p	powers
Linear	1	426.394	2.561	0.464	1
m = 1	2	425.339	1.506	0.217	2
m = 2	4	423.833	0.000	---	0 0.5

Thus, the final main effects model is defined. The six categorical covariates, along with a linear term for Term GPA, the total bill in hundreds of dollars, and days before drop are used. The fit of this model is shown previously in Table 6, and is not replicated here. Next, the model will be checked for interaction terms.

Interaction Effects

Interaction effects occur when the effects of the levels of one variable differ for levels of another. As such, pairs of variables are tested for significance in a two-way interaction. It should be noted that interaction between the three variables describing the student financial status cannot be interacted because of the way they are formulated. The variable that describes the financial need indicates if a student had financial need or not, while only those with financial need are eligible for a Pell grant and many types of loans. This structure will create empty categories that will remove cases from the model, at best, and at worst create issues with multicollinearity.

Table 11 shows the p-values associated with each two-way interaction when included in the model. Interaction between the student's preparation status and the financial status shows that success is predicted perfectly for seven observations (there is no financial information and preparation in two subjects). There are two significant interactions, one between the dorm status and the term GPA, and the other between the loan status and the preparation status. Adding these two interactions gives a final preliminary effects model which will be assessed for goodness-of-fit, and for consistency of fit across the observations. The fit of the full model including significant interaction terms is shown in Table 12.

Table 11 The p-values associated with the two-way interactions. “---“ indicates induced multicollinearity. “++” indicates an empty category. “**” indicates removed observations due to a perfect prediction of success.

		Fin. Status				Preparation Status					
	First	Has	No			NCBO	One	Two	Term	Total Bill	EA
	Gen	Need	Need	Pell	Loan	Only	Subj.	Subj.	GPA	(in hund.)	DIT
Dorm	0.81	0.14	0.25	0.08	0.21	0.31	0.55	0.62	0.02*	0.06	0.95
First Gen.		0.64	0.68	0.85	0.22	0.35	0.46	0.78	0.82	0.64	0.26
Fin. Status											
Has Need				---	0.83	0.44	0.45	0.66	0.60	0.35	0.95
No Need				++	---	---	**	0.82	0.11	0.33	0.20
Pell					0.60	0.93	0.70	0.47	0.81	0.21	0.26
Loan						0.04*	0.43	0.61	0.70	0.86	0.40
Prep. Status											
NCBO									0.92	0.58	0.68
One Subj.									0.54	0.64	0.77
Two Subj.									0.42	0.45	0.60
Term GPA										0.17	0.15
Tot. Bill (hund.)											0.28

Assessing Fit

The selected model contains 340 subjects in 339 covariate patterns. The post estimation commands available in Stata produce and store the Pearson residuals, the deviance residuals, and the relevant diagnostic statistics. Additionally, these commands produce results for testing goodness-of-fit. Several statistics and graphical displays are presented to assess the fit of the model.

The Hosmer – Lemeshow Test

The Hosmer-Lemeshow test divides the data into groups according to the percentiles of the estimated probabilities, $\hat{\pi}_j$, with 34 subjects going into one of ten groups. The classifications of the subjects are shown in Table 13. The Hosmer-Lemeshow test fails to reject the hypothesis that the model fits, chi2(8) = 5.62, p = 0.69.

It should be noted there are very small expected frequencies in some cells, though all are greater than zero.

Table 12 The results of fitting the full model with interactions.

	Interaction Model			
	LL = -123.47926		LR chi2(16) = 189.47	
	Pseudo R2 = 0.4341		Prob > chi2 = 0.0000	
	Coefficient	Std. Error	z	P > |z|
Dorm	1.726	0.598	2.89	0.004
FirstGen	-0.387	0.352	-1.10	0.272
Financial Status				
Has Need	-1.387	0.619	-2.24	0.025
No Need	-0.178	0.657	-0.27	0.787
Pell	1.753	0.428	4.10	0.000
Loan	0.631	0.451	1.40	0.162
Preparation Status				
NCBO Only	-1.642	1.232	-1.33	0.183
One Subject	0.636	0.535	1.19	0.234
Two Subjects	-1.193	0.840	-1.42	0.155
TermGPA	2.681	0.541	4.95	0.000
Total Bill (hund.)	-0.041	0.016	-2.54	0.011
Days Into Term	0.015	0.009	1.65	0.100
Dorm*GPA	-1.289	0.564	-2.28	0.022
Loan*NCBO Only	3.036	1.508	2.01	0.044
Loan*One Subj.	-0.468	0.786	-0.60	0.551
Loan*Two Subj.	0.669	1.124	0.60	0.552
Constant	-1.345	0.893	-1.51	0.132

Table 13 The classification table for the Hosmer-Lemeshow goodness-of-fit test on the full interaction model.

		Retained = Yes (1)		Retained = No (0)		
Group	Probability	Observed	Expected	Observed	Expected	Total
1	0.1021	4	1.8	30	32.2	34
2	0.2699	6	5.7	28	28.3	34
3	0.4816	11	13.3	23	20.7	34
4	0.6579	20	19.6	14	14.4	34
5	0.8011	23	24.9	11	9.1	34
6	0.8690	29	28.3	5	5.7	34
7	0.9425	31	30.8	3	3.2	34
8	0.9683	32	32.6	2	1.4	34
9	0.9867	34	33.2	0	0.8	34
10	0.9993	34	33.8	0	0.2	34

The Osius-Rojek Test

An aggregated dataset containing the predicted probabilities for 339 covariate patterns is used to calculate the relevant statistics for the Osius-Rojek test. These values compute $A = 2\left(J - \sum_{j=1}^{J} \frac{1}{m_j}\right) = 2(339 - 338.5) = 1$, $\mathrm{X}^2 = 285.49799$, and the weighted regression on c_j gives RSS = 9905.21726. The relevant test statistic on the standard normal distribution is $z_{\mathrm{X}^2} = -0.357$. The two-tailed test gives p = 0.72 and fails to reject the hypothesis that the model fits.

Stukel's Test

Stukel's test is formulated using the covariate patterns to define the two dichotomous variables added to the data set. Adding these variables to the model and using a partial likelihood ratio test gives chi2(2) = 2.73 with p = 0.26. This test is in agreement with the Hosmer-Lemeshow test and the Osius-Rojek test in that it fails to reject the hypothesis that the model fits the data.

