Proportional Hazards Model for Assessing Risk Factors in Undergraduate Calculus

V. J. DuRapau, Jr., Ph.D.

Xavier University of Louisiana
Department of Mathematics
New Orleans, LA 70125

Presented at the

17th Annual International Conference on Technology in Collegiate Mathematics
in New Orleans, LA

October 29, 2004

Abstract

The purpose of this paper is to present a methodology that may be used to investigate risk factors in undergraduate Calculus. A challenge in assessing the effectiveness of a first semester Calculus course is the relatively large number of students who withdraw during the semester. Techniques such as correlation and linear regression are hampered by the loss of data due to such withdrawals. This study applies survival analysis methodology to examine risk factors contributing to student failure to successfully complete first semester undergraduate Calculus. The concept of censored data is discussed, and the proportional hazards model

\[ h_i(t) = [h_0(t)] e^{\beta_0 x_{i1} + \ldots + \beta_p x_{ip}}, \quad i = 1, 2, \ldots, n \]

is introduced. Some of the underlying theory is discussed, including Cox regression to estimate parameters of the model. Finally, Cox regression is applied to Fall 2003 data from a private, liberal arts university to investigate risk factors in Calculus 1.
Contents

1 Introduction 3

2 Methodology 3
   2.1 Success vs. Failure ("Unsuccess") 3
   2.2 Censored Data 3
   2.3 Distribution and Density Functions 4
   2.4 Survival and Hazard Functions 4
      2.4.1 Continuous Failure Time 5
   2.5 Proportional Hazards Models 5
      2.5.1 Model Specification 6
      2.5.2 Model Assumptions 6
      2.5.3 Survival and Hazard Curves 6
      2.5.4 Parameter Estimation 7

3 Application to Survival Data in Calculus 1 7
   3.1 Event of Interest 7
      3.1.1 Failure To Earn At Least D 8
      3.1.2 Failure To Earn At Least C 8
      3.1.3 Status Variable Selection 8
   3.2 Proxy for Time 8
   3.3 Proportional Hazards Model Specification 8

4 Results 11
   4.1 Descriptive Statistics 11
   4.2 Results from Cox Regression 11
      4.2.1 Baseline Hazard Function 12
      4.2.2 Survival and Hazard Functions 12
      4.2.3 Parameter Estimates 13
      4.2.4 Interpretation of the Final Model 15

5 Discussion 16

A Appendix: Variables in Calculus Database 17

---

1 For a discussion of discrete failure time (or time to event of interest), see [3, pp. 4-5]
1 Introduction

One or more semesters of Calculus often is required of college students in such disparate majors as biology, chemistry, computer science, economics, pharmacy, physics, and pre-medicine. To be successful in their chosen careers, students may need the basic knowledge of and skills in Calculus. Other fields may use Calculus as a "gatekeeper."

A major problem in analyzing factors contributing to students’ successful completion of a first semester Calculus course is the large number of students who withdraw from the course during the semester. In any given semester, more than a third of students enrolling in Calculus may withdraw during the semester or fail to earn a passing grade. Traditional statistical approaches (e.g., correlation and linear regression) to determine factors leading to successful completion of Calculus may give biased results due to the large number of withdrawals during the semester.

Statistical techniques under the rubric of survival analysis offer ways of handling such censored data.[7, pp. 155-166] The purpose of this paper is to suggest methods from survival analysis to address these challenges.

2 Methodology

This is a longitudinal study that involves observations of students over a one-semester period (Fall 2003) at a small liberal arts university. "Several analytical problems arise when standard research tools such as ordinary least squares (OLS) regression are used to study longitudinal data. These problems include how to incorporate explanatory variables that change over time and how to deal with observations that are censored."[3, p. 2]

2.1 Success vs. Failure ("Unsuccess")

One way to address this challenge is to focus on the complementary event unsuccessful in completing the course. The "unsuccess" event occurs if the student (a) earns a final letter grade of F or FE [failure due to excessive absences] at the end of the semester or (b) withdraws during the semester with a failing average.[4, p. 6]

The variable of interest in this study is unsuccessful in completing Calculus 1, or simply "unsuccess", and when that event occurs.

2.2 Censored Data

Censoring occurs when the outcome or event of interest is not known for an individual during the period of observation. Consider four cases of students in Calculus 1 during Fall 2003 (Figure 1). Case 1 is a student who completed the semester with a failing grade F. Cases 2 and 3 represent students who withdrew during the semester, the first with a grade of C and the second with a grade of F. Case 4 represents a student who completed the semester with a grade of A. The event of interest is failure to successfully complete the course (or simply "unsuccess"). In case 1, the event occurred during the period of observation (at the end of the semester) so no censoring occurred. Case 2 is an example of censoring. The student withdrew during the semester but had a C average to date in the course. This student is no longer at risk of failure. The student in case 3 withdrew during the semester and had an F average in the course at the time of withdrawal. The event "unsuccess" occurred for this student, so no censoring occurred. The student in case 4 completed the course with an A. The event "unsuccess" did not occur. This student is no longer at risk of failure. However, in
In the context of survival analysis terminology, this case is considered censored.

