AC 2009-2218: PREDICTING POST-SECONDARY EDUCATIONAL OUTCOMES WITH SURVIVAL ANALYSIS

Gillian Nicholls, University of Pittsburgh

Gillian Nicholls is a Lecturer in Industrial Engineering at the University of Pittsburgh. Her research interests are in applying statistical analysis and optimization to engineering education and transportation management. She holds the B.S. in Industrial Engineering (Lehigh University), Masters in Business Administration (Penn State University), and M.S. in Industrial Engineering and Ph.D. in Industrial Engineering(University of Pittsburgh.) Address: 1048 Benedum Hall, University of Pittsburgh, Pittsburgh, PA 15261; telephone 412.400.8631; fax: 412.624.9831; e-mail: gmn3@pitt.edu.

Harvey Wolfe, University of Pittsburgh

Harvey Wolfe is the William Kepler Whiteford Professor of Industrial Engineering at the University of Pittsburgh. After many years working in the area of applying operations research methods to the health field, he is now active in the development of models for assessing engineering education. He is a co-author of Engineering Ethics: Balancing Cost Schedule and Risk - Lessons Learned from the Space Shuttle (Cambridge University Press, 1997). He holds the B.E.S. in Industrial Engineering, M.S.E. in Operations Research, and Ph.D. in Operations Research (Johns Hopkins University).

Mary Besterfield-Sacre, University of Pittsburgh

Mary Besterfield-Sacre is an Associate Professor of Industrial Engineering and the Fulton C. Noss Faculty Fellow at the University of Pittsburgh. Her research interests are in engineering education evaluation and empirical and cost modeling applications for quality improvement in manufacturing and service organizations. She holds the B.S. in Engineering Management (University of Missouri Rolla), M.S. in Industrial Engineering (Purdue University), and Ph.D. in Industrial Engineering (University of Pittsburgh).

Larry Shuman, University of Pittsburgh

Larry J. Shuman is Associate Dean for Academic Affairs, School of Engineering, University of Pittsburgh and Professor of Industrial Engineering. His research includes studies directed at the undergraduate engineering learning experience, assessment and the ethical behavior of engineers. Dr. Shuman has published widely in the engineering education literature and is a co-author of Engineering Ethics: Balancing Cost Schedule and Risk - Lessons Learned from the Space Shuttle (Cambridge University Press, 1997). He holds the B.S. in Electrical Engineering (University of Cincinnati) and Ph.D. in Operations Research (Johns Hopkins University).

Predicting Post-Secondary Educational Outcomes with Survival Analysis

Survival Analysis, STEM, NELS

Abstract

Identifying potential students and understanding what affects their decision to depart the track of obtaining a college degree in engineering is critical to engineering educational research. This study used data from the National Education Longitudinal Study of 1988-2000 (NELS) to develop a model for predicting post-secondary educational outcomes with particular focus on students earning a college degree in Science, Technology, Engineering, or Mathematics (STEM). The objective was to identify factors that affected the probability of a given student "surviving" on the STEM track past a key time point in the study at which most students attending college were nearing graduation. NELS provided a comprehensive set of variables for 12,144 students of which 11,128 had clearly determinable educational outcomes. The set of potential outcomes included dropping out of high school, completing education at high school graduation, dropping out of college, earning a less than four year degree, earning a degree other than STEM, earning a STEM degree, or having an incomplete degree at the study's conclusion. The modeling process utilized demographic, attitudinal, and academic performance data mainly collected at the 8th grade level as well as standardized test scores and college enrollment status variables.

Survival analysis models were fit with randomly selected data and then applied to reserved test data to determine the models' sensitivity and specificity in predicting a STEM vs. other outcome. The models performed well in distinguishing between STEM students and those who did not successfully complete a college degree. The different categories of educational outcomes exhibited markedly different hazard curves showing the periods of greatest STEM track departure risk varied between student outcomes. This suggested that the optimum times at which to offer positive interventions to keep students on the potential STEM track vary by the type of outcome they are otherwise likely to experience. This includes encouraging students to remain in high school, to apply to college, and to persist once enrolled in an STEM program.

Introduction

The process of educating students from junior high school through college is of vital importance to the field of engineering education. Producing a sufficient number of engineering graduates depends directly upon the number and quality of students that enter college and select engineering as a major. Students that are lost to engineering by dropping out of high school; choosing not to pursue a college degree; dropping out of college; or switching out of engineering represent "leaks" from the engineering education pipeline. Some of these students would not have ultimately earned an engineering degree due to greater interest in other fields of study, but others might have persisted to graduation with greater preparation, encouragement, and engagement. This study examines the factors that predict whether or not a student will persist to graduate college with a degree in Science, Technology, Engineering, or Math (STEM).

Literature Review

The identification of factors that show significant differences between STEM and Non-STEM students has been the subject of much prior study. Sax[1] studied students that achieved a bachelor's degree in a STEM subject to determine the likelihood they would go on to pursue a scientific research career. She explored differences by gender in the students' persistence in a scientific research career. Persistence in Sax's research was defined as students who achieved a bachelor's degree in a STEM major continuing their STEM education until they earned a STEM graduate degree. A study by Smyth and McArdle[2] used Cooperative Institutional Research Program (CIRP) [3] data to explore racial/ethnic and gender differences in students achieving a STEM degree from selective institutions.