Overall Assessment of Goodness-of-Fit

None of the tests for goodness-of-fit reject the hypothesis that the model fits the data. The pseudo-R^2 is 0.4341, which is fairly high. The overall conclusion is that there is no evidence to support that the model does not fit the data.

Sensitivity/Specifity

The panel of graphs in Figure 12 shows the model has excellent discrimination. The area under the ROC curve for this model is 0.91, which is classified as “excellent discrimination” by Hosmer and Lemeshow. A scatterplot of the outcome variable, Retained, against the estimated probabilities (with jitter added to the points) shows that higher estimated probabilities are more prevalent among those that did retain, and lower estimated probabilities are more prevalent among those that did not retain. The stacked histograms in the right column of the figure further accentuate this characteristic of the model by showing the densities of the estimated probabilities for each value of the outcome variable. Using a probability cutoff of 0.5 to sort the observations into a classification table shows the model has an overall rate of correct classification of 84.4%, with a sensitivity of 90.6% and a specificity of 72.4%. These figures all support a model that fits the data well.

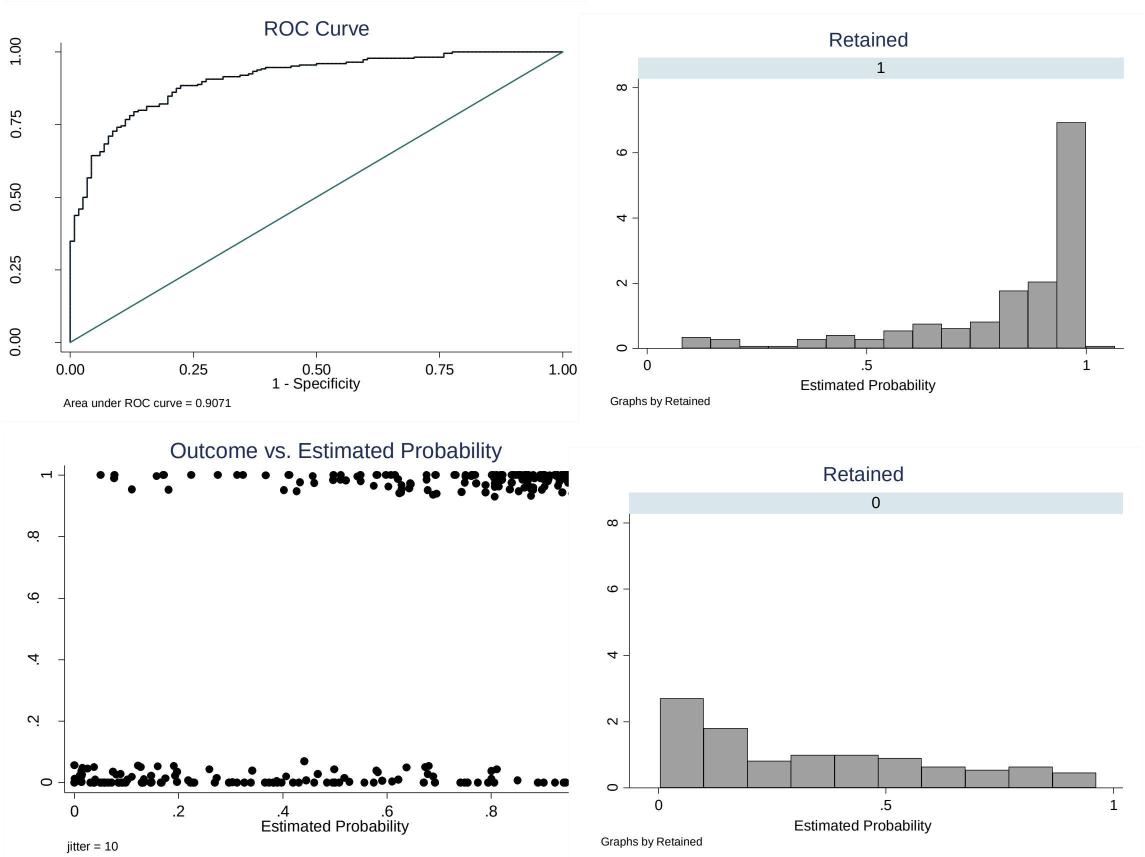


Figure 12 Top Left: The ROC curve for the full interaction model. Bottom Left: A scatter plot of the outcome variable, Retained, against the estimated probabilities. Top Left: A histogram of estimated probabilities for those who retained. Bottom Left: A histogram of estimated probabilities for those who did not retain.

MODEL DIAGNOSTICS

The model fits the data well and shows good discrimination. The covariate patterns are assessed for overall fit using several plots. The panel of four plots in Figure 13 provides a visual assessment of model fit by covariate pattern. The plot of the leverage values against the estimated probabilities show two covariate patterns that lie away from the main body of the data. None of the leverage values are large, but covariate pattern 293 corresponds to the observation containing the extreme outlying bill, so the influence of this case on the model is of particular interest.

The plot of the Pregibon's delta-beta shows four covariate patterns that are out-of-place compared to the rest of the data. One of these is the case already mentioned, covariate pattern 293, which is of concern. This case will be removed from the model and the data refit to assess the true impact of this particular observation. The plot that shows the Hosmer-Lemeshow delta-chi2 shows two cases that have particularly large values, while the delta-D plot does not really highlight any cases of interest.

Table 14 shows a summary of these covariate patterns and the relevant diagnostic statistics. The two high-leverage cases, covariate patterns 188 and 293 show a moderate estimated probability of retention at 0.57 and 0.47 respectively and neither student retained. Covariate patterns 86, 186, and 324 go against the model in the the first had a low probability of retaining (0.079) yet retained while the other two had high probabilities of retaining (0.962, 0.946 respectively) yet failed to retain. This is not problematic in itself; however, the removal of the high leverage case with the large bill is still examined.

