

ECONOMETRIC DURATION MODELS OF
COLLEGE STUDENT PERSISTENCE

By

JAMES BRIAN EELLS

Bachelor of Science
Southwest Missouri State University
Springfield, Missouri
1989

Master of Science
Oklahoma State University
Stillwater, Oklahoma
1998

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 1999

Thesis
1999D
E265e

ECONOMETRIC DURATION MODELS OF
COLLEGE STUDENT PERSISTENCE

Thesis Approved:

Lee C. Adair

Thesis Advisor

Ronald L. Monahan

James R. Fein

Wade Brown

Wayne B. Powell

Dean of the Graduate College

ACKNOWLEDGMENTS

I am grateful to a number of people who have supported and encouraged me throughout this project. In particular, I wish to thank my advisor, Dr. Lee C. Adkins. If not for his keen insights, patience, and encouragement, this project would have been nearly impossible to complete. I would also like to thank the my dissertation committee members, Dr. James R. Fain, Dr. Ronald L. Moomaw, and Dr. Wade Brorsen. Their insightful comments and suggestions about the methodological and theoretical issues surrounding college student persistence have given me a deeper appreciation as to the possibilities and limitations of research in this area.

I also would like to thank those in the institutional research community at Oklahoma State University. I'm especially grateful to those in the Office of University Assessment: Dr. Stephen Robinson, Dr. Ronald Chaney, Dr. Stacey Boyle, and Dr. Becky Johnson. While employed at the office, their support and encouragement were what peaked my interest in this area of research in the first place. I would also like to thank Lee Tarrant and Barbara Lafon of the Office of Planning, Budget, and Institutional Research. Their painstaking efforts in collecting and educating me about the data made this dissertation a feasible empirical endeavor. I owe a debt of gratitude to Leon Gust of the Oklahoma State Regent for Higher Education for providing me the needed data from the Unitized Data System. Finally, I would like to thank everyone in the Admissions Office for their help, especially Paulette Cundiff, who always helped me with innumerable issues about various processes, policies, and contact people regarding undergraduate enrollment.

I am especially grateful to Ruby Diamond of the Department of Economics and Legal Studies for selflessly assisting me in organizing the completion of this project. I am also grateful to my colleagues Timothy Bisping, Steven Petty, and Ronald Endsley for patiently allowing me to discuss my research with them and for providing insightful feedback. I especially want to thank my undergraduate mentor, Dr. John Harms, who has been a constant inspiration to me.

TABLE OF CONTENTS

1	INTRODUCTION	1
2	REVIEW OF THE LITERATURE	7
2.1	Tinto's Model	7
2.2	Extensions of Tinto's Model	12
2.2.1	A Brief Discussion of Path Analysis	13
2.2.2	Empirical Examples of Tinto's Model	14
2.3	Economics and Student Persistence	18
2.4	Chapter Summary	20
3	MODELING METHODOLOGY	23
3.1	Hypotheses	23
3.1.1	Exit Rate/Persistence Behavior	24
3.1.2	Observed Pattern of Attrition	27
3.1.3	Predictive Evaluation	28
3.2	Statistical Methodology	30
3.2.1	Hazard Functions	30
3.2.2	Duration Dependence	32
3.2.3	Product Integral Representation of the Survivor Function	32
3.2.4	Censoring	35
3.2.5	Multiple Destination Models	35
3.2.6	Covariates	38

3.2.7	Accelerated Failure Times and Proportional Hazard Models . . .	40
3.2.8	Unmeasured Heterogeneity	41
3.2.9	Parametric Hazard Specification, Estimation, and Inference . . .	43
3.2.10	Predictive Measurements	47
3.3	Chapter Summary	54
4	DESCRIPTION OF THE DATA	57
4.1	Student Records	58
4.2	Unitized Data System	66
4.3	School District Data	67
4.4	Variable Descriptions	68
4.4.1	Dependent Variables	68
4.4.2	Independent Variables	70
4.5	Data Reduction Methods	74
4.5.1	Multicollinearity Diagnostics	75
4.6	Chapter Summary	80
5	ANALYSIS AND RESULTS	82
5.1	Hypothesis Tests	82
5.1.1	The Impact of Relative Rank on Enrollment Duration	83
5.1.2	Distinguishing Between Dropout and Transfer	87
5.1.3	The Impact of Class Size on Enrollment Duration	88
5.1.4	The Impact of Graduate Teaching Assistants on Enrollment Duration	89
5.1.5	The Impact of Summer Enrollment on Enrollment Duration . . .	90
5.1.6	The Impact of Unmeasured Heterogeneity on The Hazard Func- tion	90

5.1.7	The Predictive Performance of The Weibull Hazard Model: Enrollment Duration	91
5.1.8	Predictive Performance of the Weibull Model: Exit Destinations	96
5.1.9	Other Results	99
5.2	Chapter Summary	104
6	SUMMARY, DISCUSSION, AND CONCLUSION	107
	REFERENCES	119
	APPENDIX. IRB APPROVAL FORM	123

LIST OF TABLES

1.1	Observed Annual Exit Rates of 2090 Fall 1993 Freshman at Oklahoma State University.	3
3.1	An $r \times c$ Contingency Table.	52
4.1	Variable Names and Descriptions from Oklahoma State University Student Demographics File	59
4.2	Missing Value Replacements for ACT and High School Performance Data. Fall 1990 to Spring 1997 Semesters.	61
4.3	Before and After Comparison of Missing Value Imputation.	64
4.4	Variable Names and Descriptions from The Course Data File	64
4.5	Variable Names and Descriptions from The New Freshman Student Retention Data File	65
4.6	Variable Names and Descriptions from The Unitized Data System File	66
4.7	Variable Names and Descriptions from Oklahoma School District Data	68
4.8	Summary Statistics of Persistence Related Variables. N=2,090.	69
4.9	Top Destination Schools for OSU Transfers.	70
4.10	Summary Statistics of Time-Constant Independent Variables	71
4.11	Summary Statistics of Time-Varying Independent Variables	72
4.12	Variance Inflation Factors of the Independent Variables	75
4.13	Variance Proportions for Selected Regressors with Condition Number Values over 30.	77

4.14	Variance Inflation Factors for the Final List of Independent Variables	78
4.15	Variance Proportions Corresponding to the Two Largest Condition Numbers from the Scaled Hessian of the Hazard Model.	80
5.1	Parameter Estimates for Single-Stage Hazard Model - General Attrition. Dependent Variable: Log(DURATION).	82
5.2	Parameter Estimates for Transition Intensity Model - Transfer. Dependent Variable: Log(DURATION).	84
5.3	Parameter Estimates for Transition Intensity Model - System Dropout. Dependent Variable: Log(DURATION).	84
5.4	Marginal Impact of RELRANK on Enrollment Duration.	86
5.5	Test of Equality Between Transfer and Dropout Coefficients.	87
5.6	Marginal Impact of NSTUDENT on Enrollment Duration.	88
5.7	Marginal Impact of PCTGRAD on Enrollment Duration.	89
5.8	Ordinary Least Squares Coefficients of Determinants of Student Persistence. Dependent Variable: DURATION.	91
5.9	Ordered Logit Coefficients of Determinants of Student Persistence. Dependent Variable: DURATION	92
5.10	In-Sample Predicted versus Actual Enrollment Duration - WGH Model.	93
5.11	In-Sample Predicted versus Actual Enrollment Duration - OLS Model.	93
5.12	In-Sample Predicted versus Actual Enrollment Duration - Ordered Logit Model.	94
5.13	Out-of-Sample Predicted versus Actual Enrollment Duration - WGH Model.	94
5.14	Out-of-Sample Predicted versus Actual Enrollment Duration - OLS Model.	95
5.15	Out-of-Sample Predicted versus Actual Enrollment Duration - Ordered Logit Model.	95

5.16	In- and Out-of-Sample Test and Association Results for WGH, OLS, and Ordered Logit Models	96
5.17	Multinomial Logit Coefficients of Determinants of Dropout, and Trans- fer. Normalized on Persistence	96
5.18	In-Sample Predicted versus Actual Destination - WGH Model.	98
5.19	In-Sample Predicted versus Actual Destination - Multinomial Logit Model.	98
5.20	Out-of-Sample Predicted versus Actual Destination - WGH Model.	98
5.21	Out-of-Sample Predicted versus Actual Destination - Multinomial Logit Model.	99
5.22	In- and Out-of-Sample Test and Association Results for WGH, OLS, and Ordered Logit Models	99
5.23	Marginal Impact of ACTCOMP on Enrollment Duration.	101
5.24	Marginal Impact of ENGINEER on Enrollment Duration.	101
5.25	Marginal Impact of PREPROF on Enrollment Duration.	102
5.26	Marginal Impact of PROBENR on Enrollment Duration.	102
5.27	Marginal Impact of RANKPCTL on Enrollment Duration.	103
5.28	Marginal Impact of RESCODE on Enrollment Duration.	103
5.29	Marginal Impact of SEXCODE on Enrollment Duration.	104

LIST OF FIGURES

2.1	Tinto's Longitudinal Model of Institutional Departure.	9
3.1	Illustration of Heterogeneity on Observed Hazard.	27

CHAPTER 1

INTRODUCTION

This study attempts to specify and estimate statistical models appropriate for predicting undergraduate enrollment durations. Known as failure-time or duration models, these methods offer elegant ways to account for two types of information available in persistence data; namely, the time-to-exit and the type or characteristics of exit. This approach is well suited for longitudinal enrollment data where a cohort of students are followed over time:

To see that persistence in higher education is an interesting topic of study, one need only examine the pervasiveness of student departure. Of the nearly 2.4 million students entering higher education in 1993 for the first time, over 1.5 million will depart from their first choice institution without receiving a degree. Furthermore, of the 1.5 million leaving their first institution, nearly 1.1 million will withdraw from higher education altogether (Tinto, 1993). When restricting attention to four-year institutions, at least two regularities can be observed: First, the typical institution can routinely expect to lose 25 to 30 percent of its entering freshman cohort every academic year. Second, half of the entering cohort will actually maintain continuous enrollment and attain a degree.

This sizable and continuing attrition is not without consequence, either for the individual or the institution. For individual students, attaining a degree takes longer and is more costly, if it is attained at all. Much of the monetary, occupational, and

social rewards of higher education are conditional on earning a college degree. This does not imply that students who fail to obtain a college degree do not benefit from their college experience; however, it is commonly recognized that a college degree, especially a four-year degree, is an important signal to employers and thus, a key to entry into desirable occupations.

Student attrition also affects institutions in that students represent, among other things, an important source of revenue including tuition, fees, state appropriations, and donations from graduates. Even with relatively stable attrition rates across entering student cohorts, institutions will feel budgetary pressure if overall enrollment is declining. Indeed, the projected decline in the college-going population appears to have arrived. Belated recognition of this fact has led institutions to appreciate, as never before, the necessity of retaining as many students as possible. Bean (1982) summarizes the prevailing sentiment succinctly:

In a period when demographic data suggest that freshmen enrollments will decline substantially, the importance of improving retention rates may become more a matter of institutional survival than of academic interest (p. 292).

In response to these pressures, most four-year institutions have invested, in one form or another, in marketing and recruitment activities aimed at increasing the number of applicants. Most have also expanded their efforts to attract applicants other than the traditional college-bound high school student. As a result, the composition of the student body at most four-year institutions have become increasingly heterogeneous, and this diversity has been a confounding influence in studying persistence patterns.

Partly in response to the increased demand for understanding student departure, a great deal of scholarly effort has been expended on the empirical study of student attrition. Traditional (post 1975) studies of student departure are typically institution specific, using path-analytic methods to allocate the variance of factors relating to

enrollment behavior. Student departure is usually defined as a discrete event (depart or continue) within a fixed time period (usually the freshman year). While there is good reason for focusing on the first year of enrollment (most departures occur during this time), the fixed-time approach generally does not account for the impact “fixity” or censoring has on the estimated parameters of the model. The censoring arises because some students have not dropped out by the end of the time window, and not accounting for this could result in misleading inferences, especially if the results are to be generalized to second or third year persistence. Table 1.1 depicts the exit rates of 2090 fall 1993 freshman at Oklahoma State University.

Table 1.1: Observed Annual Exit Rates of 2090 Fall 1993 Freshman at Oklahoma State University.

Year	Exit Rate
Fall 1993	0
Fall 1994	27.0
Fall 1995	13.2
Fall 1996	7.3

As can be seen in Table 1.1 most departure (27 percent) occurs in the first year, and of these survivors, 13.2 percent fail to make it to the second year. Ignoring subsequent departure behavior in research could lead to retention policies that are “front-loaded” where most or all resources devoted to retention are used in the first year. The rationale is to design policies that get students “over the hump” so that their likelihood of persisting improves. These policies may be misguided, especially if the student is highly exit-prone to begin with. In this case, such efforts may only postpone the inevitable. A second problem is that even when data are available on where students depart to when leaving an institution, traditional methodologies either ignore this information entirely or handle it in rather awkward ways. Finally, most traditional studies are not well suited for actually predicting an individual student’s

time-to-exit. The primary reason for this is that in all path-analytic approaches, the ultimate dependent variable depends on unobservable factors, and the quantification of these factors is rarely (if ever) a consideration. Indeed the strength of path analysis is that it allows investigators to make inferences about the parameters of independent variables when unobservable factors exist. "Prediction" in these studies usually refers to the amount of variance in the dependent variable explained by an independent variable. This, of course, is not the same thing as actually predicting the dependent variable.

Some important questions that are addressed in this study include: How are observed enrollment durations distributed and what influence do unmeasured student characteristics have on these observations? To what extent do relative academic standing, classroom composition and staffing, commitment indicators, previous academic experience and skills, and background characteristics influence persistence? In terms of predicting persistence, does the model developed here offer improved predictive accuracy relative to competing models? To answer these questions, this study builds upon the key consistencies found in previous research on student attrition. In particular, the conceptual model proposed by Tinto (1975) is used as a basis for selecting important independent variables to be included. The economics of relative status as proposed by Frank (1985) and extended to higher education by Heath (1993) is used to specify a key variable in predicting student persistence: the academic rank of a student in relation to his or her immediate peers. This study differs from previous studies in a number of important respects. First, a truly longitudinal approach is used where time-to-exit is taken as the dependent variable. Second, the data used in this study allow for the distinction between a student who transfers to another institution and one who drops out of the system. Third, a statistical methodology is used that is appropriate for enrollment duration analysis, and a model is specified that is general enough to allow for different types of exit and for the influence of unmea-

sured student characteristics on observed enrollment durations. Finally, prediction is assessed in terms of the model's ability to accurately predict the dependent variable. The model developed in this study is compared to potential competitors in terms of predictive accuracy where a hold-out sample is used for out-of-sample validation (a practice rarely used in previous attrition research).

The organization of this dissertation is as follows. Chapter 2 reviews both the literature conceptualizing the process of student attrition and the empirical studies of the process. Chapter 3 lays out the research design, the specific hypotheses to be tested, and the statistical methodology used in this study. Chapter 4 discusses the data used in this study. Chapter 5 provides an analysis of the modeling results. Chapter 6 provides additional discussion and conclusions. A summary of the key findings is provided below:

- A student's academic performance relative to his or her immediate classmates is directly related to persistence. A student earning D's in classes where D is the average is more likely to persist than if he or she were in classes where B is the average. This is independent of the effect of poor overall performance.
- Marginal changes in class size do not affect a student's likelihood of persisting.
- Classroom staffing has a major impact on persistence. A student whose courses are taught primarily by graduate student teaching assistants is less likely to persist than if he or she were being taught primarily by faculty.
- Dropouts behave differently than transfers.
- Student heterogeneity (unmeasured or unobserved) affects the observed dropout rates. When slow quitting students are studied together with fast quitting students, the observed average dropout rates over time are dominated initially by the fast quitters.

- The hypotheses regarding the predictive performance of the WGH model were inconclusive. The WGH model could not beat OLS or ordered logit in predicting enrollment duration according to goodness-of-fit tests; however, the WGH model had a much higher hit rate. Multinomial logit performed better than WGH in predicting departure destination based on goodness-of-fit tests and hit rates.

CHAPTER 2

REVIEW OF THE LITERATURE

A wide variety of research has emerged in response to concerns about college student attrition. Most can be classified into three categories: psychological, sociological, and economic. None are mutually exclusive and each have implications for policies dealing with attrition. Excellent surveys of the literature include Tinto (1975), Terenzini and Pascarella (1980), and Tinto (1993). Emerging from these surveys is a comprehensive and sweeping view of student attrition, now referred to as the “interactional theory” of student departure.

2.1 Tinto’s Model

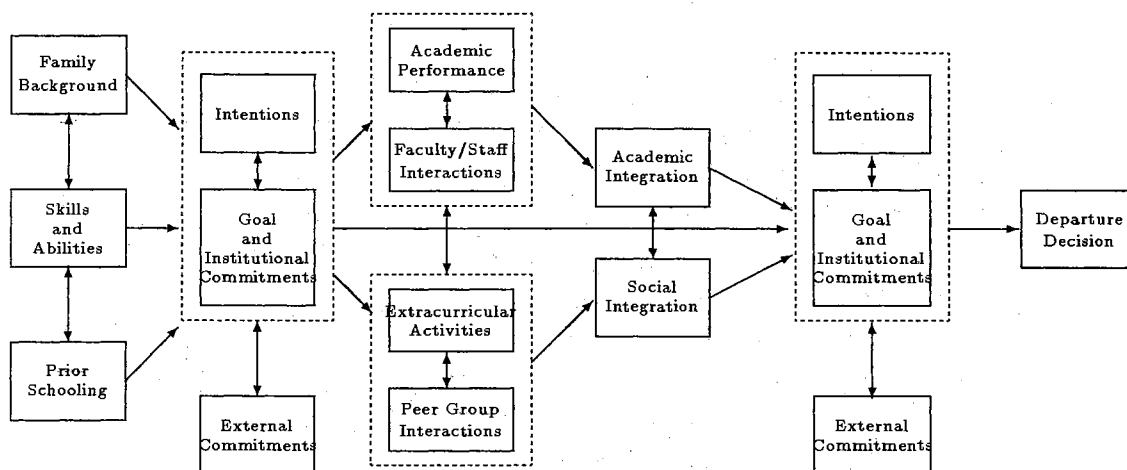
The interactional theory of student departure has gained a considerable following since its principal contributor, Vincent Tinto, published his 1975 article. Tinto’s theory is derived from Van Gennep’s rites of passage and Durkheim’s theory of suicide. The former stresses that entering college involves stages of passage, a separation from past communities, transition from high school to college, and incorporation into the society of college. The latter suggests that student withdrawal, like suicide, arises when individuals are either unable to become sufficiently integrated and establish membership within the communities of college, namely the academic and social systems that exist on campuses, or when the norms and rules on campus are not well defined. Obviously, dropout can occur if a student fails to integrate academically (poor grade

performance, for example). In this case, dropout can either be voluntary (like suicide) or forced through dismissal. Social integration, through peer group associations, extracurricular activities, and interaction with faculty and administrative personnel, plays an important part as well. The more students participate in the social life on campus, the more likely they are to continue enrollment. However, excessive social integration can inhibit academic integration, thus leading to withdrawal. For Tinto, the key to understanding the attrition process is to understand how levels of academic and social integration of the students change over time and how institutions influence these systems. Also important to the process are the decisions made by students regarding the costs and benefits of continuing enrollment. Tinto acknowledges that the ultimate decision to discontinue enrollment is a rational one; that is, at the time of dropout, the perceived costs of continuing enrollment are greater than the benefits. Unfortunately, Tinto does not develop this notion further, and it does not explicitly enter his longitudinal model.

Of primary importance for Tinto was to develop a predictive theory of dropout. He also believed that knowledge of how student perceptions and integration changed over time was the appropriate focus for validating such a theory. As a result, he developed a conceptual model of student departure from a longitudinal perspective. The original diagram of Tinto is presented in Figure 2.1.

In Figure 2.1, a decision is ultimately made to either continue enrollment at a specific institution or leave it. This decision is systematically influenced by the interplay of social, economic, and institutional forces that precede it. For example, family background, individual attributes, and precollege schooling all have an influence on the student's initial commitments to the specific institution. These commitments are manifested in terms of goal commitments (earning a college degree) and institutional commitments (attending a particular institution). For given levels of initial commitments, students begin to integrate into the academic and social systems of

Figure 2.1: Tinto's Longitudinal Model of Institutional Departure.



the institution. The former is reflected primarily in terms of grade performance and intellectual development, while the latter in terms of peer-group and faculty interactions. The degree to which a student integrates into the systems of the institution influences the extent to which goal and institutional commitments are revised. These revisions, along with the level of initial commitment, ultimately influence the decision to stay or leave the institution.

In this framework, attrition occurs primarily because of low goal commitment or low institutional commitment, both being directly related to persistence. The level of initial commitments, together with integration and commitment revision, permit a number of plausible cases in which dropout would likely occur. For example, a strong prior goal commitment to degree completion, in spite of low levels of academic and social integration, could lead to a decision to “stick it out.” It could also lead to a transfer. Another possibility is a student with moderately high prior goal commitment, and who is highly integrated socially but not academically. In this case, the student may have the drop decision forced upon him (i.e., suspension) or may elect to transfer to an institution with similar social systems but more forgiving academic

systems.

Tinto also points out that researchers frequently fail to distinguish between the various forms of dropout. The behaviors associated with voluntary withdrawal are considerably different than those for academic dismissal. While it is true that a lack of academic integration can lead to either form of dropout, voluntary withdrawal has a mismatch dimension that is much less prevalent in academic dismissal. It is also important to distinguish between a complete cessation of involvement in higher education (system dropout), a temporary break in enrollment (stopout), and discontinuing enrollment in one institution to continue in another (transfer). Each of these types should exhibit differing behaviors.

Other important dimensions given by Tinto (1975) include college quality and student composition. The higher the quality of the college, as measured by the proportion of Ph.D faculty or income per student, the higher are the graduation rates. Tinto notes that this comparison masks important interactions among institutional quality, student composition, and individual performance. He cites the "frog pond" effect where there is a direct relationship between the ability level of the student body and the expectations individual students hold for themselves. Students will tend to self-sort, perhaps transferring to institutions where their abilities and expectations are in line with that of the institution. A countervailing force that dampens the self-sorting process is what Tinto refers to as the "social status" effect. Paraphrasing Tinto (1975, p.114) the prestige of an institution is of value to the individuals within the institution and may prevent individuals of low relative rank from dropping out.

Tinto never published an empirical investigation of his theory. His contribution was to define a consistent method for research that yields testable hypothesis. In particular, "[Tinto's] theoretical model of dropout...argues that the process of dropout from college can be viewed as a longitudinal process of interactions between the individual and the academic and social systems of the college" Tinto (1975, p.94). Tinto

presents several claims that are potentially empirically testable, which are summarized as follows:

- Academic and social integration are directly related to persistence.
- Grade performance is likely to be the strongest indicator of academic integration.
- Social integration is reflected primarily through peer group associations, informal faculty interactions, and extracurricular activities.
- Goal and institutional commitments are directly related to persistence.
- Initial goal commitments are shaped by family background, individual attributes, and pre-college schooling. These initial commitments are re-evaluated as the individual begins the integration process upon entering college.
- For given levels of goal commitment, institutional commitment is directly related to persistence, and for high levels of goal commitment, varying levels of institutional commitment may indicate the difference between dropout and transfer.
- It is important to distinguish between the various types of dropout: Voluntary withdrawal versus academic dismissal, and between system withdrawal (dropout), dropout and return (stopout), and transfer.

Tinto's longitudinal model is tailor-made for path analysis. In the next section, path analysis is briefly described and examples in the literature of using path analysis to test many of Tinto's core relationships are discussed.

It should be noted that Tinto's model does not take into account certain characteristics of student departure observed over time. First, Tinto does not explicitly deal with the observed pattern of dropout experienced on many campuses; namely,

exit rates rise at first, reach a maximum, then decline over time. He does recognize that the majority of attrition occurs during the first year of enrollment. A partial explanation is that as one approaches his educational goal, the likelihood of dropout should decrease. This suggests that the college may have some impact on the drop behavior of more tenured students, and that to reduce attrition, policies need only focus on getting students “over the hump.” If the exit profile instead reflects the mobility-prone students dropping out early, leaving behind the “slow quitters”, then such policies may not produce the intended effect, but only delay the inevitable.

