Application of the Cox Regression Model to Estimate Dropout Rate in Introductory Statistics

DAVID EMMANUEL LARTEY

Department of Mathematical Sciences
Montana State University

May 3, 2018

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

DAVID EMMANUEL LARTEY

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Dr. Stacey Hancock Writing Project Advisor
Date	Mark C. Greenwood Writing Projects Coordinator

Abstract

Do students with a low number of prior math courses have a higher dropout rate in Introductory statistics than students with a high number of prior math courses? How does a student's ACT, SAT or MPLEX score affect their dropout rate? To answer these questions, one needs to perform survival analysis. Survival, or time-to-event analysis, is one of most significant advancements of mathematical statistics in recent years with broad applications in the fields of mechanical research, engineering and especially biomedical research. In this paper, review the properties and modeling methods for survival data, then fit a Cox Proportional Hazards Model for the data on time until dropout for students in the introductory statistics course (STAT 216). The data showed that the number of prior math courses taken and/or MPLEX score could have an effect on the time to drop out for students enrolled in the STAT 216 course.

Keywords: Survival analysis, Cox proportional hazards model, Kaplan-Meier estimate.

Contents

1 Introduction

As instructors or tutors, one goal is to see students succeed in the course, and thus see a considerably low dropout rate. Unfortunately, a few studies have been undertaken to determine factors that affects student's survival in course. Most studies focus on factors that affect college dropout, as opposed to course specific dropouts. For example, a study funded by the Bill and Melinda Gates foundation stated that, "More Americans are going to college than ever before, but students face unprecedented challenges. Over 44 million Americans collectively hold more than $1.4 trillion in student loan debt and only 54.8 percent of students graduate in six years." (Bill & Melinda Gates Foundation, 2017). Another study found that over 40% of full time four-year college students fail to earn a bachelor's degree within six years, and many never complete their education (E.D. Velez, 2014). More interestingly, Cheryl Miller in her thesis, "Dropped out or Pushed Out: A Case Study on Why Students Drop Out", proposed that students just do not decide one day that they are tired of school and stop attending, and that instead, there is a series of events that occur long before the student makes the announcement that he/she is planning to dropout (Miller C., 2006). Even though this study did not investigate a potential factor of a student's performance in courses taken during his or her time at college, one might expect that with all else being equal, dropout rates will be lower for students with strong academic performance compared to students with a poor academic standing. This could be due to the fact that students with a good standing would be more interested in completing their degrees.

This paper, although limited by its scope of inference, attempts to apply the technique of survival analysis to estimate, interpret and assess the relationship of several explanatory variables with a student's survival in the Introduction to Statistics course (STAT 216) at Montana State University, Bozeman (MSU). Survival in this context is defined as a student's ability to complete the STAT 216 course within a particular semester he/she has enrolled. Most disciplines require students to take at least one statistics course before they can graduate. This implies that there is a potential for students who are unable to complete this course to be frustrated, and in the long run, may dropout of college. In this paper, we introduce the basics of survival analysis, how to display survival data, followed by the necessary foundations for Cox regression models. The paper concludes with a detailed summary, analysis and discussion of results from the study.

1.1 Objectives

The objectives of this study are to:

(i) fit an appropriate Cox proportional hazards model to data obtained for students enrolled in STAT 216 for the 2014/2015 to 2016/2017 academic years,

(ii) determine which explanatory variables affect the dropout rate for STAT 216 students in a given semester,

(iii) predict time to dropout for STAT 216 students per semester, and

(iv) compare the survival probabilities of STAT 216 students with respect to the different explanatory variables.

The factors considered in this model include the prior number of Math courses taken, ACT, MPLEX, SAT scores, and the age of a student.

2 Methodology

2.1 Introduction to Survival Analysis

Survival analysis examines and models the time it takes for events to occur. The most typical of such event is death, from which the name 'survival analysis' and much of its terminology derives, but the ambit of application of survival analysis is much broader (Fox J., 2008). Other events include response to treatment, device failure, regaining mobility and, for this study, dropout.

The object of survival analysis (also know as failure time or time-to-event analysis) are data in the form of times from a well-defined time origin to an end point, where the end point could be the occurrence of some particular event or a particular time point (Ni J., 2009). In most studies, the time origin will correspond to the recruitment of the observational or experimental unit, such as a clinical trial to compare two or more treatments. The focus of this paper is on the application of survival analysis to data on the time until dropout of STAT 216 students per semester.

One feature of survival data that renders standard statistical methods inappropriate is that the distribution of survival data tend to be positively skewed, that is, most data will be bunched up toward the left and with a 'tail' stretching toward the right. Even though this characteristic could easily be resolved by first transforming the data to give a more symmetric distribution, a more satisfactory approach would be to select an alternative distributional model for the original data or use some other non-parametric approach. However, the distinguishing feature of survival data is that at the end of the study period, the event of interest may not have occurred for all units in the study. This phenomenon is defined as *censoring.*

2.2 Censoring

The survival time of an individual is said to be censored when the end-point of interest and/or starting time has not been observed for that individual. Formally, an observation is *right censored* at time C if the lifetime (T) is only known to be greater than C. Similarly, an observation is *left censored* if the lifetime is only known to be less than C (Lawless 1982, 2003). Another type of censoring is *interval censoring*, which occurs when individuals are known to have experienced an event within an interval of time. For example, say the time to dropout is recorded in days, however we only observe the number of students still enrolled in STAT 216 at the end of the week (Friday). If a students is observed to be enrolled in the course on Monday, but is found to have dropped out when the class roll was checked on Friday, the actual day that student dropped out is known to be within Monday and Friday. The observed time to dropout is said to be interval-censored. Examples of the various types of censoring are illustrated in Figure 1 below for a selection of ten STAT 216 students from our data set.

2.3 Definitions of Important Distribution Functions Used in Survival Analysis

2.3.1 Survivor Function

Let the random variable T be the survival time in days to dropout of STAT 216 in a semester and t be regarded as an observed value of T. Now suppose that T has a probability distribution with underlying *probability density function*, $f(t)$. The *distribution function* (cumulative distribution function) of T is then given by:

$$F(t) = P(T \leq t) = \int_0^t f(u)du, \tag{1}$$

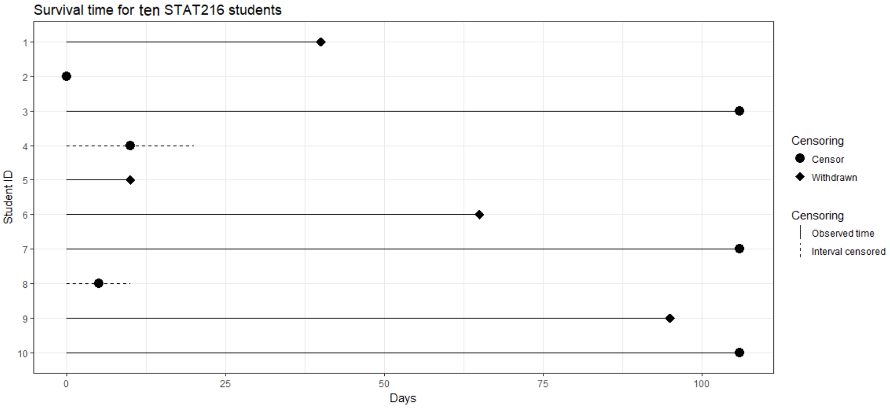


Figure 1: Study time for ten students in a semester. **Left Censored**: Student 2; **Right censored**: Students 3, 7 and 10; **Interval censored**: Students 4 and 8

which is defined as the probability that a students survival time is less than some value t. The survivor function, $S(t)$, is defined as the probability that the survival time is greater than t, and so from Equation (1),

$$S(t) = P(T > t) = 1 - F(t) \tag{2}$$

2.3.2 Hazard Function

The *hazard function* is used to express the risk or hazard of an event at some time t. This function is obtained from the probability that a student drops out at time t, conditional on he or she having survived in the course until that time. Formally, we define the hazard function $h(t)$ as:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t < T \leq t + \delta t | T > t)}{\delta t} \right\} \tag{3}$$

The function $h(t)$ is also referred to as the *hazard rate*, the *instantaneous death rate*, the *intensity rate* or the *force of mortality.*

The definition of the hazard function in Equation (3) leads to some useful relationships between the survivor and hazard functions. Based on the definition of conditional probabilities, we have:

$$\begin{aligned} P(t \leq T < t + \delta t | T > t) &= \frac{P(t < T \leq t + \delta t)}{P(T > t)} \\ &= \frac{F(t + \delta t) - F(t)}{S(t)}. \end{aligned}$$

Then,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.$$

Now,

$$\lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}$$

is the definition of the derivative of $F(t)$ with respect to t, which is $f(t)$, and so,

$$h(t) = \frac{f(t)}{S(t)}. \tag{4}$$

Altogether, Equations (1), (2) and (4) show that from any one of the three functions, $f(t)$, $S(t)$ and $h(t)$, the other two can be derived.