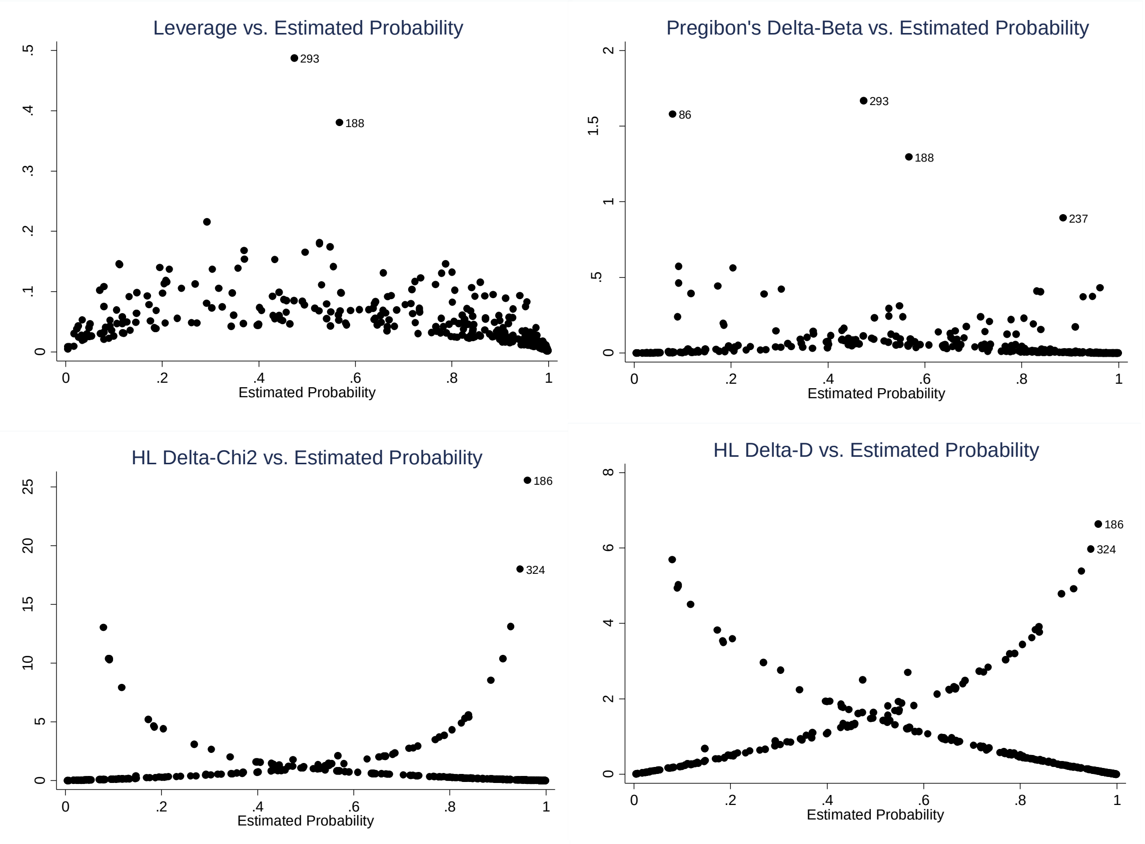


Figure 13 Diagnostic plots for the full interaction model. The top row identifies two high leverage observations. The bottom row identifies two observations with large impacts to residual-based tests for fit.

Table 14 Summaries of the covariate patterns identified as being poorly fit (186, 324, 86), or having high leverage (293, 188).

Covariate Pattern	86	186	188	293	324
Dorm	1	1	1	1	1
First Generation	1	0	0	0	1
Financial Status					
Has Need	1	1	0	0	1
No Need	0	0	1	0	0
Pell	1	1	0	0	0
Loan	0	1	0	0	1
Preparation Status					
NCBO Only	1	0	1	0	0
One Subject	0	1	0	1	0
Two Subjects	0	0	0	0	0
Term GPA	0.000	1.539	1.600	2.824	3.400
Total Bill (hundreds)	40.241	35.631	36.623	12.494	39.588
Days Into Term	32	67	66	4	35
Retained	1	0	0	0	0
$\hat{\pi}_j$	0.079	0.962	0.567	0.473	0.946
$\hat{g}(x_j)$	-2.454	3.224	0.269	-0.106	2.871
$\Delta\hat{\beta}_j$	1.578	0.430	1.295	1.669	0.373
ΔX^2	13.040	25.561	2.112	1.754	18.015
ΔD	5.686	6.637	2.700	2.503	5.973
h_j	0.108	0.017	0.380	0.487	0.020

THE ADJUSTED MODEL

With the covariate pattern associated with the extreme outlying bill excluded, repetition of the model selection procedure yields identical results with two exceptions: the variable for the total bill is not significant in the model, suggesting that the lone, extremely outlying bill (where the corresponding student did not retain) is influential in the model, and removal of the dorm status has a smaller impact on the remaining coefficients.

Appendix D presents detailed results from all steps of the procedure applied to the reduced data set. Fitting the initial multivariable model on the refined data set produces results similar to fitting on the full data set. Only the removal of the total hours from the initial multivariable model is warranted. Cycling back through the initially excluded variables does not warrant the inclusion of any of the variables, even the total bill which was significant before the removal of the extreme case. There are no noticeable changes to the shape of the two continuous variables, the term GPA and the timing of the Early Alert days into the term. Fitting the fractional polynomial term to the adjusted data set, as before, supports the linear model for the number of days into the term. Testing for interactions yields the same results as before, and the final model adopted is shown in the Table 15. It includes the six categorical covariates, the two continuous covariates, both in linear form, and the two significant interactions.

Goodness-of-Fit

The data contains 339 observations in 338 covariate patterns. The Hosmer-Lemeshow test for goodness-of-fit fails to reject the hypothesis that the model fits the data (chi2(8) = 2.01, p = 0.98). The classification of subjects for this test is in Table 16. There are no concerns above those stated previously The 338 covariate patterns used in the Osius-Rojek test compute $A = 2\left(J - \sum_{j=1}^{J} \frac{1}{m_j}\right) = 2(338 - 337.5) = 1$, $\mathrm{X}^2 =$

310.477, and the weighted regression on c_j gives RSS = 6647.27. The relevant test statistic on the standard normal distribution is $z_{\chi^2} = -0.166$. The two-tailed test gives p = 0.87 and fails to reject the hypothesis that the model fits. Stukel's test using a partial likelihood ratio test also concludes that the model fits the data, with chi2(1) = 0.05 and p = 0.83. All of these tests indicate the model fits the data.