The distinction between dropout, stopout, and transfer is important for reasons other than the behavioral ones given by Tinto. There are different processes at work generating the observed data on dropout, stopout and transfer. In particular, the statistical treatment of a stopout should be different from a dropout or transfer because it is a renewal process, exhibiting an on-again-off-again pattern absent with other forms of dropout. This difference may also show up in the behavioral variables, but to address stopout explicitly in a statistical sense requires a completely different, and more complicated analytical approach.

2.2 Extensions of Tinto’s Model

Interactional theory has enjoyed considerable attention in institutional research literature precisely because it offers a comprehensive framework in which to understand the dropout process. It is also well-suited for empirical estimation. Though most claim that the models developed are predictive, much of the discussion in the papers revolves around the explanatory (confirmatory) power of the models. The usual mode of empirical implementation has been to track an entering cohort of freshman for several months, obtaining repeated observations on responses to survey questions designed to “load” on academic and social integration, as well as to obtain information on the other elements in the path diagram. Once the data has been collected,

the structural path coefficients are estimated. The directions and magnitudes are then examined to determine the relative strengths of the hypothesized relationships. A comprehensive review of this literature is found in Pascarella (1980). The next section provides a brief, nontechnical discussion of path analysis.

2.2.1 A Brief Discussion of Path Analysis

A nontechnical discussion of path analysis is found in (Kline, 1994) and a more comprehensive presentation is in Loehlin (1987). According to Loehlin (1987) path analysis, factor analysis, and linear structural relations analysis (LISREL) are all forms of latent variable analysis because some of the variables are not directly observed. A central part of path analysis is that the process is time-ordered and this is depicted in what is called a path diagram. Figure 2.1 represents such a diagram, though it should not be interpreted as a literal representation of an estimable path model. Straight one-headed arrows represent causal relationships between variables while two-headed arrows represent simple correlations. Typically the two-headed arrows are also curved to make them more distinct than the one-headed causal arrows. Also, not shown in 2.1 are various one-headed unlabelled arrows leading to certain variables. These are known as residual arrows and represent a composite of other influences on the variables they point to.

There are essentially two types of variables encountered in path analysis: independent or source variables and dependent or downstream variables. These are analogous to exogenous or predetermined, and endogenous variables in the econometric literature. Independent variables are considered the source of causation. They do not have one-headed arrows pointing toward them; however, they can have two-headed correlation arrows connecting them. Dependent or downstream variables are causally dependent on the other variables in the path diagram. In Figure 2.1, the pre-entry attributes and external commitments are the source variables; everything

else is downstream. Residual arrows are attached to all downstream variables and never to a source variable. The “source” variables corresponding to the residual arrows are assumed to be random variables with zero means, constant variance, and are uncorrelated with the source variables. It is also assumed that the relationships indicated by arrows are linear.

Specifying a path model as a system of equations suggests that the number of equations will equal the number of downstream variables, and each equation expresses a downstream variable as a function of its causal path. The causal paths then represent the structural parameters of a simultaneous equations model and all the identification problems associated with simultaneous equations apply to path analysis as well. The usual method for ensuring identification is to be sure not to have more causal paths (unknown parameters) than downstream variables (equations). This is equivalent to the exclusion restrictions used for identification in simultaneous equations models. Many path models use the time-ordered assumption of the process for creating a *recursive* system of equations. In such a system, ordinary least squares can be used to obtain consistent estimates of model’s parameters, and the standardized estimates are called path coefficients. When some of the downstream variables are latent, factor analysis is used to obtain the path coefficients. The factor pattern from a factor analysis are the path coefficients. In Tinto’s model, goal and institutional commitments as well as academic and social integration are considered latent variables. The signs of the path coefficients independently reflect the direction of change in an upstream variable on a downstream variable and its magnitude and statistical significance the partial strength of the relationship.

2.2.2 Empirical Examples of Tinto’s Model

The literature is replete with empirical examples of Tinto’s model being estimated by path-analytic methods. Most differ slightly in the exclusion restrictions used to create

recursive models while preserving the key components of the model. These models typically rely on a combination of longitudinally tracked student records and surveys. Terenzini and Pascarella (1980) describes the findings of six studies aimed at testing the validity of Tinto's model. Bean (1980), Bean (1982), and Bean (1983) alludes to a theory of worker turnover as a theory of student attrition. The final model closely resembles the Tinto model, and Bean used a path analysis to assess the importance of goal commitment on dropout. Pascarella and Terenzini (1983) used Tinto's original path diagram in a path analysis, focusing on voluntary withdrawal. A generalization of path analysis, LISREL (LInear Structural RELationship modeling), was used by Stage (1988) and Stage (1989) to commitment levels and the integration aspect of Tinto's model. Eaton and Bean (1995) redefine academic and social integration in terms of academic and social approach/avoidance, as suggested from the psychological theory of coping. They use LISREL analysis on the expanded model.

The path-analytic methods for estimating interactional structural models have a number of limitations. Obviously, the quality of the survey instrument will influence the precision of the path coefficients. More importantly, since much interest centers on the process by which students revise their goal and institutional commitments, many of the questions in the instrument involve the intentions of the student. It is frequently assumed that statements of intent are point estimates (forecasts) of future behavior. This is too optimistic. According to Manski (1990), even when intentions are formed under the best of circumstances (i.e., rational expectations) the best a researcher can hope for is to place bounds on probable behavior. In addition to the statistical limitations, there are practical considerations. As previously mentioned, most of the studies involve repeated administration of survey instruments at key points during the period of study. There must be at least two collection points and most opt for three or more. This is likely to be prohibitively costly for most institutions to maintain on an on-going basis.

Comparing the studies using Tinto's framework is difficult because the researchers often use different statistical methodologies, apply these techniques to different populations, use different survey instruments, and expand, restrict, or redefine the original model to suit their purposes. Discussing the results of various models therefore amounts to comparing the directional impact of key variables. In spite of these difficulties, Tinto's conceptual design has demonstrated a remarkable robustness.

- **Integration Constructs:** Those studies that explicitly controlled for academic and social integration found that higher levels of each were directly related to persistence. Eaton and Bean (1995) uses self-reported responses to surveys to establish directional impacts of academic and social integration on persistence. They find that higher academic integration tended to reduce the intent to leave. Particularly important is student formal and informal interaction with faculty. This is directly related to academic integration. Pascarella (1980) finds similar results. In studies segmented by sex Pascarella and Terenzini (1983), Stage (1988) social integration is far more important in predicting dropout for women than for men. In both studies, surveys are used to obtain data for developing these constructs. The surveys differ among the studies, thus limiting the comparability of the results.
- **Commitment Constructs:** Again, where explicitly controlled for, goal and institutional commitments are directly related to persistence. Comparing the relative strengths of either across studies is difficult because of the variations in the causal model specifications.
- **Family Background:** A wide variety of variables are considered to be included in this category. In some form or another, they tend to reflect some dimension of socio-economic status, particularly parent education, income, and hometown demographics. Pascarella (1980) and Stage (1988) find that background char-

acteristics are not significantly related to persistence. Other studies find similar results and exclude such characteristics from the analysis (e.g., Bean (1982)).

- **Individual Attributes and Pre-College Experience:** Individual attributes typically include race, sex, standardized test scores, and choice of major. Pre-College experience variables are drawn from high school performance measures such as grade point average, class rank, and extracurricular involvement. Most analyses are segmented by sex in order to differentiate certain behavioral characteristics, such as the aforementioned social integration. Where explicitly accounted for (Pascarella (1980)), pre-college experience is directly related to persistence.

The merits of these studies are best understood within the context of why they were undertaken. They serve to validate Tinto's conceptual design and to provide a list of potential independent variables important in modeling student persistence, both of which appear to have been accomplished. The criticisms of these analyses are many, depending especially on the statistical orientation of those reviewing these methods. Some common criticisms emerge. First, most of the analyses claim to be involved in developing predictive models of student attrition, yet with the exception of Pascarella and Terenzini (1983), none fully explore the predictive capabilities of their models. Most interpretations of model prediction center around an independent variable's (or set thereof) ability to explain variation. Predictive performance should not be assessed solely in terms of this explanatory dimension; it should include, indeed emphasize, the model's ability, taken as a whole, to predict the dependent variable accurately. Second, models are fitted to the data in ways that may not hold up to out-of-sample validation. The use of a hold-out sample for validation purposes does not appear to be a wide-spread practice in the literature (Terenzini & Pascarella, 1980). Third, these models assume a relatively short dropout time horizon, usually the first year. Granted, the first couple of semesters are where the majority of attrition is observed, but these methods cannot be used to predict dropout "after the hump."

2.3 Economics and Student Persistence

Tinto's model is essentially a sociological model of student attrition. The key drivers of the decision process are the student's ability or inability to integrate into the social systems of college. Furthermore, attrition is presented as if it were a treatable condition. Very little emphasis is placed on the rational process by which students weigh the costs and benefits of persisting in college. Economic theory stresses this point. Early research by McKenzie and Staaf (1974), Kohn, Manski, and Mundel (1976), and Manski and Wise (1983) stressed that individual decisions about persistence are no different in substance than any other economic decision that weighs the costs and benefits of alternative ways of investing one's scarce resources.

Human capital theory has played an important role in providing a framework to model student decision making. In one line of thought, the student is both a producer and consumer of "knowledge" Levin and Tsang (1987). The student is engaged in producing a number of activities (one being attending college) requiring scarce resources. The ultimate purpose of these activities is to enter the student's utility function in a way that yields the highest utility attainable, given the various constraints facing the student. To achieve this utility level, the student must produce these activities in the most efficient way possible. Observed choices (i.e., what school to attend and whether to persist) depend on the interplay between the student's preferences and the constraints s/he faces.

An alternative method of modeling student attrition arises naturally out of the labor economics literature on search theory and matching. An excellent survey of this literature is found in Mortensen (1986). In the job search and matching models, the wage is the key decision variable. The parallel for students is grade performance. For a given freshman cohort, the chosen college represents the result of an optimal search strategy and a criterion for determining which college to accept. In essence, the choice is based on which of all admissible institutions yield the highest net bene-

fits, where admissible means the institutions are feasible (i.e., will admit the student) and achieve at least the minimum (reservation) level of expected relative grade performance. Because information about prospective colleges is imperfect, acquiring information about them involves a cost in time and resources. Because of this, no rational student will search indefinitely for the perfect college to attend, and will likely continue search after enrolling at an institution. This is reasonable since much of what is unknown about the student-college match (especially grade performance) can only be determined through experience. A student learns about these characteristics over time and re-evaluates his decision. The decision to persist, transfer to another institution, or leave higher education altogether, involves a comparison of what is learned "on the job" and the opportunities available to the student.

An offshoot of the human capital approach notes that decisions to attend specific institutions are influenced by more than lifetime earnings considerations. They are also influenced by considerations of where the student will likely fit in the academic and social hierarchy of the institution. Heath (1993, p.83) terms "global status" as the earnings a student expects, given his major, degree, and institutional choice. "Local status" is primarily reflected in terms of grade performance. To the extent that local status matters, a student faces a trade-off between global and local status and will trade one for the other according to his preferences when making college-going decisions. Frank (1985) generalized the concepts of global and local status as economic goods.

Heath (1993) explores these ideas in a utility maximization model and derives a number of implications. First, if some students prefer more local status than others, it is to be expected that equally able students will choose different calibre institutions simply because some students prefer to be "a big fish in a small pond." Second, as enrollment tenure increases, students prefer increasing global status. The skills a student learns while attending college are somewhat transferable, and as students

learn about their true abilities, they may transfer to institutions offering greater global status. Also, at the given institution, seniors would prefer measures to increase global status (for example, higher entrance standards) because such measures enhance the long-run value of their degrees without subjecting them to the consequences of greater rigor.

2.4 Chapter Summary

In order to pull together many of the results of this chapter so that important measurable variables may be identified, refer back to 2.1. The chronology of events leading to the departure decision may be categorized as follows: Pre-college Attributes, Initial Goals and Commitments, Institutional Experiences, Academic and Social integration, and Revised Goals and Commitments. The list below summarizes the types of variables used in student persistence research within these categorizations.

- *Pre-College Attributes*: These are the source variables in a path analysis. In most studies of student attrition, pre-college attributes have been found to be of secondary importance. A further breakdown of pre-college attributes includes family background variable such as parent's education, income, and student's sex and race. The student's skills and abilities are also considered and these are usually measured by a standardized tests such as the SAT or ACT. Finally, the student's prior schooling is considered and measures include high school grade point average, graduating rank and class size, and extracurricular involvement.
- *Initial Goals and Commitments*: These are latent variables in a path analysis. The primary method of obtaining data on initial goals and commitments is by using self-reported responses to surveys administered during the first semester of college. Measuring goals includes expectations about the highest degree to be earned and importance of graduating. Institutional commitments are usually

measured by considering the student's ranking of the chosen university relative to others and their relative confidence about their choice.

- *Institutional Experiences*: Several variables have been considered in measuring institutional experiences. Within the academic systems, the primary variables considered are the student's academic performance as measured by grade point average, and student-faculty interactions. Measuring student-faculty interactions usually takes the form of a self-reported count in the past semester of the number of formal contacts with faculty lasting at least 10 minutes (not counting class time). Regarding the social systems of college, informal peer group and faculty interactions are captured through self-reported extracurricular activities.
- *Academic and Social Integration*: These are also latent factors in the path/factor analysis, and as such, rely on how the measured characteristics relate to them. The preceding bullet lists some of the measures that are used in combination with survey results to form the integration factors.
- *Revised Goals and Commitments*: These are again latent variables which must necessarily be measured by follow-up surveys to the initial goals and commitments. In some cases, where a student drops before the survey is administered, and exit survey or interview can be used to obtain the data.
- *Departure Decision*: This decision is observed when a student either stops coming to all classes or fails to enroll in the following semester. This is usually defined to occur within some time window, for example, during the first year of enrollment.

Clearly, much of what is used in studying persistence involves repeated and extensive use of surveys. This poses potential problems for retention research. First, institutional budgets may be such that proper survey administration is not feasible,

and even if feasible, the administration may not receive adequate attention. Second, legal considerations may require that the students be given the option to make their responses anonymous; thus precluding the ability to track an entire cohort and raising self-selection bias problems. Finally, given the variety of survey instruments used in the literature, and the fact that the responses are self-reported, it is difficult to get a sense of the reliability of the instruments.

There are a number of variables not used in the literature that may help to fill the gap where survey data is lacking. Variables that have received little or no attention include high school background information such as expenditure per student, school district population, average home value in the district (a proxy for tax revenue resources), the student-teacher ratio in the district, and the district poverty rate. Under academic performance, the student's relative rank does not receive attention. This variable measures the academic performance of the student relative to the students in the portfolio of courses he or she is taking and is suggested by the local/global status theory discussed above. Related to student-faculty interactions are the number of students per course in a student's portfolio and the proportion of those courses being taught by graduate teaching assistants. Finally, proxies for goals and commitments include the course-load a student takes per semester and whether the student attends summer courses.

CHAPTER 3

MODELING METHODOLOGY

As discussed in Chapter 2, most research has been aimed at explaining variation in student dropout behavior using path analysis to specify a structural model and to estimate the impacts of the behavioral component of the model. Dropout is usually defined as a binary event that occurs within a fixed time window. Prediction is usually associated with the relative importance (in terms of partial R-square) of each estimated coefficient's impact on dropout. The approach used in this study is considerably different.

3.1 Hypotheses

The aim of this study is to use statistical duration methods to model student attrition, test hypotheses pertaining to attrition, and evaluate the predictive performance of the duration model relative to competitors. Like previous studies, a longitudinal approach is used where a cohort of students is followed for a given period of time, and attrition is influenced by a number of factors. Unlike other studies, the random variable of interest is the amount of time a student remains enrolled, as well as the destination of the student, once departure has occurred.

Using enrollment duration underscores the dynamic nature of student enrollment by utilizing the longitudinal data more effectively for hypothesis testing; the effects of regressors are understood not only in terms of whether a student voluntarily drops out

or not, but also in terms of the time to dropout, and the destination of the student after dropout (i.e., system dropout or transfer). Furthermore, the use of duration methods offers a way to control for the influence of unmeasured characteristics on observed exit rates. The pattern of dropout, the initial increase in exit rates early in enrollment, followed by declining rates after a certain peak, is not part of the path analysis typically undertaken. Finally, path models are seldom used in a truly predictive capacity, that is, to predict dropout. In this study the ability of a duration model, taken as a whole, to predict actual out-of-sample dropout is evaluated.

3.1.1 Exit Rate/Persistence Behavior

Control variables useful for explaining and predicting dropout have been suggested by Tinto (1975) and validated empirically by the empirical studies outlined in Chapter 2. Academic integration, as measured by grade point average (GPA), was identified as an important factor, being directly related to persistence (inversely related to dropout). A better measure in line with the status-seeking theories of Frank (1985) and Heath (1993) is the relative rank of the student to his/her immediate classmates. Relative rank of a student enrolled in $i = 1, 2, \dots, N$ courses is defined as follows:

$$REL\text{RANK} = \frac{1}{N} \sum_{i=1}^N \frac{S_i}{C_i}$$

where S_i is the student's grade point for course i and C_i is the class grade point average for course i . If the academic performance of an individual student deviates substantially from that of the average in the portfolio of courses currently taken, then the likelihood of exit should increase. Formally stated:

Hypothesis 1 *For a given portfolio of courses, and all else constant, persistence increases with relative rank to a certain point then decreases.*

In other words, the relationship between persistence and relative rank is nonlinear. Low or high relative ranks indicate a possible mismatch between the individual and course-portfolio he or she is taking. Furthermore, low relative rank applies only to the portfolio of courses, and does not necessarily imply probational enrollment. If the rank falls below some reservation level, students are more likely to seek alternatives where their abilities are more in line with that of the immediate group. In other words, local status matters. On the other hand, high ranking students are more likely to seek alternatives yielding a higher expected return (global status) while maintaining parity between themselves and that of the group. This reasoning also suggests that in terms of relative rank, and that of other behavioral characteristics, the alternatives a student chooses upon exit are distinct.

Hypothesis 2 *The behavioral characteristics, in terms of the parameters of the models, are individually and collectively different for dropouts and transfers.*

In terms of the Hypothesis 1, when distinguishing between the destinations, the hypothesized impact of relative rank for dropouts should not exhibit the nonlinearity expected for transfers. Likewise, the behavior of transfers in general should be different than that of dropouts.

In Tinto's framework, relative rank would be considered part of academic integration. Another dimension of academic integration is that of student-faculty interactions. Two proxy variables are considered for this dimension for an individual student: the number of students per class in the student's portfolio and the proportion of the portfolio taught by graduate student teaching assistants. The hypothesis for the number of students is:

Hypothesis 3 *Persistence and average class size are inversely related.*

Students in large classes are less likely to interact with faculty for a number of reasons. Competing for faculty attention is more difficult (at least not any easier)

in large classes. For a given amount of time outside of class, the average time per student available decreases as class size increases. There is also an increased sense of anonymity in large classes. Large classes are impersonal: faculty can rarely know all students by name. Each of these factors contributes to lower student-faculty interaction and academic integration. This, in turn, increases likelihood of departure.

A high proportion of a student's portfolio being taught by graduate students is also expected to have an impact on exit rates. This impact is formally stated as:

Hypothesis 4 *Within the portfolio of courses taken by a student, the higher the proportion of those courses being taught by graduate students, the less likely the student will persist.*

Hypothesis 4 follows directly from the fact that graduate student teachers are generally not considered to be faculty by the students, parents, faculty, or administration. Thus, the higher the number of graduate students teaching courses, the fewer the opportunities for students to interact with faculty, and this leads to lower academic integration. It also leads to lower social integration by reducing informal student-faculty interaction.

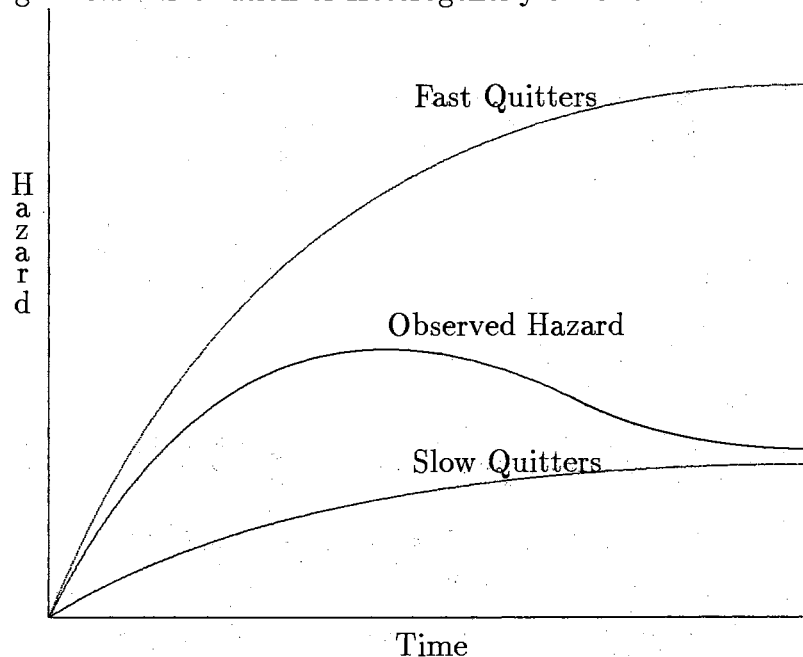
Tinto considered a student's goals and commitments to earning a degree and to the institution as important predictors of persistence. One measure of commitment is whether or not the student enrolls in summer courses. This is because summer enrollment is not required to maintain full-time student status or to graduate within four years. Stated as a hypothesis:

Hypothesis 5 *Students participating in summer courses are more likely to persist, all else constant.*

3.1.2 Observed Pattern of Attrition

The pattern of dropout typically observed is one where dropouts rise rapidly in the beginning, reach a maximum rate, and then fall over time. In duration modeling terminology, an increasing exit rate is called *positive duration dependence* and a decreasing exit rate is called *negative duration dependence*. The explanatory variables used in the model serve to characterize at least some dimensions of this behavior. However, if important variables are omitted from the model, the observed exit rate will be biased downward (i.e., toward negative duration dependence). In terms of duration models, these omitted variables are generically considered as unmeasured heterogeneity. The effect of unmeasured heterogeneity is illustrated in Figure 3.1.

Figure 3.1: Illustration of Heterogeneity on Observed Hazard.



If the student cohort is comprised of two sub-populations, “slow quitters” and “fast quitters” and we observe hazard rates for the cohort as a whole, then the observed hazard rates will reflect the fast quitting behavior first followed by that of the remaining slow quitters. In essence, the observed quit rates reflect the self-sorting behavior of a heterogeneous population.

In this study, the distribution of the unmeasured heterogeneity is parametrically specified. The estimated parameter measures the impact of unobservable/unmeasured student characteristics on the observed dropout rate. If the heterogeneity is significant, then any observed negative duration dependence is augmented by it.

Hypothesis 6 *Observed enrollment duration is affected by unmeasured heterogeneity.*

If false, then all important variables are included in the model, and a simpler model can be used (i.e., one without heterogeneity). If true, the underlying process may still exhibit negative duration dependence. Tinto (1975) hypothesized that the likelihood of dropout should diminish the closer one is to achieving one's goal. Analogous to Mortensen (1988) students may have rising reservation levels of relative rank with tenure. As a student's relative rank improves, the likelihood of being lured to another institution of similar global status is reduced, especially when that institution is offering similar relative rank prospects. On the other hand, the underlying process could exhibit positive duration dependence. The signaling effect of a college degree aside, the knowledge and skills students accumulate during their enrollment tenure is likely to have some market value, and for some the difference between expected earnings with and without degree could be negligible. Unfortunately, the data used in this study are not rich enough to test these competing hypotheses.

