IR Applications

Using Advanced Tools, Techniques, and Methodologies

Association for Institutional Research

Enhancing knowledge. Expanding networks.
Professional Development, Informational Resources & Networking

Volume 5, June 30, 2005

# Analyzing Student Learning Outcomes: Usefulness of Logistic and Cox Regression Models

Chau-Kuang Chen, Ed.D., Meharry Medical College

**Abstract**

Logistic and Cox regression methods are practical tools used to model the relationships between certain student learning outcomes and their relevant explanatory variables. The logistic regression model fits an S-shaped curve into a binary outcome with data points of zero and one. The Cox regression model allows investigators to study the duration and timeline of the critical events, which are also a binary and dichotomous measure. This paper introduces logistic and Cox regression models by illustrating examples, implementing step-by-step SPSS procedures, and further comparing the similarities and differences of the model characteristics. Logistic regression analysis was conducted to investigate the effects of the explanatory variables such as pre-admission variables, college cumulative GPAs, and curriculum tracks on student licensure examination. Moreover, logistic regression analysis was employed to quantify the effect (odds or odds ratio) of specific explanatory variables on the binary outcome holding other variables constant. With regards to Cox regression analysis, the outcome variable of interest was the timing of experiencing academic difficulty—dismissal, withdrawal, and leave of absence. The Cox regression model was used to detect when students were most likely to experience academic difficulty beyond their matriculation. The model also allowed the investigators to measure the effect (relative hazard or hazard ratio) of specific risk factors on the academic difficulty after adjusting for other factors. Identifying the occurrence of critical events along with the explanatory variables, college administrators and faculty could implement intervention strategies to ensure student success.

## Introduction

Regression analyses are statistical procedures that describe the relationship between an outcome (dependent, or response) variable and one or more explanatory (independent, or predictor) variables or risk factors. The choice of regression models depends largely on the research objectives and the measurement scales of the outcome variable in the study. Logistic regression analysis is a suitable technique to investigate the effects of the explanatory variables on the binary outcome, either success or failure. Moreover, it allows investigators to perform predictions based on the resulting model. The Cox regression model becomes the appropriate choice for studying the risk factors in relation to the duration and timeline until occurrence of the critical event, which is also a binary measure. However, the model itself does not gear up for the purpose of future predictions.

Logistic regression analysis has been widely used to study student performance, enrollment, graduation, and post-graduate employment status, which are restricted to a binary outcome such as "success or failure', 'enrolled or not enrolled', 'graduated or not graduated', and 'primary care or non-primary care specialty' (Case, et al., 1994; Sadler, et al, 1997; Strayhorn, 2000; and Hojat, 1995). The logistic regression method allows investigators to estimate and interpret the effects of the explanatory variables on the binary outcome. The maximum likelihood estimation technique is readily available to estimate the regression coefficients for the non-linear model equation. This technique maximizes the probability of getting the observed data given the fitted regression coefficients (Hosemer and Lemeshow, 1989; Eliason, 1993; and Pampel, 2000). The model equation renders itself to perform future predictions to assess the predictive power. In addition, investigators can transform the model equation into the odds and odds ratio of the event occurrence. The odds or odds ratio is a measure of the direction and strength of the relationship between the explanatory variables and the effects of such variables on the binary outcome. The logistic regression model quantifies the association between the explanatory variables and the outcome variable of interest after adjusting for other explanatory variables (Matthews and Farewell, 1996). Therefore, it is a useful tool to analyze the

dependence of a binary outcome on a set of explanatory variables.