**Censoring**

![Diagram of censoring cases](image)

**Fall 2003**

Figure 1:

Cases 2 and 4 are referred to as right censored. The data analyzed in this study do not involve left censoring. Left censoring can arise in some types of longitudinal data and presents difficulties both statistically and practically.[3, p. 3]

### 2.3 Distribution and Density Functions

For a continuous random variable $T$, the cumulative distribution function (cdf) (or distribution function) and probability density function (pdf) (or density function) are defined as follows:

**Definition 1** The cumulative distribution function of $T$, denoted by $F(t)$, is given by

$$F(t) = P(T \leq t) \quad \text{for} \quad -\infty < t < \infty \quad (1)$$

**Definition 2** Let $F(t)$ be the distribution function for a continuous random variable $T$. The probability density function of $T$ is

$$f(t) = \frac{d}{dt}F(t) = F'(t) \quad (2)$$

wherever the derivative exists and 0 elsewhere.

(See [12, pp. 151-155] for a discussion of the properties of these functions.)

### 2.4 Survival and Hazard Functions

Let $T$ be a nonnegative random variable representing the time to an event of interest (e.g., death, failure of a machine component, failure to successfully complete *Calculus 1*) for an individual in the population. The random variable $T$ often is referred to as failure time.

**Definition 3** The survival distribution function (or survival function) of $T$ is

$$S(t) = \Pr(T > t) \quad (3)$$
The survival function\(^2\) is the probability that the event of interest (e.g., failure) occurs after time \(t\), that is, the individual survives past time \(t\). The survival function may be expressed in terms of the cumulative distribution function \(F\) of the random variable \(T\).

\[
S(t) = 1 - Pr(T \leq t) = 1 - F(t) .
\]

(4)

An alternative, but less intuitive, way to express the distribution of \(T\) is by means of its hazard function \(h(t)\). The random variable \(T\) may be either continuous or discrete. However, in discussing some of the underlying theory in this paper, I consider only continuous \(T\).

### 2.4.1 Continuous Failure Time\(^3\)

**Definition 4** If \(T\) is a continuous, non-negative random variable representing failure time (or time to some event of interest), the hazard function is the instantaneous failure rate at time \(t\):

\[
h(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t} .
\]

(5)

A hazard function uniquely determines the probability density function (pdf) (or density function), \(f(t)\), of \(T\).[8, p. 407] The product of the hazard and survival functions is equal to the density function \(f\). This relationship is an immediate consequence of the following theorem.

**Theorem 5** If \(f(t)\) is the probability density function of the random variable \(T\), representing time to failure, then the hazard function is

\[
h(t) = \frac{f(t)}{S(t)} .
\]

(6)

**Proof.** Let \(F(t)\) represent the cumulative distribution function of \(T\).

\[
h(t) = \lim_{\Delta t \to 0^+} \frac{Pr(t < T \leq t + \Delta t \mid T > t)}{\Delta t}
\]

\[
= \lim_{\Delta t \to 0^+} \frac{Pr([t < T \leq t + \Delta t] \cap (T > t)]}{\Delta t \cdot Pr(T > t)}
\]

\[
= \frac{1}{Pr(T > t)} \lim_{\Delta t \to 0^+} \frac{Pr(t < T \leq t + \Delta t)}{\Delta t}
\]

\[
= \frac{1}{S(t)} \cdot \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}
\]

\[
= \frac{1}{S(t)} \cdot \frac{d}{dt} [F(t)]
\]

\[
= \frac{f(t)}{S(t)} .
\]

(See [11] for a discussion of right-hand (and left-hand) derivatives.) ■

### 2.5 Proportional Hazards Models

The Proportional Hazards Model (PHM) was first proposed by Cox in 1972[2] and has been used in a wide range of applications to study event times. Among the applications listed by Seetharaman and Chintagunta in their extensive review of recent applications are[9, p. 39]

---

\(^2\)Another name for the survival function is reliability.[6, p. 188]

\(^3\)For a discussion of discrete failure time (or time to event of interest), see [3, pp. 4-5]
time until dissolution of marriage  survival time of cancer patients
    time to stroke attack           inter-purchase time
premarital cohabitation        job turnover
unemployment spell             leadership tenure
time to spread for infectious disease  time to death for smokers
labor market transition        animal mortality.