2.3.3 Cumulative Hazard Function

From Equation (4), it follows that

$$h(t) = -\frac{d}{dt}\log S(t), \tag{5}$$

and so

$$S(t) = \exp\{-H(t)\}, \tag{6}$$

where

$$H(t) = \int_0^t h(u)du. \tag{7}$$

The function $H(t)$ features widely in survival analysis, and is called the *integrated* or *cumulative hazard function.* From Equation (6), the cumulative hazard function can also be obtained from the survivor function, since

$$H(t) = -\log S(t). \tag{8}$$

The cumulative hazard function, $H(t)$, is the cumulative risk of a student dropout occurring by time t. That is, $H(t)$ summarizes the risk of dropout up to time t, given that no dropout has not occurred before t. The survivor function, hazard function and cumulative hazard function are all estimated from the observed survival times when analyzing survival data.

2.4 Estimating the Survivor Function

For this study, we focus on *non-parametric* or *distribution-free* methods of estimating the survival and hazard functions which are conveniently used to summarize survival data. These methods do not require specific assumptions to be made about the underlying distribution of the survival times. Once we are able to estimate the survivor function, we can proceed to estimate the median and other percentiles of the distribution of survival times. These estimated survival functions can serve as an informal tool for comparing survival experience of individuals in two groups. Not surprisingly, there are formal non-parametric procedures for comparing two or more groups of survival times which will also be discussed.

Suppose we obtain data on a single sample of survival times, where none of the observations are censored Then the survival function in Equation (2) can be estimated by the *empirical survivor function* given by:

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the data set}} \tag{9}$$

Equivalently, $\hat{S}(t) = 1 - \hat{F}(t)$, where $\hat{F}(t)$ is the *empirical distribution function*, that is, the ratio of the total number of students still enrolled in STAT 216 at time t to the total number of students enrolled in STAT 216 at the beginning of the semester. Notice that the empirical survivor function is equal to unity for values of t before the first dropout, and zero at the final dropout time. We also assume that $\hat{S}(t)$ is constant between two adjacent dropout times, and so a plot of $\hat{S}(t)$ against t is a decreasing step function as shown in Figure 2.

2.4.1 Kaplan-Meier Estimate of the Survivor Function

The method of estimating the survivor function using the empirical distribution function is problematic when there are censored observations. In 1958, E. L. Kaplan and P. Meier came up with one of the most frequently used non-parametric approaches used to estimate the survivor function, the Kaplan-Meier method. The Kaplan-Meier or Product-limit estimator is defined as the probability of surviving a given length of time while considering time in many small intervals (Goel, M.K., et al., 2010). The three assumptions used are:

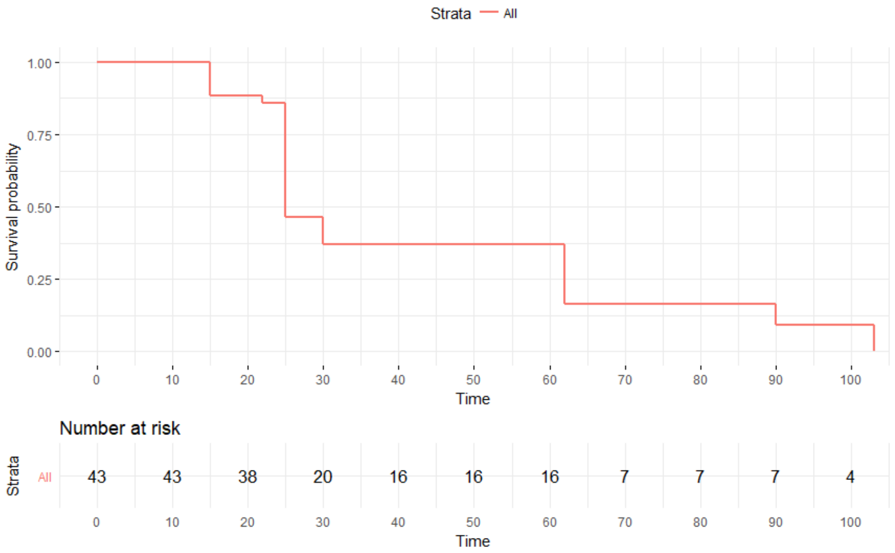


Figure 2: Estimated survivor function for a sample of 43 students in a STAT 216 course.

1. At any time, students who are censored have the same survival prospects as those who continue to be studied.
2. The survival probabilities are the same for students enrolled early and later in the semester.
3. The dropout occurs at the time (day) specified.

The Kaplan-Meier estimator involves computing probabilities of occurrence of event at a certain point of time. The successive probabilities are then multiplied by any earlier estimated computed probabilities to get the final estimate. Formally, the Kaplan-Meier estimate of the survivor function is given by:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{10}$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1,\ 2,\ ,...,\ r$. We define $\hat{S}(t) = 1$ for $t < t_{(1)}$, and $t_{(r+1)}$ is taken to be ∞, where

- $t_{(k)}$ is the time of k^{th} dropout, for $t_{(1)} < t_{(2)} < ... < t_{(r)}$,
- $t_1, ..., t_n$ are the observed survival times for n students,
- r represents the number of dropout times observed, such that $r \leq n$,
- n_j represents the number of students still enrolled just before time $t_{(k)}$, including those who are about to dropout at this time, and
- d_j represents the number of students who dropout at time $t_{(k)}$.

Other non-parametric methods used in estimating survivor function include the life-table and Nelson-Aalen estimates of the survivor function.

2.4.2 Standard Error of the Kaplan-Meier Estimate

Using the delta method, Greenwood's formula gives an asymptotic standard error for the Kaplan-Meier (KM) estimator (Rodriguez G., 2005). The variance for the KM estimate of the survivor function can be shown to be:

$$\hat{\text{var}}\{\hat{S}(t)\} \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}, \text{ for } t_{(k)} \le t < t_{(k+1)}. \tag{11}$$

Finally, the standard error of the Kaplan-Meier estimate of the survivor function, defined to be the square root of the estimated variance of the estimate, is given by

$$\text{se}\{\hat{S}(t)\} \approx \hat{S}(t) \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)} \right\}^{\frac{1}{2}}, \text{ for } t_{(k)} \le t < t_{(k+1)}. \tag{12}$$

This result is known as Greenwood's formula.

Thus, a pointwise $100(1-\alpha)\%$ confidence interval for $S(t)$, for a given value of t, is the interval given by:

$$\hat{S}(t) \pm z_{\alpha/2}\text{se}\{\hat{S}(t)\}, \tag{13}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value for the standard normal distribution. These intervals can be superimposed on a graph of the estimated survivor function.

2.5 Estimating the Median and Percentiles of Survival Times

Using the estimated survivor function, the estimated p^{th} percentile of the distribution of survival times is the observed survival time, $\hat{t}(p)$, for which

$$\hat{S}\{\hat{t}(p)\} < 1 - (p/100), \tag{14}$$

with an associated standard error given by:

$$\text{se}\{\hat{t}(p)\} = \frac{1}{\hat{f}\{\hat{t}(p)\}}\text{se}[\hat{S}\{\hat{t}(p)\}], \tag{15}$$

where $\hat{f}\{\hat{t}(p)\}$ is the KM estimate of the probability density function of the survival times at $t(p)$. Once the standard error of the estimated p^{th} percentile has been found, a $100(1-\alpha)\%$ confidence interval for $t(p)$ has limits of:

$$\hat{t}(p) \pm z_{\alpha/2}\text{se}\{\hat{t}(p)\}, \tag{16}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value for the standard normal distribution. Note that these confidence intervals are only approximations, and thus, do not have an exact $1-\alpha$ coverage probability. Alternative methods with better coverage probabilities have been developed.