Cox regression analysis is a branch of survival analyses, which is used to analyze the timeline until a critical event occurs (Cox, 1972; Cox and Oakes, 1984; and Miller 1981). Survival analyses discussed in this paper appear in diverse fields under a variety of names. In biomedical and health sciences, these are called survival analyses because the event of interest is mortality or death. In sociological research, these are often referred to as event history studies. In engineering studies, the term reliability theory is commonly applied to the lifetime of an object. Cox regression analysis allows investigators to answer two questions simultaneously: "Has the critical event of interest occurred?" and "Which risk factors contribute to the event occurrence?" (Singer and Willett, 1991). In Cox regression analysis, the hazard and survival functions are expressed as a function of time and the risk factors, which enable investigators to address research questions in education such as: "How many semesters elapse before students drop out of school?", "At what year are students likely to graduate from the college?", and "What explanatory variables contribute to students' dropout or graduation?" Numerous studies related to the timing of student departure from college were conducted during the past decade (DesJardins et al., 1997; Han and Ganges, 1995; Huff and Fang, 1999; Ronco, 1994; Singer and Willett, 1993; and Willett and Singer, 1991). Moreover, in the Cox regression model, investigators can determine the effectiveness of college programs by comparing the patterns of the survival functions. For example, if the survival function for Group A consistently lies above that of Group B, the intervention program would appear more effective for Group A during the study period.

In the Cox regression model, the hazard function is a function of a set of risk factors and baseline hazard. The risk factors are similar to independent variables of the linear regression model except they appear to be non-linear in the exponential expression. The baseline hazard is similar to the constant or intercept of the linear regression model. It describes the overall level of risk and reveals the main effect of the time variable (Singer and Willett, 1991). The hazard function displays the timeline fluctuation of a student experiencing the critical event. The student with the greater risk has a higher value of the hazard function as compared to one with the lesser risk at that same particular time. Thus, a high hazard function indicates the critical event is more likely to occur. Conversely, a low hazard function shows that the critical event is less likely to occur (Kleinbaum, 1996). For instance, investigators can detect the ineffectiveness of the support services program by comparing the hazard functions. If the hazard function is higher for Program A than Program B, Program A appears to be ineffective during the study period.

Similar to the logistic regression analysis, the Cox regression model allows investigators to estimate and interpret the effects of the risk factors on the event occurring. The partial maximum likelihood estimation technique is used to estimate the regression coefficients of the model equation (Hosemer and Lemeshow, 1989; and Kleinbaum, 1996). The interpretation of the regression coefficients in the Cox regression model is virtually the same as in logistic regression analysis. For the positive regression coefficient, the hazard of a student experiencing the critical event increases as the value of the risk factor increases. Moreover, if the regression coefficient is zero, the value of the relative hazard becomes one, indicating that the hazard of a student experiencing the critical event is not affected by the risk factor. However, for the negative regression coefficient, the hazard of a student experiencing the critical event decreases when the value of the risk factor increases. For example, if student dropout is the critical event of interest, the Cox regression model can be used to study the timeline and the risk factors associated with student dropout.

The model assumption of Cox regression is very different from that of the logistic regression model. Logistic regression assumes that residuals are normally distributed with a mean of zero and a constant variance. However, the Cox regression model assumes that the hazard ratio for different students with different values of the risk factor is independent of the time variable (Belle, et. al., 2004; and Kleinbaum, 1996). Thus, when the two hazard functions are proportional, the title 'proportional hazards model' is applied to Cox regression. The proportionality of the model assumption implies that the hazard ratio is constant over time. Therefore, the verification of the model assumption is essential to ensure the model appropriateness.

In an attempt to demonstrate the usefulness of the two methods, this paper provides a brief overview and example illustrations. In particular, it offers step-by-step SPSS PC commands used for analyzing a binary result or event occurring related to student-learning outcome. In this study, the research objectives and questions are defined as the foundation of the investigation. Next, the study variables are gathered to form a research-oriented database, which are extracted from the existing student tracking system. In addition, the principle of model strategies serves as a guideline to build the models. It comprises all relevant variables at the initial phase of the model fitting and achieves parsimony and consistency upon model completion. The paper focuses mainly on the model equations, the assessment of model fittings, the interpretation of odds ratio and hazard ratio, and the importance of the model assumptions. Moreover, the two methods are placed side by side to analyze the similarities and differences of their characteristics.