2.5.1 Model Specification

The proportional hazards model (PHM) assumes that the time to event and the covariates are related by the equation

\[ h_i(t) = [h_0(t)] e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}, \quad i = 1, 2, \ldots, n \]

where

- \( h_i(t) \) is the hazard for the \( i \)-th case at time \( t \)
- \( h_0(t) \) is the baseline hazard at time \( t \)
- \( p \) is the number of covariates
- \( \beta_j \) is the value of the \( j \)-th regression coefficient
- \( x_{ij} \) is the value of the \( i \)-th case of the \( j \)-th covariate.

The hazard function is "multiplicatively decomposable as two components – the baseline hazard and the covariate function."[9, p. 3] The hazard and survival functions also are related in the following way:[10, Cox regression case studies, p. 2]

\[ S_i(t) = e^{-\int_0^t h_i(u) du} = \exp \left\{ - \int_0^t h_0(u) e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} du \right\} \]

2.5.2 Model Assumptions

The hazard function is a measure of the potential for the event to occur at a particular time \( t \), given that the event did not yet occur. The larger the value of the hazard function, the greater the potential for the event to occur. The baseline hazard measures this potential independently of the set of covariates. Since only the baseline hazard, and not the covariates effect, is a function of time, the shape of the hazard curve over time is determined by the baseline hazard for all cases.

The ratio of the hazards for any two cases \( i \) and \( i^* \) at any time is

\[ \frac{h_i(t)}{h_{i^*}(t)} = \frac{[h_0(t)] e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{[h_0(t)] e^{\beta_0 + \beta_1 x_{i^*1} + \ldots + \beta_p x_{i^*p}}} = \frac{e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{e^{\beta_1 x_{i^*1} + \ldots + \beta_p x_{i^*p}}} \]

the ratio of their covariate effects. This is the proportional hazards assumption.

2.5.3 Survival and Hazard Curves

Figures 2 and 3 show basic survival and hazard curves from a breast cancer survival study of 1,266 women.[10, sample file Brest_cancer_survival.sav] The time variable \( T \) is in months and the event of interest is died. The horizontal axis for both curves shows time to the event died. In Figure 2, the survival curve, the vertical axis shows the probability of survival. A point on the survival curve shows the probability that the "average" subject will survive past that time. For example, the probability of survival beyond 40 months was approximately 0.95. Beyond approximately 65
months, the survival curve is less smooth. There are fewer subjects who have survived to that point, so there is less information available. The curve begins to look more like a stepwise function.

Figure 2: Survival Function at mean of covariates

Figure 3: Hazard Function at mean of covariates

The basic hazard curve (Figure 3) is a visual display of the cumulative model-predicted potential for the "average" subject to die. The horizontal axis shows the time in months. The vertical axis shows the cumulative hazard. By 8, the cumulative hazard is

\[ \int_0^t h(u) \, du = - \ln S(t) , \]  

(9)

the negative log of the survival probability. Beyond approximately 65 months, the (cumulative) hazard curve, like the survival curve, becomes less smooth and more stepwise for the same reason (less information available).

2.5.4 Parameter Estimation

The method of maximum likelihood is a commonly used procedure to estimate unknown parameters when a representative (e.g., random) sample of data is available. The underlying probability density function (pdf), \( f(t) \), for the population must be known or assumed. The likelihood function is a function of known (from the sample) values \( t_1, t_2, \ldots, t_n \) of the random variable \( T \) and the unknown parameters \( \beta_0, \beta_1, \ldots, \beta_p \). The likelihood function is expressed in terms of \( f(t) \) or \( \ln f(t) \).

Cox regression uses partial maximum likelihood, a variation of the maximum likelihood method.[5, p. 1] The procedure does not require that any parametric structure be imposed on the baseline hazard and thus is considered semiparametric.[9, p. 3]

3 Application to Survival Data in Calculus 1

I used a proportional hazards model to examine risk factors influencing student success or failure in a first semester undergraduate Calculus course (Calculus 1).

3.1 Event of Interest

Final letter grades of A, B, C, D, F, FE [for failure due to excessive absences], or W were recorded by teachers in 10 sections of Calculus 1. The first four (A through D) are passing grades. Some students,
however, need to obtain a grade of at least a C to successfully complete the course. Therefore, I defined two distinct status variables to indicate the event of interest: unsuccessful in completing the course.

3.1.1 Failure To Earn At Least D

\[ Evnt_{WF} = \begin{cases} 
0, & \text{if "no"} \\
1, & \text{if "yes"} 
\end{cases} \] (10)

indicating whether or not the individual (a) withdrew during the semester with an F average to date or (b) earned a final grade of F or FE.