2.6 Comparison of Two Groups of Survival Data

A visual assessment of the survival times obtained from two groups of individuals can be made by plotting the corresponding estimates of the two survivor functions on the same axes. Any differences observed between the two survival curves could be due to two possible explanations. The first is random chance; that is, there is no real difference between the distribution survival times in each group, and that the difference observed is merely the result of random variation. The second is that there is a real difference between the distribution of survival times for the two groups of individuals, so that our data reflects this difference. We can conduct a *hypothesis test* to help distinguish between the two possible explanations. In other words, an hypothesis test enables us to assess the extent to which an observed set of data are consistent with a particular hypothesis.

2.6.1 Hypothesis Testing

The three basic steps taken to conduct an hypothesis test are given as:

(a) Assume that the null hypothesis is true. The *null* or *working* hypothesis is usually viewed as the data-generating process, and represents the hypothesis that there is no difference between two survival distributions.

(b) Formulate a *test statistic.* The test statistic is use to measure the extent to which the observed data departs from the null hypothesis. Generally, large test statistics suggest evidence against the null hypothesis.

(c) Calculate the *probability value (p-value).* This value represents the probability of obtaining a test statistic value as extreme as or more extreme than the observed value, when the null hypothesis is true. It is used to summarize the strength of evidence in the sample data against the null hypothesis. Generally, when the p-value is large, we conclude that the observed data is likely to be obtained when the null hypothesis is true, and that there is no evidence against the null hypothesis. On the contrary, for small p-values, we can conclude there is strong evidence against the null hypothesis.

2.7 The Log-rank Test

The log-rank test is a non-parametric procedure that can be used to quantify the extent of between-group differences for survival data. The log-rank test statistic can be obtained under the null hypothesis as:

$$W_L = \frac{U_L^2}{V_L} \sim \chi_1^2, \tag{17}$$

where:

$$e_{1j} = \frac{n_{1j} d_j}{n_j}; \quad U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}); \quad \text{and} \quad V_L = \sum_{j=1}^{r} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}.$$

The larger the value of W_L, the greater the evidence against the null hypothesis. Given that W_L is approximately chi-squared with one degree of freedom if the null hypothesis is true, the p-value can be obtained from the distribution function of a chi-squared random variable.

2.8 The Cox Proportional Hazards Model

We may expect survival times to depend on the outcome of several explanatory variables. Accordingly, the values of these variables would be recorded at the outset of the study. In 1972, Sir David Cox developed what has come to be known as the *Cox Regression Model*, which both unifies and extends the non-parametric procedures earlier discussed.

Unlike many linear models one may have encountered in regression analysis and in the analysis of data from designed experiments, here, the hazard function is modeled directly. The objective of the modeling process is to determine which combination of potential explanatory variables affect the form of the hazard function. Another reason is to obtain an estimate of the hazard function itself for an individual, which would in turn aid in estimating the survivor function from the relationships described in Equation (5). However, principles and procedures used in linear modeling carry over to the modeling of survival data.

In the model proposed by D. Cox is based on the assumption that hazards are proportional. This means that the hazard functions for two individuals are proportional to each other. In other words, the hazard of dropout at any given time for a student in one group (for e.g. low ACT score) is proportional to the hazard at that time for a similar student in another group (for e.g. high ACT scores).

2.8.1 The General Proportional Hazards Model

The Cox proportional hazards model is given as:

$$h_i(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)h_0(t), \tag{18}$$

where;

- $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})'$ is a vector representing the set of values of the explanatory variables for the i^{th} individual,
- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ represents the vector of unknown regression coefficients, and
- $h_0(t)$ is the hazard function for an individual for whom $\boldsymbol{x} = \mathbf{0}$. The function is also known as the *baseline hazard function.*

The Cox model proposed in Equation (18) does not make any assumptions about the actual form of $h_0(t)$, hence, it is known as a *semi-parametric model.* We are only interested in obtaining estimates for $\boldsymbol{\beta}$. These parameters can be estimated using the method of *maximum likelihood.* This approach allows estimating $\boldsymbol{\beta}$ by using the ranks of the dropout and uncensored times. Therefore, if $t_{(1)}, ..., t_{(k)}$ are k ordered dropout times and $R(t_{(j)})$, the risk set, is the set of students which have survived until $t_{(j)}$, immediately prior to the j^{th} survival time, then the relevant likelihood function for the Cox model is:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l\in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})} \tag{19}$$

The likelihood function in Equation (19) is however not a true likelihood, since it does not make direct use of the actual censored and uncensored survival times, hence, it is known as a *partial likelihood function.* The likelihood function only depends on the ranking of the dropout times, since this the determines the risk set at each dropout time. Consequently, inferences about the effect of explanatory variables on the hazard function depend only on the rank order of the survival times.

Now, suppose that the data consist of n observed survival times, denoted by $t_1, .., t_n$, and that δ_i is an event indicator, which is zero if the i^{th} survival time t_i, $i = 1, ..., n$, is right censored, and unity otherwise. Then we express the partial likelihood function of Equation (19) as:

$$\prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{\sum_{l\in R(t_i)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right\}^{\delta_i}, \tag{20}$$

where $R(t_i)$ is the risk set at time t_i. We then obtain the partial log-likelihood function from Equation (20) as:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}'\boldsymbol{x}_i - \log \sum_{l\in R(t_{(j)})} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l) \right\}. \tag{21}$$

Given that there are no closed-form solutions to the partial log-likelihood function, the *Newton-Raphson procedure* is generally used to obtain maximum likelihood estimates of the β-parameters.

Once we are able to obtain these estimates, we can also estimate standard errors and construct an approximate $(1-\alpha)100\%$ confidence interval for $\boldsymbol{\beta}$, and estimate the hazard ratio, $exp(\boldsymbol{\beta})$.

3 Analysis and Results

3.1 Data Description

We apply the methods discussed to a retrospective cohort study for STAT 216 students enrolled at Montana State University, Bozeman, between the 2014/2015 to the 2016/2017 academic years. Secondary data was obtained from the Office of the Registrar, and combined with data from the Department of Mathematical Sciences. In the process of merging the two datasets, some information was unfortunately lost, and we thus proceed with the combined dataset as a convenience sample of STAT 216 students enrolled in a particular semester. We only focus on the Fall and Spring semesters for each academic year. The following variables are studied:

T: the number of days until a students drops or withdraws from STAT216 per semester.

X_0: Status of censoring

0- student is right-censored, that is, student completed the course with a grade.
1- student dropped out.

X_1: Number of prior math courses taken.

0- at most 2 math courses (Low).
1- at least 3 math courses (High).

X_2: Age of student at the start of the semester.

0- less than or equal to 21 years old (Young).
1- greater than 21 years (Old).

X_3: Math placement exam (MPLEX) score.

0- $0.0 \leq$ MPLEX ≤ 2.5 (Low)
1- $2.5 <$ MPLEX ≤ 5.0 (High)

X_4: ACT score.

0- $0 \leq$ ACT ≤ 20 (Low)
1- $20 <$ ACT ≤ 30 (Moderate)
2- $30 <$ ACT ≤ 40 (High)

We assume that all students started the course at the beginning of each semester, and that all covariates were recorded prior to the commencement of the semester. Table (1) and Figure (3) summarize the number of students per semester who dropped out. A summary of the observed data for each of the explanatory variables by status of survival (dropout or completed) are also summarized in the Tables constructed in the Appendix.