## Literature Review

Using logistic regression analysis (Sadler et al., 1997), the second semester enrollment was discovered to be significantly associated with the explanatory variables such as tuition benefits, state residency, race, high school GPA, and summer orientation. The results of this study could help the institution implement appropriate interventions that target students at-risk for leaving. To investigate student retention at Historical Black Colleges and Universities, a logistic regression model was constructed (McDaniel and Graham, 1999). The research findings showed that students were more likely to return for the second year if they earned higher ACT test scores. In addition, students aspiring to achieve doctoral and professional degrees were more likely to persist than students with lower degree aspirations. These study results could assist college administrators and faculty to select students who are more likely to succeed at their institution.

By using logistic regression analysis, one study found that student performance in the college program predicted whether or not students would apply to medical school, get accepted by medical school, and graduate from medical school (Strayhorn, 2000). In constructing this logistic regression model, the significant explanatory variables for the probability of passing the United States Medical Licensure Examination (USMLE) Step 1 were medical college admission test scores, medical school freshman GPAs, sophomore course performances, and financial aid support (Chen et al., 2001). These study results could be used to document the effectiveness of academic programs and applicant screening. Logistic regression analysis (Case et al., 1994) was also conducted to investigate the relationship between the initial performance of identifying examinees and the ultimate pass rates of the USMLE Steps 1 and 2. The study results showed that the probability of ultimately passing both Steps 1 and 2 was significantly related to the initial score achieved. Based on the availability of student performance measures, professional activities, satisfaction results, and research productivities, a logistic regression model was able to predict primary care and non-primary care status from the significant predictors, which include specialty interest, professional plan, and interests expressed in medical school (Hojat, 1995). The study results could be used to document the different tracks of physician training and education.

By constructing hazard models for different college exiting modes (graduation, withdrawal, transferal, dismissal, and leave of absence), investigators could better understand the different factors that influence student behavior (Singer and Willett, 1991). A survival analysis study indicated that the academic resource index significantly influenced graduation (DesJardins and Moye, 2000). Using survival analysis, another study (Han and Ganges, 1995) revealed the occurrence of crisis for a selected group of students who persisted for four or more years and still left their university prior to graduation. The results from this study suggest that students in the risk group participate in the intense and sustained intervention program.

The result of survival analysis indicated that some students experiencing academic difficulty remained at risk throughout the first three years of medical school (Fang, 2000). The research finding implied that academic support programs focusing only on the entering year might not be sufficient to fully address this extended period of risk. The study results of survival analysis (Huff and Fang, 1999) demonstrated that the increase of the relative risks of students experiencing academic difficulty were associated with low MCAT scores, low science GPA, low undergraduate institutional selectivity, being a woman, being a member of a racial-ethnic underrepresented minority, or being older. Clearly, investigators in higher education can use survival analysis to identify students who are most likely to experience the occurrence of critical events—dropout, withdrawal, dismissal, and delay of graduation. Knowing the time-to-event occurrence and related risk factors, college administrators and faculty can effectively implement the intervention strategies to increase the likelihood of student success.

## Logistic Regression Equation

The logistic regression model is primarily written as $Y = P(X) + E$, where Y is the binary outcome—event occurring coded as 1 or event not occurring coded as 0 (Hosemer and Lemeshow, 1989). The probability P(X) of obtaining the binary outcome is considered to be the estimated value given the explanatory variables (X) are known observations. The error term (E) also called the residual, represents the difference between the actual binary outcome (Y) and the estimated probability P(X). The model is commonly written as $P(X)=e^Z/(1+e^Z)$ or the equivalent form of $P(X) = 1/(1+e^{-Z})$, where Z stands for a linear combination of $\beta_o + \beta_1X_1 + \beta_2X_2 +...+ \beta_pX_p$ (Hosemer and Lemeshow, 1989). The "e" term in the equation is the base of the natural logarithm, which is approximately 2.718. The regression coefficients (β) are unknown parameters to be estimated. Moreover, the model assumes that residuals have a mean of zero and a constant variance of P(X)[1-P(X)], which are statistically independent of one another.