3.1.2 Failure To Earn At Least C

\[ Evnt_{WDF} = \begin{cases} 
0, & \text{if "no"} \\
1, & \text{if "yes"} 
\end{cases} \] (11)

indicating whether or not the individual (a) withdrew during the semester with a D or F average to date or (b) earned a final grade of D, F, or FE.

3.1.3 Status Variable Selection

The analysis of survival data for Calculus 1 was performed twice, first using Evnt_{WF} as the status variable and then using Evnt_{WDF}. Only results of the analysis with Evnt_{WDF} are included in this paper. Thus, the event of interest (failure to successfully complete Calculus 1) was defined as follows:

**Definition 6** The event "failure to successfully complete Calculus 1" (or simply "unsuccess") is defined as (a) withdrew during the semester with a D or F average at the time of withdrawal or (b) earned a final course grade of D, F, or FE [failure due to excessive absences].

3.2 Proxy for Time

Determining an exact time in days that a student withdraws from a class is difficult. Some students simply stop attending class but do not officially withdraw until near the end of the semester. A solution to this time problem is to use a proxy for time to the event of interest (Evnt_{WDF}).

There were five major tests (Mod1 to Mod5) administered during the semester plus the final exam. I used these six events to define the variable Time to Event (Mods), the proxy for time:

\[ T_{Mods} = \begin{cases} 
1, & \text{if event occurred (Evnt_{WDF} = 1) after Mod1 test} \\
2, & \text{if event occurred (Evnt_{WDF} = 1) after Mod2 test} \\
... \\
6, & \text{if event occurred (Evnt_{WDF} = 1) after final exam} 
\end{cases} \] (12)

Although T_{Mods} is discrete, I treated it as a continuous random variable. The implications of this decision need further exploration.

3.3 Proportional Hazards Model Specification

The proportional hazards model used in this study is

\[ h_i(t) = h_0(t) e^{\beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \beta_4x_{i4}}, \quad i = 1, 2, ..., n \] (13)

where
\( h_i(t) \) is the hazard for the \( i \)-th case at time \( t \)
\( h_0(t) \) is the baseline hazard at time \( t \)
\( \beta_j \) is the value of the \( j \)-th regression coefficient \( (j = 0, 1, 2, 3, 4) \)
\( x_{ij} \) is the value of the \( i \)-th case of the \( j \)-th covariate.

There were \( n = 303 \) students and \( p = 4 \) covariates:

- \( x_{i1} = DVMath \) Took Developmental Math \((0 = \text{no}; 1 = \text{yes})\)
- \( x_{i2} = Atmp1070 \) Number of previous attempts of Calculus 1 \((0, 1, \ldots)\)
- \( x_{i3} = PreTstN \) Pre-Test score \((\text{max} = 25)\)
- \( x_{i4} = PCbypass \) By-passed Pre-Calculus \((0 = \text{no}; 1 = \text{yes})\)

The pretest ("Assessment of Basic Mathematical Knowledge and Skills for Calculus 1") was a 25 item multiple choice test administered on the first day of classes. The pretest was authored by mathematics faculty teaching Calculus 1 in Fall 2003. Each question had five choices, one correct and four incorrect. Two-hundred eighty-four (284) students took the pretest. Some of these students dropped the course before the first major test, and those students are not included in the data. Students who were not present on the first day of classes do not have a pretest score. Cronbach's Alpha reliability coefficient\(^4\) for the pretest was 0.616.

I used Statistical Package for the Social Sciences (SPSS) 12.0 for Windows to perform a stepwise Cox regression with \( T_{-}\_Mods \) as the time to event, \( Evnt\_WDF \) as the event indicator\(^5\), and with the entry criteria for the first three covariates set to "Forward:LR", for forward likelihood ratio test. (See Figure 4.) Covariates that are not statistically significant will not appear in the final model.

![Figure 4](image)

Since I wanted the third covariate (\( PCbypass \)) to appear in the final model, I entered it in a separate block with entry method "Enter" specified:

\(^4\)A measure of internal consistency based on the average inter-item correlation. A coefficient of 1 would indicate perfect consistency.

\(^5\)Only output with \( Evnt\_WDF \) as the event of interest is included in this paper. Output with \( Evnt\_WF \) is available upon request.
Both *DV Math* (took *Developmental Math*?) and *PC bypass* (by-passed *Pre-Calculus*?) are categorical variables (1 for "yes" and 0 for "no"). Figure 6 displays the specification of these two variables as categorical. Figure 7 shows the settings in the Cox regression *Plots* window to generate separate survival and hazard functions for each of the two values of *PC bypass* at the mean of any other covariates remaining in the model at the end of the stepwise procedure.