3.1.1 Dealing with Missing ACT and SAT Scores

Note that most students only took either the ACT or SAT exam. The data set, therefore, was incomplete for such students. Given that both standardized tests are organized by the College Board, we were able to obtain an ACT-SAT concordance table that allows us to convert scores. With this, we were able to obtain equivalent ACT scores for students who took the SAT (Compassprep, 2016).

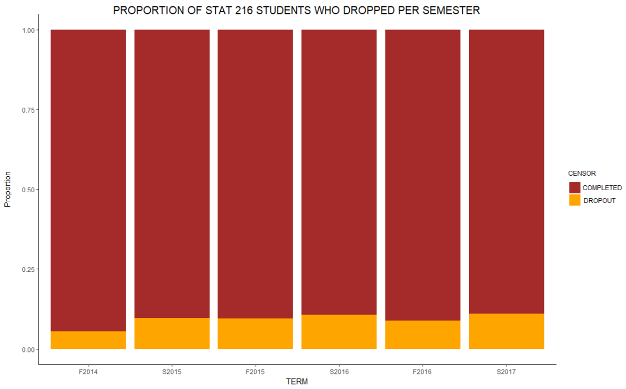


Figure 3: Status of students per semester.

TERM	No. of Students	No. of dropout (%)	No. of censored (%)
F2014	296	16 (5.41)	280 (94.59)
S2015	240	23 (9.58)	217 (90.42)
F2015	820	78 (9.51)	742 (90.49)
S2016	640	68 (10.63)	572 (89.37)
F2016	865	76 (8.79)	789 (91.21)
S2017	641	70 (10.92)	571 (89.08)
Total	3502	331 (9.45)	3171 (90.55)

Table 1: Data summary for STAT 216 for each semester.

3.1.2 Exploratory Data Analysis

From Figure (3) and Table (1), we observe that the proportions of students that dropout are approximately equal across all semesters (8.79%-10.92%) except for the Fall 2014 semester, where there is a very low dropout rate (5.41%). A summary of the association between each of the explanatory variables and the dropout rates can also be seen in Tables 14 to 17 (refer to Appendix). At a glance, there appears to be no practical difference between levels of each explanatory variable across semesters.

3.1.3 Kaplan-Meier Survival estimates

The Kaplan-Meier estimates of the survivor function for each semester are given in Figures (4) - (6).

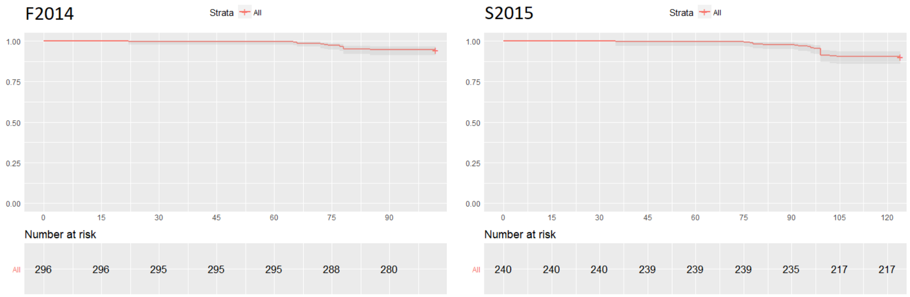


Figure 4: Kaplan-Meier estimates of survival for the 2014/2015 Academic year.

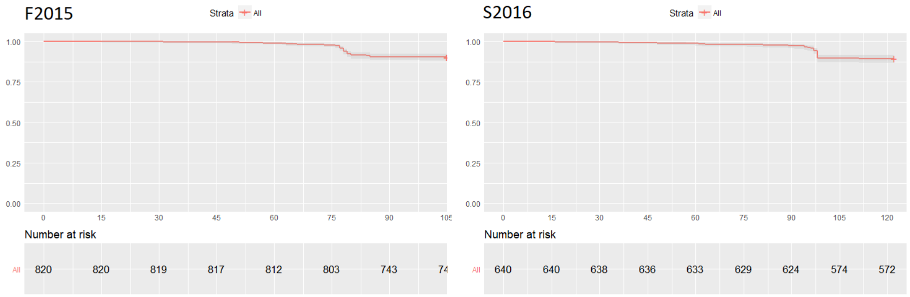


Figure 5: Kaplan-Meier estimates of survival for the 2015/2016 Academic year.

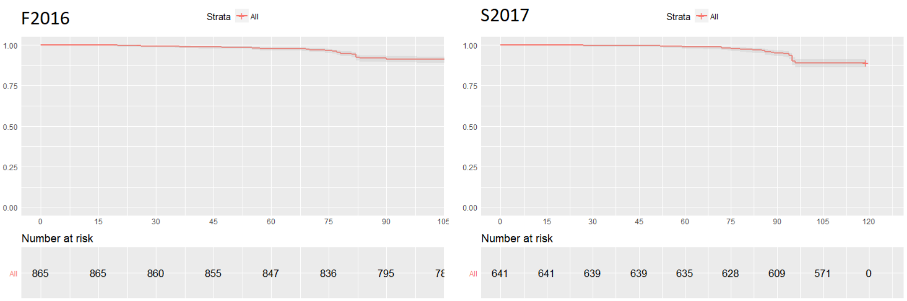


Figure 6: Kaplan-Meier estimates of survival for the 2016/2017 Academic year.

We observe that the KM estimates appear to be similar across all semesters. Due to the low number of withdrawals in each semester, a majority of the survival probabilities range within 0.875 to 1. Also in the Fall semester, students are most likely to withdraw within the 60^{th} and 90^{th} day, whereas students withdraw between the 80^{th} and 100^{th} day in the Spring semesters. The number at risk tables below each KM plot also displays the number of students who have still not dropped out on a particular day within that semester.

3.1.4 Estimation of Cox Regression Parameter

(a) **Fall 2014 Semester:**
We first fit the proportional hazards model using each covariate only and produce a log-cumulative hazard plot (Figure 7) to assess the assumption of proportional hazards. The lines in all four log cumulative hazard plots are close to parallel, so there does not seem to be an obvious violation of the proportional hazards assumption. This is confirmed by the Log-rank test p-values displayed on each plot. We then proceed to fit the cox proportional hazards model for the i^{th} student, given

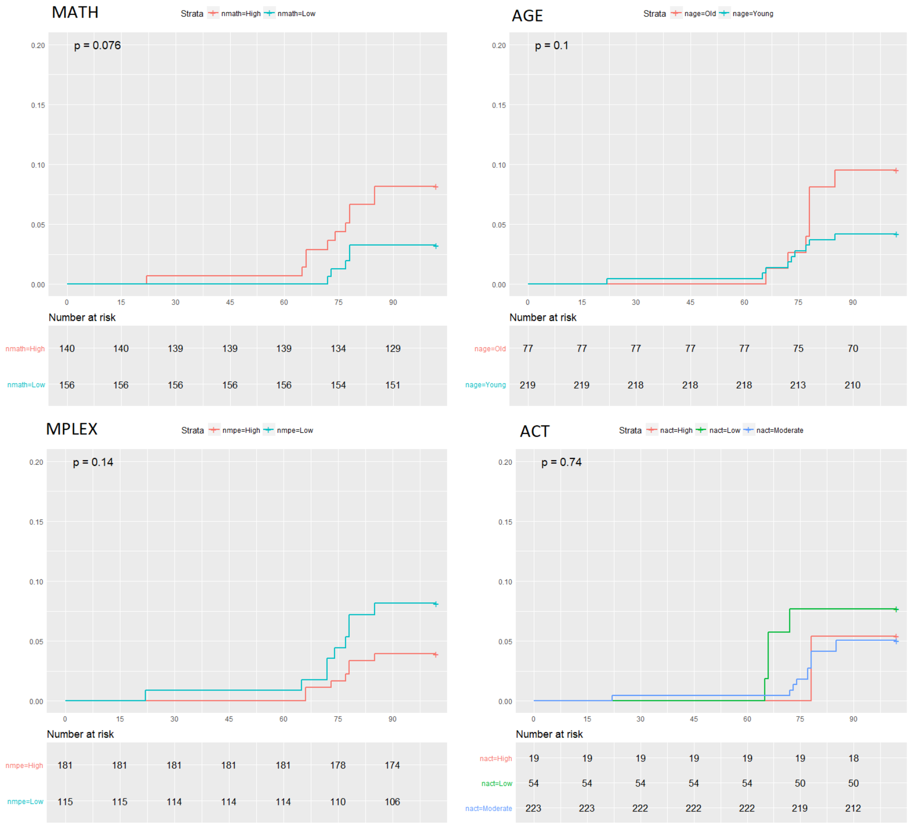


Figure 7: Log Cumulative Hazard Plot for each covariate in the F2014 semester

as:

$$\hat{h}_i(t) = exp\{-0.704X_{1i} - 0.526X_{2i} + 0.477X_{3i} - 0.22X_{41i} - 0.263X_{42i}\}h_0(t)$$

where $h_0(t)$ is the hazard rate for a student with High number of prior math courses, old age, high MPLEX and ACT scores.