In a logistic regression model, the probability of a student obtaining a binary outcome is always in the range of zero to one, regardless of the value of Z. When Z is negative infinity for the model equation $P(X) = 1/(1+e^{-Z})$, the probability of a student obtaining a binary outcome is virtually zero ($e^{-Z}$ becomes positive infinity because of minus negative infinity of Z; and one divided by one plus positive infinity equals zero). Also, when Z increases from negative infinity to zero for the model equation $P(X)=\mathbf{e}^Z/(1+\mathbf{e}^Z)$, the probability of a student obtaining a binary outcome increases from zero to one half ($\mathbf{e}^Z$ equals one

when Z is zero; and the model equation shows that one divided by two equals one half). When Z increases from zero to positive infinity for the model equation $P(X)=\mathbf{e}^{Z}/(1+\mathbf{e}^{Z})$, the probability of a student obtaining a binary outcome increases from one half to one (positive infinity of $e^{Z}$ divided by one plus positive infinity equals one). The logistic regression equation is the steepest when the probability equals 0.5 and flattens out on both top and bottom tails as the probability value approaches zero and one (Dey and Astin, 1993). All of the characteristics mentioned above allow investigators to fit the S-shaped curve into a data set of binary outcomes containing values of zero and one.

The logistic regression equation is a non-linear curve rather than a straight-line. Thus, the parameter estimation method for the logistic regression model is called the maximum likelihood estimation, which is different from the least-square estimation method for the linear regression model. Therefore, it is necessary to use the iterative process via the computer software to find better approximations of the logistic parameters that satisfy the log likelihood equation. When the log likelihood equation is satisfied or maximized, the probability of an individual obtaining the observed data is maximized (Eliason, 1993; Pampel, 2000; and Hosemer and Lemeshow, 1989). In other words, the solution of the log likelihood equation implies that the effect of the explanatory variables on the probability of event occurrence is also maximized.

When all regression coefficients are estimated, the values of the explanatory variables can be plugged into the logistic equation to perform predictions. For example, if the probability of a student obtaining a binary outcome is greater than or equals to a cut-off point (defaulted value of one half) that student is placed into the success group. On the other hand, if the probability of a student obtaining a binary outcome is less than one half, then that student is categorized into the failure group (SPSS, Inc, 2002). Therefore, by comparing the predictive results and actual observations, investigators can calculate the prediction accuracy for the success and failure groups, as well as for the combined success and failure group. The cut-off point for predicting the binary outcome can be adjusted either upward or downward in order to increase specificity (accuracy of the prediction results for the failure group) or sensitivity (accuracy of the prediction results for the successful group) of the model equation.

To assess the overall model fitting and the significance of specific explanatory variables, the logistic regression model allows investigators to perform the likelihood ratio and Wald tests. The likelihood ratio test compares the likelihood for the intercept only model to the likelihood for the model with the explanatory variables (Eliason, 1993; Pampel, 2000; and Hosemer and Lemeshow, 1989). The logic of hypothesis testing for the likelihood ratio test in logistic regression is similar to the F test in the linear regression model. Investigators may conclude that at least one of the explanatory variables contributes to the probability of the student obtaining a binary outcome if the p value is less than the predetermined significance level ($\alpha$=.01, .05, or .001). In logistic regression analysis, the Wald statistic is commonly used to test the significance of the individual logistic regression coefficients (Hosemer and Lemeshow, 1989). Moreover, the logic of hypothesis testing for the Wald test in logistic regression is similar to the t test in the linear regression model. In the case that a specific regression coefficient is significantly different from zero, the corresponding explanatory variable significantly contributes to the probability of a student obtaining the binary outcome.

### Interpretations of Odds, Log Odds, Odds Ratio, and Delta P

By means of the mathematical transformation, investigators can transform an estimated probability into odds, log odds, odds ratio, and delta p, respectively. An example of a simple logistic regression model containing only one explanatory variable is used to illustrate this transformation process and interpret the resulting statistics. The odds can be derived as a ratio of the two probabilities, which is written as odds = P(event occurring)/P(event not occurring) = $P(X)/[1-P(X)] = e^{(\beta o + \beta X)}$, where $P(X) = e^{(\beta o + \beta X)}/[1 + e^{(\beta o + \beta X)}]$ (Hosemer and Lemeshow, 1989). The term $e^{\beta o}$ refers to the value of the odds for the constant and $e^{\hat{a}}$ represents the value of the odds related to the explanatory variable. The odds ($e^{\beta}$) can be interpreted as a stand-alone statistic without the involvement of the change (increase or decrease) of the explanatory variable. For instance, when the value of the odds equals three (0.75/0.25), it means that the odds of student obtaining a binary outcome (event occurring) is three times higher than that of the same student not obtaining a binary outcome. However, if the value of the odds equals one (0.5/0.5=1), the odds of obtaining a binary outcome is equivalent to the chance of obtaining a head when flipping a fair coin.