Nicholls, Wolfe, Besterfield-Sacre, Shuman, and Larpkiattaworn[4] created a methodology for rapidly analyzing CIRP data to identify variables that predicted interest in a STEM major. Leslie, McClure, and Oaxaca[5] developed models using data from CIRP and the National Longitudinal Survey of Youth (NLSY) [6] to predict achieving an engineering or science degree. Besterfield-Sacre, Atman, and Shuman[7] developed a methodology for assessing the attitudes, self-confidence, and expectations of freshmen engineering students and identified interest in engineering as a significant factor in persistence in vs. switching out of engineering.

Among the different studies, certain findings stood out. A student that is stronger academically, has greater personal motivation to study STEM, and that has greater confidence in his or her academic ability is more likely to pursue a STEM degree. Since many of the prior studies have found a core set of significant variables that predicted success or failure in earning a STEM degree a logical extension was to determine if students at risk of dropping off the STEM degree track could be identified at an early point in their education and offered supportive intervention programs. Analysis of longitudinal data from students' progression through high school and college offered a means to model the likelihood of students departing the STEM track by a fixed point in their educational careers.

Survival analysis techniques have been applied to longitudinal data in order to identify factors predictive of students ultimately experiencing a general event of interest. Chimka, Reed-Rhoads, and Barker[8] used proportional hazards models to identify variables that showed significant differences in engineering students persisting to college graduation. SAT math scores, science ACT scores, and gender proved valuable in examining variation in student graduation when controlling for other covariates such as major, hometown population, and in-state residence. Zwick and Sklar[9] constructed discrete-time survival models[10] to estimate the conditional probability of a student graduating with a bachelors degree based on students SAT scores and high school grade point averages by ethnicity (White, Black, or Hispanic) and first language (English or Spanish). They found significant differences in the probability of graduation between the different ethnicity/language groups with White/English students more likely to graduate within 15 quarters of enrollment than Black/English, Hispanic/English, or Hispanic/Spanish students, in that order.

Mensch and Kandel[11] utilized the National Longitudinal Survey of Youth (NLSY) dataset to analyze the impact of drug involvement upon dropping out of high school. Civian[12] used survival analysis techniques to explore the time to complete a doctorate at the Harvard

University Graduate School of Education (HGSE). Willett and Singer[13] stated that educational researchers should employ survival analysis techniques in order to study topics such as student persistence and teacher attrition. The article maintained that one of the best reasons to apply survival analysis is that standard statistical techniques require knowledge of when the event occurred (the outcome) for each sample member. This is a standard unlikely to be met in studying event times. Regardless of the length of the study, it is probable that some sample members will not experience the event of interest prior to the end of data collection.

The prior education research indicates that the use of survival analysis techniques can be quite powerful in modeling educational event occurrences. The ability to test time-varying predictors as well as time invariant predictors is a particularly valuable benefit of applying survival analysis techniques. The research to date has employed single statistical techniques or a series of nonintegrated single techniques to explore these complex problems. This limits the degree of insight that can be obtained and the potential for decision-making about intervention methods. Ideally, analysis should be able to pinpoint the most critical time to initiate educational interventions as well as the set of predictors that describe which students would benefit most.

Methodology

Developing a model to predict between a STEM outcome vs. another outcome for a given student involves using data to discriminate between the two potential results. A valuable tool in assessing the accuracy of the discrimination is Receiver Operating Characteristics (ROC) curve[14] analysis. The prediction accuracy is a tradeoff between sensitivity and specificity. Sensitivity is the probability of correctly identifying a STEM student while specificity is the probability of correctly identifying a student having an outcome other than STEM. Such "Not-STEM" students were considered to be part of the "All Else" category. Classifying a student outcome as STEM vs. All Else is based upon the value of a prediction threshold. Consider a threshold value between [0, 1] where a prediction represents the probability of a STEM outcome. The threshold value or "cutpoint" determines which of two outcomes the model predicts. If the cutpoint is set to 0.5 then records for which the model estimates a probability of a STEM outcome ≥ 0.5 will be classified as a STEM prediction.

Records for which the model estimates a probability of a STEM outcome < 0.5 will be classified as a Not-STEM prediction. If the cutpoint is set to a low value, then the model will predict more students to have a STEM outcome. As a result more true STEM students will be correctly predicted to have a STEM outcome but correspondingly, more true Not-STEM students will be incorrectly predicted to have a STEM outcome. If the cutpoint is set to a high value, then fewer students will reach that value and be predicted to have a STEM outcome. Thus fewer true STEM students will be correctly predicted to have a STEM outcome and correspondingly fewer true Not-STEM students will be incorrectly predicted to have a STEM outcome. The choice of the cutpoint value determines the results in discriminating between the two potential outcomes.