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\|z\|)$	lower(.95)	upper(.95)
X_{11}	-0.704	0.495	0.572	-1.23	0.22	0.16118	1.518
X_{21}	-0.526	0.591	0.538	-0.98	0.33	0.20570	1.698
X_{31}	0.447	1.564	0.529	0.84	0.40	0.55406	4.415
X_{41}	-0.220	0.802	1.154	-0.19	0.85	0.08357	7.704
X_{42}	-0.263	0.769	1.052	-0.25	0.80	0.09786	6.044

Table 2: Results of Cox-regression model estimation for F2014

We can interpret the coefficient for X_1 as; for old students with high MPLEX and ACT scores, the estimated hazard rate for students with low number of prior math courses is 50.5% lower compared to students in the high group. However, we observe a weak evidence that survivor distributions differ between combinations of treatments (Table 3). In other words, there is little to no evidence that at least one covariate has an effect on the time to student dropout in Fall 2014 semester.

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	5.32	5	0.3786
Wald	5.13	5	0.4005
Score(logrank)	5.52	5	0.3554

Table 3: Testing Global Null Hypothesis: BETA=0 (F2014)

(b) **Spring 2015 Semester:**
We also fit the proportional hazards model using each covariate only and produce a log-cumulative hazard plot (Figure 8) to assess the assumption of proportional hazards. Similar to what was observed in F2014, the lines in all four log cumulative hazard plots are close to parallel. This is confirmed by the Log-rank test p-values displayed on each plot. The cox proportional hazards model for the i^{th} student is then given as:

$$\hat{h}_i(t) = exp\{-0.606X_{1i} + 0.956X_{2i} + 0.123X_{3i} + 16.4X_{41i} + 16.4X_{42i}\}h_0(t)$$

where $h_0(t)$ is defined as before.

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\|z\|)$	lower(.95)	upper(.95)
X_{11}	-0.6055	0.5458	0.4624	-1.310	0.190	0.2205	1.351
X_{21}	0.9556	2.600	0.5618	1.701	0.089	0.8646	7.820
X_{31}	0.1234	1.131	0.4315	0.286	0.775	0.4856	2.635
X_{41}	16.45	1.388×10^7	4878	0.003	0.997	0.0000	∞
X_{42}	16.38	1.298×10^7	4878	0.003	0.997	0.0000	∞

Table 4: Results of Cox-regression model estimation for S2015

We can interpret the coefficient for X_2 as; for students with a high number of prior math courses with high MPLEX and ACT scores, the hazard rate for younger students is 160% higher compared to older students. However, we observe a weak evidence that survivor distributions differ between combinations of treatments (Table 5). In other words, there is little to no evidence that at least one covariate has an effect on the time to student dropout in Spring 2015 semester.

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	5.38	5	0.3709
Wald	4.25	5	0.514
Score(logrank)	4.85	5	0.4342

Table 5: Testing Global Null Hypothesis: BETA=0 (S2015)

(c) **Fall 2015 Semester:**
Figure 9 displays the log-cumulative hazard plot used to assess the assumption of proportional hazards. Even though the lines in all four log cumulative hazard plots are close to parallel, the log-rank test provides evidence that the survival distributions does not differ between levels of all covariates, except for the number of prior math courses where there is strong evidence that the survival distributions differs for students in the low and high groups. We then estimate the cox proportional hazards model for the i^{th} student as:

$$\hat{h}_i(t) = exp\{-0.0203X_{1i} - 0.3022X_{2i} + 0.5748X_{3i} + 1.40082X_{4i} + 1.1841X_{42i}\}h_0(t)$$

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\|z\|)$	lower(.95)	upper(.95)
X_{11}	-0.0203	0.9799	0.2403	-0.08	0.933	0.6119	1.569
X_{21}	-0.3022	0.7392	0.2763	-1.09	0.274	0.4301	1.27
X_{31}	0.5748	1.7769	0.2369	2.43	0.015	1.1168	2.827
X_{41}	1.4008	4.0585	1.0508	1.33	0.183	0.5175	31.829
X_{42}	1.1841	3.2677	1.0150	1.17	0.243	0.4469	23.8914

Table 6: Results of Cox-regression model estimation for F2015

We can interpret the coefficient for X_{41} as; for old students with high number of prior math courses and MPLEX score, the time to dropout for students with low ACT scores is 305.85% higher compared to students with high ACT scores.

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	14.71	5	0.01169
Wald	13.86	5	0.01654
Score(logrank)	15.05	5	0.01016

Table 7: Testing Global Null Hypothesis: BETA=0 (F2015)

We also observe a strong evidence that survivor distributions differ between combinations of treatments (Table 7). In other words, there is strong evidence that at least one covariate has an effect on the time to student dropout in Fall 2015 semester. From Table 6, we observe that the MPLEX score has an effect on the survival probabilities for students in F2015 semester.

(d) **Spring 2016 Semester:**
The log-cumulative hazard plot in Figure 10 is used to assess the assumption of proportional hazards and the log-rank test provides evidence that the survival distributions differ between levels of number of prior math courses and MPLEX scores, but for ACT scores and age of student, there is no evidence that the survival distributions differs for students. The estimated cox proportional hazards model for the i^{th} student given as:

$$\hat{h}_i(t) = exp\{-0.804X_{1i} - 0.228X_{2i} + 0.562X_{3i} - 0.431X_{41i} - 0.625X_{42i}\}h_0(t)$$

After adjusting for Age, MPLEX and ACT scores, we are 95% confident that the dropout rate of the population of STAT 216 students in S2016 semester with low number of prior math courses is between about 11.72% to 77.31% lower than the hazard rate if student has a high number of prior

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\|z\|)$	lower(.95)	upper(.95)
X_{11}	-0.804	0.448	0.347	-2.32	0.020	0.2269	0.8828
X_{21}	-0.228	0.796	0.266	-0.86	0.391	0.4729	1.3406
X_{31}	0.562	1.754	0.258	2.18	0.029	1.0585	2.9067
X_{41}	-0.431	0.650	0.573	-0.75	0.452	0.2112	1.9985
X_{42}	0.625	0.535	0.535	-1.17	0.243	0.1877	1.5275

Table 8: Results of Cox-regression model estimation for S2016

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	16.26	5	0.006149
Wald	14.72	5	0.01163
Score(logrank)	15.6	5	0.008069

Table 9: Testing Global Null Hypothesis: BETA=0 (S2016)

math courses. We also observe a very strong evidence that survivor distributions differ between combinations of treatments (Table 9). In other words, there is strong evidence that at least one covariate has an effect on the time to student dropout in Fall 2014 semester. From Table 8, we observe that the MPLEX score and the number of prior math courses taken both have an effect on the survival probabilities for students in S2016 semester.

(e) **Fall 2016 Semester:**
From the log-cumulative hazard plot in Figure 11, the log-rank test provides evidence that the survival distributions differs between levels of number of prior math courses and AGE of a student, but for ACT and MPLEX scores, there is little to no evidence that the survival distributions differs for students. The cox proportional hazards model for the i^{th} student is given as:

$$\hat{h}_i(t) = exp\{-0.620X_{1i} - 0.408X_{2i} + 0.169X_{3i} + 16.8X_{41i} + 16.9X_{42i}\}h_0(t)$$

where $h_0(t)$ is the hazard rate for a student with High number of prior math courses, old age, high MPLEX and ACT scores.