3.1.3 Predictive Evaluation

In Chapter 2, the models reviewed discussed validity in terms of the agreement between Tinto's hypothesized directional impacts and that of the estimated coefficients. This was sometimes presented as "predictive validity." This study differs in the interpretation of predictive validity. The model's predictive validity is evaluated by its ability to predict dropout using out-of-sample data. Models that fit well in-sample

often do poorly out-of-sample, and ones that perform relatively well out-of-sample are considered to be more robust. Of course, this is only true when pretesting is used variable selection or if there is structural change in the data generating process. The variable selection procedures in this study differ from other research in that the procedures here are primarily based on collinearity diagnostics and not statistical selection methods, such as stepwise regression.

A specific parametric model of student attrition is considered in this study: enrollment duration is assumed to be distributed as a Weibull random variable and that unmeasured heterogeneity enters multiplicatively as a unit Gamma random variable with constant variance, or Weibull with Gamma Heterogeneity (WGH) model for short. The complete statistical specification is presented below; however, a brief explanation of the choice is in order. The Weibull specification allows flexibility in the determination of duration dependence; positive, negative, or constant. This is controlled parametrically and the parameter estimate provides insight into the process. The unmeasured heterogeneity enters multiplicatively as a unit Gamma with constant variance precisely because the estimate of the variance yields the degree to which unmeasured heterogeneity affects the observed enrollment duration. The duration models estimated in this study are compared to two competing models: Ordinary Least Squares (OLS) regression of dropout time on the independent variables, and an ordered logit model, which is considered to be a semiparametric method for estimating duration models Greene (1995). OLS is a methodology closely tied to the models reviewed in Chapter 2. Ordered logit has not been used in retention research and is considered because here it estimates the probabilities of dropout in a theoretically consistent way, and it circumvents the problem of unmeasured heterogeneity. Stated in terms of a hypothesis,

Hypothesis 7 *The WGH model outperforms OLS or ordered logit in terms of out-of-sample enrollment duration prediction.*

The predictive validity is evaluated by considering two components of the problem: enrollment duration itself, and the destination after enrollment is terminated. The destination of the student is categorical and a number of models exist to deal with a categorical dependent variable (for example, the multinomial logit). This offers another testable hypothesis regarding the WGH model:

Hypothesis 8 *The WGH model outperforms multinomial logit in terms of out-of-sample destination prediction.*

The WGH model generalized to multiple destinations uses what are called transition intensities. These are similar to the multinomial counterparts and can be compared to them in terms of classification accuracy.

3.2 Statistical Methodology

The material presented in this section draws heavily from Lancaster (1990), Amemiya (1985), Heckman and Singer (1986), and Petersen (1986). Notational styles are adopted from Greene (1993) and Lancaster (1990).

3.2.1 Hazard Functions

Assume that the time to departure is a continuous random variable, T , and that a large number of students enroll for the first time at a given university, identified as $T = 0$. Thus, T measures the duration of stay at the university. For the moment, students are assumed to be homogeneous with respect to the systematic factors that affect the distribution of T . This implies that everyone's duration of stay t will be a realization of a random variable from the same probability distribution.

Let dt be a short interval of time after t . The probability that a student departs within an interval dt at or after t is $P(t \leq T < t + dt | T \geq t)$. Dividing by dt yields

the average probability of departure per unit of time over the interval after t , and taking the limit of shorter and shorter intervals formally defines the hazard function:

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}.$$

Thus $h(t)$ measures the instantaneous rate of departure per unit of time at t .

Denote the duration distribution function as $F(t) = P(T < t) = \int_0^t f(z)dz$ and the probability density function as $f(t) = dF/dt$. Then

$$h(t) = f(t|T > t) = \frac{f(t)}{1 - F(t)}$$

by the law of conditional probability where $1 - F(t) = P(T \geq t)$. The denominator is known as the survivor function and measures the probability a student will be enrolled to period t . Denote this as $S(t) = 1 - F(t)$. Note that $f(t) = -dS(t)/dt$.

There is no requirement that $\lim_{t \rightarrow \infty} \int_0^t h(z)dz \rightarrow \infty$ or equivalently that $\lim_{t \rightarrow \infty} 1 - F(t) \rightarrow 0$. If these conditions are satisfied, the duration distribution is termed *non-defective*; otherwise, it is *defective*. A defective distribution implies that in the limit, there is a positive probability of survival.

Given the initial condition $S(0) = 1$, $h(t) = f(t)/S(t)$ is a differential equation in t . This is seen by noting that $h(t)S(t) = -dS/dt$. Thus, $dS/dt + h(t)S(t) = 0$ is a homogeneous first order differential equation. The solution is given by

$$S(t) = \exp \left[- \int_0^t h(z)dz \right]. \quad (3.1)$$

Aside from the negative sign, the term in the exponent of equation 3.1 is known as the *integrated hazard*. The above arguments establish a fundamental relationship between the hazard and survivor functions; if either are known, then the other can easily be derived. This relationship is used repeatedly in this section.

3.2.2 Duration Dependence

Positive duration dependence is said to occur if $dh(t)/dt > 0$, whereas negative duration dependence occurs if $dh(t)/dt < 0$. The former implies that the rate of exit increases over time while the latter means that the rate of exit decreases with time. The condition $dh(t)/dt = 0$ defines a memoryless system which is uniquely identified with the exponential distribution.

3.2.3 Product Integral Representation of the Survivor Function

Another way to consider the fundamental relationship between the hazard and survivor function is to consider the product integral representation of the survivor function. Consider the event $T \geq t$ with probability $S(t)$. Divide the interval from zero to t into $n-1$ subintervals with $s_1 = 0, s_2, \dots, s_{n-1}, s_n = t$. To have $T \geq t$ it is necessary and sufficient to survive each subinterval, and that the event $T \geq t$ is equivalent to the event $T \geq s_1, T \geq s_2, \dots, T \geq s_n$. Thus according to (Lancaster, 1990, p.11),

$$\begin{aligned} P(T \geq t) &= \prod_{j=2}^n P(T \geq s_j | T \geq s_{j-1}) \\ &= \prod_{j=2}^n [1 - P(T < s_j | T \geq s_{j-1})] \\ &= \prod_{j=2}^n [1 - P(s_{j-1} \leq T < s_j | T \geq s_{j-1})] \\ &= \prod_{j=2}^n [1 - h(s_{j-1})(s_j - s_{j-1})] + R_n \end{aligned} \tag{3.2}$$

by the product law of probability, where R_n goes to zero as the difference $(s_j - s_{j-1})$ goes to zero. Equation 3.2 is true for any n , thus

$$P(T \geq t) = \varphi_0^t [1 - h(s) ds]$$

where

$$\varphi_0^t [1 - h(s) ds] \equiv \lim_{n \rightarrow \infty} \prod_{j=2}^n [1 - h(s_{j-1})(s_j - s_{j-1})] \quad (3.3)$$

Equation (3.3) defines the *product integral* of the function $h(s)$ from 0 to t . By (3.1) $S(t) = P(T \geq t)$; hence, it follows that

$$\varphi_0^t [1 - h(s) ds] = \exp \left[- \int_0^t h(z) dz \right]. \quad (3.4)$$

Another property of the product integral representation of the survivor function is that it factors into products of conditional survivor functions, that is $\varphi_0^t = \varphi_0^{t_1} \cdot \varphi_{t_1}^t$ for $0 \leq t_1 \leq t$. This follows because

$$\begin{aligned} \exp \left[- \int_0^t h(z) dz \right] &= \exp \left[- \int_0^{t_1} h(z) dz - \int_{t_1}^t h(z) dz \right] \\ &= \exp \left[- \int_0^{t_1} h(z) dz \right] \cdot \exp \left[- \int_{t_1}^t h(z) dz \right]. \end{aligned} \quad (3.5)$$

In general, the survivor function factors into the product of conditional survivor functions for nonoverlapping adjacent segments of time. This is an especially convenient property when dealing with time-varying covariates, and this topic is taken up below.

Essentially in the product integral representation of the survivor function at t , the survival to t is considered to be the survival through a sequence of Bernoulli trials where success is surviving through the interval $[s_{j-1}, s_j)$. The probability of success, given the survival to the start of the interval, is one minus the product of the hazard rate for that interval and the interval length, as the interval length goes to zero.

A discrete time hazard model is defined when $h(t)$ is zero except at a finite or

countably infinite number of points t_j , where the hazard function takes values $h_j(t_j)$. This implies that $P(s_{j-1} \leq T < s_j | T \geq s_{j-1}) = h_t$ if $[s_{j-1}, s_j)$ contains the point t_k and is zero if $[s_{j-1}, s_j)$ contains none of the points t_j . In this case, (3.3) becomes

$$\begin{aligned} P(T \geq t) &= \int_0^t [1 - h(s)] ds \\ &= \prod_{j|t_j < t} (1 - h_j). \end{aligned} \quad (3.6)$$

The above considerations illustrate a connection between discrete time (Markov) models and continuous time (duration) models. In fact, Amemiya (1985, p.433) begins his discussion of duration models as the limit of discrete time Markov models and proceeds to derive many of the above results from that perspective. He further states that one may want to consider using a continuous time Markov model (i.e., a duration model) in situations where observations are observed discretely over irregular intervals.

[I]n many practical situations a researcher may be able to observe the state of an individual only at discrete times. If the observations occur at irregular times, it is probably more reasonable to assume a continuous-time Markov model rather than a discrete-time model (Amemiya, 1985, pp.440-441).

Lancaster (1990, pp.12-13) provides other reasons for considering continuous-time models over their discrete counterparts.

First it is often mathematically simpler and more elegant. Second, there is rarely in economics a natural discrete-time unit. And third, if different investigators each work with a continuous-time model they will report estimates of parameters that are at least dimensionally comparable even when their data may be grouped or aggregated over time in different ways.

These considerations are especially relevant when dealing with enrollment data, which are almost always recorded discretely (i.e., semesters or quarters) and often in irregular intervals (e.g., summer versus fall semester).

3.2.4 Censoring

Most duration studies involve some sort of censoring mechanism. Kalbfleisch and Prentice (1980) describe many forms of censoring that can occur. Notationally, d_k is a censoring indicator, assuming a value of 1 if the failure event k occurs and 0 if censored. The type of concern in this study is *right* censoring. This occurs because some students are still enrolled when the sampling period is terminated. The censoring just describes is sometimes referred to as Type I censoring. Also considered as Type I censoring is the case when censoring times vary between individuals but are known in advance. If censoring times vary and are not known in advance, this is referred to as random censoring. In contrast, Type II censoring occurs when the experiment is terminated after observing a certain failure time (after the earliest) with the remaining surviving times censored. The importance of censoring is that censored observations are incomplete; that is, their failure times have not been observed. Essentially, all that can be estimated from these observations is the probability of being censored. Estimation is also more complicated because the log-likelihood function now contains the survivor function in the equation.

3.2.5 Multiple Destination Models

The type of process considered in this study is a single cycle model with multiple destinations. Single cycle refers to the passage of a person from entry into a state to exit from it. Thus, the cycle of student enrollment will end when the student transfers to another institution or drops out of the system.

Multiple destinations can be introduced simply by subscripting the hazard and

survivor functions, though the interpretation of each changes somewhat. Suppose there are K possible destinations $k = 1, 2, \dots, K$ and let the set d_k contain dummy variables where d_k equals 1 if destination k is entered and zero otherwise. Then the *transition intensities* are written as

$$h_k(t)dt = P(\text{depart to state } k \text{ in the interval, } (t, t + dt) \text{ given survival to } t).$$

Lancaster (1990) shows that the hazard function is the sum of the transition intensities over the destination states:

$$h(t) = \sum_{k=1}^K h_k(t). \quad (3.7)$$

Obviously, when there is only one destination, the transition intensity is the hazard function.

The marginal probabilities of the destinations are defined as

$$\pi_k = P(\text{departure to destination } k), \quad k = 1, 2, \dots, K.$$

The connection between the marginal probabilities and the transition intensities is established by first noting that

$$\begin{aligned} S(t)h_k(t)dt &= P(\text{survival to } t) \times P(\text{departure to } k \text{ in } (t, t + dt) | \text{ survival to } t) \\ &= P(\text{departure to } k \text{ in } t, t + dt). \end{aligned} \quad (3.8)$$

In essence, (3.8) specifies the proportion of an entering cohort that departs to destination k in $(t, t + dt)$. Integrating over t gives the proportion of the cohort departing for destination k ,

$$\pi_k = \int_0^{\infty} S(s)h_k(s)ds. \quad (3.9)$$

The survivor function conditional on departure to destination k is

$$S_k(t) = P(\text{survival to } t, \text{ given that departure is to } k)$$

with $F_k(t)$ and $f_k(t)$ being the corresponding distribution and density functions. The probability of surviving to t and departing to k is $\pi_k S_k(t)$. It follows that summing these probabilities over the number of destinations gives the probability of surviving to t , that is, the survivor function:

$$S(t) = \sum_{k=1}^K \pi_k S_k(t). \quad (3.10)$$

The probability of departure to k in $(t, t + dt)$ is

$$S(t)h_k(t)dt = \pi_k f_k(t)dt. \quad (3.11)$$

To see this, integrate (3.11) over t

$$\begin{aligned} \int_0^{\infty} S(t)h_k(t)dt &= \int_0^{\infty} \pi_k f_k(t)dt \\ &= \pi_k \int_0^{\infty} f_k(t)dt \\ &= \pi_k \end{aligned}$$

which is the marginal probability as specified in (3.9). From (3.11) the transition intensity $h_k(t)$ is equal to $f_k(t)\pi_k/S(t)$, which shows that the conditioning event for the transition intensity is survival to t , not survival to t and departure to k .

The joint probability density function of the destination indicators d_k and T is derived from (3.1), (3.7), and (3.11). First note that

$$P(\text{departure to } k \text{ at } t) = h_k(t)S(t)dt$$

$$P_k(t) = h_k(t) \exp \left[- \int_0^t \sum_{k=1}^K h_k(u) du \right] dt. \quad (3.12)$$

Then the joint probability density function for the d_k and t is

$$\begin{aligned} P(d_1, d_2, \dots, d_K, t) &= \prod_{k=1}^K P_k(t)^{d_k} \\ &= \prod_{k=1}^K \left[h_k(t)^{d_k} \exp \left(- \int_0^t \sum_{k=1}^K d_k h_k(u) du \right) \right] \\ &= \exp \left(- \int_0^t \sum_{k=1}^K h_k(u) du \right) \prod_{k=1}^K h_k(t)^{d_k} \\ &= \exp \sum_{k=1}^K \left[d_k \log h_k(t) - \int_0^t h_k(u) du \right]. \end{aligned} \quad (3.13)$$

Once a functional form is specified for the transition intensities, (3.13) leads immediately to the likelihood function.

3.2.6 Covariates

Introducing covariates (or explanatory variables) into the analysis allows for systematic differences between students to condition the duration distribution. Thus, the hazard may be written as

$$h(t|\mathbf{x}) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t + dt | T \geq t, \mathbf{x})}{dt}$$

where \mathbf{x} is a vector of covariates. The covariates are assumed to be time invariant. If this is not the case, then special estimation problems arise, with implications for hazard function models.

Kalbfleisch and Prentice (1980) identify two broad classifications for covariates; external and internal. External covariates are either considered fixed over time or if they vary, they are not directly related to the observed durations or exit rates. Time-varying external covariates do not functionally depend on stochastic process that

generates the durations. They are exogenous (Lancaster, 1990). Examples of fixed external covariates in this study include ACT scores, race, and sex. Time-varying external covariates would include, for example, the student's age, grade point average, or marital status. Internal, time-varying covariates are observed so long as the individual survives and is not censored. Its observed value carries information about the survival time of the corresponding individual. Examples of internal covariates in this study include the relative rank of a student or enrollment status. These variables are considered to be endogenous, though not necessarily in the same sense as simultaneous equations. The endogeneity in duration models affects the interpretation the hazard and the relationship between the hazard and survivor function. Internal covariates preclude the probabilistic interpretation of the hazard and survivor functions because the conditional probability of exit at time t is conditioned by $\mathbf{x}(t)$, itself a function of t . Inferences about the hazard conditional on $\mathbf{x}(t)$ can be made if certain assumptions about $\mathbf{x}(t)$ are made.

Denote a vector of time-varying covariates as $\mathbf{x}(s)$ whose value at time t is $\mathbf{x}(t)$. The process may be stochastic (grade point average) or deterministic (age). If stochastic, the state space may be discrete or continuous. Petersen (1986) has shown that if time can be divided into nonoverlapping adjacent time segments such that the time-varying covariates are constant in each segment, then the likelihood function can be factored into a step-like function which can be maximized according to the parameters of the model. For simplicity, assume only one destination and no censoring. To formulate the model, let t be divided into n exhaustive, nonoverlapping intervals $s_0 < s_1 < \dots < s_n$, where $s_0 = 0$ and $s_n = t$. The covariates are assumed to stay constant within each interval, but may vary between intervals. The hazard for the interval (s_{j-1}, s_j) is $h(t|\mathbf{x}_j)$. Then from the relationship between the hazard and

survivor function,

$$P(T \geq s_j | T \geq s_{j-1}, \mathbf{x}_j) = \exp \left(- \sum_{k=1}^K \int_{s_{j-1}}^{s_j} h(u | \mathbf{x}_j) du \right).$$

Thus, the survivor and probability density functions are

$$\begin{aligned} S(t | \mathbf{x}(t)) &= \prod_{j=1}^n P(T \geq s_j | T \geq s_{j-1}, \mathbf{x}_j) \\ &= \exp \left(- \sum_{j=1}^n \int_{s_{j-1}}^{s_j} h(u | \mathbf{x}_{j-1}) du \right) \end{aligned} \quad (3.14)$$

and

$$\begin{aligned} f(t | \mathbf{x}(t)) &= h(t | \mathbf{x}(t)) S(t | \mathbf{x}(t)) \\ &= h(t | \mathbf{x}(t)) \times \exp \left(- \sum_{j=1}^n \int_{s_{j-1}}^{s_j} h(u | \mathbf{x}_j) du \right). \end{aligned} \quad (3.15)$$

Using the results from (3.13) for the case of multiple destinations, (3.15) becomes

$$p(d_1, d_2, \dots, d_K, t | \mathbf{x}(t)) = \exp \sum_{k=1}^K \left[d_k \log h_k(t) - \sum_{j=1}^n \int_{s_{j-1}}^{s_j} h(u | \mathbf{x}_{j-1}) du \right]. \quad (3.16)$$

3.2.7 Accelerated Failure Times and Proportional Hazard Models

The assumption behind both the accelerated failure time and proportional hazard model is the ability to separate the hazard functions into two parts. The proportional hazard model assumes the hazard can be expressed in the following form:

$$h(\mathbf{x}, t) = k_1(\mathbf{x}) k_2(t),$$

where k_1 and k_2 are the same functions for all individuals. The function k_2 is called the baseline hazard. Covariates affect the hazard multiplicatively. The ability to factor the hazard into two parts is a great simplification in estimation, especially for

log-transformation of the hazard.

The accelerated failure time model expresses the duration of an individual as

$$T = \frac{T_0}{\lambda(\mathbf{x}'\beta)},$$

where T_0 is a random variable not involving \mathbf{x} or β and λ is some function. The duration or failure time of an individual is accelerated or decelerated with \mathbf{x} relative to T_0 depending on whether $\lambda > 0$ or $\lambda < 0$ (hence the name). Taking logs of both sides yields

$$\log T = \log \lambda(\mathbf{x}'\beta) + \log T_0 + U$$

where if $\lambda(\mathbf{x}'\beta) = \exp(\mathbf{x}'\beta)$, the model would resemble a linear regression model and be estimable via least squares, depending on the assumptions about the error term U . If some observations were censored, then least squares would not be appropriate; instead, limited dependent variable procedures such as the tobit model could be used. In any case, though it is possible to use simpler estimation techniques in duration modeling, Kalbfleisch and Prentice (1980) shows that they are inefficient relative to maximum likelihood, especially the ordinary least squares estimator.

3.2.8 Unmeasured Heterogeneity

When regressors are used in the hazard function, it is sometimes assumed that those covariates (1) completely capture the systematic differences between individuals, and (2) they are measured without error, denoted v . If either of these assumptions fail to hold, the models will contain unmeasured heterogeneity. In standard regression, assumptions about the error term are made to alleviate the effects of unmeasured heterogeneity. In duration models, the effect of unmeasured heterogeneity is to bias the hazard function toward negative duration dependence. To see this consider the argument given by Heckman and Singer (1986, p.53) which is based on an application

of the Cauchy-Schwartz theorem. Let $h(t|\mathbf{x}, v)$ be the hazard conditional on \mathbf{x}, v and $h(t|\mathbf{x})$ be the hazard conditional only on \mathbf{x} . The conditional distributions for these hazards are respectively $F(t|\mathbf{x}, v)$ and $F(t|\mathbf{x})$. Then by the definition of the hazard function

$$\begin{aligned} h(t|\mathbf{x}, v) &= \frac{f(t|\mathbf{x}, v)}{1 - F(t|\mathbf{x}, v)} \\ h(t|\mathbf{x}) &= \frac{\int_v f(t|\mathbf{x}, v) du(v)}{\int_v [1 - F(t|\mathbf{x}, v)] du(v)} \\ \text{and} \\ \frac{\partial h(t|\mathbf{x}, v)}{\partial t} &= \frac{\frac{\partial f(t|\mathbf{x}, v)}{\partial t}}{1 - F(t|\mathbf{x}, v)} + \left[\frac{f(t|\mathbf{x}, v)}{1 - F(t|\mathbf{x}, v)} \right]^2. \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial h(t|\mathbf{x})}{\partial t} &= \frac{\int_v [1 - F(t|\mathbf{x}, v)] \frac{\partial h(t|\mathbf{x}, v)}{\partial t} du(v)}{\int_v [1 - F(t|\mathbf{x}, v)] du(v)} \\ &+ \frac{[\int_v f(t|\mathbf{x}, v) du(v)]^2 - \int_v \frac{f(t|\mathbf{x}, v)}{1 - F(t|\mathbf{x}, v)} du(v) \int_v [1 - F(t|\mathbf{x}, v)] du(v)}{[\int_v [1 - F(t|\mathbf{x}, v)] du(v)]^2}. \end{aligned}$$

The numerator of the second term can be rearranged further to show

$$\int_v [1 - F(t|\mathbf{x}, v)] du(v) \times \left[\left(\frac{\int_v f(t|\mathbf{x}, v) du(v)}{\sqrt{\int_v [1 - F(t|\mathbf{x}, v)] du(v)}} \right)^2 - \int_v \frac{f^2(t|\mathbf{x}, v)}{(\sqrt{1 - F(t|\mathbf{x}, v)})^2} du(v) \right].$$

The bracketted term is always nonpositive by the Cauchy-Schwartz inequality.

Incorporating unmeasured heterogeneity into the duration model is usually accomplished by conditioning the hazard function, $h(t|\mathbf{x}(t), v(t))$. If $v(t) = v$ for all t , this is referred to as unmeasured scalar heterogeneity, and is often used in practice. Furthermore, unmeasured heterogeneity is assumed to enter the hazard multiplicatively; $h(t|\mathbf{x}, v) = vh(t|\mathbf{x})$.

3.2.9 Parametric Hazard Specification, Estimation, and Inference

Now consider specific forms of the duration distribution. In particular, consider the Weibull model:

$$F(t) = 1 - \exp(-\lambda t)^\alpha \quad (3.17)$$

$$S(t) = \exp(-\lambda t)^\alpha \quad (3.18)$$

$$f(t) = \alpha \lambda t^{\alpha-1} \exp(-\lambda t)^\alpha \quad (3.19)$$

$$h(t) = \alpha \lambda (\lambda t)^{\alpha-1} \quad (3.20)$$

where $\alpha, \lambda > 0$ and equations 3.17 through 3.20 describe the distribution, survivor, density, and hazard functions, respectively. Models with covariates typically specify $\lambda = \exp(-\mathbf{x}'\beta)$. This notation will be suppressed in the following discussion. Depending on whether α is less than (greater than) 1, the hazard will be monotonically decreasing (increasing), and in the case where α equals 1, the hazard is constant. Thus, the Weibull model by itself would not be an appropriate specification for the observed duration of enrollment. Accounting for unmeasured heterogeneity results in a mixture model specification that allows non-monotonic hazard rates.