ROC curves visually depict the tradeoff by plotting sensitivity vs. (1 – specificity) for a range of cutpoint values. Plotting (1 – specificity) on the horizontal axis and sensitivity on the vertical axis produces a curve line. Ideally, the ROC curve should resemble a vertical line at a low value of (1 – specificity) which transitions to a horizontal line over the remaining values of (1 –

specificity). ROC curves with a steep gradient mean that a very high value of sensitivity with a correspondingly low value of (1 – specificity) is attainable. They indicate that a cutpoint can be chosen which offers a high probability of true positive detection and correspondingly low probability of false positive prediction. If an ROC curve contained the point (0, 1) it would indicate the model could be set to a cutpoint that would provide perfect predictive ability.

ROC curve analysis allows prediction models to be evaluated by examining the resulting ROC curve for a wide range of cutpoints. Models which result in an ideally shaped ROC curve have better ability to discriminate between two potential outcomes than those with a flatter ROC curve. A shallow ROC curve that resembles a 45 degree line between the axes implies that the model has negligible discrimination value. Such a model is as likely to predict a true positive as a false positive and has no useful predictive ability. A visual estimate of the model's predictive ability may be gained by comparing the ROC curve to a 45 degree line and determining how much space lies between the two curves.

The area under the ROC curve ("AUC" or "c") provides an estimate of the model's predictive ability. The sensitivity and (1 – specificity) range from 0 to 1 so the ROC curve is plotted within the unit square formed by the points (0, 0), (0, 1), (1, 0), and (1, 1). Therefore, the area under the ROC curve is a portion of the unit's square area of 1, and c is a value between 0 and 1. A model with a high value for AUC is judged to better discriminate between the outcomes. Hosmer and Lemeshow[15] indicate that a value for AUC of 0.5 indicates that the model is of little use since it is as likely to correctly predict a binary outcome as flipping a fair coin. A result of $0.7 \leq AUC < 0.8$ represents "acceptable" ability to discriminate between potential outcomes. A result of $0.8 \leq AUC < 0.9$ represents "excellent" ability to discriminate between potential outcomes. A result of $AUC \geq 0.9$ represents "outstanding" ability to discriminate between potential outcomes.

The ROC curve can be used to improve a predictive model by selecting a cutpoint for outcome discrimination that provides a good tradeoff between sensitivity and specificity. The selection of a preferred cutpoint to use in discriminating between outcomes is based on the objectives of the analyst in developing the model. If the goal is to optimize sensitivity and specificity then both are plotted against the range of cutpoint values and the cutpoint value at the intersection of the curves is selected. If the goal is to maximize the correct prediction of the outcome of interest, then the cutpoint can be chosen without regard to the probability of incorrect predictions. If the goal is to maximize the correct prediction of the outcome of interest subject to a constraint, then the cutpoint may be chosen to achieve the highest sensitivity probability without violating the constraint. For example, the constraint may be budgetary. If a cost is associated with false positive predictions, then the cutpoint is limited by the probability of a false positive applied to the population of interest.

Data Collection

The selection of data for this study was narrowed to a choice between the CIRP surveys of incoming freshmen and the U.S. Department of Education's National Center for Education Statistics (NCES). NCES has conducted a series of extensive longitudinal studies to collect data about selected students, their families, their schools, and their teachers in each decade since the 1970's. Among them is the National Education Longitudinal Study of 1988 (NELS:88)[16]

conducted between 1988 and 2000. The NELS:88 study consisted of collecting demographic, attitudinal, experiential, educational, and vocational data about a representative cohort of American students at specific stages in their scholastic progression. The goal of the study was to be able to draw conclusions about the factors that could affect the student's progression and achievement by 2000. Academic performance was validated by obtaining transcripts from post-secondary school attended and by conducting cognitive learning tests in three waves of data collection during high school. Parents, teachers, and school administrators were also invited to complete surveys of questions regarding specific students participating in the study. In contrast to the CIRP survey, NELS:88 data was collected at periodic intervals during the high school years, during the likely midpoint of college, and after most students had completed their post-secondary education. There were selected follow-up surveys of the CIRP data, but the scope was more limited than that offered by the NELS dataset. Ultimately, the NELS dataset was chosen for this study.

Analysis

The survival analysis model was constructed to examine the "survival" of students on the STEM track. In this design, the educational event of interest was the point at which students depart the STEM track. This departure could have occurred at any point along the educational progression that began in 1988 with the 8th grade and continued until the study concluded in 2000. The hazard function in this sense was the conditional failure rate or the "approximate" probability that a student with a given profile at a point in time departed the STEM track in the next moment. What was measured was the time to failure with departing the STEM track considered failure.

Classifying Student Outcomes by STEM Track Departure Type

There are a number of ways in which students departed the STEM track. For example, students departed the track by dropping out of high school, by not continuing on to college, by dropping out of college, by switching out of a STEM major, by graduating college with a degree in a subject other than STEM, or by pursuing a college degree without completing it by the study's end. Other theoretical departures included dying or declining to participate further in the study, but these departure types were eliminated by the design of the data collection. Students were purged from the study if they died, were not selected for further sample inclusion, could not be located for the fourth and final follow up (F4), or declined to participate in the fourth follow up. The 12,144 records in the NELS:88/2000 dataset reflected all students that were chosen for inclusion in F4 and responded to the survey. Students who actually achieved a STEM degree never experienced the event of interest because they did not depart the STEM track. The 12,144 records were sorted to remove from analysis those students that did not participate in all waves of data collection for the NELS study. This reduced the dataset to a group of 11,328 records.