Figure 8 shows the settings in the Cox regression *Options* window: 95% confidence interval for $\exp(\beta) = e^\beta$, correlation of the estimated $\beta$'s, probability for entering a covariate into the model (.05), probability for removal of a covariate from the model (.10), maximum number of iterations in the stepwise procedure (20). Also, the baseline hazard function will be estimated.
4 Results

This section contains (a) descriptive statistics of several variables, including some that are not in the PHM, and (b) results of the Cox regression.

4.1 Descriptive Statistics

Table 1 gives the minimum, maximum, mean, and standard deviation for ACT Math, SAT Math, the university’s Math Placement Test, the Pretest, each of the major (module) tests, the final exam, and the final course grade based on 100. Table 2 displays correlation coefficients for some of the variables.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACT: ACT Math Score</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>SAT: SAT Math Score</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Tot: XU Math Placement Test: Total</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>PreTest: Assessment (Pre-) Test (based on 25)</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Mod1: Mod 1 Test Grade</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Mod2: Mod 2 Test Grade</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Mod3: Mod 3 Test Grade</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Mod4: Mod 4 Test Grade</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Mod5: Mod 5 Test Grade</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Exam: Final Exam Grade (based on 100)</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>NGrd: Final Course Grade (based on 100)</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
</tr>
</tbody>
</table>

Table 1:

<table>
<thead>
<tr>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACT: ACT Math Score</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>SAT: SAT Math Score</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Tot: XU Math Placement Test: Total</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>PreTest: Assessment (Pre-) Test (based on 25)</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>Exam: Final Exam Grade (based on 100)</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td><strong>NGrd: Final Course Grade (based on 100)</strong></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Sig (2-tailed)</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2:

4.2 Results from Cox Regression

Results of the Cox regression are reported in this section.
4.2.1 Baseline Hazard Function

Table 3 shows, for each value of the time variable $T_{Mods}$ the baseline cumulative hazard function. The values of the survival function and cumulative hazard function (not the baseline hazard) at the mean of the covariates also is displayed.

<table>
<thead>
<tr>
<th>Time</th>
<th>Baseline Cum Hazard</th>
<th>Survival</th>
<th>SE</th>
<th>Cum Hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.837</td>
<td>0.990</td>
<td>0.006</td>
<td>0.110</td>
</tr>
<tr>
<td>2</td>
<td>0.219</td>
<td>0.943</td>
<td>0.013</td>
<td>0.059</td>
</tr>
<tr>
<td>3</td>
<td>0.505</td>
<td>0.849</td>
<td>0.021</td>
<td>0.183</td>
</tr>
<tr>
<td>4</td>
<td>0.729</td>
<td>0.821</td>
<td>0.023</td>
<td>0.197</td>
</tr>
<tr>
<td>5</td>
<td>0.762</td>
<td>0.814</td>
<td>0.023</td>
<td>0.206</td>
</tr>
<tr>
<td>6</td>
<td>1.575</td>
<td>0.654</td>
<td>0.028</td>
<td>0.425</td>
</tr>
</tbody>
</table>

Table 3:

4.2.2 Survival and Hazard Functions

The survival and hazard curves are for the "average" student, that is, for the mean values of the covariates as shown in Table 4.

<table>
<thead>
<tr>
<th>Coverate Means and Pattern Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>DVMath</td>
</tr>
<tr>
<td>Agep107</td>
</tr>
<tr>
<td>PreTest</td>
</tr>
<tr>
<td>PCbypass</td>
</tr>
</tbody>
</table>

Table 4:

Survival Functions Figure 9 shows the model-predicted survival for the "average" student, that is, at the mean of the covariates. Figure 10 shows separate model-predicted survival curves for values of $PCbypass$. Students who by-passed Pre-Calculus and went directly into Calculus 1 ($PCbypass =$
1) have a survival curve that is higher than those who did not by-pass Pre-Calculus ($PC_{bypass} = 0$).

Figure 9: Survival Function at mean of covariates

Figure 10: Survival Function for patterns 1 - 2

Hazard Functions  Figure 11 displays the model-predicted cumulative hazard curve for the "average" student, that is, for average values of the covariates. Figure 11 shows separate model-predicted cumulative hazard curves for values of $PC_{bypass}$. Students who by-passed Pre-Calculus and went directly into Calculus 1 ($PC_{bypass} = 1$) have a cumulative hazard curve that is lower (less risk) than those who did not by-pass Pre-Calculus ($PC_{bypass} = 0$).