Table 15 The selected model.

	Interaction Model			
	LL = -123.16064		LR chi2(15) = 187.95	
	Pseudo R2 = 0.4328		Prob > chi2 = 0.0000	
	Coefficient	Std. Error	z	P > |z|
Dorm	1.405	0.562	2.50	0.012
FirstGen	-0.414	0.351	-1.18	0.238
Financial Status				
Has Need	-1.204	0.602	-2.00	0.045
No Need	-0.033	0.656	-0.05	0.960
Pell	1.682	0.424	3.97	0.000
Loan	0.581	0.451	1.29	0.197
Preparation Status				
NCBO Only	-1.573	1.139	-1.38	0.167
One Subject	0.864	0.559	1.55	0.122
Two Subjects	-1.219	0.821	-1.48	0.138
TermGPA	2.440	0.506	4.83	0.000
Days Into Term	0.013	0.009	1.45	0.146
Dorm*GPA	-1.011	0.540	-1.87	0.061
Loan*NCBO	3.039	1.417	2.14	0.032
Loan*One Subj.	-0.699	0.806	-0.87	0.386
Loan*Two Subj.	0.694	1.111	0.62	0.532
Constant	-2.585	0.694	-3.73	0.000

Table 16 The classification of observations for the Hosmer-Lemeshow goodness-of-fit test.

		Retained = Yes (1)		Retained = No (0)		
Group	Probability	Observed	Expected	Observed	Expected	Total
1	0.1101	2	2.1	32	31.9	34
2	0.2486	7	5.8	27	28.2	34
3	0.4609	12	12.9	22	21.1	34
4	0.6630	21	20.1	13	13.9	34
5	0.8003	24	25.2	10	8.8	34
6	0.8674	28	28.5	6	5.5	34
7	0.9385	31	30.8	3	3.2	34
8	0.9697	32	32.6	2	1.4	34
9	0.9867	34	33.3	0	0.7	34
10	0.9989	33	32.8	0	0.2	33

Sensitivity/Specificity

As with goodness-of-fit testing, there are no results regarding the sensitivity of the model that were different from the first formulation. The panel of graphs in Figure 14 confirm the model achieves excellent discrimination by the Hosmer-Lemeshow standards outlined previously. A scatterplot of the outcome variable, Retained, against the estimated probabilities (with jitter added to the points) shows that higher estimated probabilities are more prevalent among those that did retain, and lower estimated probabilities are more prevalent among those that did not retain. The histograms in the right column of the figure further accentuate this characteristic of the model by showing the densities of the estimated probabilities for each value of the outcome variable. Using a probability cutoff of 0.5 to sort the observations into a classification table shows the model has an overall rate of correct classification of 83.8%, which is slightly lower than before, with a sensitivity of 90.2% and a specificity of 71.3%. The specificity with this adoption of the model is also slightly lower that the model that contained the total bill (see Table 17). These figures all support a model that fits the data well.

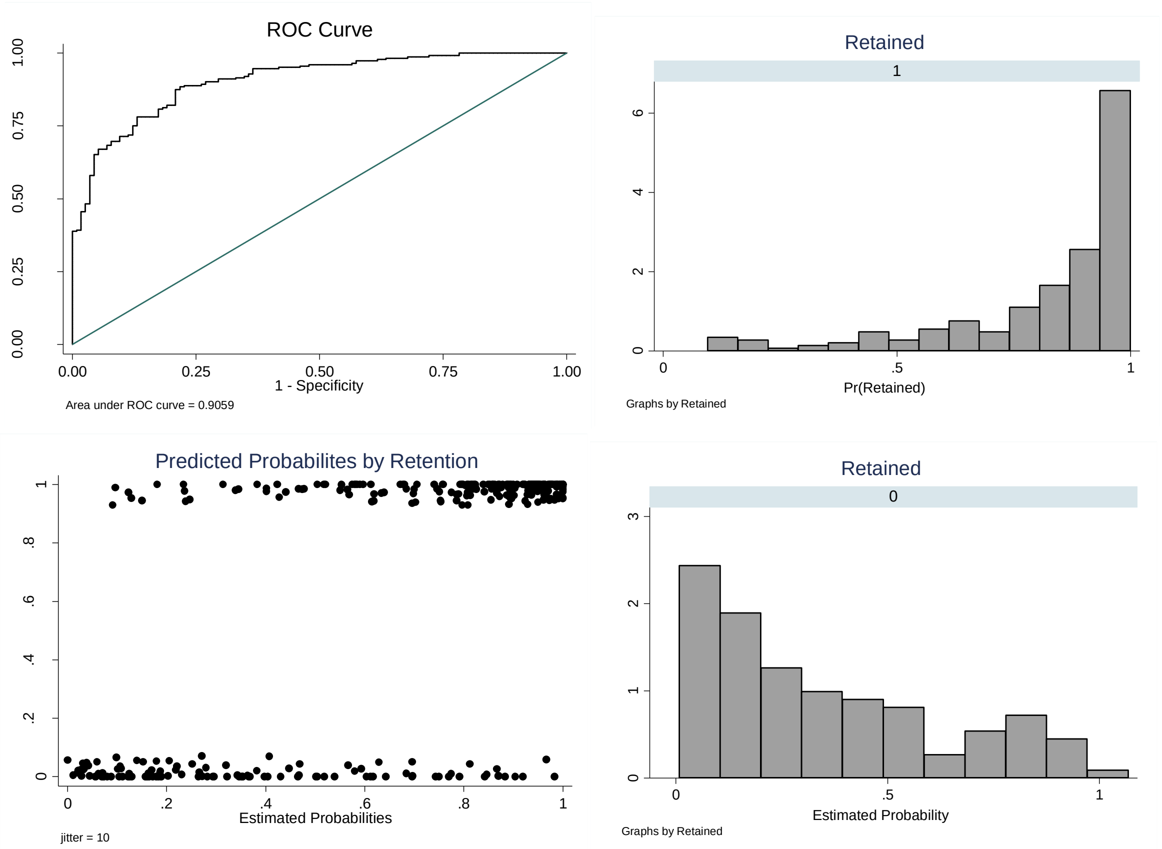


Figure 14 Top left: The ROC curve; Bottom left: the estimated probabilities according to the outcome variable; Right column: histograms of the estimated probabilities by outcome.