With survival analysis, if a subject does not experience the event of interest prior to the study's end date, then the data is said to be "right censored." The term means that if the event of interest occurred, it had to have happened after the study's data collection ended. This was the case for students that were still pursuing a STEM degree at the time the study ended. Since the NELS study ended in 2000, the records for students who had not yet experienced the event of dropping out of STEM were right censored as of December 31, 2000. Another class of right censoring is

"competing risks" censoring in which some subjects experience the event of interest for different reasons. For example, students departed the STEM track by never graduating from high school or by graduating from college with a major other than STEM. These two departure scenarios were competing risks and each was of interest. "Random censoring" is a special case of competing risks in which a student experiences a competing risk that precludes further participation. A student that graduated with a STEM degree was classified as randomly right censored because the competing "risk" of graduating with a STEM degree made it impossible for the person to later experience the event of interest: departing the STEM track.

Another censoring category is "interval censoring." This is utilized when the event of interest is known to have occurred within a fixed time interval, but the exact time cannot be precisely determined. This was the case for students who dropped out of high school or college by a certain point but for whom the exact departure date was unknown. The NELS study was designed to try to elicit information to gauge when students dropped out of high school or stopped attending college; however, it was not always possible to obtain this information. For some of the records, the student's educational status changed within a time interval and the exact time was not determinable.

The classification of students by STEM track departure type utilized the post secondary educational transcript (PETS) dataset that accompanied the restricted version of the NELS dataset. The PETS data contained additional variables related to the timing and nature of educational events from high school through college attendance that were necessary for obtaining the STEM departure time event data needed for survival analysis. The use of this data allowed the student outcomes to be more precisely identified so that students earlier classified as having No Degree could be further sorted by whether they had graduated high school or ever attended college. The process of sorting the students by STEM track departure type resulted in shifting some students between categories as degrees that were reported were in some cases not confirmed by the transcript data. In several other cases, students that had not reported four year degrees were found to have earned them. Each student's record was tested by SAS™ code designed to consult both the main NELS dataset and the PETS dataset to ascertain the final educational outcome. If students reported degrees but did not have valid dates of graduation in one dataset, but they had valid dates in the second dataset they were categorized as having earned a degree. If the information from the two datasets conflicted in a way that could not be resolved, the records were excluded from further analysis. This was the case for 200 of the 11,328 records so the data used for the survival analysis model contained 11,128 records. A summary of the different types of STEM track departures and how they were handled in the model is detailed in Table 1.

Table 1 Determining the Time of STEM Track Departure by Type of Departure

STEM Track Departure Type	Number of Students	Time of STEM Track Departure
Drop Out of High School	567	Final date enrolled in high school or 1991 if missing data
Conclude Education at High School	2,369	High School Graduation date
Drop out of College	2,164	Date of last post-secondary educational enrollment
Incomplete Degree	433	Date of study end: December 31, 2000
Graduate College with a Sub 4 Year Degree	1,703	Date of college graduation
Graduate College with Other 4 Year Degree	3,156	Date of college graduation
Obtain a STEM Degree	736	Date of graduation with STEM degree
Total	11,128	

Deriving the Time to Event Occurrence

The time to the event of interest was determined from the STEM track departure time and the starting or "origin" point in time. Two logical choices for this origin point were the study start time and the students' date of birth. The choice of the origin point affects the estimates of any coefficients in a model as well as its fit. Since the NELS study began in 1988 when the students were in 8th grade and concluded on December 31, 2000 the maximum time to event could range from approximately 12 years using the study start time as the origin to approximately 28 years using the students' individual birth dates. The argument in favor of using the study start time is that it focused the analysis on the time period during which the data was being collected. Factors which affected the students prior to the study were not determinable and automatically lengthening the time to event by 12-14 years might obscure subtle variations in the data during the study time. More importantly, if we assume that each student is a potential STEM graduate, the opportunity to depart the STEM track in this analysis does not start until the study's initiation. While a student could theoretically have a particularly weak academic preparation prior to 8th grade such that the risk of departing STEM was elevated prior to the study's start, there is no data from earlier time periods available to quantify this risk. Based on this reasoning, the study start time was used as the survival analysis "origin" time.