Figure 11: Hazard Function at mean of covariates

Figure 12: Hazard Function for patterns 1 - 2

4.2.3 Parameter Estimates

Table 5 shows the case processing summary. The status variable ($Event\_WDF$) identifies whether the event (failure to successfully complete Calculus 1) occurred for a given case. If the event did not occur, the case is said to be censored. Censored cases are not used in the computation of the regression coefficients but are used to compute the baseline hazard. Table 5 shows that 169 cases
were censored.

Table 5: Table 6:

Table 6 shows that SPSS recoded the two categorical variables. This recoding is useful for interpreting the regression coefficients for these categorical variables. By default, the reference category is the "last" category of the covariate. Thus, for example, even though students who bypassed Pre-Calculus have PCbypass = 1 in the data file, they are coded as 0 for the purposes of regression.

As specified in the dialogs in Figures 4 and 5, the model-building process takes place in two blocks. In block 1, a forward stepwise algorithm is employed to determine which of DVMath, Atmp1070, and PreTstN will be entered into the model and remain in the model. In block 2, the variable PCbypass is entered.

Table 7 shows statistics needed prior to the start of the stepwise procedure in block 1.

Table 7:
The omnibus tests are measures of how well the model performs. Table 8 displays the omnibus tests for the first block. At any stage in the stepwise procedure, the change in the Chi-square ($\chi^2$) statistic from the previous step is the difference between $-2 \ln L$, where $L$ is the likelihood function, at the previous step and the current step. If the step is to add a variable, as PreTstN was added in step 1 and DVMath in step 2, the inclusion makes sense if the significance of the change is less than 0.05. The variable Atmp1070, number of previous attempts of Calculus1, was never added to the model since the significance in Table 7 is 0.204.

Table 8:
However, if a step was to remove a variable, the exclusion makes sense if the significance of the change is greater than 0.10.

Table 9 gives estimates for the coefficients of the variables in the final model of block 1.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PreTstN</td>
<td>-1.22</td>
<td>0.036</td>
<td>14.210</td>
<td>1</td>
<td>0.00</td>
<td>4.884</td>
<td>(3.34, 6.93)</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DVMath</td>
<td>-1.56</td>
<td>0.526</td>
<td>3.778</td>
<td>1</td>
<td>0.05</td>
<td>0.574</td>
<td>(0.28, 1.095)</td>
</tr>
<tr>
<td>PreTstN</td>
<td>-1.13</td>
<td>0.031</td>
<td>13.315</td>
<td>1</td>
<td>0.00</td>
<td>3.93</td>
<td>(3.04, 5.09)</td>
</tr>
</tbody>
</table>

Table 9:

These estimates change after $PCbypass$ is added to the model in block 2.

4.2.4 Interpretation of the Final Model

Table 10 displays omnibus tests for the final model after $PCbypass$ is added to the model in block 2.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients$^b$</th>
<th>Overall (score)</th>
<th>Change From Previous Step</th>
<th>Change From Previous Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log Likelihood</td>
<td>Ch-square</td>
<td>df</td>
<td>Sig</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1010.744</td>
<td>24.016</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 10:

The significance levels of 0.041 for change in $\chi^2$ from previous step and 0.000 for overall $\chi^2$ indicate that the final model performs very well. Table 11 shows the parameter estimates for the variables in the final model.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVMath</td>
<td>-2.28</td>
<td>0.314</td>
<td>6.677</td>
<td>1</td>
<td>0.01</td>
<td>0.172</td>
<td>(0.03, 0.85)</td>
</tr>
<tr>
<td>PreTstN</td>
<td>-1.69</td>
<td>0.031</td>
<td>12.193</td>
<td>1</td>
<td>0.00</td>
<td>0.197</td>
<td>(0.09, 0.44)</td>
</tr>
<tr>
<td>PCbypass</td>
<td>-1.72</td>
<td>0.187</td>
<td>12.193</td>
<td>1</td>
<td>0.00</td>
<td>0.197</td>
<td>(0.09, 0.44)</td>
</tr>
</tbody>
</table>

Table 11:

The variables considered for inclusion in the model were

- $x_{1i} = DVMath$ Took Developmental Math (0 = no; 1 = yes)
- $x_{12} = Atmp1070$ Number of previous attempts of Calculus 1 (0, 1, ...)
- $x_{13} = PreTstN$ Pre-Test score (max = 25)
- $x_{14} = PCbypass$ By-passed Pre-Calculus (0 = no; 1 = yes)

From Table 9, we have the final model predicated hazard for the Cox regression model.

$$h_i(t) = [h_0(t)]e^{\beta_0 + (-0.258)DVMathi + (-0.109)PreTstNi + (0.471)PCbypassi}$$

The statistic $\exp (\beta) = e^\beta$ is the predicted change in the hazard for a unit increase in the predictor. The value $e^{\beta_1} = 0.772$ for $DVMath$ means that the "unsuccess" hazard for a student who did not
take Developmental Math is 0.772 times that of a student who did not take Developmental Math. The model-predicted hazard for a student who did not take Developmental Math is less than, by a factor of 0.772, for a student who took Developmental Math. However, since the coefficient for DVMath is not significant (0.411) in the final model, the effect from the variable DVMath is probably due to chance.