Table 17 The classification of subjects using a probability cutoff of 0.50.

	Outcome		
	Retained = 1	Retained = 0	Total
$\Pr(Retained) \geq 0.50$	202	33	235
$\Pr(Retained) < 0.50$	22	82	104
Total	224	115	339

Model Diagnostics

The panel of plots in Figure 15 show that the removal of the outlying bill leaves no covariate patterns with high leverage values, and none that will impact the coefficients greatly when removed. Though labeled with different identifiers, the same three covariate patterns show large changes to model summaries when removed. These cases will be examined in the results chapter.

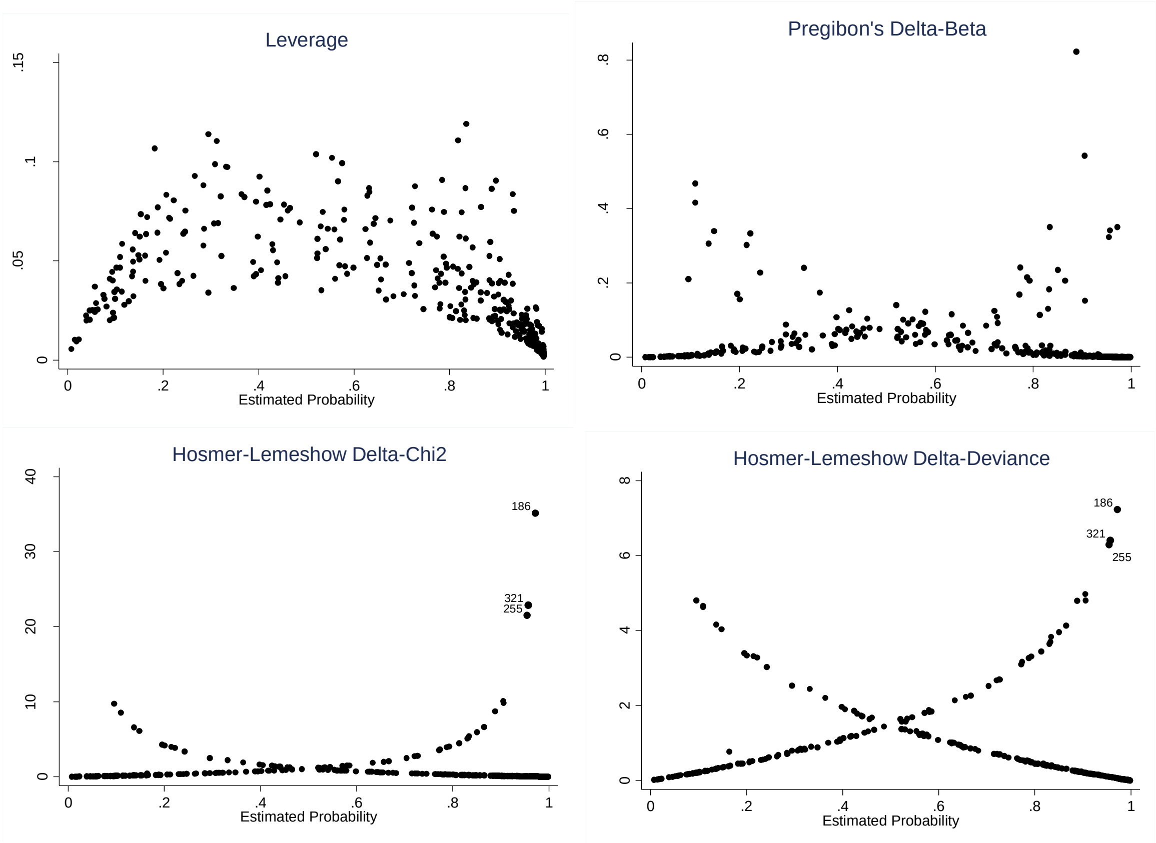


Figure 15 The diagnostic plots for the adjusted model.

SUMMARY

From the study variables selected, the dorm status, first generation status, the financial status, the pell status, the loan status, and the preparation status are significant categorical covariates, or were retained due to impacts to coefficients if removed. The term GPA, as well as the research variable describing the timing of the Early Alert are significant continuous covariates. Additionally, two significant interactions, between dorm status and GPA and between the loan status and the preparation status are included in the model. Goodness-of-fit testing shows the selected model is an adequate fit for the data.

CHAPTER VI

RESULTS AND DISCUSSION

Table 18 presents the odds ratios and confidence intervals calculated from fitting the final selected model to the data. The research variable included in the model is in bold type in the table. Notably in the table, the odds-ratio for Pell recipients suggests that the odds of retaining these students are about 5 times the odds of retaining a non-Pell student, and the confidence interval suggests they are at least twice as likely to retain. Also notably, the odds-ratio for students found to have financial need suggests that these students retain at a lower rate than those that did not apply for financial aid. The odds of retaining a student with financial need are about 70% lower than the odds of retaining a student that did not apply for financial aid, and this difference can be as much as 91% lower, or as little as 2% lower. The odds-ratio for first-generation students suggests that they retain at a lower rate than their non-first-generation counterparts, though this is not a significant variable in the model.

For students that have loans, the odds ratio that compares them to those that do not have loans depends upon the preparation status of the student. Table 18 indicates that for students participating in only NCBO preparation, the odds of successful retention for a student with loans retaining are 37 times higher than the odds of a student without loans retaining. Overall, the table suggests that the odds of a student with a loan retaining are higher than the odds of a student without a loan retaining.

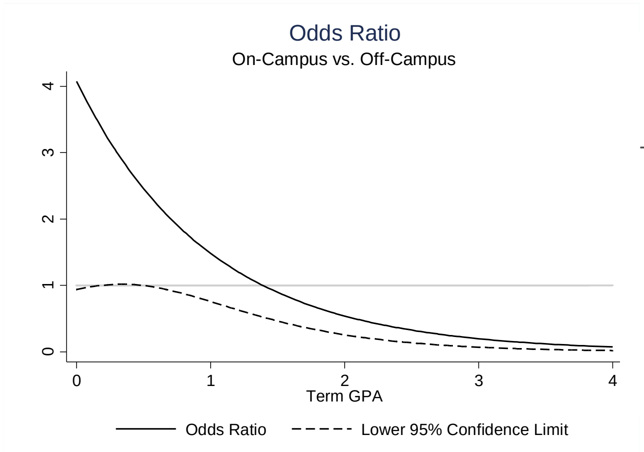
Table 18 Odds ratios and confidence intervals based on the final selected model.