Constructing Fit and Test Data Samples

The samples for the survival analysis model were created by using a set of eleven random number seeds to select separate fit and test samples. The first seed was used as the "original" seed for analysis, and the others were used to assess the consistency of the model when applied to different groups of records. The 11,128 records selected for analysis were first sorted by the students' study assigned ID number. Then a variable, "STEM_Outcome," was created to use in stratifying the population into STEM and All Else outcomes. The STEM_Outcome variable was set to "1" if the STEM track Departure_Type variable = 7 for a STEM degree outcome and "0"

for all other departure types. Then the random number seeds were used to create fit data samples by randomly selecting 70% of the data from the STEM and All Else strata. Similarly, the test data samples were created from the 30% of the records not previously selected for the fit data samples. The result of this was eleven randomly chosen samples containing 7,791 records to fit models with and eleven randomly chosen samples containing 3,337 records to test the models developed. Table 2 provides an example of the breakdown of the student outcomes within a test sample created using the first random number seed.

Table 2 Numbers of Students Classified by Departure in Test Sample for the Original Seed

Category	Number of Records
STEM	220
All Else:	3,117
Other 4 Yr Degree	935
Sub-4 Yr Degree	525
Incomplete Deg.	141
College Dropout	629
H.S. Graduate	720
H.S. Dropout	167
Total Records	3,337

Model Construction

The Proc LIFEREG[17] procedure in SAS was used to fit log-normal, log-logistic, gamma, exponential, and Weibull models using a set of 76 covariates previously found to be significant predictors when constructing a logistic regression model to predict the STEM vs. Non-STEM student outcomes. The log-logistic models consistently provided values with a smaller negative magnitude indicating better models. The LIFEREG procedure of SAS was used to build log-logistic models for each of the 11 fit data samples. Covariates whose estimated coefficients were not significantly different than 0 according to a chi-square test at the $\alpha = 0.05$ level were dropped from the model and the modeling was repeated with the subset of previously significant covariates. Subsequent iterations continued until all model covariates were found to be significant and the likelihood ratio statistic confirmed that the global test of model significance was met.

Once a final model for each fit data sample had been created, it was applied directly to the records to estimate the probability of survival on the STEM track past time 7.25 years. This point in time was chosen after examining the estimated hazard functions for the different sub-populations of students categorized by STEM track departure type. Selecting a time earlier in the study would have resulted in a high probability of continuing to remain on the STEM track in the next instant for most of the students. Even selecting a time after most students had graduated high school and begun college would not have improved the ability to discriminate the STEM vs. All Else student sub-populations since most would still have had a high probability of surviving then. The cumulative hazard functions for the four year degree students were closely related with an intersection at approximately year 7. The high school dropouts had a markedly different hazard function with all the students in this group departing by year 5 of the study. The high school graduates departed the STEM track at an increasing pace with a sharp jump during year 4 when most earned their high school diplomas. The college dropouts and students who earned a

sub-4 year degree had very similar hazard functions that were dissimilar to the other hazard functions until after year 8.

Defining the departure times for college graduates as the graduation date meant that most of the students that completed bachelor's degrees in any subject earned them in the year 1997. The mean departure time for the STEM students was 1997.42 compared to 1997.00 for the Other Degree students and 1997.04 for all four year degree students. The standard deviation of the departure times was 4.36 years for the STEM students, 1.04 years for the Other Degree students, and 4.89 for all four year degree students.

Analysis of the hazard functions suggested that most of the students that departed the STEM track tended to do so by 6.5 years past the origin time of 1987.92. By the early months of year 8 most of the other four year degree students and STEM students had graduated from college. The analysis suggested the points at which the probability of survival on the STEM track was the highest for the STEM students and correspondingly lower for the other students was within the window of 6.33 to 8.17 years. The fit data for the original seed was repeatedly modeled to estimate the survival probability beyond time points 6.33, 6.41, 6.67, 7.25, 7.33, 7.41, 7.5, 7.67, 8.0, and 8.17 years. Based on the probability of survival past a given point, the student was predicted to have a STEM outcome for higher vs. lower probability values. Various cutpoints were used. The sensitivity and specificity of the predictions did not vary much, but slight improvements were found using 7.25 years. Thus this was the point in time used to discriminate between the STEM and All Else students based on the probability of remaining on the STEM track. The final log-logistic model fitted via Proc LIFEREG for the original seed fit data was applied to the original seed test data. The final models for the other ten randomly chosen fit datasets were applied to their associated test datasets in the same manner. Predicted outcomes were made in response to different cutpoints within the interval (0, 1).

Findings and Conclusions

A survival analysis model was constructed for each of the eleven random fit data samples and then tested with the reserved test data samples. Overall, the survival analysis provided good predictive accuracy in distinguishing between STEM and All Else students. The ROC curve for the predictive accuracy of the survival analysis model constructed with the original seed sample is shown in Figure 1. For example, approximately 80% of the 220 STEM students in the sample were correctly identified in conjunction with 40% of the 3,117 All Else students incorrectly predicted to have a STEM outcome. The results for the other ten random seed samples were similar.