The value $e^{\beta_2} = 0.897$ for PreTstN means that the "unsuccess" hazard is multiplied by 0.897 for each unit increase in PreTstN. In other words, as expected, the higher the Pretest score, the lower the hazard and the higher the survival rate, that is, the more likely the student successfully completes Calculus 1.

The value $e^{\beta_3} = 1.602$ for PCbypass means that the "unsuccess" hazard for a student who did not bypass Pre-Calculus is 1.602 times that of a student who did bypass the course.

5 Discussion

The main objective of this paper was to present a methodology to explore, in a longitudinal study, risk factors in undergraduate Calculus. A proportional hazards model was used to describe the relationships between (a) an event of interest (failure to successfully complete the course), (b) the time this event occurs during the period of observation (a full semester), and (c) several covariates. A status variable (Evnt WDF) was defined to indicate whether or not the event occurred (1=yes, 0=no). Because of difficulties inherent in determining an exact time, measured in days, that the event occurs for a student, a proxy for time (T Mods) was defined and used in the model. Four covariates were considered (DVMath, Atmp1070, PreTstN, PCbypass).

Cox regression was used to estimate parameters (regression coefficients) of the model. A forward stepwise procedure was used to include or exclude covariates in block 1 (DVMath, Atmp1070, PreTstN). Based on criteria set in the stepwise procedure, only DVMath and PreTstN were included in the model at the end of block 1. Since I wanted the covariate bypassed Pre-Calculus (PCbypass) to appear in the final model, I included this variable in block 2 of the model building process. No criteria was set for inclusion of PCbypass. The coefficient for DVMath in the final model was not statistically significant after the variable PCbypass was included.

Although the time variable T Mods as defined in this study is discrete, I treated it as a continuous random variable. The implications of this decision need to be investigated. For one of the 10 sections of the course, I did have a measure in days of time to event (T Days). Using the same proportional hazards model described in this paper with T Days as the time variable, only PreTstN was significant in the final model. Of the 22 students in that one section, the event Evnt WDF occurred for 7 while the other 15 students were censored. The value of $e^{\beta}$ for PreTstN in this model was 0.791, somewhat lower than the 0.897 value for the model with T Mods for time.

A grade of at least D may be acceptable for some students while others may need C or higher. In other words, the criteria for success or "unsuccess" is dependent on the student’s major. The status variable Evnt WDF was used to indicate whether or not the event of interest (failure to successfully complete the course). An area for further research is how a conditional status variable can be modeled.

The 303 subjects included in this study were a convenience sample, namely, the students enrolled in Calculus 1 for Fall 2003 at the private liberal arts university. Inferences to other populations may not be appropriate. However, the methodology proposed in this paper for investigating risk factors can be applied to other courses in which there is a relatively large number of students withdrawing during the semester. If a time to event variable is not defined, or one wishes to analyze the data independent of time, binary logistic regression with Evnt WDF as dependent variable may be used.[4] (See [1] for an example of logistic regression in college regression.)
A Appendix: Variables in Calculus Database

Several variables used in data verification are omitted. Values of some variables are present only for of the 10 sections (e.g., Days Absent, responses to multiple choice questions on final exam).