Assume that v is distributed as a Gamma random variable. The density of v is

$$f(v) = \frac{b^a v^{a-1} e^{-bv}}{\Gamma(a)}, \quad a, b > 0, v \geq 0$$

with $E(v) = a/b$, $Var(v) = a/b^2$, and the denominator is the Gamma function. If v is a realization of a unit Gamma random variable with mean 1 and variance σ^2 , then $a = b = \theta$, $E(v) = 1$, and $Var(v) = \sigma^2 = 1/\theta$. The density of v now becomes

$$f(v) = \frac{\theta^\theta v^{\theta-1} e^{-\theta v}}{\Gamma(\theta)}.$$

Lancaster (1990) shows that a mixture of the Gamma with Weibull distributions yields what is called the *Burr distribution*, the survivor function may be written as (Greene, 1995, p. 738)

$$\begin{aligned}
 S(t) &= \text{expected value over } v \text{ of } S(t|v) \\
 &= \int_0^\infty v S(t|v) dv \\
 &= [1 + \theta(\lambda t)^\alpha]^{-1/\theta}
 \end{aligned}$$

where $S(t|v) = v \exp -(\lambda t)^\alpha$. Recall that the hazard is the product of the survivor and density functions, and that the density is minus the derivative of the survivor function, it is fairly simple to show that the hazard function is

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1}[1 + \theta(\lambda t)^\alpha]^{-1} \quad (3.21)$$

$$= \alpha\lambda(\lambda t)^{\alpha-1}S(t)^\theta. \quad (3.22)$$

The first term, $\alpha\lambda(\lambda t)^{\alpha-1}$, is the Weibull hazard and $[1 + \theta(\lambda t)^\alpha]^{-1}$ is the mixture survivor function. The parameter θ captures the effect of unmeasured heterogeneity. The log-logistic model emerges as a special case when $\theta = 1$, and the Weibull model results when $\theta = 0$. The expected survival time is given by

$$E(T) = \lambda^{1/\alpha} \frac{\Gamma(1 + \frac{1}{\alpha})\Gamma(\frac{1}{\theta} - \frac{1}{\alpha})}{\theta^{1+\frac{1}{\alpha}}\Gamma(\frac{1}{\theta} + 1)}$$

Lancaster (1990, pp. 195-197) provides a score statistic test to determine the existence of an interior maximum of θ , which occurs for non negative values. This has implications for the appropriateness of the mixture model. A quick method of checking this by computing the following

$$S = \frac{1}{2} \left[\sum_{n=1}^N t_n^2 - 2 \sum_{n=1}^N t_n \right]$$

where t_n is the duration of individual n . If S is negative, then potential computing problems are likely in trying to fit the mixture model to data.

In specifying the likelihood function, let $n = 1, 2, \dots, N$ denote the n th individual, $k = 1, 2, \dots, K$ be the k th destination, and γ be a J dimensional vector of parameters. Based on (3.22), this vector includes α, β , and θ where $\lambda = \exp(-\mathbf{x}'_n \beta)$. The data consist of a duration t and a vector \mathbf{d} of $K - 1$ binary destination indicators of which exactly one is unity and the rest are zero. (The origin state is excluded). The log likelihood contribution of an individual is (Lancaster, 1990)

$$L_i = \sum_{k=1}^K [d_{nk} \log h_{nk}(t_n) - \int_0^{t_i} h_{nk}(u) du] \quad (3.23)$$

where $h_{nk}(t)$ is the transition intensity of individual n out of the origin state, the summation is over all possible states, excluding the origin, and the k th element of the vector \mathbf{d}_n is d_{nk} . The full log likelihood is given by

$$L = \sum_{n=1}^N \sum_{k=1}^K [d_{nk} \log h_{nk}(t_n) - \int_0^{t_i} h_{nk}(u) du] \quad (3.24)$$

or, by interchanging the order of summation

$$L = \sum_{k=1}^K L_k \quad (3.25)$$

where

$$L_k = \sum_{n=1}^N [d_{nk} \log h_{nk}(t_n) - \int_0^{t_i} h_{nk}(u) du]. \quad (3.26)$$

L is in part the sum of the contributions from each of the $K - 1$ destinations. If $K^* \leq K$ are specified parametrically and the remaining unspecified transition intensities are functionally independent of γ , their contribution in (3.25) becomes an additive constant. Since adding a constant to the log likelihood does not affect maximization

with respect to γ , (3.25) may be written as

$$L = \sum_{k \in K^*} L_k. \quad (3.27)$$

Following Lancaster, simplify the notation by writing

$$z_{nk} = \int_0^{t_n} h_{nk}(u) du.$$

The first order conditions are

$$\frac{\partial L}{\partial \gamma_j} = \sum_{n=1}^N \sum_{k \in K^*} \left[d_{nk} \frac{h_{nk}^j}{h_{nk}} - z_{nk}^j \right] = 0, j = 1, 2, \dots, J, \quad (3.28)$$

where

$$h^j = \frac{\partial h}{\partial \gamma_j}, z^j = \frac{\partial z}{\partial \gamma_j},$$

and J is the dimensionality of γ . The Hessian is given by

$$\frac{\partial^2 L}{\partial \gamma_j \partial \gamma_l} = \sum_{n=1}^N \sum_{k \in K} \left[d_{nk} \left\{ \frac{h_{nk}^{jl}}{h_{nk}} - \frac{h_{nk}^j h_{nk}^l}{(h_{nk})^2} \right\} - z_{nk}^{jl} \right], \quad (3.29)$$

$$j, l = 1, 2, \dots, J.$$

The information matrix is

$$I_{jl} = -E \left[\frac{\partial^2 L}{\partial \gamma_j \partial \gamma_l} \right]. \quad (3.30)$$

The maximum likelihood estimates, $\tilde{\gamma}$, are found by solving equation (3.28) for γ . The well-known asymptotic properties of the MLE are (1) the MLE is consistent, (2) the MLE is asymptotically normally distributed, and (3) the MLE is asymptotically efficient.

The MLEs $\tilde{\gamma}$ are substituted into (3.30) to derive an estimate of the asymptotic

covariance matrix of $\tilde{\gamma}$

$$\tilde{V}(\tilde{\gamma}) = I(\tilde{\gamma})^{-1}/T$$

which may be used for hypothesis testing about the elements of γ .

3.2.10 Predictive Measurements

The Weibull model is compared to competing models in terms of both predicting enrollment duration and the destination upon departure. The models used for predicting enrollment duration are ordinary least square (OLS) and ordered logit. The model used for predicting destination (i.e., continue, transfer, or dropout) is a multinomial logit. The competing models are estimated using data from the last semester enrolled (just prior to departure). In contrast, the WGH model is estimated using the longitudinal data for each student which includes data observed each semester the student is enrolled. However, all models use the final semester's enrollment data for making predictions. This is done so that predictions from all models are compared on the same information sets.

If longitudinal dynamics matter in the departure decision, then the WGH model should have a predictive advantage because it uses this information in the estimation process. On the other hand, if only the most recent information matters, the WGH model may not have an advantage and could perform worse than the other models. The WGH model may also perform worse because it is overparameterized; that is, it imposes too much structure on the problem and does not fit the data well. Before discussing how prediction comparisons are made, each of the competing models are briefly described. For a textbook discussion of these models, see Greene (1993). To avoid confusion, the parameter vector β is used generically to describe a vector of unknown parameters to be estimated and is not intended to be specific to any one model.

Ordinary Least Squares

Consider the linear model, $y = \mathbf{X}\beta + e$ where y is a $T \times 1$ vector of observations on enrollment durations, \mathbf{X} is a $T \times K$ matrix of independent variables, β is a $K \times 1$ vector of unknown parameters, and e is the disturbance term assumed to be an independently, identically distributed random variable with zero mean, constant variance, and no correlation with the independent variables. Then the OLS estimator is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y. \quad (3.31)$$

If the assumptions about the error term are correct, then the OLS estimator is unbiased, consistent, and has smaller sampling variance than any other linear unbiased estimator. If the error term is normally distributed, then the OLS estimator is also the maximum likelihood estimator and is asymptotically efficient.

Because enrollment durations are nonnegative, the zero mean assumption of the error term is questionable. However, if the log of duration is used instead, the accelerated lifetime model described in the above section applies, and though inefficient relative to maximum likelihood, OLS can be used.

Ordered Logit

Suppose the linear model for the i th individual ($i = 1, 2, \dots, T$) is now given by $y_i^* = \beta'x_i + e_i$ where y_i^* is unobserved enrollment duration; however, what is observed is

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ 1 & \text{if } 0 < y_i^* \leq \mu_1 \\ 2 & \text{if } \mu_1 < y_i^* \leq \mu_2 \\ \vdots & \\ J & \text{if } \mu_{J-1} < y_i^* \end{cases}$$

where the μ 's are unknown parameters to be estimated along with β . This is the process that describes the data in this study: observed durations are either zero semesters, one semester, two semesters, etc.

Let $\mu_0 = -\infty$ and $\mu_J = \infty$. Then define the following

$$z_{ij} = \begin{cases} 1 & \text{if } \mu_{j-1} < y_i \leq \mu_j \\ 0 & \text{otherwise} \end{cases}$$

and

$$P(z_{ij} = 1) = F(\mu_j - \beta' \mathbf{x}_i) - F(\mu_{j-1} - \beta' \mathbf{x}_i)$$

where F is the standard logistic cumulative distribution function (cdf), $1/(1+\exp(-\beta' \mathbf{x}_i))$.

The likelihood function is given by

$$L = \prod_{i=1}^T \prod_{j=1}^J [F(\mu_j - \beta' \mathbf{x}_i) - F(\mu_{j-1} - \beta' \mathbf{x}_i)]^{z_{ij}}. \quad (3.32)$$

Equation 3.32 can be maximized with respect to the μ_j and β using iterative methods.

Multinomial Logit

Suppose there are m unordered categories for each individual $i = 1, 2, \dots, T$ with corresponding probabilities $P_{i1}, P_{i2}, \dots, P_{im}$ and F is the standard logistic cdf. Let

$$\begin{aligned} \frac{P_{i1}}{P_{i1} + P_{im}} &= F(\beta'_1 \mathbf{x}_i) \\ \frac{P_{i2}}{P_{i2} + P_{im}} &= F(\beta'_2 \mathbf{x}_i) \\ &\vdots \\ \frac{P_{i,j-1}}{P_{i,j-1} + P_{im}} &= F(\beta'_{i,j-1} \mathbf{x}_i) \end{aligned}$$

These imply that

$$\frac{P_{ij}}{P_{im}} = \frac{F(\beta'_j \mathbf{x}_i)}{1 - F(\beta'_j \mathbf{x}_i)} = \exp(\beta'_j \mathbf{x}_i)$$

Note that

$$\sum_{j=1}^{m-1} \frac{P_{ij}}{P_{im}} = \frac{1 - P_{im}}{P_{im}} = \frac{1}{P_{im}} - 1$$

so that

$$P_{im} = \frac{1}{1 + \sum_{j=1}^{m-1} \exp(\beta'_j \mathbf{x}_i)}$$

and

$$P_{ij} = \frac{\exp(\beta'_j \mathbf{x}_i)}{1 + \sum_{j=1}^{m-1} \exp(\beta'_j \mathbf{x}_i)}$$

If we consider the P_{ij} and P_{im} as multinomial probabilities and a dummy category indicator is defined as

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ is observed in category } j \\ 0 & \text{otherwise} \end{cases}$$

then the multinomial logit model likelihood function can be written as

$$L = \prod_{i=1}^T P_{i1}^{y_{i1}} P_{i2}^{y_{i2}} \dots P_{im}^{y_{im}}. \quad (3.33)$$

Equation 3.33 can be maximized with respect to the unknown parameters β'_j using iterative methods.

Predictive Evaluation Methods

The problem of evaluating the predictive performance of the WGH model is divided into two parts: (1) evaluate the ability of the WGH model to predict enrollment duration and (2) evaluate its ability to predict departure destination. Regarding enrollment duration, the WGH model is compared to OLS and ordered logit. With respect to departure destination, the WGH model is compared to multinomial logit.

All models are compared using statistical methods suitable for contingency tables.

To be consistent with the discrete nature of the dependent variable, the integer value of each model's duration predictions are used for comparison to the actual number of semesters completed. That is if the WGH, OLS, or ordered logit yielded a prediction of 6.8 semesters, the integer part, 6, would be used for predicted enrollment duration. This is consistent with how the dependent variable is defined: if a student's actual (unobserved) departure occurred at 6.8 semesters, the student would have been observed to complete 6 semesters and fail to enroll in the 7th.

The WGH and OLS model yield predictions of enrollment duration whereas ordered logit predicts the probability of departing at a particular semester. To convert the ordered logit probability predictions into enrollment durations, the following formula is used:

$$ETIME = \sum_{t=1}^T t * P_t$$

where t is the semester and P_t is the predicted probability of departing at semester t . Both the modeling and validation data are censored at the 7th semester; therefore, any prediction exceeding the 7th semester is censored as well.

The discrete enrollment predictions are cross-tabulated with the actual enrollment durations, forming a contingency table. A model is said to predict well if there is a strong, positive linear association between the predicted and actual enrollment durations. Various tests and measures of goodness-of-fit are available for contingency tables and those used in this study will be described shortly.

The categorical dependent variable designating departure destination is defined as follows: If a student survives to the 7th semester, then the destination variable is coded 0 for "continue." If the student drops out, the destination variable is coded 1 and if the student transfers, a 2 is coded. To compare the WGH model's ability to predict departure destinations, the predicted enrollment duration and the transition intensities are used to define a categorical destination prediction variable. If the pre-

dicted enrollment duration is at least 7 semesters, then the destination prediction is coded zero to designate “continue.” If the predicted enrollment duration is less than 7 semesters, then the maximum of the transition intensities are used to determine whether the student drops out or transfers. If this maximum is the dropout intensity, then the destination prediction is coded 1 for “dropout”, otherwise the destination prediction is coded 2 for “transfer.” For the multinomial predictions, the maximum of the three probabilities is chosen as the predicted destination and coded accordingly. In the multinomial model, continued enrollment is set to be the normalizing category. The WGH and multinomial destination predictions are cross-tabulated with the actual destinations to form contingency tables. Again, a good fit is indicated by a strong, positive linear association between the actual and predicted destinations. The general form of an $r \times c$ contingency table is presented in Table 3.1:

Table 3.1: An $r \times c$ Contingency Table.

Predicted Outcome	Actual Outcome				Total
	1	2	...	c	
1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
\vdots	\vdots	\vdots		\vdots	\vdots
r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$...	$n_{.c}$	n

Notes

n_{ij} is the number of i predictions that were actually j .

$n_{i.}$ is the sum of row i .

$n_{.j}$ is the sum of column j .

r_i is the i th row number (rank order).

c_j is the j th column number (rank order).

$$\bar{r} = \sum_i \sum_j n_{ij} r_i / n$$

$$\bar{c} = \sum_i \sum_j n_{ij} c_j / n$$

The Kruskal-Wallis nonparametric test is used to test the null hypothesis that the predictions of the k models are independent samples from identical populations. For large samples, the statistic is approximately distributed as a chi-square random

variable with $k - 1$ degrees of freedom under the null hypothesis and is computed as follows:

$$h = \frac{12}{n(n+1)} \sum_{i=1}^k k \frac{r_i^2}{n_i} - 3(n+1)$$

where r_i is the sum of the ranked predictions of model i . Failure to reject the null hypothesis suggests that the models predict similarly. Rejecting the null hypothesis provides a basis for comparing the relative differences between the models.

Within each model, a simple (corrected for continuity) chi-square test is used for testing the independence of the actual and predicted durations and destinations. The test is computed as follows:

$$C = \sum_i \sum_j \frac{[\max(0, |n_{ij} - e_{ij}| - 0.5)]^2}{e_{ij}}$$

where n_{ij} and e_{ij} are the observed and expected frequencies, respectively, for cell ij . The expected frequency is calculated as follows: $e_{ij} = (n_{i.} n_{.j})/n$. This statistic is distributed as a chi-square random variable with $(r-1)(c-1)$ degrees of freedom under the null hypothesis of independence. Failure to reject the null hypothesis implies that the model's predictions are independent of the actual outcomes. Rejecting the null hypothesis provides a basis for measuring the degree of association between the predictions and actual outcomes.

When $e_{ij} \leq 5$, the chi-square test above may not be valid, and test results are only approximations. Fisher's exact test can be used to confirm the suggested significance from the chi-square approximation. To perform Fisher's exact test, all possible tables of nonnegative integers are computed consistent with fixed $n_{i.}$ and $n_{.j}$. For each table, a hypergeometric probability (p-value) is computed using the following formula:

$$P = \frac{(\prod_{i=1}^r n_{i.}!) (\prod_{j=1}^c n_{.j}!)}{n! \prod_{ij} n_{ij}!}$$

The p-value for the observed table, P_{crit} , is also computed. The p-value of the test is the sum of all p-values less than or equal to P_{crit} , and if this sum is less than or equal to a chosen significance level, a significant association between the rows and columns exists. Unfortunately, the test is not computationally practical when $n/(r-1)(c-1) > 5$ (SAS Institute, 1990, pp.333-34). In this study, this fraction is $593/49 \approx 12$ using the smaller validation sample, implying Fisher's exact test is infeasible.

Association measures for enrollment duration are straightforward because of the ordinal nature between actual and predicted durations. Pearson's correlation coefficient is used to evaluate the degree of association between predicted and actual enrollment durations. This statistic is computed as follows:

$$r = \frac{\sum_i \sum_j n_{ij} (r_i - \bar{r})(c_j - \bar{c})}{\sqrt{\sum_i \sum_j n_{ij} (r_i - \bar{r}) \sum_i \sum_j n_{ij} (c_j - \bar{c})}}$$

It has range $-1 \leq r \leq 1$. When comparing model predictions, the model with the highest correlation coefficient is presumed to be the best performer of those being compared.

A more restrictive measure of association is the overall hit rate: $\sum_{i=j} n_{ij}/n$. It measures the degree of exact agreement between the predicted and actual outcomes. High hit rates can be a misleading indicator of overall fit if there is an especially high concentration of hits in a particular diagonal cell.

Pearson's correlation coefficient is not an appropriate measure of association for the destination models (Stokes, Davis, & Koch, 1995, p.95). This is because the rows and columns of these tables are not ordinal. The measure used instead is a conditional entropy coefficient. The entropy coefficient measures the uncertainty (entropy) of the actual outcomes explained by the predictions. Its range is $0 \leq U_{C|R} \leq 1$ and is

computed as follows:

$$U_{C|R} = \frac{-\sum_i (n_{i.}/n) \ln (n_{i.}/n) - \sum_j (n_{.j}/n) \ln (n_{.j}/n) + \sum_i \sum_j (n_{ij}/n) \ln (n_{ij}/n)}{-\sum_j (n_{.j}/n) \ln (n_{.j}/n)}$$

The model with the higher entropy coefficient is presumed to predict better.

For a general discussion of the the tests and associations measures discussed above, see Stokes et al. (1995) and SAS Institute (1990).

3.3 Chapter Summary

This chapter began by describing the major hypotheses to be addressed in the analysis. Four hypotheses were presented relating to the longitudinal character of student enrollment, as well as to the relative merits of using statistical duration methods to model persistence. Theoretical considerations suggest that the effect of academic integration is nonlinear; that is, low levels of academic integration (measured here by relative rank) are associated with a low likelihood of persistence. As relative rank increases, the likelihood of persistence increases, but after a point, higher relative rank decreases this likelihood. Distinguishing between system dropout and transfer is important, and it is hypothesized that the behavioral characteristics of each are different. Related to student-faculty interaction and academic integration, two variables are considered to be inversely related to persistence: the average class size in a student's portfolio and the proportion of the portfolio being taught by graduate teaching assistants. Two measures that show higher goal commitment and should be directly related to persistence are the student's willingness to take a heavier course load and the willingness to enroll in summer courses.

Many characteristics that are theoretically important to predicting student persistence are unobservable. Social integration, goal commitments, and others all reflect to a certain degree the students tastes and preferences, and given that these tastes

and preferences vary considerably from student to student, it is hypothesized that this unmeasured heterogeneity is the primary reason for the observed pattern of student departure. In other words, students self-sort. The declining dropout rates with enrollment tenure reflect this self-selection process, and not that continued enrollment profoundly changes preferences in favor of persistence. Finally, the model used in this study is compared to competing models in terms of out-of-sample predictive accuracy. This type of validation has not been the norm in previous studies. It is hypothesized that the model developed here will outperform the comparison models in out-of-sample predictive accuracy.

The statistical approach was also described in this chapter. A general approach to modeling duration data using hazard functions was described and various properties were illustrated. The method of incorporating covariates into the analysis was illustrated and linkages between linear models and hazard models were made. It was shown how the effect of unmeasured heterogeneity influences the hazard, and how, given assumptions about the hazard and unmeasured heterogeneity, a parametric model of duration could be specified and estimated using maximum likelihood. Under certain conditions, these techniques can be used to estimate cause- or destination-specific hazard models (also called transition intensities).

CHAPTER 4

DESCRIPTION OF THE DATA

The data in this study consisted of a cohort of new freshmen beginning enrollment in the Fall 1993 semester at Oklahoma State University (OSU). There were 2,188 new freshmen officially recorded as enrolled at this time. Of these, 98 were excluded from analysis: 14 were erroneously included with the freshman class, 8 never attended their first semester, 62 were missing background data, and 14 were actually transfer students. Thus there were 2,090 true new freshmen available for analysis and the summary statistics for these students are presented in this chapter. For modeling purposes, an additional 258 students were excluded because of an academic suspension in the subsequent semesters. This was done to be consistent with the existing methods of studying *voluntary attrition*. Of the remaining 1,832 students, 799 (43.6 percent) voluntarily departed by the Fall 1996 semester. Various background characteristics as well as regular semester course and performance data were used to estimate the models. Background information included the student's high school rank as a percentage of the graduating class size, the composite ACT score, the student's sex and ethnicity, and the residency status of the student. Longitudinal data included the relative rank of the student within the current portfolio of courses taken, the average number of students in the portfolio, the proportion of instruction conducted by graduate student teaching assistants, financial aid, summer enrollment, problematic enrollment (i.e., academic notice or probation), and preprofessional and engineering

major indicators. The dependent variables measure the number of semesters of continuous enrollment and whether the student, immediately upon departure, transfers to another institution or drops out of the system. These data are randomly assigned to two samples: two-thirds of the data are used to estimate the models and one-third are used to provide out-of-sample predictive validation.

There are two primary data sources: Student Records maintained at OSU, and the Unitized Data System (UDS) maintained by the Oklahoma State Regents for Higher Education. A third source, School District data obtained from the National Center for Educational Statistics, was initially considered but ultimately not used. The following sections describe the data available from each source. Obviously, not all data are directly used in the analysis. Many of the variables are included for matching purpose or to facilitate the creation of analysis variables. A final list of analysis variables is provided in the chapter summary.

4.1 Student Records

Student records are maintained and updated by OSU on a per-semester basis, and the fall semester marks the beginning of a new school year. For analysis purposes, any summer enrollment activity is usually combined with the results of the following fall semester, so that fall and spring semesters are the tracking units. This is done primarily because summer enrollment is not required to maintain “continuing student” status for financial aid and reporting purposes.

Student records data for this study were extracted from several sources and three files were created: the student demographics, course data, and student retention files. The student demographics file elements are described in Table 4.1.