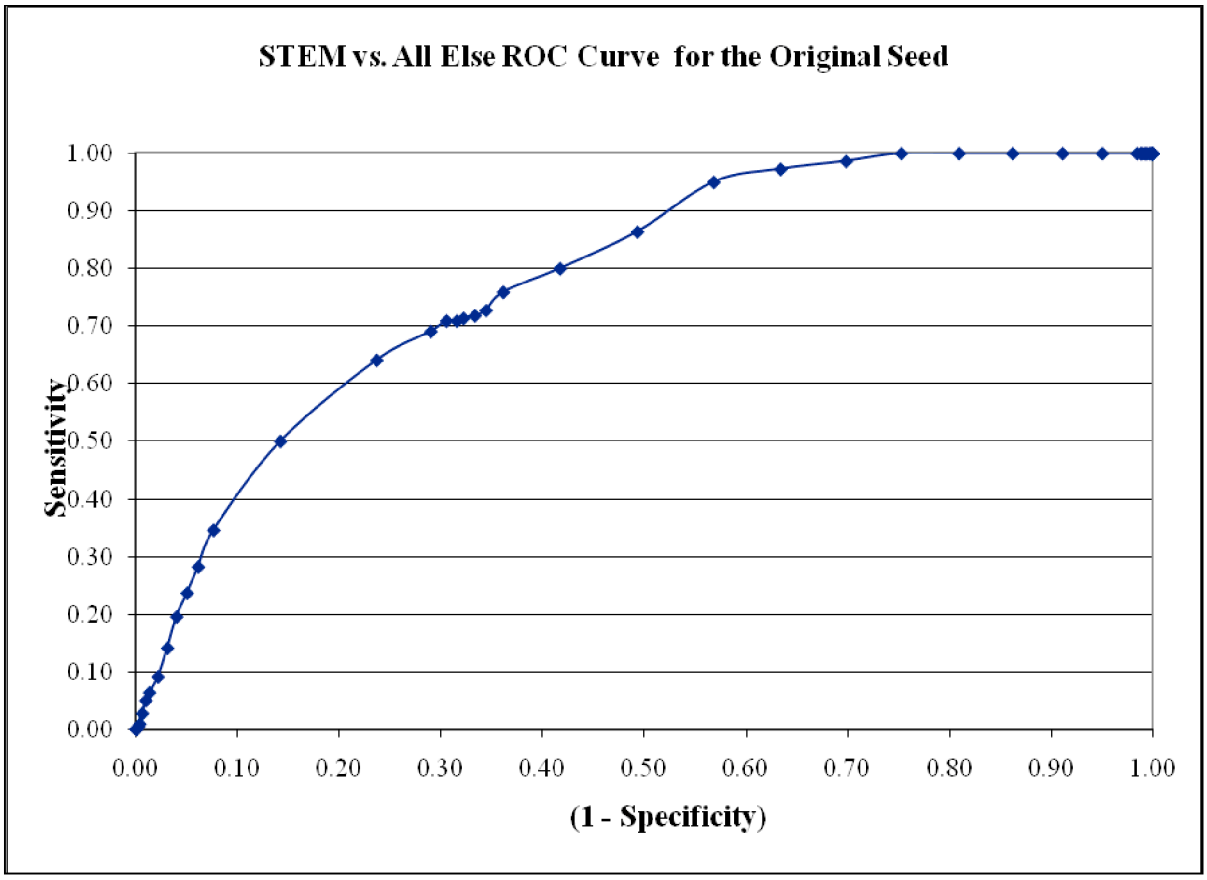

Figure 1 ROC Curve for the STEM vs. All Else Survival Analysis Model with the Original Seed

One of the advantages of examining the data with ROC Curve analysis is the ability to "tune" the model by selecting different cutpoint threshold values for the prediction of a STEM vs. All Else outcome. Figure 2 shows the tradeoff between sensitivity and specificity achieved at different cutpoint threshold values. Equivalence between sensitivity and specificity was achieved at approximately 70%. This provides a means of targeting potential STEM students for a pro-STEM intervention and selecting the cutpoint at which the greatest number of potential STEM students can be targeted. Limitations on the funding for potential intervention programs can be used to select a cutpoint that maximizes the student audience without exceeding the budget. A portion of the students targeted for such an intervention may be encouraged to pursue a STEM vs. Non-STEM degree.