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\|z\|)$	lower(.95)	upper(.95)
X_{11}	-0.620	0.538	0.255	-2.43	0.015	0.3264	0.8863
X_{21}	-0.408	0.665	0.251	-1.62	0.105	0.4064	1.0888
X_{31}	0.169	1.18	0.241	0.70	0.482	0.7389	1.8992
X_{41}	16.8	2.02×10^7	2360	0.01	0.994	0.0000	∞
X_{42}	16.9	2.25×10^7	2360	0.01	0.994	0.0000	∞

Table 10: Results of Cox-regression model estimation for F2016

After adjusting for the number of prior math courses, MPLEX and ACT scores, we are 95% confident that the dropout rate of the population of STAT 216 students in S2016 semester who are young is between about 59.36% lower to about 8.88% higher than the hazard rate of a student who is old. From Table 10, we observe that the number of prior math courses taken has an effect on the survival probabilities for students in F2016 semester.

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	21.26	5	0.0007243
Wald	11.53	5	0.04179
Score(logrank)	17.09	5	0.004333

Table 11: Testing Global Null Hypothesis: BETA=0 (F2016)

(f) **Spring 2017 Semester:**
Finally, Figure 12 shows the log-cumulative hazard plot to assess the assumption of proportional hazards for the S2017 semester. The log-rank test provides evidence that the survival distributions differs between young and old students, but for the number of prior math courses taken, ACT and MPLEX scores, there is little to no evidence that the survival distributions differ for students. The cox proportional hazards model for the i^{th} student is given as:

$$\hat{h}_i(t) = exp\{-0.117X_{1i} - 0.433X_{2i} + 0.249X_{3i} + 0.164X_{41i} + 0.154X_{42i}\}h_0(t)$$

where $h_0(t)$ is the hazard rate for a student with High number of prior math courses, old age, high MPLEX and ACT scores.

Covariate	coef	exp(coef)	se(coef)	z	$Pr(>\lvert z\rvert)$	lower(.95)	upper(.95)
X_{11}	-0.117	0.890	0.258	-0.45	0.651	0.5365	1.476
X_{21}	-0.433	0.649	0.260	-1.67	0.096	0.3899	1.079
X_{31}	0.249	1.283	0.254	0.98	0.327	0.7798	2.110
X_{41}	0.164	1.179	0.635	0.26	0.796	0.3397	4.090
X_{42}	0.154	1.167	0.599	0.26	0.797	0.3604	3.778

Table 12: Results of Cox-regression model estimation for S2017

After adjusting for the number of prior math courses, Age and ACT scores, we are 95% confident that the dropout rate of the population of STAT 216 students in S2016 semester with low MPLEX scores is between about 22.02% lower to about 111% higher than the hazard rate of a student with high MPLEX scores.

Test	Chi-Square	DF	Pr>Chi-Sq
Likelihood Ratio	5.49	5	0.3592
Wald	5.7	5	0.3363
Score(logrank)	5.82	5	0.3245

Table 13: Testing Global Null Hypothesis: BETA=0 (S2017)

We also observe that there is little to no evidence that at least one covariate has an effect on the time to student dropout in Spring 2017 semester. From Table 12, we observe that the none of the covariates appear to have an effect on the survival probabilities for students in S2017 semester.

3.1.5 Estimating Percentiles of Survival Times

Given that the time to dropout was low for each semester, we estimate percentiles of survival times with their associated 95% confidence intervals for each semester after controlling significant covariates.

(a) **Fall 2014**:

We are 95% confident that the 5^{th} percentile of survival times is at least 74 days. In other words, 95% of students survived past at least 74 days, not adjusting for any of the factors considered.

(b) **Spring 2015**:

We are 95% confident that the 10^{th} percentile of survival times is at least 99 days. That is, 90% of students survived past at least 99 days, not adjusting for any of the factors considered.

(c) **Fall 2015**:

i. For students in the population with high MPLEX scores, we are 95% confident that 95% of students will survive at least 78 days to at most 80 days, not adjusting for other factors considered.

ii. For students in the population with low MPLEX scores, we are 95% confident that 95% of students will survive at least 63 days to at most 78 days, not adjusting for other factors considered.

(d) **Spring 2016**:

i. For students in the population with high MPLEX scores and high number of prior math courses, we are 95% confident that 90% of students will survive at least 98 days, not adjusting for other factors considered.

ii. For students in the population with low MPLEX scores and high number of prior math courses, we are 95% confident that 90% of students will survive at least 97 days to at most 98 days, not adjusting for other factors considered.

iii. For students in the population with high MPLEX scores and low number of prior math courses, we are 95% confident that 95% of students will survive at least 91 days, not adjusting for other factors considered.

iv. For students in the population with low MPLEX scores and low number of prior math courses, we are 95% confident that 90% of students will survive at least 81 days, not adjusting for other factors considered.

(e) **Fall 2016**:

i. For students in the population with high number of prior math courses taken, we are 95% confident that 95% of students will survive at least 60 days to at most 81 days, not adjusting for other factors considered.

ii. For students in the population with low number of prior math courses taken, we are 95% confident that 95% of students will survive at least 78 days, not adjusting for other factors considered.

(f) **Spring 2017**:

We are 95% confident that the 10^{th} percentile of survival times is at least 95 days. That is, 90% of students survived past at least 95 days, not adjusting for any of the factors considered.

4 Summary and Conclusions

This study introduces some theories and modeling methods in survival analysis and applies the Cox Proportional Hazards Model to analyze the time (days) to dropout for students enrolled in Introduction to Statistics course at Montana State University between 2014/2015 - 2016/2017 academic years.

Based on this data, we observe that the Age and ACT scores of a student do not have an effect on the time to dropout. However, for some semesters we were able to observe some association between the number of prior math courses taken and/or the students MPLEX score. The

Further studies could therefore be done to determine which math courses are relevant to improve on a students ability to survive int he course, and whether or not the MPLEX score required for students to be eligible to take the STAT 216 course be adjusted.

Finally, given that the data obtained limited the ability to assess the effect of other factors on the time to dropout, we recommend that other factors need to be examined. Results of other methods such as Bayesian survival analysis could be compared to these results.

Limitations of the study are as follows: The study is was conducted based on secondary data from two sources which might have incomplete and biased information. Also information might have been missed during the merging process and in the case of many censored observations, given that a very high proportion of students completed the course in given semester. Thus the cause of student dropout may not be determined accurately. Results may also only be inferred to the students in the study and other similar students at Montana State University, Bozeman.

5 Appendix

TERM	Category	No. of dropout	No. of censored	Total	% dropout
F2014	Low	5	151	156	3.21
	High	11	129	140	7.86
S2015	Low	8	100	108	7.41
	High	15	117	132	11.36
F2015	Low	40	431	471	8.49
	High	38	311	349	10.89
S2016	Low	10	180	190	5.26
	High	58	392	450	12.89
F2016	Low	24	395	419	5.73
	High	52	394	446	11.66
S2017	Low	25	239	264	9.47
	High	45	332	377	11.94

Table 14: Number of prior math courses taken by students and status of survival.

TERM	Class	No. of dropout	No. of censored	Total	% dropout
F2014	Young	9	210	219	4.11
	Old	7	70	77	9.09
S2015	Young	19	149	168	11.31
	Old	4	68	72	5.56
F2015	Young	56	605	661	8.47
	Old	22	137	159	13.84
S2016	Young	47	426	473	9.94
	Old	21	146	167	12.57
F2016	Young	50	617	667	7.50
	Old	26	172	198	13.13
S2017	Young	44	427	471	9.34
	Old	26	144	170	15.29

Table 15: Age classes of students in years by status of survival.

TERM	Category	No. of dropout	No. of censored	Total	% dropout
F2014	Low	9	106	115	7.83
	High	7	174	181	3.87
S2015	Low	10	83	93	10.75
	High	13	134	147	8.84
F2015	Low	37	219	256	14.45
	High	41	523	564	7.27
S2016	Low	33	181	214	15.42
	High	35	391	426	8.22
F2016	Low	33	275	308	10.71
	High	43	514	557	7.72
S2017	Low	29	183	212	13.68
	High	41	388	429	9.56

Table 16: MPLEX scores for students according to status of survival.