Name (Position) Label

Faculty (1)
TYPE (2) Course Type
  0  Basic Statistics 1
  1  Calculus 1
COURSE (3)
SF (4) Suffix
SC (5) Section
line (6) Database Line Number
CRN (7)
ID (8)
sem (9)
STUDENT (10)
LEVEL (11)
DEG (12)
DEPT (13)
PROG (14)
MAJOR (15)
CL (16)
SH (17)
Abs (18) Days Absent
Evnt_WDF (19) Unsuccessful (D, F, FE, WD, WF)
  0  Censored
  1  D or F or FE or WF
Evnt_WF (20) Unsuccessful (F, FE, WF)
  0  Censored
  1  F or FE or WF
T_Days (21) Time to Event (Days)
T_Mods (22) Time to Event (Mods)
PCbypass (23) By-Passed Pre-Calculus
  0  no
  1  yes
PCgrade (24) Highest Pre-Calculus Grade
PCtakes (25) No. of Times Pre-Calculus Taken
DVtakes (26) No. of Times Developmental Math Taken
DVMath (27) Took Developmental Math
  0  no
  1  yes
Atmp1070 (28) No. of Previous Attempts of Calculus 1
LastTest (29) Score on Last Test Taken
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod1</td>
<td>Mod 1 Test Grade</td>
</tr>
<tr>
<td>Mod2</td>
<td>Mod 2 Test Grade</td>
</tr>
<tr>
<td>Mod3</td>
<td>Mod 3 Test Grade</td>
</tr>
<tr>
<td>Mod4</td>
<td>Mod 4 Test Grade</td>
</tr>
<tr>
<td>Mod5</td>
<td>Mod 5 Test Grade</td>
</tr>
<tr>
<td>Exam</td>
<td>Final Exam Grade (based on 100)</td>
</tr>
<tr>
<td>NGrd</td>
<td>Final Course Grade (based on 100)</td>
</tr>
<tr>
<td>LGrd</td>
<td>Final Course Letter Grade</td>
</tr>
<tr>
<td>ACT</td>
<td>ACT Math Score</td>
</tr>
<tr>
<td>SAT</td>
<td>SAT Math Score</td>
</tr>
<tr>
<td>Prt1</td>
<td>XU Math Placement Test: Part 1</td>
</tr>
<tr>
<td>Tot</td>
<td>XU Math Placement Test: Total</td>
</tr>
<tr>
<td>PreTstPc</td>
<td>Assessment (Pre-) Test %</td>
</tr>
<tr>
<td>PreTstN</td>
<td>Assessment (Pre-) Test (based on 25)</td>
</tr>
<tr>
<td>PTQ01L</td>
<td>PT q01 (correct = [suppressed])</td>
</tr>
<tr>
<td>PTQ25L</td>
<td>PT q25 (correct = [suppressed])</td>
</tr>
<tr>
<td>PTQ01N</td>
<td>Linear eq var both sides</td>
</tr>
<tr>
<td>PTQ25N</td>
<td>Solve log equation</td>
</tr>
<tr>
<td>COMMENT</td>
<td></td>
</tr>
<tr>
<td>Verified</td>
<td></td>
</tr>
<tr>
<td>OrigTot</td>
<td></td>
</tr>
<tr>
<td>VerTot</td>
<td></td>
</tr>
<tr>
<td>FN01</td>
<td>[final exam, 1st mult.choice ques, letter]</td>
</tr>
<tr>
<td>FN02</td>
<td></td>
</tr>
<tr>
<td>FN18</td>
<td></td>
</tr>
<tr>
<td>FNc01</td>
<td>[final exam, 1st mult.choice ques, 1 or 0]</td>
</tr>
<tr>
<td>FNc02</td>
<td></td>
</tr>
<tr>
<td>FNc18</td>
<td></td>
</tr>
<tr>
<td>filter_</td>
<td>TYPE = 1 &amp; T_Mods &gt; 0 (FILTER)</td>
</tr>
<tr>
<td>0</td>
<td>Not Selected</td>
</tr>
<tr>
<td>1</td>
<td>Selected</td>
</tr>
</tbody>
</table>
References


Index

ACT Math, 11
baseline hazard, 7
baseline hazard function, 6, 12
by-pass Pre-Calculus, 9, 15
categorical variable, 10
censored data, 3
Chi-square, 14, 15
correlation matrix, 11
covariate(s), 6, 12
Cox regression, 9–11
Cox, D. R., 5
cumulative distribution function (cdf)
distribution function, 4, 5
cumulative hazard function, 12
descriptive statistics, 11
Developmental Math, 10
event WDF, 8
event WF, 8
failure time, 4
failure to successfully complete Calculus, 8
hazard curve, 6, 7, 12, 13
hazard function, 4, 5
instantaneous failure rate, 5
least squares (OLS) regression, 3
left censoring, 4
likelihood function, 14
logistic regression
binary logistic regression, 16
Math Placement Test, 11
maximum likelihood, 7
model assumptions, 6
multiplicatively decomposable, 6
negative log of survival, 7
omnibus tests, 14, 15
parameter estimation, 7
parametric, 7
partial maximum likelihood, 7
pre-test score, 9, 15
probability density function (pdf)
density function, 4, 5
proportional hazards assumption, 6
proportional hazards model
PHM, 5, 7
proxy for time, 8
references, 19
right censoring, 4
risk factors in Calculus, 7
SAT Math, 11
semiparametric, 7
status variable, 8
stepwise, 10
stepwise function, 7
survival analysis, 3
survival curve, 6, 7, 12, 13
survival function, 4
time proxy, 8
time to event (Mods), 8
unsuccess, 3, 15