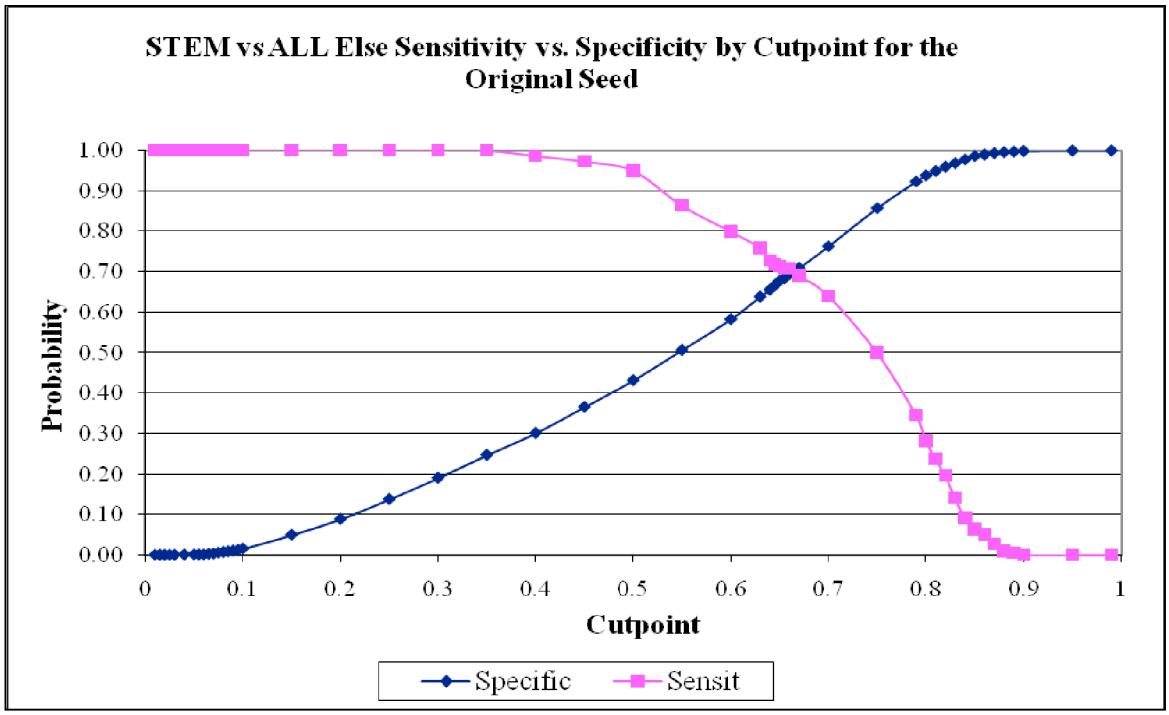

Figure 2 Sensitivity vs. Specificity for STEM vs. All ELSE Survival Analysis Model using the Original Seed

At the point of equivalence between sensitivity and specificity, the model correctly predicted 168 of the 220 STEM students while it incorrectly predicted 952 of the 3,117 All Else students to have a STEM outcome. Table 3 indicates the breakdown of the false positives by the actual student outcome and indicates how these outcomes were represented in the test sample. The All Else students that were incorrectly predicted to have a STEM outcome were mainly the students that earned another four year degree, those who earned a less than four year degree, and those who had dropped out of college. This suggests that the model was picking up on factors about those students that would have made them good candidates to pursue a STEM degree. Such students would be good candidates for a pro-STEM intervention program to encourage their interest in and motivation to study a STEM major in hopes of expanding the pool of potential STEM students.

Table 3 False Positive Breakdown by STEM Track Departure Category for the Original Seed

STEM Track Departure Category	# of Dep. Category in Sample	% of Other Than STEM	# of False Positives	% Type of False Positive	% of False Positive by Category
H.S. Dropout	167	5.36%	0	0.00%	0.00%
H.S. Graduate	720	23.10%	51	5.36%	7.08%
College Dropout	629	20.18%	146	15.34%	23.21%
Incomplete Deg.	141	4.52%	57	5.99%	40.43%
Sub-4 Yr Degree	525	16.84%	101	10.61%	19.24%
Other 4 Year Degree	935	30.00%	597	62.71%	63.85%
Total	3,117	100.00%	952	100.00%	30.54%

One aspect of the data that may offer a potential for improving the survival analysis model's predictive accuracy in future research is that the survival times of the 736 STEM students and the 3,156 students who earned bachelors degrees in other subjects were very similar in this study. The similarity of the survival times may have hampered the survival analysis model by making it harder to discriminate between the STEM and All Else students. The similar mean values for the departure times of the bachelor degree holders made it more difficult for the model to distinguish between the groups despite treating the departure times for the STEM students as censored rather than observed events.

One approach that might improve the power of the survival analysis model would be changing the STEM track departure time for students that earned other college degrees. If the time at which the students declared a major other than STEM were defined as the departure time, there would be a greater disparity in the departure times between STEM and Other Degree students. In addition, this approach would better reflect the phenomena of students starting a STEM major and later switching to a different major. In future research, it would be worthwhile to explore the feasibility of collecting this data. In the absence of information showing when a student declared a Not-STEM major or switched from a STEM major, the date the Other Degree students began attending college could be tested as a revised interpretation of the departure time.

This study found that survival analysis could be applied to the STEM degree acquisition process, and it provided valuable insights into variations between different groups of students over time in the probability of earning a STEM degree. These insights were obtained in the process of addressing the following questions: "Will Survival Analysis of the NELS:88 data reveal that the probability of a student achieving a STEM degree differ over time for students in different outcome groups?" and "Are there key time points in the educational process where distinct decreases or slight increases in the probability of achieving a STEM degree occur as students developed academically? If so, are these key time points at which students were most likely to depart the STEM track sufficiently common for different student profiles that they could suggest the timing for delivery of pro-STEM intervention?"

Differences in the STEM probabilities by outcome group were found. As discussed earlier, the hazard functions for the STEM and Other Degree students were similar at some time points and divergent at others. These two groups had hazard functions that were clearly different from the other departure types.