Table 4.1: Variable Names and Descriptions from Oklahoma State University Student Demographics File

Variable	Description
ID	Student's OSU identification number*
LNAME	Student's last name
FNAME	Student's first name
SEX	Student's sex (M or F)*
RESIDENT	Student's residency status (In- or Out-of-State)*
CLASS	Classification code (Freshman, Sophomore, etc.)*
BIRTHMON	Student's birth month (MM)*
BIRTHDAY	Student's birth day (DD)*
BIRTHYR	Student's birth year (YY)*
MARITAL	Marital status indicator*
STARTOSU	Year and Semester started OSU*
DORM	Dorm code*
HSCODE	High school code*
HSGPA	High school grade point average (4.0 scale)*
HSRANK	High school rank*
CLASSIZE	High school graduating class size*
HSGRAD	Indicator of high school graduation*
HSENGL	OSRHE high school english units requirement
HSMATH	OSRHE high school mathematics units requirement
HSHIST	OSRHE high school history units requirement
HSSCI	OSRHE high school science units requirement
HSOTHER	OSRHE other high school units requirements
ENGLSTAT	Indicator: met OSRHE english requirement
MATHSTAT	Indicator: met OSRHE mathematics requirement
HISTSTAT	Indicator: met OSRHE history requirement
SCISTAT	Indicator: met OSRHE science requirement
NATMERIT	Indicator of being a national merit scholar
ACTENGL	ACT English sub-score*
ACTMATH	ACT Mathematics sub-score*
ACTREAD	ACT Reading sub-score*
ACTSCI	ACT Science sub-score*
ACTCOMP	ACT composite score*
SATVERB	SAT Verbal score*
SATQUANT	SAT Quantitative score*
ETHNIC	Ethnicity code*
MAJCODE	Major code*
COLLEGE	College enrolled in*
CURHRS	Current semester hours attempted*
ACCUMGPA	Accumulated grade point average (4.0 scale)*

Note: An asterisk * indicates the element was extracted for use in this study

continued on next page

Table 4.1 *continued from previous page*

Variable	Description
ACCUMHRS	Accumulated hours*
ACADSTAT	Academic status code
SEMHOURLS	Current hours completed*
SEMGRDPT	Current grade points earned*
HRSPASS	Hours earned "pass"*
HRSI	Hours earned incomplete*
TRNCRHR	Hours transferred to OSU*
TRNGPTS	Grade points transferred to OSU*
TRNHRSP	Hours transferred "pass"*
TRNHRSF	Hours transferred "fail"*
LASTCOL	College code of last college attended
EXPGRAD	Expected semester and year of graduation
WITHDRDA	Formal withdraw date
WDREASON	Reason code for formal withdrawal
OSUEMP	Indicator of OSU student employee
FINAID	Indicator of financial aid recipient*
ATHTYPE	Athlete type code
ATHSCH	Athletic scholarship indicator*
STREET	Current street address
CITY	Current city
STATE	Current state
ZIP	Current zip code
PERMSTR	Permanent street address
PERMCITY	Permanent city
PERMSTAT	Permanent state
PERMZIP	Permanent zip code

Note: An asterisk * indicates the element was extracted for use in this study

With the exception of student ID, all other variables that could uniquely identify the student were excluded. The reason for retaining ID was to provide a match criterion for other data sources. Once matching was complete and the data sets constructed, ID was removed. The remaining excluded variables were excluded either because they lacked variation (e.g., the high school units variables) or because they were insufficiently populated (e.g., withdraw reason and date).

Missing data for key variables such as ACT scores and high school performance posed a significant problem. Approximately 15 percent (275 observations) had one or more missing values per student for these variables. However, when these missing val-

ues were compared with the variables that had no missing values (i.e., ethnicity, sex, and residency status) no particular pattern emerged. Rather than drop the observations, a number of methods were examined to replace the missing values. A large sample of 21,532 entering freshmen between the fall 1990 and spring 1997 semesters was constructed. The data were then segmented according to ethnicity, sex, and residency status, and the segment means were used to replace missing values in the current sample. This is essentially the same method suggested in Greene (1993, pp.276-7). The advantage of using this method over simply replacing the blanks with sample means is that information in the form of covariation between the regressors is used to estimate the missing values. The primary reason for using this approach here was to preserve as many degrees of freedom as possible, so that ultimately the data could be split into relatively large estimation and validation samples. The means used for substitution are presented in Table 4.2. The before- and after-substitution means and standard deviations are presented in Table 4.3. In Table 4.3 the variable RANKPCTL is computed as follows: $RANKPCTL = (CLASSIZE - HSRANK) / CLASSIZE$.

Table 4.2: Missing Value Replacements for ACT and High School Performance Data. Fall 1990 to Spring 1997 Semesters.

Category	Female		Male	
	Mean	N	Mean	N
<u>In-State Asian</u>				
ACTENGL	22.21	132	21.23	151
ACTMATH	22.31	132	24.68	152
ACTREAD	23.23	121	22.82	146
ACTSCI	21.41	121	22.84	146
CLASSIZE	327.04	126	331.90	145
HSGPA	3.53	120	3.32	141
HSRANK	62.85	126	85.88	145
<u>Out-of-State Asian</u>				
ACTENGL	22.67	9	22.77	13
ACTMATH	21.22	9	23.54	13
ACTREAD	24.11	9	24.33	12

continued on next page

Table 4.2 continued from previous page

Category	Female		Male	
	Mean	N	Mean	N
ACTSCI	23.22	9	24.17	12
CLASSIZE	425.29	7	270.62	13
HSGPA	3.37	5	3.28	15
HSRANK	86.29	7	80.69	13
<u>In-State Black</u>				
ACTENGL	20.20	244	19.32	217
ACTMATH	18.71	243	19.51	218
ACTREAD	20.95	232	20.86	210
ACTSCI	19.19	232	20.75	210
CLASSIZE	266.85	241	221.94	216
HSGPA	3.20	236	3.02	217
HSRANK	75.53	241	83.80	215
<u>Out-of-State Black</u>				
ACTENGL	20.41	27	17.68	53
ACTMATH	20.00	26	18.64	53
ACTREAD	21.75	24	19.10	50
ACTSCI	19.71	24	18.86	50
CLASSIZE	315.64	39	321.82	76
HSGPA	3.15	32	2.61	66
HSRANK	91.26	39	164.43	76
<u>In-State Hispanic</u>				
ACTENGL	22.86	130	21.74	111
ACTMATH	20.82	130	22.12	111
ACTREAD	24.52	125	23.40	104
ACTSCI	21.67	125	23.66	104
CLASSIZE	328.27	128	250.17	104
HSGPA	3.39	123	3.19	105
HSRANK	71.12	129	87.53	104
<u>Out-of-State Hispanic</u>				
ACTENGL	22.63	8	19.25	24
ACTMATH	20.25	8	21.42	24
ACTREAD	22.71	7	20.92	24
ACTSCI	20.29	7	21.75	24
CLASSIZE	389.33	12	365.61	28
HSGPA	3.27	10	2.90	24
HSRANK	78.33	12	138.64	28
<u>In-State Native American</u>				
ACTENGL	22.66	696	21.81	583
ACTMATH	20.25	693	22.13	583
ACTREAD	24.00	664	24.00	582
ACTSCI	21.94	664	23.76	547

continued on next page

Table 4.2 continued from previous page

Category	Female		Male	
	Mean	N	Mean	N
CLASSIZE	205.43	661	214.36	547
HSGPA	3.42	655	3.23	555
HSRANK	54.97	661	73.52	534
<u>Out-of-State Native American</u>				
ACTENGL	21.86	28	21.50	22
ACTMATH	20.61	28	24.23	22
ACTREAD	23.52	27	22.60	20
ACTSCI	22.11	27	24.10	20
CLASSIZE	319.35	31	329.66	29
HSGPA	3.33	27	3.11	26
HSRANK	93.71	31	117.75	28
<u>In-State Other</u>				
ACTENGL	23.46	7034	22.70	6616
ACTMATH	21.28	7000	22.88	6602
ACTREAD	24.26	6544	24.49	6035
ACTSCI	22.36	6546	24.29	6035
CLASSIZE	269.82	6702	270.66	6206
HSGPA	3.42	6520	3.24	6028
HSRANK	69.08	6702	89.13	6203
<u>Out-of-State Other</u>				
ACTENGL	23.85	671	23.09	649
ACTMATH	22.39	671	23.71	648
ACTREAD	25.36	652	24.91	619
ACTSCI	23.33	652	24.44	619
CLASSIZE	295.36	718	298.52	822
HSGPA	3.40	634	3.17	704
HSRANK	74.41	716	102.58	821
<u>Non Resident Alien</u>				
ACTENGL	21.33	9	19.65	17
ACTMATH	20.33	9	24.88	17
ACTREAD	19.67	9	21.53	17
ACTSCI	20.44	9	23.82	17
CLASSIZE	331.80	5	275.00	3
HSGPA	3.32	5	2.97	3
HSRANK	73.00	5	108.33	3

Table 4.3: Before and After Comparison of Missing Value Imputation.

Variable	Before		After	
	Mean	St Dev	Mean	St Dev
ACTENGL	23.37	4.37	23.32	4.20
ACTMATH	22.71	4.33	22.69	4.17
ACTREAD	25.10	5.46	25.00	5.22
ACTSCI	23.62	4.28	23.56	4.09
HSGPA	3.08	1.14	3.41	0.50
RANKPCTL	0.76	0.21	0.75	0.20

Note that the mean of HSGPA demonstrated the most dramatic change. This is because they were originally coded as zero if missing. Note also that the variability of HSGPA decreased considerably. This may have implications for obtaining stable parameter estimates when estimating the models.

Like the student demographic file, the course data file is maintained by OSU on a per-semester basis. Each student in the file will have multiple observations, one for each course enrolled in. The course data files contains data on all enrolled students so it was necessary to match the fall 1993 cohort IDs against the IDs in the course data files to obtain the relevant records. This was done for the fall and spring semesters from fall 1993 to fall 1996. The file elements are described in Table 4.4.

Table 4.4: Variable Names and Descriptions from The Course Data File

Variable	Description
ID	Student's OSU ID number*
PREFIX	Course prefix (e.g., ECON)*
NUMBER	Course number (e.g., 2013)*
SECTION	Course section
GRADE	Final letter grade earned in course*
NUMSTUDS	Number of students enrolled in course*
NUMA	Number of As granted in course*
NUMB	Number of Bs granted in course*
NUMC	Number of Cs granted in course*

Note: An asterisk * indicates the element was extracted for use in this study
continued on next page

Table 4.4 *continued from previous page*

Variable	Description
NUMD	Number of Ds granted in course*
NUMF	Number of Fs granted in course*
NUMI	Number of incompletes granted in course*
NUMP	Number of passes granted in course*
NUMW	Number of withdraws in courses*
NUMWF	Number of withdraw-failing in course*
INSTNAME	Instructor's name
INSTETH	Instructor's ethnicity code
INSTGEND	Instructor's sex code (M or F)
TEACHER	Instructor's title*

Note: An asterisk * indicates the element was extracted for use in this study

Variables were excluded either to protect privacy, such as instructor name, or because they were not likely to be useful in the analysis. A potentially important variable, instructor's ethnicity, was excluded because upon examination, 85 percent of the responses were reported as the catchall category "other".

The student retention file is simply a per-semester tracking file where new fall semester freshman and transfer cohorts are followed longitudinally. Several of the elements previously described are used to populate the fields in this file. The elements of the new freshman file are described in Table 4.5.

Table 4.5: Variable Names and Descriptions from The New Freshman Student Retention Data File

Variable	Description
ID	Student's OSU ID*
NAME	Student's full name
STARTOSU	Starting year and semester of the student*
ACTENGL	ACT English sub-score*
ACTMATH	ACT Mathematics sub-score*
ACTREAD	ACT Reading sub-score*
ACTSCI	ACT Science sub-score*
HSGPA	High school grade point average*
CLASS	Student classification*
SEX	Sex code (M or F)*

Note: An asterisk * indicates the element was extracted for use in this study
continued on next page

Table 4.5 continued from previous page

Variable	Description
ETHNIC	Ethnicity code*
MAJCODE	Major code*
ACCHRS	Accumulated hours*
ACCGPA	Accumulated grade point average*
ACADSTAT	Enrollment status*
DEGREE	Degree earned*
YRGRANT	Semester and year granted*

Note: An asterisk * indicates the element was extracted for use in this study

These three files form the basis of the data used in the analysis.

4.2 Unitized Data System

The second primary data source is the Unitized Data System (UDS), which is maintained by the Oklahoma State Regents for Higher Education (OSRHE). All institutions in the Oklahoma state system of higher education are required to submit student-level and faculty- and staff-level data each semester to OSRHE in a specific format. This format forms the file layout that OSRHE ultimately constructs. The UDS provides a longitudinal picture of the performance and movement of students, faculty, and staff within Oklahoma's higher education system. Of primary interest in this study is the tracking of student enrollment between institutions. OSRHE used the fall 1993 cohort IDs to construct a longitudinal data set for the fall 1993 to fall 1996 semesters. The elements of this data set are described in Table 4.6.

Table 4.6: Variable Names and Descriptions from The Unitized Data System File

Variable	Description
ID	Student's ID
INST	Institution code*
LASTCOL	Institution FICE code of last college attended
ENRACT	Enrollment status code*

Note: An asterisk * indicates the element was extracted for use in this study
continued on next page

Table 4.6 *continued from previous page*

Variable	Description
CLASS	Student classification
WDRAW	Formal withdrawal indicator
CHRS	Current hours attempted
RGPA	Retention grade point average (4.0 scale)
EDGOAL	Immediate educational goal
HIDEG	Highest college degree/certificate earned
PGMCODE	Current instructional program code
DEG1	First degree awarded code
DEG2	Second degree awarded code
PGM1	Instructional program code for degree 1
PGM2	Instructional program code for degree 2
IOR	Institution of record*
FINAID1	Financial aid code for grants
FINAID2	Financial aid code for loans
FINAID3	Financial aid code for scholarships
FINAID4	Financial aid code for student employment
FINAID5	Financial aid code for other support

Note: An asterisk * indicates the element was extracted for use in this study

Examining the content of these data raised serious questions about their integrity. For example, it was discovered that codes unique to FINAID4 were being used to populate the other FINAID variables. For students who remained enrolled at OSU throughout the analysis period, there were discrepancies between RGPA and current hours in the UDS and those maintained in OSU student records. Because of these considerations, the UDS data was only used to determine if a student transferred, given that OSU student records indicated a termination of enrollment.

4.3 School District Data

Primary and secondary school district data was obtained from the National Center for Educational Statistics web-site. The data were compiled for all states in 1989 and are available in both summary and detail form. Except for major metropolitan areas, Oklahoma school districts are closely tied to the counties in which they reside. The original intent for this data was to provide proxies for previous educational resources

and the demographics of the area where the student attended high school. The elements initially considered are presented in Table 4.7.

Table 4.7: Variable Names and Descriptions from Oklahoma School District Data

Variable	Description
ZIPCODE	School district zip code
SCHOOL	High school name
HSCODE	High school code
DISTPOP	School district population
POVRATE	Poverty rate for the district
VALHOME	Median home value in the district
INCOME	Median family income in the district
STRATIO	District student-teacher ratio
EXPSTUD	Total expenditure per student
DISTRICT	School district name

A distinction was also made as to whether the school was private, public, or magnet. Some of the high schools were too new to be included in the 1989 district data. Zip codes were used to supply values for these fields, and in instances where a match could not be established, the Oklahoma average values were used. High school codes of “999999” were given the United States average values.

4.4 Variable Descriptions

To prepare the data for analysis, both the dependent and independent variables were created and ultimately arranged into a longitudinal data set. Each are described in turn.

4.4.1 Dependent Variables

Time enrolled used the number of consecutive fall and spring semesters completed. The variable TIME was created to indicate the total number of semesters completed.

A break in enrollment was identified based on the OSU retention file data, and the UDS data were used to determine if the student enrolled in another institution in the Oklahoma State system. If so, a transfer dummy variable (TRANSFER) was created, and was set to 1 at the time of transfer and zero otherwise. If transfer could not be identified, a dropout dummy variable (DROPOUT) was created, and set to 1 at the time of dropout and zero otherwise. If the student dropped out for at least one semester, did not attend anywhere else, and ultimately resumed enrollment, the indicator variable STOPOUT was created and accordingly assigned a value of one. If the student graduated, an indicator variable GRADUATE was created. Finally, if the student remained continuously enrolled through fall 1996, a censoring or end-of-sampling-period indicator (CENSOR) was created. Summary statistics of these variables are presented in Table 4.8.

Table 4.8: Summary Statistics of Persistence Related Variables. N=2,090.

Variable	Mean	Description
CENSOR	.50	Censoring indicator (0 if censored, 1 otherwise)
DROPOUT	.36	Dropout indicator (1 if dropout, 0 otherwise)
GRADUATE	.01	Graduation indicator (1 if graduated, 0 otherwise)
STOPOUT	.07	Stopout indicator (1 if stopout, 0 otherwise)
TIME	4.95	Enrollment duration (1, 2, ..., 7 semesters. SD=2.39)
TRANSFER	.06	Transfer indicator (1 if transfer, 0 otherwise)

The proper treatment of stopout is as a renewal process and is beyond the scope of this study. Therefore, stopout and dropout are treated equivalently. Also, the handful (29) of students who graduated did so at the end of the sampling period and were also enrolled at that time. For these individuals, the censoring and graduate indicators were treated equivalently. Table 4.9 ranks the top destination schools for students transferring from OSU.

Table 4.9: Top Destination Schools for OSU Transfers.

Rank	School
1	University of Central Oklahoma
2	Tulsa Junior College
3	University of Oklahoma
4	OSU Technical Branch - Oklahoma City
5	East Central University
5	Oklahoma City Community College
6	Langston University
7	Northern Oklahoma College
8	Cameron University
8	Mid-American Bible College
9	Northwestern Oklahoma State University
9	Southeastern Oklahoma State University
9	Rose State College
9	Southern Nazarine University
10	Oklahoma Panhandle State University
10	Eastern Oklahoma State College
10	Northeastern Oklahoma A&M College
10	Western Oklahoma State College
10	Seminole Junior College
10	Oral Roberts University

In the present sample, the top three schools in Table 4.9 account for over 61 percent of the transfers between the fall 1993 and fall 1996 semesters.

4.4.2 Independent Variables

A number of potential independent variables were considered for the analysis. Some of the variables remain constant over the student's enrollment while others vary while the student is enrolled. Table 4.10 provides summary statistics and descriptions of the variables that remain constant and Table 4.11 does the same for the time-varying independent variables.

Table 4.10: Summary Statistics of Time-Constant Independent Variables

Variable	Mean	Std Dev	Description
ACTENGL	23.11	4.21	ACT English score
ACTMATH	22.49	4.17	ACT Math score
ACTREAD	24.81	5.20	ACT Reading score
ACTSCI	23.43	4.08	ACT Science score
ALIEN	.01	.07	Foreign student indicator (0,1)
ASIAN	.02	.13	Asian student indicator (0,1)
BLACK	.03	.16	Black student indicator (0,1)
CLASSIZE	259.90	213.3	Graduating class size in high school
DISTPOP	54.19	83.83	HS district population (00)
DROPOUT	.36	.48	Dropout indicator (0,1)
EXPSTUD	3.69	.84	Expenditure per student (000)
HISP	.02	.13	Hispanic student indicator (0,1)
HSGPA	3.35	.53	HS grade point average (4.0 Scale)
HSRANK	75.53	96.74	HS graduating rank
INCOME	26.20	6.49	HS district median income (000)
NATAM	.08	.27	Native American indicator (0,1)
OTHER	.85	.35	White student indicator (0,1)
POVRATE	14.36	5.75	HS district poverty rate
PRIVATE	.05	.21	Private/Magnet school indicator (0,1)
RANKPCTL	.73	.21	HS rank relative to HS class size
RESCODE	.13	.34	Non-resident indicator (0,1)
SEXCODE	.51	.50	Female student indicator
STRATIO	17.08	2.28	HS district student-teacher ratio
VALHOME	54.50	19.06	HS district median home value (000)

An average ACT score, ACTCOMP, was computed for each student. Also, a combined ethnic indicator, NONWHITE, was computed by summing the ASIAN, BLACK, HISP, NATAM, and ALIEN indicators. The variable RANKPCTL is computed as $(CLASSIZE - HSRANK)/CLASSIZE$. Values approaching unity indicate a top high school graduate.

Table 4.11: Summary Statistics of Time-Varying Independent Variables

Variable	Mean	Std Deviation	Description
CHF93	13.62	2.76	Current hours attempted
CHS94	14.14	2.59	
CHF94	14.27	2.01	
CHS95	14.16	2.12	
CHF95	14.19	2.04	
CHS96	14.26	2.20	
CHF96	14.16	2.39	
CHS97	14.01	2.58	
CRANKF93	1.02	.37	
CRANKS94	1.01	.38	
CRANKF94	1.03	.37	
CRANKS95	1.04	.34	
CRANKF95	1.04	.33	
CRANKS96	1.06	.29	
CRANKF96	1.02	.30	
CRANKS97	1.02	.25	
F93GPA	2.63	.99	Current grade point average
S94GPA	2.61	1.00	
F94GPA	2.66	.96	
S95GPA	2.84	.76	
F95GPA	2.81	.90	
S96GPA	2.94	.82	
F96GPA	2.93	.87	
S97GPA	3.05	.80	
FAIDF93	.43	.50	
FAIDS94	.44	.50	
FAIDF94	.26	.44	
FAIDS95	.27	.44	
FAIDF95	.47	.50	
FAIDS96	.49	.50	
FAIDF96	.49	.50	
FAIDS97	.51	.50	
NSTUDF93	69.10	28.50	Average class size
NSTUDS94	68.31	25.64	
NSTUDF94	83.41	36.74	
NSTUDS95	75.34	33.07	
NSTUDF95	74.11	35.62	
NSTUDS96	61.32	29.29	
NSTUDF96	53.63	29.00	
NSTUDS97	48.89	25.81	

continued on next page

Table 4.11 continued from previous page

Variable	Mean	Std Deviation	Description
PCTGF93	.39	.23	Proportion graduate student TAs
PCTGS94	.45	.27	
PCTGF94	.30	.22	
PCTGS95	.28	.22	
PCTGF95	.20	.21	
PCTGS96	.19	.21	
PCTGF96	.16	.19	
PCTGS97	.13	.18	
PROBF93	.15	.35	Problematic enrollment indicator (0,1)
PROBS94	.06	.23	
PROBF94	.05	.22	
PROBS95	.03	.18	
PROBF95	.02	.13	
PROBS96	.01	.11	
PROBF96	.01	.11	
PROBS97	.02	.12	

To calculate the class rank variables (CRANK) the student's grade average was computed for the courses completed. This is different from the standard grade point average in at least two respects. First only the grade earned is considered with the standard coding of A=4, B=3, C=2, D=1, and F=0. The number of credits earned was not factored in. Second, if the course was "pass/fail", pass was assigned 2 and fail was assigned 0. A similar grade average was computed for each course a student was enrolled in. Their individual grade average is divided by the course average, and this ratio is averaged across the portfolio of courses taken that semester. The problematic enrollment indicator variables (PROB) assumes values of one when a student is either put on academic notice, probation, or suspension (notice or probation in the modeling data). This condition is evaluated for each semester. To assess the impact of graduate student teaching on student persistence, the fraction of the student's portfolio taught by a graduate student was computed. This fraction is calculated for each semester.

Three other variables not presented in the list were the summer enrollment, engineering, and pre-professional indicators, SUMMER, ENGINEER, and PREPROF

respectively. The summer indicator assumed a value of one for the fall semester if a student was enrolled in summer courses prior to that fall semester. The engineer indicator assumes a value of one for each semester a student claims an engineering type major (e.g, electrical engineering). Likewise, the pre-professional indicator assumes a value of one for each semester a student claims a pre-professional major (e.g., pre-law). These were determined from the official list of major codes provided by the OSU Office of Admissions.

The full list of independent variables considered for analysis is as follows. Time-constant variables include: ACTCOMP, ALIEN, ASIAN, BLACK, DISTPOP, EXPSTUD, HISP, HSGPA, INCOME, NATAM, POVRATE, PRIVATE, RANKPCTL, RESCODE, SECCODE, STRATIO, and VALHOME. Time-varying covariates include: CHRS (time series of CHF93 - CHS97), RELRANK (time series of CRANKF93 - CRANKS97), RELRNK2 (RELANK squared), CURGPA (time series of F93GPA - S97GPA), FINAID (time series of FAIDF93 - FAIDS97), NSTUDNT (time series of NSTUDF93 - NSTUDS97), PCTGRAD (time series of PCTGF93 - PCTGS97), PROBENR (time series of PROBF93 - PROBS97), SUMMER, ENGINEER, and PREPROF. The Dependent variables include DURATION (cumulative enrollment duration), DROPOUT, and TRANSFER.