TERM	Category	No. of dropout	No. of censored	Total	% dropout
F2014	Low	4	50	54	7.41
	Moderate	11	212	223	4.93
	High	1	18	19	5.26
S2015	Low	5	42	47	10.64
	Moderate	18	171	189	9.52
	High	0	4	4	0.00
F2015	Low	18	111	129	13.95
	Moderate	59	558	647	9.12
	High	1	43	44	2.27
S2016	Low	22	135	157	14.01
	Moderate	42	411	453	9.27
	High	4	26	30	13.33
F2016	Low	18	152	170	10.59
	Moderate	58	587	645	8.99
	High	0	50	50	0.00
S2017	Low	19	133	152	12.5
	Moderate	48	406	454	10.57
	High	3	32	35	8.57

Table 17: ACT scores for students according to status of survival.

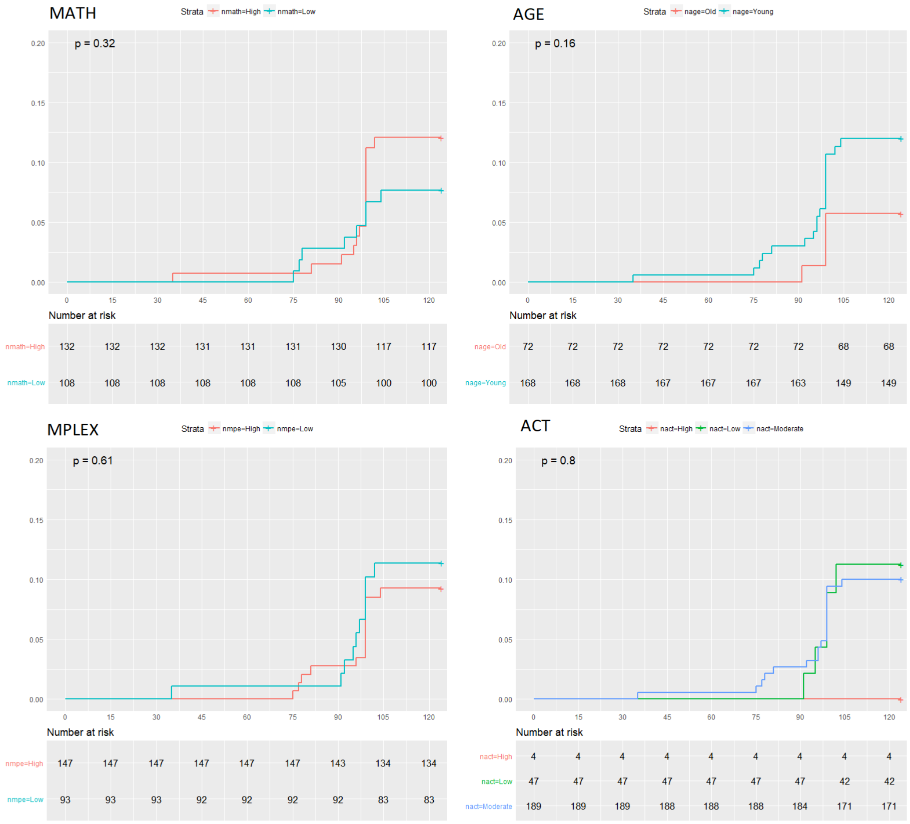
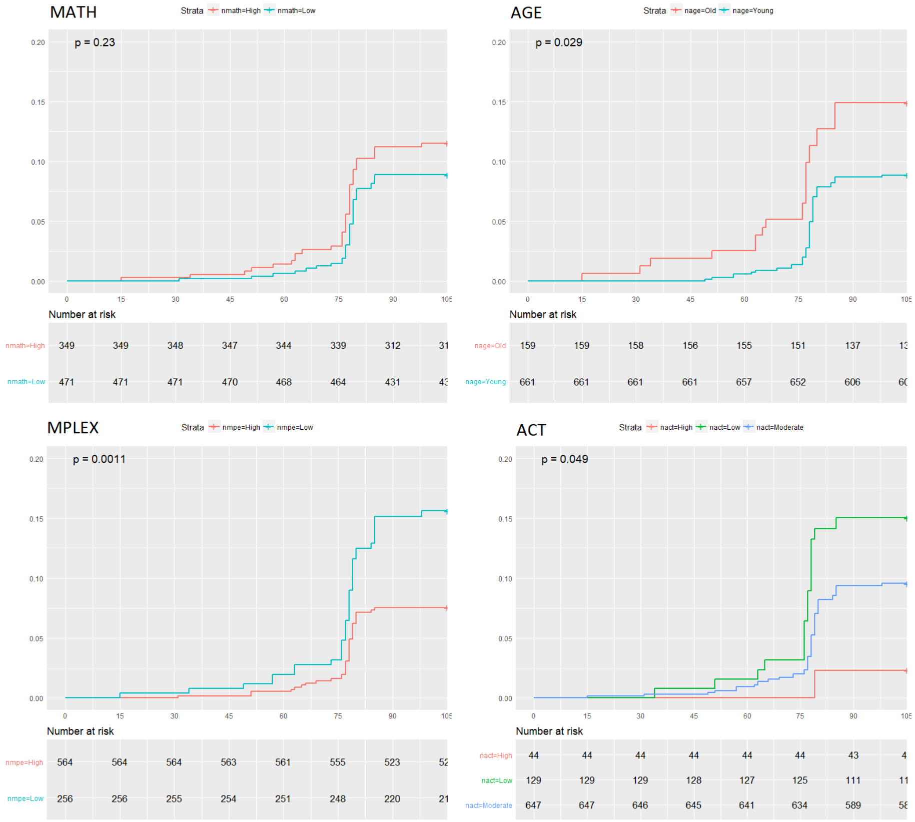


Figure 8: Log Cumulative Hazard Plot for each covariate in the S2015 semester



Number at risk

nmath=High	349	349	348	347	344	339	312	31
nmath=Low	471	471	471	470	468	464	431	43

Number at risk

nage=Old	159	159	158	156	155	151	137	13
nage=Young	661	661	661	661	657	652	606	60

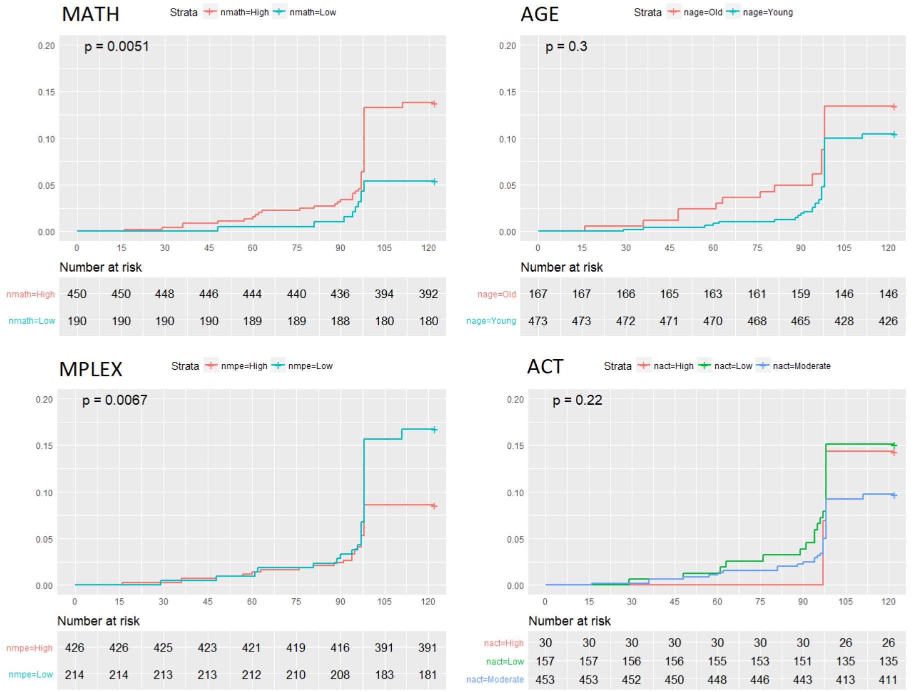
Number at risk

nmpe=High	564	564	564	563	561	555	523	52
nmpe=Low	256	256	255	254	251	248	220	21

Number at risk

nact=High	44	44	44	44	44	44	43	4
nact=Low	129	129	129	128	127	125	111	11
nact=Moderate	647	647	646	645	641	634	589	58