It was envisioned that key time points in the educational process could be found to exhibit distinct changes in the probability of earning a STEM degree over time across different groups of students that could indicate the best timing for a pro-STEM intervention. Again, examination of the cumulative hazard functions for each group provides evidence to answer this research question affirmatively. The probability of earning a STEM degree drops as the hazard function increases. The students that graduate high school and go on to attend college did not experience increasing hazard function values until 3 years past the study start in approximately 1991. This suggests that a pro-STEM intervention conducted in high school would be able to target these students prior to their leaving the STEM track. Year 3-5 represents the period at which the

probability of departing the STEM track rises the most sharply for the students that drop out of college or complete less than 4 year degrees.

The hazard functions for the high school dropout and students completing their education by graduating high school exhibit different track departure patterns. The high school dropouts experienced the sharpest increase in the probability of STEM departure at the study's start with a more gradual increase until midway through their junior year. The students that graduated high school experienced a less steep but steady increase in the probability of departure until year 4.

The conclusion reached from examining the hazard functions is that to target potential STEM degree students successfully, the pro-STEM intervention must occur before 8th grade. To reach all of these students, the intervention may have to occur in the 7th grade or earlier. The curves for the students whose educations did not go beyond high school were sufficiently dissimilar to those of the other students that it may be worthwhile to consider developing more than one intervention program. The first would occur prior to 8th grade and would focus on assisting students that would not otherwise be predicted to go on to college. The second intervention program would occur prior to 11th grade and focus on encouraging students that are predicted to attend college to consider pursuing a STEM degree. A potential third intervention program would take place after the first year of college for students as STEM students consider switching to a major outside STEM.

Bibliography

[1] Sax, Linda J., "The impact of college on post-college commitment to science careers: gender differences in a nine-year follow-up of college freshmen," Proceedings, Annual Meeting of the Association for the Study of Higher Education, Memphis, TN, November 1996.

[2] Smyth, Frederick L. and John J. McArdle, "Ethnic and gender differences in science graduation at selective colleges with implications for admission policy and college choice," Research in Higher Education, Vol. 45, No. 4, June 2004, pp. 353-381.

[3] Astin, Alexander W., "Studying how college affects students: a personal history of the CIRP," About Campus, July-August 2003, pp. 21-28.

[4] Nicholls, Gillian M., Harvey Wolfe, Mary Besterfield-Sacre, Larry J. Shuman, and Siripen Larpkiattaworn, "A method for identifying variables for STEM intervention," Journal of Engineering Education, Vol. 96, No. 1, pp. 33-44, January 2007.

[5] Leslie, Larry L., Gregory T. McClure, and Ronald L. Oaxaca, "Women and minorities in science and engineering: a life sequence analysis," The Journal of Higher Education, Vol. 69, No. 3, May/June 1998, pp. 239-276.

[6] United States Department of Labor, Bureau of Labor Statistics, National Longitudinal Survey of Youth 1979 (NLSY79), http://www.bls.gov/nls/nlsy79.htm.

[7] Besterfield-Sacre, Mary, Cynthia J. Atman, and Larry J. Shuman, "Characteristics of freshman engineering students: models for determining student attrition in engineering," Journal of Engineering Education, Vol. 86, No. 2, April 1997, pp. 139-149.

[8] Chimka, Justin R., Teri Reed-Rhoads, & Kash Barker, "Proportional hazards models of graduation," Journal of College Student Retention: Research, Theory, & Practice, Vol. 9, No. 2, pp. 221-232, 2007-2008.

[9] Zwick, Rebecca & Jeffrey G. Sklar, "Predicting college grades and degree completion using high school grades and SAT scores: the role of student ethnicity and first language," American Educational Research Journal, Vol. 42, No. 3, Autumn, 2005, pp. 439-464.

[10] Cox, D.R., "Regression models and life tables," Journal of the Royal Statistical Society, Series B, No. 34, pp. 187-202.

[11] Mensch, Barbara S. & Denise B. Kandel, "Dropping out of high school and drug involvement," Sociology of Education, Vol. 61, April 1988, pp. 95-113.

[12] Civian, Janet Trabucco, "Examining duration of doctoral study using proportional hazards models", unpublished dissertation, Harvard University Graduate School of Education, 1990.

[13] Willett, John B. & Judith D. Singer, (1991), "From whether to when: new methods for studying student dropout and teacher attrition", Review of Educational Research, Winter, 1991, Vol. 61, No. 4, pp. 407-450.

[14] Fawcett, Tom, "An introduction to ROC analysis," Elsevier B.V, 2005, http://www.csee.usf.edu/~candamo/site/papers/ROCintro.pdf.

[15] Hosmer, David W. and Stanley Lemeshow, Applied Logistic Regression 2nd edition, John Wiley & Sons, Inc., New York, NY, 2000, pgs. 160-164.

[16] National Center for Education Statistics, National Education Longitudinal Study of 1988, Project Officers: Peggy Quinn and Jeffrey T. Owings (Washington, DC), http://nces.ed.gov/surveys/nels88/index.asp.

[17] Allison, Paul D., Survival Analysis Using SAS: A Practical Guide, SAS Institute Inc., Cary, NC, 1995, pages 61-109.