Odds Ratios for Covariates

Covariate	Odds Ratio	95% Confidence Interval	
First Generation	0.66	0.33	1.32
Pell	5.38	2.34	12.34
Financial Status			
No Application	1.00		
Has Need	0.30	0.09	0.98
No Need	0.97	0.27	3.50
Loan*Preparation Status			
None Required	1.79	0.74	4.34
NCBO Only	37.37	3.52	397.32
One Subject	4.01	0.61	26.36
Two Subjects	3.58	0.42	30.39
Days Into Term	**1.10***	**0.97***	**1.25***

*For a seven day difference

The odds of retaining a dorm student compared to a student living off campus are dependent upon the GPA of the student. Figure 16, along with the accompanying table, gives a picture of how the odds of retaining a student living in the dorm compare to the odds of retaining a student living off-campus. The figure shows that the odds are equalized at a GPA of 1.389. The upper confidence interval is not shown on the graph as the values are too large for the scale of the odds ratios to accommodate it comfortably, as can be seen in the table. The plot suggests that at lower GPA values, the odds of retaining a student living in the dorms are higher than the odds of retaining a student not living in the dorm, by as much as four times.



GPA	OR	Std. Error	95% CI
0.00	4.1	0.75	(0.94, 17.71)
0.75	1.9	0.77	(0.91, 18.34)
1.50	0.9	1.10	(0.47, 35.31)
2.25	0.4	1.56	(0.19, 86.90)
3.00	0.2	2.06	(0.07, 232.69)
3.75	0.1	2.58	(0.03, 645.59)

Figure 16 The odds of retaining a dorm student compared to the odds of retaining an off-campus student according to the term GPA.

THE TIMING OF THE EARLY ALERT

Overall, the Early Alert does not seem to improve retention in terms of it being a critical and immediate early intervention. The frequency of early alerts increases as the drop date approaches, and so does the retention rate. It is perhaps the case that students who were the subjects of warning notifications early in the term were not engaged in the college endeavor both prior to and just after entry, while those that are warned later in the term began engaged, somehow withdrew, and then returned to recover after prompting.

32% (*n* = 108) of the sampled students were first submitted for Early Alert referrals before 30 days of the term had elapsed. Nearly all of these students were indicated to be passing or able to pass the course that was referred (*n* = 105), yet only 35% of this group went on to achieve a term GPA of at least 2.0 while 57% (*n* = 62) re-enrolled in the spring. The other 65% of the sampled students were first submitted for Early Alert referrals 30 or more days into the term (*n* = 232). Among these referrals, 72% were indicated to be passing or could pass. 36% of the students in this group went on to achieve a GPA of at least 2.0 (*n* = 85), while 70% re-enrolled in the spring.

These comparisons demonstrate how the retention rate is higher for students that receive Early Alerts later in the term. While it is important to identify students that are at-risk for attendance in hopes of supplying a redeeming intervention, the model supports that among entering full-time, first-time, degree-seeking freshmen, a referral submitted early in the term identifies students that are not engaged at the outset, but does not improve the odds of retention.

Overall, of the 340 students selected in the initial sample, 66% had an Early Alert referral for a single course submitted a single time during the term. For another 14% of the sampled students, multiple Early Alert referrals were submitted for a single course. The remaining 20% of the sample experienced Early Alert referrals in multiple core courses. In terms of course outcomes, of the 108 students referred before 30 days into the term, 32% passed the referred courses, 23% completed but failed the referred courses, 30% dropped the referred courses, and the remaining 15% experienced a combination of those three outcomes among the referred courses. Of the students that received the first Early Alert 30 or more days into the term, 28% passed all of the referred courses, 40% completed but failed all of the referred courses, 28% dropped all of the referred courses, and the remaining four percent had mixed course outcomes. This suggests that the timing of the Early Alert referral does not positively impact course performance in terms of successful completion.

This may be due to a lack of specific intervention, such as ensuring that a student who is referred for tutoring actually receives the tutoring. Such a high failure rate for students receiving an Early Alert later in the semester suggests that students are not taking measures in response to the notification, either to drop the course or to take measures leading to successful course completion.

It does not seem to matter if the instructor recommended a drop or indicated that the student was passing or could still pass. About two-thirds of the students were indicated to be able to pass, if they were not already, on their first Early Alert, but the same proportion of students with this option retained as those that retained without this option.

Further research should be conducted including faculty information and course section information to make this analysis more robust. There was no way to account for variance in faculty, as some faculty are more prone to submit and Early Alert than others, and some may have more of a “hair-trigger” than others. Overall, the retention of the student seems to not depend upon the timing of the Early Alert or the possibility of passing the course.

Recommendations moving forward would be to ensure that Advising Services checks the Early Alert status of a student before green-lighting the registration for the next semester to make appropriate support referrals. Use of the services to which a student was referred should also be tracked, so that students who retain to the next spring, but do not connect with the resources are identified for additional intrusive advising. Additionally, increased communication between Advising Services and Residential Living may be appropriate for dorm students.

References

Asby, S. (2015). Early Alert and Intervention Systems and Student Persistence: An Exploration of Student Perceptions.

Bai, H. a. (2009-2010). A Multilevel Approach to Assessing the Interaction Effects on College Student Retention. *Journal of College Student Retention, 11*(2), 287-301.

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Howard, J. a. (2015). A Comparison of Student Retention and First Year Programs Among Liberal Arts Colleges in the Mountain South. *Journal of Learning in Higher Education*, *11*(1), 67 - 84.

Hudson, W. E. (2006). Can an Early Alert Excessive Absenteeism Warning System Be Effective in Retaining Freshman Students. *Journal of College Student Retention*, 217-226.

Kuh, G. D. (2007). What student engagement data tells us about college readiness. *AACU Peer Review, 9*(1), 4-8.