4.5 Data Reduction Methods

Because the models in this study are nonlinear, it is important to determine the degree of multicollinearity in the data. Highly collinear data may pose convergence problems for nonlinear optimization routines because the parameters are unstable and affect the precision with which parameters of the model can be estimated. Indeed, the model would not converge using the full list of regressors. As an initial step, a linear model is used where the dependent variable is the log of enrollment duration. This is equivalent to an accelerated lifetime regression in the absence of censoring and

ordinary least squares could be used. For the moment, censoring is ignored. With a reduced specification that still allows the major hypotheses to be analyzed, the hazard models are re-estimated. A final collinearity analysis is conducted using the Hessian based on the MLEs.

4.5.1 Multicollinearity Diagnostics

Three basic diagnostic tools are used to detect multicollinearity: the Variance Inflation Factors (VIFs), the Condition Index (CI), and Variance Proportions (VPs). VIFs are essentially the multiple by which the variance of the corresponding estimates are increased, the increase being attributable to multicollinearity. A rule of thumb is to consider VIFs exceeding 2 to 5 as indicating serious multicollinearity. The CI is the square root of the ratio of the largest to smallest eigenvalue in scaled $(\mathbf{X}'\mathbf{X})$. The rule of thumb for the CI is that severe multicollinearity exists for CIs greater than 30. Finally VPs measure the proportion of variance associated with the each eigenvalue in the scaled $(\mathbf{X}'\mathbf{X})$. Combinations of variables with high VPs for a small eigenvalue (large CI) indicate near linear dependencies between those variables. The VIFs are presented in Table 4.12.

Table 4.12: Variance Inflation Factors of the Independent Variables

Variable	VIF
RELRANK	22.1167
RELRNK2	12.4388
INCOME	9.9645
CURGPA	8.5613
VALHOME	5.6509
POVRATE	5.2776
HSGPA	4.4948
RANKPCTL	4.3400
STRATIO	3.4987
EXPSTUD	2.8082
RESCODE	2.6307

continued on next page

Table 4.12 *continued from previous page*

Variable	VIF
DISTPOP	2.4454
PRIVATE	1.4716
ACTCOMP	1.3625
PROBENR	1.3421
NSTUDNT	1.1868
ENGINEER	1.1628
CURHRS	1.1588
SEXCODE	1.1564
PCTGRAD	1.1556
FINAID	1.0975
ALIEN	1.0656
NATAM	1.0616
BLACK	1.0583
PREPROF	1.0578
HISP	1.0328
ASIAN	1.0300
SUMMER	1.0293
INTERCEP	1.0000

From Table 4.12 the following parameters are potentially affected by multicollinearity: RELRANK, RELRNK2, INCOME, CURGPA, VALHOME, POVRATE, HSGPA, RANKPCTL, STRATIO EXPSTUD, RESCODE, and DISTPOP. The CIs and VPs for selected regressors are presented in Table 4.13.

Table 4.13: Variance Proportions for Selected Regressors with Condition Number Values over 30.

Variable	Condition Number					Sum
	35.90723	56.06765	57.94555	61.2944	114.18853	
INTERCEP	0.0148	0.0003	0.0000	0.0009	0.9819	0.9979
POVRATE	0.0001	0.1531	0.2637	0.0509	0.2975	0.7653
VALHOME	0.3461	0.0021	0.0570	0.0989	0.2135	0.7176
INCOME	0.0701	0.2004	0.3565	0.0836	0.2827	0.9933
STRATIO	0.0862	0.3258	0.1242	0.0129	0.4278	0.9769
EXPSTUD	0.0133	0.1167	0.0253	0.0262	0.2880	0.4695
HSGPA	0.0103	0.0455	0.0097	0.7803	0.1427	0.9885
RANKPCTL	0.0212	0.0400	0.0013	0.5884	0.0794	0.7303
RELRANK	0.0037	0.3520	0.5395	0.0921	0.0033	0.9906
RELRNK2	0.0144	0.2347	0.3282	0.0641	0.0122	0.6536
CURGPA	0.0715	0.1671	0.2740	0.0281	0.0021	0.5428

The immediate conclusion from Table 4.13 is that severe multicollinearity exists. The six smallest eigenvalues produce condition numbers from 35.91 to a condition index of 114.19. The VPs in Table 4.13 suggest several near linear dependencies. This is shown by examining the sum of the VPs for high condition numbers, and to determine which variables are involved, a VP of 0.45 or greater is used. Because the intercept is involved, the linear combination of these variables exhibits little variation.

Because the linear combination of the district data appears to have little variation, each variable has a high VIF. Since the literature suggests that these variables are, at best, of secondary importance, they are excluded from the analysis. RELRANK and RELRNK2 are central to a major hypothesis to be tested, so they are retained. CURGPA is excluded because of the near linear relationship to RELRANK and RELRNK2. These relationships are in Table 4.13. Because it is part of Tinto's specification that prior school experience be included, RANKPCTL is retained. This decision is based on the knowledge that to graduate in the top of one's class, he or she must necessarily have a high grade point average. Furthermore, RANKPCTL deflates one's ordinal class rank by the graduating class size. Therefore, RANKPCTL

contains more information than HSGPA. Finally, Tinto's framework specifies that prior skills are important to persistence.

Considerations other than multicollinearity were important in determining what to include and exclude from the analysis. The only other variables explicitly excluded from the analysis were PRIVATE and CURHRS. PRIVATE is an indicator variable of whether a student attended a private or public high school. It was excluded because of low occurrence (only 98 of 2090 students attended private high schools). CURHRS was excluded because it is used in the definition of the dependent variable and is likely to be jointly determined with persistence; hence, including it could introduce simultaneity bias. The individual ethnicity indicators (ASIAN, BLACK, HISP, NATAM, and ALIEN) were excluded, and in their place, a non-white/white indicator (NONWHITE) was used. This variable is merely the sum of the ethnicity indicators. Also included SUMMER which relates to the commitment hypotheses (Hypothesis 5) to be tested. NSTUDNT and PCTGRAD are included because they are central to Hypotheses 3 and 4. Other variables included primarily for control purposes are RESCODE, SEXCODE, FINAID, ENGINEER, PRE-PROF, and PROBENR. Thus, the final list of variables included for analysis are: ACTCOMP, ENGINEER, FINAID, NONWHITE, NSTUDNT, PCTGRAD, PRE-PROF, PROBENR, RANKPCTL, RELRANK, RELRNK2, RESCODE, SEXCODE, and SUMMER. The VIFs for the final list of modeling variables are presented in Table 4.14.

Table 4.14: Variance Inflation Factors for the Final List of Independent Variables

Variable	VIF
RELRANK	12.6257
RELRNK2	11.8919
RANKPCTL	1.4279
PROBENR	1.2881

continued on next page

Table 4.14 continued from previous page

Variable	VIF
ACTCOMP	1.2709
PCTGRAD	1.1320
ENGINEER	1.1202
SEXCODE	1.1121
NSTUDNT	1.0580
FINAID	1.0425
RESCODE	1.0409
PREPROF	1.0369
NONWHITE	1.0348
SUMMER	1.0190

The VIFs for RELRANK and RELRNK2 are still extremely high; however, this is because when one is taken as the dependent variable, and regressed on the other independent variables, the other is included as an independent variable. The R-Squares from these regressions are very high (approximately 0.93). This tends to drive up the VIFs for either variable. When each one is excluded from the regression, the VIFs for RELRANK and RELRNK2 are 1.27 and 1.25, respectively. Thus the high VIFs on these two variables reflect the fact that they are functionally related.

The specification in Table 4.14 converged and the scaled Hessian evaluated at the MLEs is used to detect collinearity in the nonlinear model. The Hessian is computed using the matrix of second partial derivatives of the log-likelihood function evaluated at the MLEs. In general, this is not proportional to the linear model Hessian, $(\mathbf{X}'\mathbf{X})$, and therefore the VPs do not necessarily indicate which variables are involved in the collinearity. The condition index still provides a measure of how ill-conditioned the Hessian is. These results are presented in Table 4.15.

Table 4.15: Variance Proportions Corresponding to the Two Largest Condition Numbers from the Scaled Hessian of the Hazard Model.

Variable	Condition Number		Sum VPROP
	27.2100	38.3306	
INTERCEP	0.1833	0.7639	0.9472
ACTCOMP	0.5989	0.2879	0.8868
ENGINEER	0.0408	0.0024	0.0432
FINAID	0.0001	0.0000	0.0001
NONWHITE	0.0080	0.0005	0.0084
NSTUDNT	0.0205	0.0050	0.0256
PCTGRAD	0.0279	0.0663	0.0942
PREPROF	0.0035	0.0001	0.0036
PROBENR	0.0000	0.0972	0.0972
RANKPCTL	0.0009	0.0000	0.0009
RELRANK	0.3028	0.6846	0.9874
RELRNK2	0.2376	0.6747	0.9123
RESCODE	0.0058	0.0008	0.0066
SEXCODE	0.0336	0.0020	0.0357
SUMMER	0.0061	0.0011	0.0072

The condition index is 38.33 which indicates that there is still a collinearity problem. This seemed evident in that the model needed 23 iterations to converge. Even though the VPs here do not necessarily indicate the variables involved in the collinearity, they suggest that a linear combination of ACTCOMP, RELRANK, and RELRNK2 are collinear with the intercept. Variables or combinations of variables collinear with the intercept suggest low variability. In light of what these variables are measuring, it seems reasonable to expect this relationship and that there may not be much independent variation in these variables. This may impact either the signs or statistical significance of these coefficients.

4.6 Chapter Summary

This chapter provides a detailed description of the data used in this analysis. All data sources are described, along with a discussion of the time-frames of sampling, variable

definitions, overlaps in the data, as well as limiting factors and problems encountered. A description of both the dependent and independent variables is provided along with summary statistics. Data reduction methods are explained, the intermediate results presented, and the final analysis list of variables presented.

CHAPTER 5

ANALYSIS AND RESULTS

This chapter presents the main results of the analysis and tests the hypotheses of the study. In the previous chapter, variable descriptions and summary statistics were presented. This chapter concentrates specifically on the model estimation results and predictive assessments as they related to the primary hypotheses. The following chapter summarizes and discusses these results in greater detail. Statistical software used included *LIMDEP* 7.0 for estimation and *SAS* 6.12 for validation.

5.1 Hypothesis Tests

To begin analyzing and testing the hypotheses, a single stage hazard model is estimated where no distinction is made between system dropout and transfer. The results are presented in Table 5.1.

Table 5.1: Parameter Estimates for Single-Stage Hazard Model - General Attrition. Dependent Variable: Log(DURATION).

Variable	Coefficient	Std Error	T-Stat	P-Value
CONSTANT	-0.8317	0.3732	2.2285	0.0258
ACTCOMP	-0.2264	0.1247	-1.8151	0.0695
ENGINEER	0.3704	0.1760	2.1052	0.0353
FINAID	-0.0112	0.0811	-0.1381	0.8902
NONWHITE	0.0484	0.1123	0.4313	0.6662
NSTUDNT	-0.1441	0.1251	-1.1523	0.2492

continued on next page

Table 5.1 continued from previous page

Variable	Coefficient	Std Error	T-Stat	P-Value
PCTGRAD	-0.8835	0.1507	-5.8644	0.0000
PREPROF	-0.5066	0.1721	-2.9440	0.0032
PROBENR	-0.4575	0.1437	-3.1838	0.0015
RANKPCTL	0.0623	0.0216	2.8880	0.0039
RELRANK	2.9658	0.3779	7.8475	0.0000
RELRNK2	-0.8188	0.1889	-4.3340	0.0000
RESCODE	-0.3257	0.1105	-2.9489	0.0032
SEXCODE	-0.1904	0.0845	-2.2543	0.0242
SUMMER	0.2859	0.2238	1.2774	0.2015
Theta	0.4764	0.2041	2.3341	0.0196
Alpha	1.4490	0.1167	12.4165	0.0000
Log-Like	-1524.528			
N Obs	6389			
N Iter	23			

5.1.1 The Impact of Relative Rank on Enrollment Duration

Recall Hypothesis 1 states that the likelihood of persistence should be low for students of low relative rank, increase as rank increases, and decrease again as rank increases. In other words, persistence should exhibit quadratic behavior in relative rank. Testing this condition involves examining the sign and statistical significance of the RELRNK2 coefficient as well as the marginal behavior of enrollment duration with respect to changes in RELRANK in Table 5.4; support for the hypothesis is suggested by a statistically significant negative coefficient. Examination of RELRNK2 in Table 5.1 suggests that the hypothesis is supported. It should be noted that the model estimated in Table 5.1 is a single destination model where no distinction is made between system dropout and transfer. The multiple destination models are presented in Tables 5.2 and 5.3. Table 5.2 provides estimates for the transition intensity where transfer is the destinations. Table 5.3 provides estimates for the system dropout transition intensity.

Table 5.2: Parameter Estimates for Transition Intensity Model - Transfer. Dependent Variable: Log(DURATION).

Variable	Coefficient	Std Error	T-Stat	P-Value
CONSTANT	3.9068	1.1158	3.5012	0.0005
ACTCOMP	-0.1285	0.3744	-0.3431	0.7315
ENGINEER	0.7106	0.6645	1.0693	0.2849
FINAID	0.2695	0.2563	1.0515	0.2930
NONWHITE	0.7623	0.5501	1.3858	0.1658
NSTUDNT	-0.6775	0.3876	-1.7478	0.0805
PCTGRAD	-0.7414	0.5133	-1.4444	0.1486
PREPROF	-0.0013	0.5859	-0.0023	0.9982
PROBENR	-0.8024	0.4503	-1.7819	0.0748
RANKPCTL	-0.0743	0.0765	-0.9702	0.3319
RELRANK	1.7769	0.8745	2.0320	0.0422
RELRNK2	-0.1371	0.4578	-0.2995	0.7645
RESCODE	0.8763	0.6183	1.4173	0.1564
SEXCODE	-0.1162	0.2675	-0.4343	0.6641
SUMMER	-0.1592	0.5002	-0.3182	0.7503
Theta	0.3527	1.0782	0.3271	0.7436
Alpha	1.3694	0.2917	4.6953	0.0000
Log-Like	-260.357			
N Obs	6389			
N Iter	21			

Table 5.3: Parameter Estimates for Transition Intensity Model - System Dropout. Dependent Variable: Log(DURATION).

Variable	Coefficient	Std Error	T-Stat	P-Value
CONSTANT	0.7542	0.4009	1.8814	0.0599
ACTCOMP	-0.2330	0.1328	-1.7547	0.0793
ENGINEER	0.3260	0.1810	1.8008	0.0717
FINAID	-0.0490	0.0853	-0.5746	0.5656
NONWHITE	-0.0043	0.1165	-0.0373	0.9703
NSTUDNT	-0.0542	0.1327	-0.4081	0.6832
PCTGRAD	-0.8893	0.1596	-5.5735	0.0000
PREPROF	-0.5508	0.1788	-3.0803	0.0021
PROBENR	-0.4325	0.1494	-2.8947	0.0038
RANKPCTL	0.0739	0.0229	3.2339	0.0012
RELRANK	3.0524	0.4243	7.1943	0.0000

continued on next page

Table 5.3 *continued from previous page*

Variable	Coefficient	Std Error	T-Stat	P-Value
RELRNK2	-0.8688	0.2164	-4.0155	0.0001
RESCODE	-0.4076	0.1163	-3.5053	0.0005
SEXCODE	-0.1967	0.0895	-2.1983	0.0279
SUMMER	0.3840	0.2452	1.5661	0.1173
Theta	0.6282	0.2431	2.5839	0.0098
Alpha	1.4799	0.1280	11.5620	0.0000
Log-Like	-1411.432			
N Obs	6389			
N Iter	28			

The RELRNK2 coefficient in the transfer transition intensity is not statistically significant; however, it is significant in the dropout intensity. The sign of RELRNK2 is negative for both the transfer transition intensity and dropout intensity.

The marginal effect of RELRANK on enrollment duration was computed at standard deviation units below and above its mean with all other independent variables at their respective sample means. In general, the marginal effect used here is defined as a change in enrollment duration due to an x-unit standard deviation change in the independent variable, all other independent variables constant at their means. This situation depicted the persistence of the “average” student. To assess differences in above- and below-average students, ACTCOMP and RANKPCTL are varied accordingly. A “gifted” student is defined here as an average student with ACTCOMP and RANKPCTL two standard deviations above the mean. Intuitively, this student’s ACT composite score was at least 32 and he or she also graduated in the top of his or her high school class. Likewise, a “challenged” student is defined as an average student with ACTCOMP and RANKPCTL two standard deviations below the mean (a 16 composite score and bottom third rank, respectively). Also evaluated is the ratio of the transition intensity to the dropout intensity (TD Ratio). Numbers less than 1 indicate that should exit occur, dropout is more likely than transfer; likewise numbers greater than 1 indicate transfer is more likely. All marginal effects had TD

ratios less than 1 and as will be seen when evaluating predictive performance, the WGH model did not predict transfers. These effects for RELRANK are presented in Table 5.4.

Table 5.4: Marginal Impact of RELRANK on Enrollment Duration.

SD Unit	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
-3	-2.5946	0.0954	-2.7024	0.1567	-2.4910	0.0580
-2	-1.8827	0.1101	-1.9610	0.1779	-1.8075	0.0680
-1	-0.9715	0.1349	-1.0119	0.2139	-0.9327	0.0848
0	0.0000	0.1601	0.0000	0.2494	0.0000	0.1023
1	0.8094	0.1735	0.8430	0.2657	0.7771	0.1128
2	1.2285	0.1674	1.2796	0.2524	1.1795	0.1105
3	1.1270	0.1434	1.1738	0.2135	1.0820	0.0959
Mean X	1.0826					
Std Dev X	0.3153					
Mean T	3.5874					

Interpreting the results in Table 5.4 is straightforward. A one standard deviation unit decrease (-1) in RELRANK results in enrollment duration decreasing by 0.9715 semesters (i.e., slightly under a semester) for the average student, 1.0119 semesters for the gifted student, and 0.9327 semesters for the challenged student. Also, all students are more likely to depart the system than transfer within the system, as indicated by a relative transfer intensity of 0.1349, which is less than 1. As RELRANK increases from low to high, enrollment duration indeed increases up to a point, then begins to decrease; however, this turning point appears to be for exceptional performers, ranking at least 3 standard deviations above the mean. This suggests that for the majority of students, RELRANK is directly related to persistence. Any student falling 2 standard deviations below the mean in RELRANK is likely to exit nearly two semesters earlier than the at-par performer. Finally, the general pattern of persistence appears to be consistent between the average, gifted, and challenged students. Gifted students

appear to be less likely to leave when faced with below-par relative performance and less likely to persist when enjoying above-par performance; however, these differences are fairly small.

5.1.2 Distinguishing Between Dropout and Transfer

Hypothesis 2 states that dropouts and transfers behave differently. The coefficients from each transition intensity are tested for equality and the results are presented in Table 5.5.

Table 5.5: Test of Equality Between Transfer and Dropout Coefficients.

Variable	Transfer	Dropout	Std Error	T-Stat	P-Value
ACTCOMP	-0.1285	-0.2330	0.2536	0.4123	0.3401
ENGINEER	0.7106	0.3260	0.4228	0.9098	0.1815
FINAID	0.2695	-0.0490	0.1708	1.8648	0.0311
NONWHITE	0.7623	-0.0043	0.3333	2.3002	0.0107
NSTUDENT	-0.6775	-0.0542	0.2602	2.3958	0.0083
PCTGRAD	-0.7414	-0.8893	0.3364	0.4397	0.3301
PREPROF	-0.0013	-0.5508	0.3823	1.4371	0.0753
PROBENR	-0.8024	-0.4325	0.2999	1.2335	0.1087
RANKPCTL	-0.0743	0.0739	0.0497	2.9815	0.0014
RELRANK	1.7769	3.0524	0.6494	1.9642	0.0248
RELRNK2	-0.1371	-0.8688	0.3371	2.1706	0.0150
RESCODE	0.8763	-0.4076	0.3673	3.4956	0.0002
SEXCODE	-0.1162	-0.1967	0.1785	0.4508	0.3261
SUMMER	-0.1592	0.3840	0.3727	1.4575	0.0725

Five of 14 variables were not significantly different at the $\alpha = 0.10$ level: ACTCOMP, ENGINEER, PCTGRAD, PROBENR, and SEXCODE. A joint test where the behavior coefficients in the dropout transition intensity were assumed equal to those in the transition intensity yielded a Wald chi-square statistic of 2133.39, which is significant at any desired level. Thus, the hypothesis is supported and it is important to distinguish between dropouts and transfers.

5.1.3 The Impact of Class Size on Enrollment Duration

Hypothesis 3 states that larger class sizes impede student-faculty interaction and thus, academic integration. It is expected that an increase in class size would tend to reduce the likelihood of persistence. To test this, the coefficient on NSTUDNT should be negative and statistically significant. Examining Tables 5.1, 5.2, and 5.3 it can be seen that the coefficients are negative in the single stage model, transfer intensity, and the dropout intensity model. Statistical significance is achieved only in the transfer intensity model. The marginal impact of NSTUDNT on enrollment duration is presented in Table 5.6.

Table 5.6: Marginal Impact of NSTUDNT on Enrollment Duration.

SD Unit	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
-2	0.2396	0.1020	0.2495	0.1709	0.2300	0.0609
-1	0.1178	0.0972	0.1227	0.1629	0.1131	0.0579
0	0.0000	0.0924	0.0000	0.1551	0.0000	0.0551
1	-0.1141	0.0878	-0.1188	0.1475	-0.1095	0.0523
2	-0.2246	0.0834	-0.2339	0.1401	-0.2156	0.0496
Mean X	0.7016					
Std Dev X	0.3250					
Mean T	3.5874					

In general, duration decreases as NSTUDNT increases; however, even at the extremes, the change in duration is well below a full semester. For any student, cutting the average class size in half (from 70 to 35) increases enrollment duration by only about 1/10th of a semester. Therefore, though there is support for the hypothesized direction of NSTUDNT, the independent impact on enrollment duration appears to be fairly small.

5.1.4 The Impact of Graduate Teaching Assistants on Enrollment Duration

Similar to Hypothesis 3, Hypothesis 4 states that a higher proportion of a student's portfolio being taught by graduate teaching assistants reduces student-faculty interaction and academic integration. This in turn leads to lower levels of persistence. To test the hypothesis, the sign of PCTGRAD should be negative and statistically significant. Referring back to Tables 5.1 and 5.3, this is indeed the case. The transfer intensity was correct in sign but lacked statistical significance. Indeed looking ahead to the OLS and ordered logit results in Tables 5.8 and 5.9, PCTGRAD is negative and significant as well. The signs of PCTGRAD for the multinomial model in Table 5.17 are positive and significant. This is not a contradiction because the multinomial probabilities are for the destination after departure. Positive signs here are also consistent with Hypothesis 4. Thus, Hypothesis 4 is supported and appears robust to different model specifications. The marginal impact of PCTGRAD on enrollment duration is presented in 5.7.

Table 5.7: Marginal Impact of PCTGRAD on Enrollment Duration.

SD Unit	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
-1	0.5751	0.1745	0.5990	0.2923	0.5522	0.1041
0	0.0000	0.1766	0.0000	0.2963	0.0000	0.1053
1	-0.4957	0.1791	-0.5163	0.3009	-0.4759	0.1066
2	-0.9228	0.1819	-0.9612	0.3059	-0.8860	0.1082
3	-1.2910	0.1850	-1.3447	0.3113	-1.2394	0.1099
Mean X	0.2985					
Std Dev X	0.2466					
Mean T	3.1879					

Enrollment duration decreases as PCTGRAD increases, consistent with the hypothesis. For any student, a standard deviation increase in PCTGRAD from the mean

appears to decrease enrollment duration by about 1/2 of a semester. Intuitively, in a six-course portfolio with four faculty and 2 graduate teaching assistants, replacing one faculty member with a graduate teaching assistant can reduce enrollment duration by nearly 1/2 of a semester. Similarly, replacing one of the graduate assistants with regular faculty can increase enrollment duration by over 1/2 of a semester.