Figure 9: Log Cumulative Hazard Plot for each covariate in the F2015 semester



Number at risk

nmath=High	450	450	448	446	444	440	436	394	392
nmath=Low	190	190	190	190	189	189	188	180	180

Number at risk

nage=Old	167	167	166	165	163	161	159	146	146
nage=Young	473	473	472	471	470	468	465	428	426

Number at risk

nmpe=High	426	426	425	423	421	419	416	391	391
nmpe=Low	214	214	213	213	212	210	208	183	181

Number at risk

nact=High	30	30	30	30	30	30	30	26	26
nact=Low	157	157	156	156	155	153	151	135	135
nact=Moderate	453	453	452	450	448	446	443	413	411

Figure 10: Log Cumulative Hazard Plot for each covariate in the S2016 semester

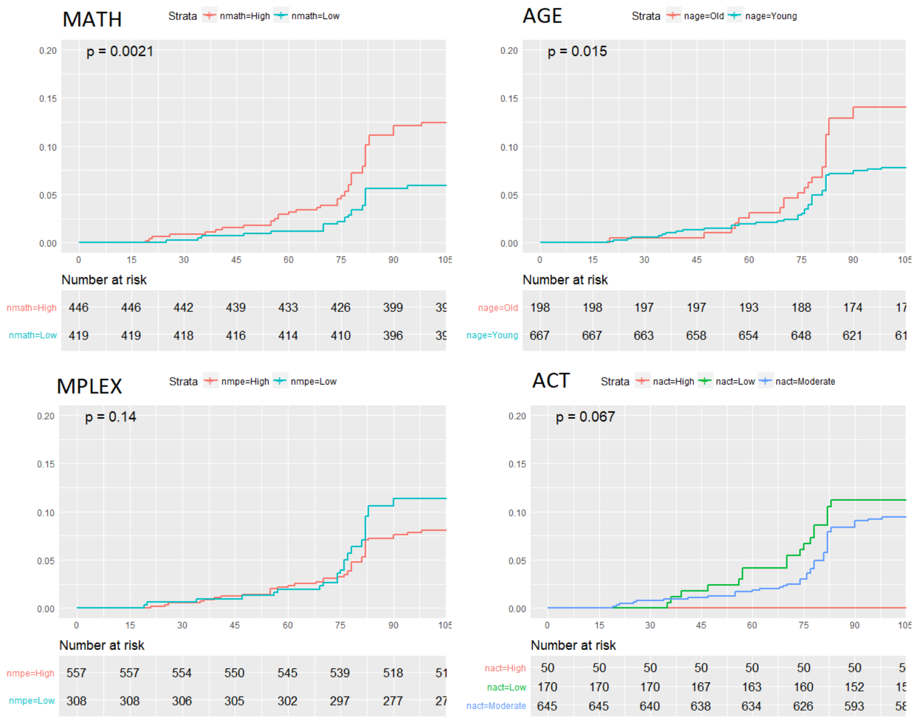


Figure 11: Log Cumulative Hazard Plot for each covariate in the F2016 semester

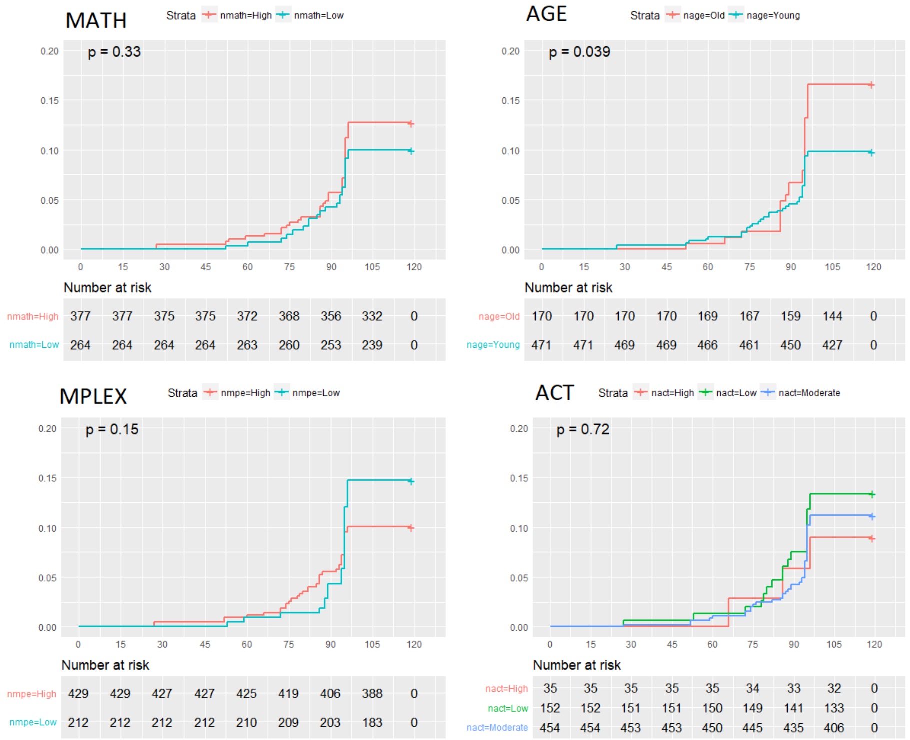


Figure 12: Log Cumulative Hazard Plot for each covariate in the S2017 semester

References

[1] C. Bungău, A. P. Pop, and A. Borza. Dropout of first year undergraduate students: A case study of engineering students. 3(1).

[2] Q. Chen, H. Wu, L. B. Ware, and T. Koyama. A bayesian approach for the cox proportional hazards model with covariates subject to detection limit. 3(1):32–43.

[3] D. Collett. *Modelling survival data in medical research*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Taylor & Francis Group, third edition edition.

[4] A. EKMAN, L. NILSSON, and H. LINDKVIST. A simulation study comparing the stepwise, lasso and bootstrap approach. page 84.

[5] J. Fan, Y. Feng, and Y. Wu. High-dimensional variable selection for cox's proportional hazards model.

[6] J. Fan, G. Li, and R. Li. *An Overview on Variable Selection for Survival Analysis*, pages 315–336. WORLD SCIENTIFIC.

[7] J. FAN and R. LI. VARIABLE SELECTION FOR COX's PROPORTIONAL HAZARDS MODEL AND FRAILTY MODEL. page 26.

[8] V. T. Farewell. An application of cox's proportional hazard model to multiple infection data. 28(2):136.

[9] J. Fox. Cox proportional-hazards regression for survival data. page 18.

[10] R. I. Garcia, J. G. Ibrahim, and H. Zhu. Variable selection in the cox regression model with covariates missing at random. 66(1):97–104.

[11] K. J. Jager, P. C. van Dijk, C. Zoccali, and F. W. Dekker. The analysis of survival data: the kaplan–meier method. 74(5):560–565.

[12] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. 53(282):457.

[13] Z. Ma and A. W. Krings. Survival analysis approach to reliability, survivability and prognostics and health management (PHM). pages 1–20. IEEE.

[14] D. M. A. Mohammed. Survival analysis by using cox regression model with application. 3(11):7.

[15] W. W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Taylor & Francis Group.

[16] M. Tableman and J. S. Kim. *Survival analysis using S: analysis of time-to-event data*. Texts in statistical science. Chapman & Hall/CRC.

[17] K. Tolosie and M. K. Sharma. Application of cox proportional hazards model in case of tuberculosis patients in selected addis ababa health centres, ethiopia. 2014:1–11.