Nora, A. B. (2005). Student persistence and degree attainment beyond the first year in college: The need for research. In A. Seidman (Ed.), *College Student Retention: Formula for Student Success* (pp. 129-153). Westport, CT: Praeger.

Raisman, N. (2013). *The Cost of College Attrition at Four-Year Colleges and Universities.* Virginia Beach, VA: Educational Policy Institution.

Rugaber, C. S. (2017, 01 12). *Pay gap between college grads and everyone esle at a record.* Retrieved from USA Today: https://www.usatoday.com/story/money/2017/01/12/pay-gap-between-college-grads-and-everyone-else-record/96493348/

Seidman, A. (2005). Where we go from here: A retention formula for student success. In A. (. Seidman, *College student retention* (pp. 295-316). Westport, CT: Praeger.

Simons, J. (2011, December). *A National Study of Student Early Alert Models at Four-Year Institutions of Higher Education.* Arkansas State University.

Spady, w. (n.d.).

Spady, W. (1970). Dropouts from Higher Education: An Interdisciplinary Review and Synthesis. *Interchange*, 64-85.

Spady, W. G. (1971). Dropouts from Higher Education: Toward an Empirical Model. *Interchange*, 38 - 62.

Summerskill, J. (1962). Dropouts From College. In *The American College* (pp. 627-657). New York: Wiley.

Tinto, V. (2013 - 2014). Isaac Newton and Student College Completion. *Journal of College Student Retention*, 1 - 7.

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 89 - 125.

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition 2nd Ed.* Chicago: The University of Chicago Press.

APPENDIX A

VARIABLE CODING AND DESCRIPTION

Variable	Type	Code	Description
Year	Dichotomous	1	Fall 2014
		2	Fall 2015
Dorm	Dichotomous	0	Not living in dorm
		1	Living in dorm
Gender	Dichotomous	0	Female
		1	Male
Race/Ethnicity	Polychotomous	0	White
		1	Hispanic
		2	African American/Black
		3	Other
First Generation	Dichotomous	0	Not first generation
		1	First generation
Financial Status	Polychotomous	0	No Information
		1	Has financial need
		2	No financial need
Pell	Dichotomous	0	No Pell received
		1	Pell received
Loan	Dichotomous	0	No loan received
		1	Loan received
Preparation Status	Polychotomous	0	No preparation required
		1	NCBO Only
		2	One subject required
		3	Two subjects required
Early Alert Recommendation**	Dichotomous	0	Drop recommended/no possible pass
		1	Passing or possible pass/no drop recommended
Age	Continuous		Age at the time of enrollment
ACT Score	Continuous		ACT score on recored (conversd SAT)
Total Bill	Continuous		Total tuition and fees charged
Total Hours	Continuous		Total credit hours attempted
Total Online Hours	Continuous		Total online hours attempted
Term GPA	Continuous		GPA earned in the fall term
Early Alert Days in Term**	Continuous		Days into the term the first Early Alert was received

APPENDIX B

SAT TO ACT CONCORDANCE

For all students present with an ACT score in the data, the ACT composite score was used. For students with no ACT scores, the sum of the SATCR and SATM test was converted to a composite ACT score according to the convention in the table below[1].

SAT CR+M	ACT
1600	36
1540 – 1590	35
1490 – 1530	34
1440 – 1480	33
1400 – 1430	32
1360 – 1390	31
1330 – 1350	30
1290 – 1320	29
1250 - 1280	28
1210 - 1240	27
1170 - 1200	26
1130 - 1160	25
1090 - 1120	24
1050 - 1080	23
1020 - 1040	22
980 - 1010	21
940 - 970	20
900 - 930	19
860 - 890	18
820 - 850	17
770 - 810	16
720 - 760	15
670 - 710	14
620 - 660	13
560 - 610	12
510 - 550	11

[1] Source: ACT Research and Policy: College Board and ACT Joint Statement, October 2009. Appropriate for scores obtained after January 2005 and before March 2016.

APPENDIX C

ENTERING COHORT AND SAMPLE COMPARISONS

Categorical Covariates

Variable	Level	% of Cohort	% of Sample
Year	2014	48.9	45.6
	2015	51.1	54.6
Gender	Male	47.3	60.6
	Female	52.6	39.4
Race/Ethnicity	White	56.1	49.1
	Hispanic	29.3	31.2
	African American/Black	7.0	11.8
	Other	7.6	7.9
Dorm Resident	On-Campus	83.5	78.5
	Off-Campus	16.5	21.5
First Generation	First Generation	44.4	49.1
	Not First Generation	55.6	50.9
Pell	Pell Received	10.3	12.6
	Pell Not Received	89.7	87.4
Loan	Loan Received	22.7	15.6
	No Loan Received	77.3	84.4
Preparation Status	No Preparation Required	73.0	57.4
	NCBO Only	4.9	9.4
	One Subject Required	16.3	22.4
	Two Subjects Required	5.8	10.9
Financial Status	No Information	10.3	12.6
	Has Financial Need	67.0	71.8
	No Financial Need	22.7	15.6
Retained	Continued in Spring	86.7	65.9
	Did Not Continue in Spring	13.3	34.1

Continuous Covariates

Variable	Group	Min	Q1	Med	Q3	Max	Mean	Std.Dev.
Age	All Freshmen	17	18	18	18	42	18.2	0.93
	Sample	17	18	18	18	39	18	1.3
ACT Score	All Freshmen	9	18	20	23	34	21	4
	Sample	12	17	19	22	31	19.5	3.4
Total Bill	All Freshmen	0.00	3633.55	3733.30	3869.80	12493.89	3684.44	932.99
	Sample	0.00	3633.55	3724.80	3869.80	12493.89	3624.85	962.99
Total Hours	All Freshmen	12	13	14	16	19	14.4	1.4
	Sample	12	13	14	15	19	14.1	1.4
Online Hours	All Freshmen	0	0	0	0	15	0.4	1.2
	Sample	0	0	0	0	15	0.39	1.30
Term GPA	All Freshmen	0.00	2.18	2.88	3.50	4.00	2.67	1.08
	Sample	0	0	1.35	2.385	4.00	1.37	1.18