5.1.5 The Impact of Summer Enrollment on Enrollment Duration

Hypothesis 5 states that enrolling in summer courses demonstrates an educational commitment, and this should translate into higher levels of persistence. To test this hypothesis, the sign of the SUMMER coefficient should be positive and statistically significant. Examining Tables 5.1, 5.2, and 5.3, only the single stage model and the dropout intensity model support the hypothesis; however, neither model achieves statistical significance. The transfer intensity model has the wrong sign; however, the coefficient is not significant by conventional standards. Thus, Hypothesis 5 does not appear to be empirically supported.

5.1.6 The Impact of Unmeasured Heterogeneity on The Hazard Function

Hypothesis 6 states that the existence of unmeasured heterogeneity significantly contributes to the negative duration dependence observed in the sample. Support for the hypothesis is found by examining the statistical significance of the parameter "Theta" in Table 5.1. Heterogeneity is significant in the single-stage and dropout intensity models. Examining the shape parameter "Alpha" suggests that the processes exhibits positive duration dependence. The coefficient is 1.449, 1.369, and 1.479 in the single-stage, transfer, and dropout models, respectively. This would produce a

Weibull hazard function that rises rapidly at first, then tends to flatten out over time; that is, the increase in the rate of exit with each unit of time becomes smaller and smaller. If this is a reasonable characterization for the enrollment hazard function, then heterogeneity is the primary reason for the observed negative duration dependence. Had the “Alpha” coefficients been less than unity, the hazard would exhibit negative duration dependence and heterogeneity would make it more pronounced. Because the heterogeneity coefficient is significant and the shape parameters are greater than unity, the observed negative duration dependence arises primarily from unmeasured student characteristics.

5.1.7 The Predictive Performance of The Weibull Hazard Model: Enrollment Duration

Hypothesis 7 states that the Weibull model with gamma heterogeneity offers better predictive performance than the standard ordinary least squares (OLS) or the ordered logit model. The coefficients for the OLS and ordered logit models are presented in Tables 5.8 and 5.9, respectively.

Table 5.8: Ordinary Least Squares Coefficients of Determinants of Student Persistence. Dependent Variable: DURATION.

Variable	Coefficient	Std Error	T-Stat	P-Value
CONSTANT	5.3907	0.4693	11.4870	0.0000
ACTCOMP	-0.1510	0.1433	-1.0530	0.2922
ENGINEER	-0.0079	0.1938	-0.0410	0.9675
FINAID	0.1138	0.1052	1.0810	0.2796
NONWHITE	0.2632	0.1295	2.0320	0.0421
NSTUDENT	-1.4584	0.1706	-8.5490	0.0000
PCTGRAD	-3.7790	0.2203	-17.1510	0.0000
PREPROF	-1.0963	0.2923	-3.7500	0.0002
PROBENR	-1.3993	0.2160	-6.4780	0.0000
RANKPCTL	0.1344	0.0296	4.5490	0.0000
RELRANK	1.3981	0.3345	4.1800	0.0000

continued on next page

Table 5.8 continued from previous page

Variable	Coefficient	Std Error	T-Stat	P-Value
RELRNK2	-0.1119	0.1815	-0.6160	0.5377
RESCODE	-0.4230	0.1541	-2.7440	0.0061
SEXCODE	-0.2231	0.1083	-2.0590	0.0395
SUMMER	-1.4774	0.4902	-3.0140	0.0026

Table 5.9: Ordered Logit Coefficients of Determinants of Student Persistence. Dependent Variable: DURATION

Variable	Coefficient	Std Error	T-Stat	P-Value
CONSTANT	1.5473	0.3115	4.9670	0.0000
ACTCOMP	-0.0763	0.1067	-0.7160	0.4743
ENGINEER	0.1742	0.1633	1.0670	0.2860
FINAID	0.0684	0.0730	0.9360	0.3493
NONWHITE	0.1648	0.0973	1.6930	0.0905
NSTUDNT	-0.9026	0.0982	-9.1910	0.0000
PCTGRAD	-2.0488	0.1524	-13.4470	0.0000
PREPROF	-0.5781	0.2195	-2.6350	0.0084
PROBENR	-0.7108	0.1364	-5.2130	0.0000
RANKPCTL	0.0869	0.0193	4.4980	0.0000
RELRANK	0.8782	0.2386	3.6810	0.0002
RELRNK2	-0.0819	0.1273	-0.6440	0.5197
RESCODE	-0.3081	0.1068	-2.8840	0.0039
SEXCODE	-0.1830	0.0752	-2.4330	0.0150
SUMMER	-0.4235	0.3789	-1.1180	0.2638
Mu(1)	0.6145	0.0451	13.6140	0.0000
Mu(2)	0.8403	0.0514	16.3480	0.0000
Mu(3)	1.0288	0.0544	18.9190	0.0000
Mu(4)	1.1404	0.0562	20.2910	0.0000
Mu(5)	1.3392	0.0582	22.9940	0.0000
Log-Like	-1390.304			
N Obs	1239			
N Iter	30			

The Kruskal-Wallis test is used to determine if the predictions from each model are independent samples from identical populations. Failure to reject the null hypothesis implies that each model yields similar predictions, so similar in fact that each cannot be distinguished from the other models. The statistic assumed a value of 1,210.0 for in-

sample and 584.67 for out-of-sample validation and are significant at all conventional levels. Pair-wise Kruskal-Wallis tests were conducted to see if any pair yielded similar predictions. The statistics for in-sample results are as follows: WGH vs. OLS = 812.41, WGH vs. ordered logit = 1,035.7, and OLS vs. ordered logit = 48.45. These are significant at all conventional levels. The out-of-sample statistics are as follows: WGH vs. OLS = 397.95, WGH vs. ordered logit = 495.73, and OLS vs. ordered logit = 24.06. Again, these are significant at all conventional levels.

Three sets of contingency tables are presented below: WGH versus actual, OLS versus actual, and ordered logit versus actual. The in-sample results are presented in Tables 5.10, 5.11, and 5.12, respectively.

Table 5.10: In-Sample Predicted versus Actual Enrollment Duration - WGH Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
0	11	1	2	1	1	1	0	17
1	21	11	7	3	4	4	8	58
2	11	7	6	3	2	1	2	32
3	12	7	2	1	1	2	5	30
4	6	13	5	1	0	2	4	31
5	7	10	3	5	0	4	10	39
6	11	11	0	4	2	1	11	40
7	69	87	38	36	23	47	675	975
Total	148	147	63	54	33	62	715	1222

Table 5.11: In-Sample Predicted versus Actual Enrollment Duration - OLS Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
0	5	2	2	0	1	0	0	10
1	14	7	8	1	0	1	0	31
2	25	21	14	3	2	3	6	74

continued on next page

Table 5.11 continued from previous page

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
3	40	46	14	14	6	5	21	146
4	36	47	12	20	7	13	76	211
5	23	16	6	10	9	18	207	289
6	3	6	7	5	7	13	308	349
7	2	2	0	1	1	9	97	112
Total	148	147	63	54	33	62	715	1222

Table 5.12: In-Sample Predicted versus Actual Enrollment Duration - Ordered Logit Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
1	6	1	0	0	1	0	0	8
2	27	17	17	2	1	4	3	71
3	48	57	13	16	7	5	21	167
4	46	51	23	22	8	16	108	274
5	19	18	8	13	14	26	421	519
6	2	3	2	1	2	11	162	183
Total	148	147	63	54	33	62	715	1222

The out-of-sample results are presented in Tables 5.13, 5.14, and 5.15, respectively.

Table 5.13: Out-of-Sample Predicted versus Actual Enrollment Duration - WGH Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
0	4	1	2	1	2	0	0	10
1	8	8	3	1	1	1	2	24
2	7	3	3	0	0	1	4	18
3	6	2	3	1	1	3	4	20
4	4	4	0	2	1	0	4	15
5	3	4	1	0	1	1	2	12
6	3	5	1	2	0	1	4	16
7	40	49	17	28	14	25	305	478
Total	75	76	30	35	20	32	325	593

Table 5.14: Out-of-Sample Predicted versus Actual Enrollment Duration - OLS Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
0	1	0	1	1	0	0	0	3
1	6	5	2	0	0	1	0	14
2	11	10	7	1	6	2	3	40
3	23	18	6	5	3	1	7	63
4	18	27	11	11	4	7	36	114
5	11	10	2	11	6	10	96	146
6	5	6	1	6	1	10	133	162
7	0	0	0	0	0	1	50	51
Total	75	76	30	35	20	32	325	593

Table 5.15: Out-of-Sample Predicted versus Actual Enrollment Duration - Ordered Logit Model.

Predicted	Actual Semesters							Total
	1	2	3	4	5	6	7	
1	1	0	1	1	0	0	0	3
2	9	7	6	1	4	2	2	31
3	26	25	9	4	5	2	7	78
4	25	32	10	13	5	8	51	144
5	14	12	4	16	6	17	198	267
6	0	0	0	0	0	3	67	70
Total	75	76	30	35	20	32	325	593

Table 5.16 presents the chi-square, Pearson's correlation, conditional entropy, and hit rate statistics for both in-sample and out-of-sample data sets.

Table 5.16: In- and Out-of-Sample Test and Association Results for WGH, OLS, and Ordered Logit Models .

Statistic	In-Sample			Out-of-Sample		
	WGH	OLS	Ordered Logit	WGH	OLS	Ordered Logit
Chi-Square	336.9	611.2	597.1	138.8	304.0	271.5
Correlation	0.249	0.635	0.624	0.377	0.610	0.588
Entropy	0.094	0.196	0.184	0.077	0.191	0.174
Hit Rate	0.579	0.154	0.068	0.545	0.167	0.066

According to the chi-square statistics, each model's prediction bears a statistically significant relationship with the actual enrollment durations. The strength of the relationship, as measured by Pearson's Correlation, is least impressive with the WGH model. Both OLS and ordered logit's predictions are more strongly related to actual durations than those of the WGH model. OLS appears to be the best predictor according to the correlation and entropy statistic, followed by ordered logit, and finally the WGH model. However, according to the hit rates, the WGH model strongly outperforms the others. This is because the WGH model successfully predicted a large number of students persisting to the censoring semester. Therefore, these results are inconclusive regarding predictive hypothesis 7 of enrollment.

5.1.8 Predictive Performance of the Weibull Model: Exit Destinations

Hypothesis 8 states that the WGH model offers better predictive performance than multinomial logit when predicting departure destination. The estimated coefficients for the multinomial model are presented in Table 5.17.

Table 5.17: Multinomial Logit Coefficients of Determinants of Dropout, and Transfer. Normalized on Persistence

Variable	Coefficient	Std Error	T-Stat	P-Value
Characteristics for Prob(Dropout)				
CONSTANT	-0.2594	0.7462	-0.348	0.7281
ACTCOMP	0.4343	0.2393	1.815	0.0695
ENGINEER	-0.5404	0.3259	-1.658	0.0972
FINAID	-0.4340	0.1597	-2.718	0.0066
NONWHITE	0.0455	0.2183	0.209	0.8347
NSTUDNT	2.0909	0.2637	7.928	0.0000
PCTGRAD	3.9607	0.3503	11.307	0.0000
PREPROF	1.3931	0.4348	3.204	0.0014
PROBENR	1.4400	0.4020	3.582	0.0003
RANKPCTL	-0.1705	0.0453	-3.761	0.0002
RELRANK	-4.1610	0.8709	-4.778	0.0000
RELRNK2	1.3323	0.4829	2.759	0.0058
RESCODE	0.8288	0.2196	3.775	0.0002
SEXCODE	0.6131	0.1656	3.702	0.0002
SUMMER	30.577	1052178.5	0.000	1.0000
Characteristics for Prob(Transfer)				
CONSTANT	-3.9034	1.41100	-2.766	0.0057
ACTCOMP	0.3713	0.47150	0.787	0.4311
ENGINEER	-1.0382	0.80039	-1.297	0.1946
FINAID	-0.8862	0.33659	-2.633	0.0085
NONWHITE	-0.9654	0.62635	-1.541	0.1233
NSTUDNT	2.7592	0.44642	6.181	0.0000
PCTGRAD	3.6501	0.66092	5.523	0.0000
PREPROF	0.7337	0.83610	0.878	0.3802
PROBENR	1.8179	0.57569	3.158	0.0016
RANKPCTL	0.0353	0.09630	0.366	0.7142
RELRANK	-3.2811	1.40142	-2.341	0.0192
RELRNK2	0.6705	0.85236	0.787	0.4315
RESCODE	-0.8027	0.75226	-1.067	0.2859
SEXCODE	0.5312	0.33767	1.573	0.1157
SUMMER	31.0978	1052178.5	0.000	1.0000

As above, the Kruskal-Wallis test is used to determine if the predictions from each model are independent samples from identical populations. The statistic assumed a value of 62.747 for in-sample and 31.205 for out-of-sample validation and each is significant at all conventional levels.

The WGH and multinomial logit contingency tables are presented below: The in-sample results for the WGH and multinomial logit models are presented in Tables 5.18 and 5.19, respectively. The out-of-sample results for the WGH and multinomial logit models are presented in Tables 5.20 and 5.21, respectively.

Table 5.18: In-Sample Predicted versus Actual Destination - WGH Model.

Predicted	Actual Destination			Total
	Continue	Dropout	Transfer	
Continue	675	267	33	975
Dropout	40	189	17	246
Transfer	0	1	0	1
Total	715	457	50	1222

Table 5.19: In-Sample Predicted versus Actual Destination - Multinomial Logit Model.

Predicted	Actual Destination			Total
	Continue	Dropout	Transfer	
Continue	636	147	17	800
Dropout	79	310	33	422
Transfer	0	0	0	0
Total	715	457	50	1222

Table 5.20: Out-of-Sample Predicted versus Actual Destination - WGH Model.

Predicted	Actual Destination			Total
	Continue	Dropout	Transfer	
Continue	308	147	23	478
Dropout	20	87	8	115
Transfer	0	0	0	0
Total	328	234	31	593

Table 5.21: Out-of-Sample Predicted versus Actual Destination - Multinomial Logit Model.

Predicted	Actual Destination			Total
	Continue	Dropout	Transfer	
Continue	300	83	10	393
Dropout	28	151	21	200
Transfer	0	0	0	0
Total	328	234	31	593

Table 5.22 presents the chi-square, conditional entropy, and hit rate statistics for both in-sample and out-of-sample data sets.

Table 5.22: In- and Out-of-Sample Test and Association Results for WGH, OLS, and Ordered Logit Models .

Statistic	WGH	In-Sample		Out-of-Sample	
		Multinomial	Logit	WGH	Multinomial
Chi-Square	230.3	420.5		85.3	208.5
Entropy	0.120	0.222		0.077	0.222
Hit Rate	0.707	0.774		0.666	0.761

Based on the chi-square results, there is a significant relationship between the predicted and actual destinations for both models, in both in-sample and out-of-sample validation. The conditional entropy statistic suggests that more uncertainty in the actual destinations is predicted by the multinomial predictions than the WGH predictions. Also, the multinomial model has a higher hit rate than WGH. This is the case in both in-sample and out-of-sample data. These results do not support the hypothesis that the WGH model is a better predictor of departure destination.

5.1.9 Other Results

A number of variables were included in the analysis for purposes of control. The variables are ACTCOMP, ENGINEER, FINAID, NONWHITE, PREPROF, PROBENR,

RANKPCTL, RESCODE, and SEXCODE. According to the persistence literature, the variables ACTCOMP and RANKPCTL reflect the student's pre-college schooling performance and skills. In previous studies, these have been found to be of secondary importance in predicting persistence, especially when various aspects of the freshman year experience are considered. Each are expected to be directly related to persistence. ENGINEER indicates that a student is an engineering major. Getting into the engineering program is a competitive, selective process, and this would reflect a commitment to persistence. PREPROF indicates the student is enrolled in a pre-professional program and is included to account for exit rates between the third and sixth semesters. While pre-professional indicates a certain commitment, the impact on persistence is expected to be negative because, upon completing the required course-work, the student usually transfers to another institution to complete their training. In this study, this should occur sometime prior to the censoring date. NONWHITE is a composite ethnicity indicator and SEXCODE indicates the student is female, both of which have been found to be negatively related to persistence. PROBENR indicates the student is a problem-enrollment in that he or she is under academic notice or probation, and it is expected to be negatively related to persistence. FINAID indicates the student is receiving financial aid. This indicates a willingness to enter into a financial contract for education and should be positively related to persistence. Finally, RESCODE indicates the student is not a resident. This is a proxy for long distance from home and has been found in other studies to be negatively related to persistence.

Using a standard level of significance of 0.10 in the single-stage model, the only significant variables from the list above are ACTCOMP, ENGINEER, PREPROF, PROBENR, RANKPCTL, RESCODE, and SEXCODE. None of these variables were significant in the transition intensity model. With the exception of ACTCOMP, each had the expected signs. The marginal impacts of these variables are presented in

Tables 5.23, 5.24, 5.25, 5.26, 5.27, 5.28, and 5.29, respectively.

Table 5.23: Marginal Impact of ACTCOMP on Enrollment Duration.

SD Unit	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
-3	0.7037	0.2293	0.8259	0.4232	0.5996	0.1244
-2	0.4550	0.2322	0.5340	0.4288	0.3877	0.1258
-1	0.2207	0.2352	0.2590	0.4346	0.1880	0.1274
0	0.0000	0.2383	0.0000	0.4406	0.0000	0.1290
1	-0.2079	0.2415	-0.2440	0.4467	-0.1771	0.1306
2	-0.4038	0.2448	-0.4739	0.4531	-0.3440	0.1323
3	-0.5883	0.2482	-0.6905	0.4596	-0.5012	0.1341
Mean X	2.4036					
Std Dev X	0.3821					
Mean T	3.5874					

A clear negative relationship is exhibited between ACTCOMP and enrollment duration across all duration models. For exceptional students (3 standard deviations above the mean) expected enrollment duration drops by half a semester. The TD ratios indicate that students are more likely to depart the system than to transfer within the system. This result runs counter to intuition; however, it does suggest that admitting bright students does not guarantee persistence. This also suggests that raising admissions standards via the ACT scores may not improve persistence.

Table 5.24: Marginal Impact of ENGINEER on Enrollment Duration.

Indicator	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
0	0.0000	0.1731	0.0000	0.2896	0.0000	0.1034
1	1.0144	0.1873	1.0566	0.3111	0.9739	0.1127
Mean X	0.1163					
Std Dev X	0.3206					
Mean T	3.4824					

Engineering majors exhibit a stronger propensity to persist than the average student by about a semester. Again, should departure occur, it appears more likely the student will drop out of the system than transfer.

Table 5.25: Marginal Impact of PREPROF on Enrollment Duration.

Indicator	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
0	0.0000	0.1698	0.0000	0.2841	0.0000	0.1014
1	-1.0757	0.1951	-1.1204	0.3270	-1.0328	0.1164
Mean X	0.0460					
Std Dev X	0.2095					
Mean T	3.6456					

Students enrolled in pre-professional fields are less likely to persist and when departure occurs, they are more likely to drop out of the system than to transfer.

Table 5.26: Marginal Impact of PROBENR on Enrollment Duration.

Indicator	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
0	0.0000	0.1622	0.0000	0.2714	0.0000	0.0969
1	-0.9852	0.1488	-1.0262	0.2501	-0.9459	0.0885
Mean X	0.0449					
Std Dev X	0.2071					
Mean T	3.6387					

Students who are placed on academic notice or probation are expected to withdraw about one semester sooner than the average student, and according to the TD ratios, are more likely to drop out of the system than transfer.

Table 5.27: Marginal Impact of RANKPCTL on Enrollment Duration.

SD Unit	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
-3	-0.7661	0.0431	-0.6799	0.0395	-0.8632	0.0471
-2	-0.5309	0.0407	-0.4711	0.0374	-0.5982	0.0443
-1	-0.2761	0.0384	-0.2450	0.0354	-0.3111	0.0418
0	0.0000	0.0364	0.0000	0.0336	0.0000	0.0395
1	0.2991	0.0346	0.2654	0.0319	0.3370	0.0375
Mean X	7.8221					
Std Dev X	1.8634					
Mean T	3.5874					

Students graduating in the top of their high school class are more likely to persist than the average student. This is indicated by a one-standard deviation increase in RANKPCTL and adds nearly a third of a semester to expected enrollment duration.

Table 5.28: Marginal Impact of RESCODE on Enrollment Duration.

Indicator	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
0	0.0000	0.1964	0.0000	0.3285	0.0000	0.1173
1	-0.7426	0.2600	-0.7735	0.4321	-0.7130	0.1562
Mean X	0.1237					
Std Dev X	0.3292					
Mean T	3.6886					

Nonresident students are expected to persist about 0.75 semesters less than the average, and the TD ratios indicate they are more likely to drop out of the system than to transfer.

Table 5.29: Marginal Impact of SEXCODE on Enrollment Duration.

Indicator	Average		Gifted		Challenged	
	Δ Duration	TD Ratio	Δ Duration	TD Ratio	Δ Duration	TD Ratio
0	0.0000	0.1731	0.0000	0.2896	0.0000	0.1033
1	-0.4722	0.2300	-0.4918	0.3826	-0.4533	0.1381
Mean X	0.5077					
Std Dev X	0.5000					
Mean T	3.8349					

The marginal impact of being a female student reduces expected enrollment duration by nearly half a semester relative to the average male student. Similar to other results, female student are more likely to drop out of the system than to transfer.

5.2 Chapter Summary

This chapter presented the analysis and empirical evidence used to test the four main hypotheses of this study. A summary of the findings follows:

- Hypothesis 1 that relative rank was quadratically related to persistence was supported by the evidence. When no distinction is made regarding where the student exits to, the hypothesis was supported. The signs of RELRNK2 in each model were negative; however, the coefficient in the transfer intensity was not statistically significant. In terms of marginal impact on enrollment duration, relative rank is the single most influential variable examined.
- Hypothesis 2 regarding the behavior of dropouts and transfers was supported. Dropouts and transfers are behaviorally distinct.
- Hypothesis 3 regarding the relationship between class size and persistence was not supported. Statistical significance was lacking to support the notion that

larger class sizes were an impediment to academic integration and persistence. The marginal impact on enrollment duration was also weak.

- Hypothesis 4 regarding the relationship between classroom staffing and persistence was strongly supported, both within the hazard models considered and in competitor models as well. The hypothesis states that the likelihood of persistence is lower the higher the proportion of graduate student teaching assistants in a student's portfolio. Marginal analysis also shows changes in PCTGRAD to be moderately influential to changes in enrollment duration; adding a third teaching assistant by substituting for a faculty member in a six-course portfolio can potentially reduce the average student's persistence by half a semester.
- Hypothesis 5 regarding the relationship between summer enrollment and persistence was generally supported. Students who take summer courses are at least as likely or more likely to persist as those who don't (by about 0.75 semesters according to marginal analysis).
- Hypothesis 6 of the relationship between unobserved student heterogeneity and the observed pattern of departure was supported. The shape parameter estimate suggests that the enrollment duration hazard exhibits positive duration dependence, and unmeasured heterogeneity, as parameterized in the model, was found to be a significant contributor of the observed negative duration dependence. Thus, unmeasured personal differences in preference for persistence are primarily responsible for the observed pattern of student departure.
- The hypotheses regarding the predictive performance of the WGH model were inconclusive. The WGH model could not beat OLS or ordered logit in predicting enrollment duration according to goodness-of-fit tests; however, the WGH model had a much higher hit rate. Multinomial logit performed better than

WGH in predicting departure destination based on goodness-of-fit tests and hit rates.

- Seven control variables were found to be significant: ACTCOMP, ENGINEER, PREPROF, PROBENR, RANKPCTL, RESCODE, and SEXCODE. None of these variables were significant in the transition intensity model. With the exception of ACTCOMP, each had the expected signs. However, ACTCOMP's marginal impact was not particularly large.

The following chapter discusses some of the other findings in the analysis as well as offers concluding remarks.

CHAPTER 6

SUMMARY, DISCUSSION, AND CONCLUSION

The emphasis of this study was to specify and estimate statistical models appropriate for predicting undergraduate enrollment durations. Known as failure-time or duration models, these methods offer elegant ways to account for two types of information available in persistence data; namely, the time-to-exit and the type or characteristics of exit. This approach is well suited for longitudinal enrollment data where a cohort of students are followed over time.