Because the base of the natural logarithm is applied to both sides of the odds equation mentioned above, the log odds known as logit can be written as $\log_e(\text{odds}) = \log_e\{P(X)/[1-P(X)]\} = \beta_o + \beta X$ (Hosemer and Lemeshow, 1989). Given the linear relationship exists between the dependent variable (log odds) and the independent variable (X), the interpretation of the effect of a specific explanatory variable is quite similar to linear regression analysis. The log odds can be interpreted as a unit change in the explanatory variable leads to the magnitude change (â units) of the log odds of a student obtaining a binary outcome. The log odds provides only the direction of the relationship, but is limited in providing meaningful information about the effect of the explanatory variable. As a result, it is difficult to understand the meaning of log odds. Therefore, the odds ratio should be used to interpret the effect of the explanatory variable.

The definition of odds ratio is different from that of the odds. Recall from the previous paragraph, that the odds is a ratio of the two probabilities when one is confined to the event occurring and the other refers to the event not occurring. The odds ratio is the ratio of two odds when the different values of the explanatory variable apply to them. The odds ratio is known as odds change, which describes the proportionate change in the odds for one-unit difference in the explanatory variable (Hosmer and Lemeshow, 1989; and Menard, 1995). It is a measure of the association between the explanatory variable and the odds of the individual obtaining a binary outcome.

An odds ratio of greater than one shows that the odds of an event occurring increases when the value of the explanatory variable increases (Menard, 1995). It demonstrates the existence of a positive relationship between the explanatory variable and the effect of that particular variable. The odds ratio is simplified to be an exponential expression ($e^1$). One can use the odds of $e^{\beta o + \beta X}$, where $\beta > 0$ to illustrate the positive relationship and interpret the meaning of the odds ratio. For the odds of $e^{\beta o + \beta}$ when X=1 versus the odds of $e^{\beta o}$ when X=0, the resulting odds ratio is $e^1$, which is a ratio of $e^{\beta o + \beta}$ and $e^{\beta o}$. Again, for the odds of $e^{\beta o + 2\beta}$ when X=2 versus the odds of $e^{\beta o + \beta}$ when X=1, the resulting odds ratio is still $e^{\beta}$, which is a ratio of $e^{\beta o + 2\beta}$ and $e^{\beta o + \beta}$. For example, this ratio implies that an average one-unit of increase in the explanatory variable leads to an increase in the odds of obtaining a binary outcome by a factor of $e^{\beta}$. Also, using tutorial program and student success as an example, if the odds ratio equals 2, it can be interpreted that an average one-month (time frame) of increase in the implementation of a tutorial program contributes to an increase in the odds of a student's success by a factor of 2.

However, an odds ratio of less than one indicates that the odds of an event occurring decreases when the value of the explanatory variable increases (Menard, 1995). In this case, the odds ratio is simplified to be an exponential expression ($e^{-\beta}$). One can use the odds of $e^{\beta o - \beta X}$, where $\beta > 0$ to illustrate the negative relationship and interpret the meaning of the odds ratio. When the explanatory variable (X) increases by one unit (from X=0 to X=1), the odds change increases by a factor of $e^{-\beta}$, which is the ratio of $e^{\beta o - \beta}$ and $e^{\beta o}$. Moreover, when the explanatory variable (X) increases by one unit (from X=1 to X=2), the odds change increases by a factor of $e^{-\beta}$, which is the ratio of $e^{\beta o - 2\beta}$ and $e^{\beta o - \beta}$. It is difficult to convey the concept of odds change for a fraction between zero and one. Therefore, a better way of interpreting a negative regression coefficient is to say an average one-unit of increase in the explanatory variable leads to a decrease in the odds for obtaining a binary outcome by a factor of $e^{\beta}$, which is an inverse of $e^{-\beta}$ ($0 < e^{-\beta} < 1$) Using anxiety score and student success as an example, if the odds ratio equals 1/2, it can be interpreted that an average one-point scale of increase in anxiety score leads to a decrease in the odds of a student's success by a factor of 2, which is the reciprocal of 1/2.