APPENDIX D

RESULTS FROM FITTING ADJUSTED MODEL

Univariate Analysis

Categorical Covariates				Continuous Covariates			
Variable	Pearson chi-squared	p-value	Carry forward	Variable	LRT chi2(1)	p-value	Carry forward
Year	0.620	0.431	No	Age	0.000	0.946	No
Dorm Status	18.080	0.000	Yes	ACT Score	1.13	0.287	No
Gender	0.117	0.732	No	Total Bill (in hund.)	0.75	0.387	No
Race/Ethnicity	3.740	0.291	No	Total Hours	3.89	0.049	Yes
First Generation	2.843	0.092	Yes	Total Online	0.65	0.420	No
Financial Status	3.610	0.164	Yes	Term GPA	141.14	0.000	Yes
Pell Status	2.527	0.112	Yes	Days Into Term	9.26	0.002	Yes
Loan Status	20.625	0.000	Yes				
Preparation Status	7.064	0.070	Yes				
Poss. Pass	0.357	0.550	no				

Initial Multivariable Model

	Initial Multivariate Model				
	LL = -129.30573		LR chi2(12) = 175.66		
	Pseudo R2 = 0.4045		Prob > chi2 = 0.0000		
Variable	Coefficient	Std. Err.	Z	p-value	Remove
Dorm student	0.625	0.377	1.66	0.097	4
First Generation	-0.275	0.337	-0.82	0.414	2
Financial Status					
Has Financial Need	-1.231	0.575	-2.14	0.032	
No Financial Need	-0.129	0.641	-0.20	0.841	
Pell	1.499	0.407	3.68	0.000	
Loan	0.729	0.353	2.06	0.039	
Total Hours	0.052	0.123	0.42	0.673	1
Preparation Status					
NCBO Only	0.304	0.552	0.55	0.582	3
One Dev. Subject	0.557	0.401	1.39	0.164	3
Two Dev. Subjects	-0.842	0.537	-1.57	0.117	3
Term GPA	1.599	0.197	8.13	0.000	
EA Days Into Term	0.018	0.009	2.02	0.044	
Constant	-2.907	1.850	-1.57	0.116	

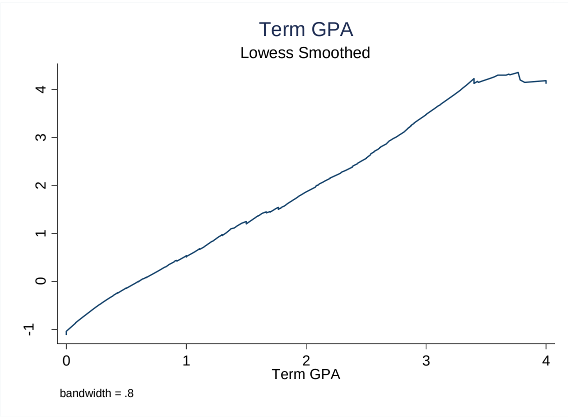
Change in Coefficients

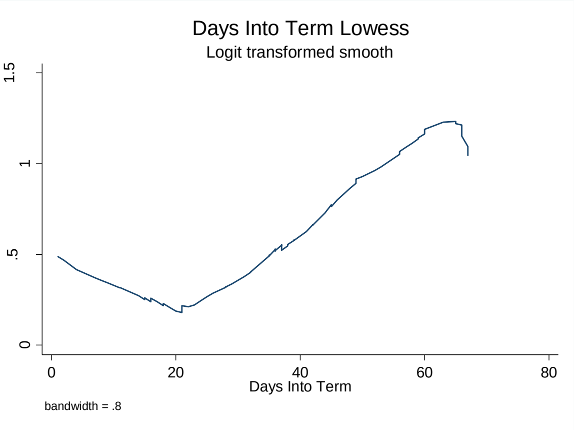
Variable Removed	Coeff. Impacted	% Change (if greater than 10%)	LR Test (df)	P-value
Total Hours	None	N/A	0.18 (1)	0.67
First Generation	Financial Status – No Need	46	0.63 (1)	0.43
	Preparation – NCBO Only	114	0.70 (1)	0.40
Preparation Status	First Generation	-25	6.40 (3)	0.09
	Financial Status – No Need	-47		
Dorm Status	Loan Status	13	2.90 (1)	0.08

Initially Excluded Variables

Variable	Coefficient	Std. Error	z	p-value	Retain
Year	0.238	0.332	0.72	0.473	No
Gender	-0.459	0.344	-1.33	0.183	No
Race/Ethnicity – Hispanic	0.461	0.3750	1.23	0.218	No
Race/Ethinicity – African American	0.494	0.564	0.88	0.381	
Race/Ethnicity – Other	0.753	0.601	1.25	0.210	
Possible Pass	-0.276	0.459	-0.60	0.548	No
Age	-0.039	0.141	-0.27	0.785	No
ACT Score	0.041	0.051	0.79	0.430	No
Total Bill	-0.019	0.019	-0.99	0.323	No
Online Hours	-0.078	0.131	-0.60	0.550	No

Assessing Linearity





Fractional Polynomials - Results of the fp procedure

	Df	Deviance	Dev. Diff	p	powers
Linear	1	425.019	2.895	0.408	1
m = 1	2	423.768	1.645	0.439	2
m = 2	4	422.124	0.000	---	0.5 0.5

Interaction Tests

		Fin. Status				Preparation Status				
	First	Has	No			NCBO	One	Two	Term	EA
	Gen	Need	Need	Pell	Loan	Only	Subj.	Subj.	GPA	DIT
Dorm	0.33	0.18	0.61	0.09	0.20	0.24	0.56	0.57	0.04*	0.89
First Gen.		0.52	0.75	0.81	0.27	0.31	0.51	0.85	0.71	0.27
Fin. Status										
Has Need				---	0.34	0.31	0.34	0.68	0.98	0.66
No Need				++	---	---	**	0.74	0.28	0.39
Pell					0.56	0.93	0.72	0.42	0.87	0.20
Loan						0.03*	0.30	0.58	0.56	0.67
Prep. Status										
NCBO									0.87	0.46
One Subj.									0.74	0.40
Two Subj.									0.40	0.61
Term GPA										0.06

“---“ indicates induced multicollinearity. “++” indicates an empty category. “**” indicates removed observations due to a perfect prediction of success.