A hazard regression model was specified assuming that enrollment durations were distributed as Weibull random variables, that observable student characteristics influence the *scale* of the distribution, and that unobservable student characteristics influence the *shape* of the distribution. In particular, these unobservable characteristics were assumed to enter the hazard model multiplicatively as Gamma distributed random variables with unit mean and constant variance. This type of model has become known in the econometric literature as a Weibull hazard model with Gamma heterogeneity. Two specifications were estimated: a single-stage model of non-specific student departure, and multiple destination transition intensity models where transfer and system dropout are made distinct. The sample was right-censored; any student not departing by the sixth semester (Fall 1996) had covariate observations up to that point. Whether they departed in the future was not known. The parameters of the model were estimated using maximum likelihood.

The data used in this study consisted of records on 1,832 Fall 1993 entering freshman at Oklahoma State University. Of these, 799 (43.6 percent) voluntarily departed by the Fall 1996 semester. Various background characteristics as well as semester course data were used to estimate the models. Background information included the student's high school rank as a percentage of the graduating class size, the composite ACT score, the student's sex and ethnicity, and the residency status of the student. Longitudinal data included the relative rank of the student within the current portfolio of courses taken, the average number of students in the portfolio, the proportion of instruction conducted by graduate student teaching assistants, financial aid, summer enrollment, problematic enrollment (i.e., academic notice or probation), and pre-professional and engineering major indicators. The dependent variables measure the number of semesters of continuous enrollment and whether the student, immediately upon departure, transferred to another institution or dropped out of the system. These data were randomly assigned to two samples: two-thirds of the data were used to estimate the models and one-third were used to provide out-of-sample predictive validation.

This study contains a number of hypotheses about how various observable characteristics affect persistence. The relationship of how unmeasured characteristics influences observed exit rates is also considered. Finally, the predictive performance of the proposed models relative to competitors is evaluated.

As expected, it was found that the behavior of dropouts is different than transfers. This result is consistent with Tinto (1993) and the empirical literature where the distinction was made (Horn, 1998). The coefficients from each specification were individually tested and jointly tested under the null hypothesis that they were the same. Five of the 14 independent variables had parameter estimates that were not statistically different: ACTCOMP, ENGINEER, PCTGRAD, PROBENR, and SEX-CODE. A joint test of all 14 variables found transfers and dropouts to be statistically

different.

Of those factors affecting persistence, Tinto (1975) and later in Tinto (1993) claimed that academic and social integration were especially important in a student's persistence-departure decision. There is general agreement that the current academic performance is a good indicator of academic integration, and that student-faculty interaction is an important component of social integration (Terenzini and Pascarella (1980), Pascarella (1980), Bean (1982), Pascarella and Terenzini (1983), Stage (1988), Stage (1989), Bean and Metzner (1985), Eaton and Bean (1995), and Pascarella, Edison, Hagedorn, Nora, and Terenzini (1996)). It has been consistently found that these integration measures are directly related to persistence. The measure of academic integration used in this study is called relative rank and it measures the average grade performance of the individual student relative to the class over the portfolio of courses taken per semester. This variable allows for the hypothesis that a student performing as well as his or her classmates is more likely to persist than a student performing well below or above the rest of class, especially if this performance is consistent across the portfolio of courses taken. The intuitive reason to expect this for high performing students is that they will likely move on to better alternatives in terms of prestige and expected income upon graduating (see Frank (1985) and Heath (1993)). The results of this study support this hypothesis, though it was found that only the extremely high performing students (i.e., at least three standard deviations above the mean) exhibit decreasing persistence. For the vast majority of students, the higher one's performance is relative to one's peers, the higher the likelihood of persistence.

This result has as much to say about the character of the institution as it does about the student. For example, this hypothesis is not likely to hold at an elite private institution, if for no other reason than there are very few better alternatives for high performing students. The evidence supported the hypothesis for dropouts

but not for transfers. That is, high performing students are more likely to persist or dropout than transfer. This implies two things; first, because high performing students are less likely to transfer to other institutions in the state system, Oklahoma State University is a relatively high ranking institution within the system. Second, while high performers are more likely to drop out of the system, this does not imply they are dropping out of higher education altogether. Data were not available on students who transfer out-of-state. These considerations suggest that the predictions involving relative rank for high performing students be interpreted with caution.

Another key finding was the relationship between persistence and the instructional composition of a student's course portfolio. In particular, a consistent finding in the proposed and competitor models alike is a negative relationship between persistence and the proportion of instruction conducted by graduate student assistants. It does not appear that any previous study has considered this; however, by reinterpreting Tinto's social integration framework, this result is plausible. This is because graduate teaching assistants are not considered to be regular university faculty by anyone involved, with the possible exception of the graduate students themselves. As a result, if more instruction is conducted by graduate students, there will necessarily be less opportunity for students to interact with the faculty. Furthermore, Pascarella and Terenzini (1983) found that low faculty interaction resulted in lower social integration and persistence. The immediate policy implication, then, is that as part of a retention program, instruction conducted by graduate students should be limited.

This conclusion is not warranted. First, the evidence presented here does not provide a complete picture of the complexity surrounding graduate students as instructors. For instance, data were not available or in reliable form to determine whether student evaluations or graduate instructor ethnicity played a part in lower persistence. Additionally, nothing is known about the level of teaching experience of the graduate instructor. Also, low performing students may simply use graduate teaching

assistants as convenient scapegoats for their poor academic performance. Second, an institution would need to consider the expected benefits of increased undergraduate retention against the expected costs of achieving it. Such costs could include, for example, reduced grant money and research production from faculty, higher faculty and graduate student attrition, and general loss of academic prestige because of lower volume or quality of research. On the other hand, it may be possible to shift many of the research responsibilities to graduate students, minimizing the impact on research production and freeing faculty to undertake more instruction. Whether this strategy does a disservice to graduate students in terms of their career development would also need to be considered. In short, the complexity of graduate students as instructors is an area that deserves careful attention and research, and the results of such an undertaking should be a part of an informed retention policy.

Other significant findings include whether the student enrolls in summer school, and whether the student is or is not a state resident. Each can be considered proxies of commitment in Tinto's framework; the first is expected to be directly related and the second negatively related to persistence. These effects were supported by the empirical evidence. A literal interpretation of Tinto's framework would not have a problem with the first result. Higher levels of commitment are indicated by a willingness to enroll in summer courses since summer enrollment is not required to maintain full-time-student status or to earn a degree within four years. The second is less obvious. Non-residents face a tough decision; an example might be whether to attend an in-state institution of lesser prestige, but also relatively inexpensive and closer to home, or attend a more expensive and prestigious out-of-state institution further from home. However, distance and prestige constant, out-of-state students should exhibit lower commitment than in-state students based on the cost differential alone: Out-of-state, non-legacy undergraduates paid an additional \$115.50 tuition per credit hour during the 1997 academic year. A legacy student is an out-of-state student,

but treated as in-state by virtue of their alumni parents. Unfortunately, data were not available on whether a student is legacy or not. Furthermore, a key requirement for establishing residency is that the student relocate to Oklahoma for at least one year not for the sole purpose of education. This is usually accomplished through full-time employment. Obviously, if this is attempted while maintaining enrollment, there will necessarily be less time available for academic studies, all else constant. Indeed, the average relative rank for in-state students was 1.12 compared to 1.07 for out-of-state students. Not only do out-of-state students have higher tuition expenses, they also pay a high price for establishing residency. Unfortunately, employment data were not available to formally test this notion.

The pattern of exit was found to be influenced by unobservable characteristics of the students or unmeasured heterogeneity. It was shown in Chapter 3 that unmeasured heterogeneity biased the hazard function toward negative duration dependence, where the probability of exit decreases over time. This study allowed for heterogeneity through a parametric specification of its distribution; more specifically, unmeasured heterogeneity was assumed to be a random variable from a Gamma distribution with unit mean and constant variance. When combined with a parametrically specified hazard function, the size and significance of the Gamma variance indicates heterogeneity is affecting observed exit rates. A positive, significant estimate suggests that observed negative duration dependence is attributable to unobserved differences between individuals and not to state dependence. Intuitively, the mobility prone students are the first to leave, and increasingly the persisters are composed of students with lower and lower chances of departure.

From a practical standpoint, these results suggest that policies intended to improve retention in the first year may not work as well as expected. Tinto (1993, pp. 145-53) provides some guidelines or "principles of effective retention," one of which is to "front-load" retention efforts during the first year. The economic rationale is

straightforward; given a fixed budget and resource for retention purposes, they should be expended where they are likely to have the most impact. The implicit assumption is that they will have a lasting impact, that is, if students can make it over the first year hump, their probability of persisting improves. This has been echoed in the recent attrition literature (e.g., (Berkner, Alamin, McCormick, & Bobbitt, 1996) and (Horn, 1998)). Such policies may have the effect of reducing attrition in the first year, and it may be tempting to take this as a sign of a successful policy. Based on the results in this study, such a conclusion may be premature: On the one hand, a necessary condition for reducing overall attrition is to reduce it during the first year. After all, students cannot persist to the third year if they don't persist through the first and second. On the other hand, front-loading may only increase the chances a student persists: the policy may be postponing the mobility-prone student's departures to later semesters where less is invested in retention.

One potential limitation of these results is that the unmeasured heterogeneity could be due to limited data on social integration. Granted, student-faculty formal and informal interactions are an important component of social (and academic) integration; however, there are other dimensions as well. The frequency and quality of peer group interactions and "buy-in" to the institution's culture via school apparel and novelties are just two examples. These data are unavailable in this study and are generally difficult and costly to obtain. Most empirical studies have used survey instruments designed to measure the differing dimensions of social integration, repeatedly administering them to a given cohort over time. The administrative logistics of this process make it difficult to capture and maintain a reliable source of data, and not surprisingly, relatively few institutions invest in this process.

Other control variables were considered as well. Six were found to be statistically significant at the 0.10 level in the single-stage model. These were the student's ACT composite score, engineering major, pre-professional degree, problematic enrollment,

high school rank, and female student. The composite ACT score was negatively related to persistence in the WGH, OLS, and ordered logit models. Based on marginal analysis, small differences in this score may not be important. Everything else constant, a student scoring three standard deviations above the mean (a nearly perfect score) is expected to persist half a semester less than the average student. Furthermore, they appear destined to drop out of the system rather than transfer. Based on earlier discussions, this may be reasonable. ACT scores are a key admissions component in most universities, and students with high ACT scores may be treating Oklahoma State University as an intermediate step to a better out-of-state alternative. Again, data were not available to test this assertion. It also suggests that admitting bright students does not guarantee persistence. Furthermore, raising admission standards via higher ACT requirements will not likely have the expected impact on retention.

The other results were in line with expectations, and based on marginal analysis, some had particularly strong effects. For example, average engineers are expected to persist about one semester longer than an equivalent non-engineer. On the other hand, those in pre-professional degrees (e.g., veterinary medicine) are likely to stay one semester less than the average student. Likewise, students on academic notice or probation are likely to (voluntarily) stay one semester less than the average student. Female students are likely to persist 1/2 a semester less than the average male. This last result is also supported in the literature. Pascarella and Terenzini (1983) found that persistence between males and females differed greatly. They attributed the difference primarily to the differences between social and academic integration, academic integration affecting males more strongly, and social integration more strongly affecting females. As mentioned earlier, measures of social integration were not readily available in this study and this may be influencing the impact of being a female student on persistence.

The models in this study were also evaluated in terms of predictive performance compared to competing models on a hold-out sample. The results were inconclusive. In both in- and out-of-sample validation, the Weibull model could not beat OLS in predicting enrollment duration based on goodness-of-fit tests; however, WGH had a much higher hit rate. The WGH model was much more successful in predicting persisting students. This result implies that even though the WGH model may not predict the exact departure time very well, it is able to identify the persisters better than OLS or ordered logit. On the other hand, the WGH model is not as well suited as multinomial logit for identifying where students are likely to go once they leave. The WGH model could not beat multinomial logit in predicting departure destination according to goodness-of-fit tests or hit rates.

Throughout this study, Tinto's theoretical framework was taken as given. Tinto's theory of student attrition is first and foremost a sociological one. In particular, Durkheim's theory of suicide is used as a basis for understanding the dropout process where students dropout primarily because they are unable to integrate into the social and academic systems of college. For Durkheim, these forms are egoistic and anomic suicide, where the social part of an individual's nature is insufficiently developed or the social setting lacks the needed rules to constrain individuals by integrating them into the collective whole. These forms of suicide were thought to be most prevalent in the transition period to modern society (Ashley & Orenstein, 1985). The transition from high school to college, the institutional structures, and the relative normlessness a student typically experiences are seen as the primary drivers of student departure. The process flow by which a student integrates into the social and academic systems has been depicted by Tinto in his well-known diagram of the longitudinal process of student departure (see Figure 2.1 on page 9).

Clearly, student departure does not have the same consequence as suicide; otherwise, dropout would be a rare event. The fact that student attrition is a sizable

phenomenon suggests that the suicide model is not entirely adequate for explaining dropout. A serious weakness of linking suicide to student departure is that it frames such departure in terms of self-destructive behavior. Tinto (1993, pp. 1, 37-45) readily points out that students who attend college and fail to obtain a degree may well receive some benefits from the experience and that attending college is as much about personal discovery as earning a degree. Even so, the general theme of his work is aptly expressed in the title of his 1993 book, "Leaving College: Rethinking the Causes and Cures of Student Attrition," where half of the book diagnoses the student departure problem and the remaining third of the book prescribes the treatments available to institutions for reducing or containing attrition. Student departure is seen as a treatable condition.

A different perspective would be to assume that a student's decision to persist or withdraw arises from rational choice. That is, a student will persist at an institution if the present value of expected net benefits is nonnegative. Once negative, a departure occurs. Whether or not a student continues at a different institution depends on the opportunities and constraints the student faces. Students continually evaluate the net benefit condition based on the arrival of new information, thus preserving the longitudinal character of departure. Students are at once producers and consumers of their educational experience. They combine pre-college skills and experiences as well as background characteristics and current experiences to produce educational outcomes. These outcomes enter as arguments in the utility function for persistence. To the extent academic and social integration are important in the decision process, the factors affecting either will enter into the student's utility function as well. In this framework, it is possible that a student's willingness to trade social integration for academic integration is feasible, thus leaving the persistence decision unchanged (assuming the student maintains the minimum standards on academic performance). Institutional factors affecting social integration, such as policies to increase student-

faculty interaction, can leave academic integration constant and retain more students by virtue of lowering the relative price of social integration. In this way, the dimensions and extent of institutional actions regarding student departure are seen in terms of responding to demand and not in terms of treating a condition. An implication here is that institutions will vary in their level of concern over student attrition. Two institutions with the same attrition rates can vary greatly in how attrition is viewed; one can view it as a condition to be treated and the other can view it as a marketing tool to attract (the best) students. The variance depends on the goals and objectives of the institution.

Because attrition is ultimately about choice, it seems that economics would have a great deal to offer institutional researchers. Tinto's work has been extremely important in aligning institutional research around a common model of student departure. What it lacks, and what economics can provide, is an analytical theory capable of mathematically specifying structural relationships in the dropout decision process. The literature on optimal job search, matching, and turnover in the labor market is a promising starting point for developing a theoretical model of the student departure decisions (in particular, see Mortensen (1986), Mortensen (1988), Jovanovic (1979), and Jovanovic (1984)). Such a theoretical model can aid in the specification of the duration model, potentially improving the precision of the estimated parameters and improving its predictions. Unfortunately, these theories are very complex and extremely difficult to estimate empirically.

Two variables that were especially important in predicting persistence were the relative rank of the student and the classroom staffing composition of the student's course portfolio. Relative rank should prove especially useful for evaluating the impact of "learning communities" Tinto (1993). Learning communities are an integrated approach for placing small cohorts of students in the same courses with the same instructors. These "teams of learners" presumably benefit from the shared experience

and find the systems of college less alienating. When evaluating the effectiveness of learning communities on persistence, a student's relative rank should prove more useful than overall grade point average, since relative rank will measure a student's performance relative to his or her teammates. Future studies may also benefit in accounting for classroom staffing, especially when classes are staffed with regular full-time faculty, temporary faculty, and graduate teaching assistants. Temporary faculty include professionals who teach a class on the side, full-time visiting, or part-time instructors. The classroom staffing question offers several interesting area for further research; for example, (1) whether graduate teaching assistants have a different impact on persistence than part-time instructors, or (2) whether, at four-year institutions, students from feeder schools staffed with part-time instructors are less likely to persist than students from feeder school staffed predominantly with full-time staff.

REFERENCES

- Amemiya, T. (1985). *Advanced econometrics*. Cambridge: Harvard University Press.
- Ashley, D., & Orenstein, D. M. (1985). *Sociological theory: Classical statements*. Boston: Allyn and Bacon.
- Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2), 155-187.
- Bean, J. P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, 17(4), 291-320.
- Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *The Review of Higher Education*, 6(2), 129-148.
- Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4), 485-540.
- Berkner, L. K., Alamin, S. C., McCormick, A. C., & Bobbitt, L. G. (1996). *Descriptive summary of 1989-90 beginning postsecondary students: Five years later*. (Tech. Rep. No. NCES 96-155). Washington D.C.: U.S. Department of Education. National Center for Educational Statistics.
- Eaton, S. B., & Bean, J. P. (1995). An approach/avoidance behavioral model of college student attrition. *Research in Higher Education*, 36(6), 617-645.

- Frank, R. H. (1985). *Choosing the right pond: Human behavior and the quest for status*. New York: Oxford University Press.
- Greene, W. H. (1993). *Econometric analysis* (Second ed.). New York: Macmillan Publishing Company.
- Greene, W. H. (1995). *LIMDEP version 7 user's manual*. New York: Econometrics Software, Inc.
- Heath, W. C. (1993). Choosing the right pond: College choice and the quest for status. *Economics of Education Review*, 12(1), 81-88.
- Heckman, J. J., & Singer, B. (1986). Social science duration analysis. In J. J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data*. Cambridge: Cambridge University Press.
- Horn, L. (1998). *Stopouts or stayouts? Undergraduates who leave college in their first year*. (Tech. Rep. No. NCES 1999-087). Washington D.C.: U.S. Department of Education. National Center for Educational Statistics.
- Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy*, 87(5), 972-990.
- Jovanovic, B. (1984). Matching, turnover, and unemployment. *Journal of Political Economy*, 92(1), 108-122.
- Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley and Sons.
- Kline, P. (1994). *An easy guide to factor analysis*. New York: Routledge.
- Kohn, M. G., Manski, C. F., & Mundel, D. S. (1976). An empirical investigation of the factors which influence college going behavior. *Annals of Economic and Social Measurement*, 5(4), 391-419.

- Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge: Cambridge University Press.
- Levin, H. M., & Tsang, M. C. (1987). The economics of student time. *Economics of Education Review*, 6(4), 357-364.
- Loehlin, J. C. (1987). *Latent variable models*. Hillsdale: Erlbaum.
- Manski, C. F. (1990). The use of intentions data to predict behavior: A best-case analysis. *Journal of the American Statistical Association*, 85(412), 934-940.
- Manski, C. F., & Wise, D. A. (1983). *College choice in America*. Cambridge: Harvard University Press.
- McKenzie, R. B., & Staaf, R. J. (1974). *An economic theory of learning: Student sovereignty and academic freedom*. Blacksburg, VA: University Publications.
- Mortensen, D. T. (1986). Job search and labor market analysis. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (Vol. 2). New York: North-Holland.
- Mortensen, D. T. (1988). Wages, separations, and job tenure: On-the-job specific training or matching? *Journal of Labor Economics*, 6(4), 445-471.
- Pascarella, E. T. (1980). Student-faculty informal contact and college outcomes. *Review of Educational Research*, 50(4), 545-595.
- Pascarella, E. T., Edison, M., Hagedorn, L. S., Nora, A., & Terenzini, P. T. (1996). Influences on students' internal locus of attribution for academic success in the first year of college. *Research in Higher Education*, 37(6), 731-756.
- Pascarella, E. T., & Terenzini, P. T. (1983). Predicting voluntary freshman year persistence/withdrawal behavior in a residential university: A path analytic validation of Tinto's model. *Journal of Educational Psychology*, 75(2), 215-226.

- Petersen, T. (1986). Fitting parametric survival models with time-varying covariates. *Applied Statistics*, 35(3), 281-288.
- SAS Institute. (1990). The FREQ procedure. In *SAS procedures guide, version 6* (Third ed.). Cary, NC: SAS Institute, Inc.
- Stage, F. K. (1988). University attrition: LISREL with logistic regression for the persistence criterion. *Research in Higher Education*, 29(4), 343-357.
- Stage, F. K. (1989). An alternative to path analysis: A demonstration of LISREL using students' commitment to an institution. *Journal of College Student Development*, 30(?), 129-135.
- Stokes, M. E., Davis, C. S., & Koch, G. G. (1995). *Categorical data analysis using the SAS system*. Cary, NC: SAS Institute, Inc.
- Terenzini, R. T., & Pascarella, E. T. (1980). Toward the validation of Tinto's model of college student attrition: A review of recent studies. *Research in Higher Education*, 12(3), 271-282.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89-125.
- Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (Second ed.). Chicago: University of Chicago Press.

APPENDIX.
IRB APPROVAL FORM

OKLAHOMA STATE UNIVERSITY
INSTITUTIONAL REVIEW BOARD
HUMAN SUBJECTS REVIEW

Date: 05-06-97

IRB#: GU-97-007

Proposal Title: ECONOMETRIC DURATION MODELS OF COLLEGE
STUDENT PERSISTENCE

Principal Investigator(s): Lee C. Adkins, James B. Eells, Becky Johnson

Reviewed and Processed as: Exempt

Approval Status Recommended by Reviewer(s): Approved

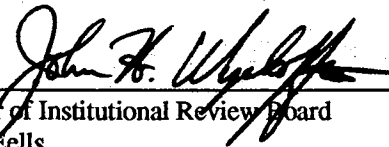
ALL APPROVALS MAY BE SUBJECT TO REVIEW BY FULL INSTITUTIONAL REVIEW BOARD
AT NEXT MEETING, AS WELL AS ARE SUBJECT TO MONITORING AT ANY TIME DURING
THE APPROVAL PERIOD.

APPROVAL STATUS PERIOD VALID FOR DATA COLLECTION FOR A ONE CALENDAR YEAR
PERIOD AFTER WHICH A CONTINUATION OR RENEWAL REQUEST IS REQUIRED TO BE
SUBMITTED FOR BOARD APPROVAL.

ANY MODIFICATIONS TO APPROVED PROJECT MUST ALSO BE SUBMITTED FOR
APPROVAL.

Comments, Modifications/Conditions for Approval or Disapproval are as follows:

Signature:



Chair of Institutional Review Board

cc: James B. Eells

Date: May 7, 1997

VITA

James Brian Eells

Candidate for the Degree of

Doctor of Philosophy

Thesis: ECONOMETRIC DURATION MODELS OF COLLEGE STUDENT PERSISTENCE

Major Field: Economics

Biographical

Education: Graduated from Glendale High School, Springfield, Missouri in May 1984; received a Bachelor of Science degree in Economics from Southwest Missouri State University, Springfield, Missouri in December 1989; received a Master of Science degree in Economics from Oklahoma State University, Stillwater, Oklahoma in December 1998; completed requirements for the Doctor of Philosophy degree in Economics from Oklahoma State University, Stillwater, Oklahoma in July 1999.

Professional Experience: Research Assistant for the Department of Economics at Southwest Missouri State University from August 1990 to May 1991. Research Assistant for the Department of Economics at Oklahoma State University from August 1991 to May 1992. Research Assistant for a funded grant headed by Dr. Lee C. Adkins at Oklahoma State University from August 1992 to May 1993. Taught Macroeconomics at Oklahoma State University from August 1993 to May 1994. Research Associate for University Assessment, Oklahoma State University from August 1994 to December 1996. Taught Economics of Social Issues in the Spring of 1995. Specialist at University Assessment, Oklahoma State University, from January to June 1998. Econometrician at American Express from June 1997 to April 1999. Manager at American Express from May 1999 to present.

Professional Memberships: Omicron Delta Epsilon, Southern Economic Association, Missouri Valley Economic Association.