Although odds and odds ratio are commonly used for interpreting logistic regression results, the delta-p statistic can be used to measure the effect of a specific explanatory variable on the probability of obtaining the binary outcome (Peng et al., 2002). Using financial aid and student success as an example, a delta-p of .10 for a student receiving financial support is interpreted as increasing the probability of a student's success by 10% as compared to student who does not receive financial support. The delta-p is transformed from the regression coefficient using a method originally recommended by a researcher, Peterson, in 1984 (Peng et al., 2002). The odds ratio can be found in SPSS printouts when logistic regression analysis is performed, however, the delta-p statistic cannot be generated by the SPSS PC Version 12.0 commands.

**Example of Logistic Regression Analysis**

A logistic regression model was constructed for a sample of 200 matriculated students randomly selected from the population in 1993-1997. The study attempted to answer the research question, "How well can the USMLE Step 1 pass status be predicted by the explanatory variables: demographics, admission test scores, undergraduate GPA, medical school cumulative GPA, course grades, and financial aid support?" Thus, the research objectives of this study were: (a) to identify explanatory variables that significantly contribute to the probability of passing the USMLE Step 1; and (b) to provide insight into the measure (odds or odds ratio) of the effects of the explanatory variables.

In this study, the outcome or response variable was binary, i.e. USMLE Step 1 pass status, which was (labeled as p-f-grp for the SPSS command), coded 1 for the pass group and 0 for the fail group. The probability of a student passing USMLE Step 1 was a continuous scale between zero and one. The explanatory variables were a combination of the discrete measure (gender, race, and medical school curriculum track) and the continuous measure (undergraduate basic sciences average, undergraduate GPA, medical college admission test scores—MCAT physical sciences, MCAT biological sciences, MCAT verbal reasoning scores, medical school freshman GPA, number of sophomore courses failed, and financial aid loan amount). The coding scheme for the discrete measurement was gender (1 for male and 0 for female), ethnicity (1 for African American and 0 for Non-African American), historical black colleges and universities status (1 for HBCU graduate and 0 for Non-HBCU graduate), and medical school curriculum track (labeled as curr_grp for the SPSS command, 1 for four-year curriculum track and 0 for five-year curriculum track).

The logistic regression model was constructed using the forward selection procedure. At each step, the explanatory variable with the smallest significance level for the Wald statistic was entered into the model. The default entry criterion for the explanatory variables was a p value of .05. The Wald statistics for all variables in the model were examined and the explanatory variable with the largest p value for the Wald statistic was removed from the model. The default removal criterion was p=.10. If no explanatory variables met the removal criterion, the next eligible variable was entered into the model. The iteration process for selecting explanatory variables continued until no additional variables met the entry or removal criterion.

### SPSS PC Commands for Logistic Regression Analysis

The eight steps of the SPSS PC Version 12.0 commands required to produce logistic regression analysis are as follows:

Step 1 - Click *Analyze*, click *Regression*, and click *Binary Logistic*; Step 2 – Click on dependent variable (*p_f_grp*), and click <right arrow> sign to move it to the dependent box; Step 3 - Hold down the *CTRL* key, click all independent variables (*ung_bsa, ung_gpa, mcat_vr, mcat_ps, mcat_bs, curr_grp, gender, ethnic, hbcu, course2f, fresh_gp,* and *loan_amt*), and click <right arrow> sign to move them to the covariates box; Step 4 - Click <down arrow> sign to display the method options and select *Forward-Wald*; Step 5 - Click C*ategorical* button; Step 6 - Hold down the *CTRL* key, and click on categorical variables (*curr_grp*, *gender, ethnic,* and *hbcu*) and click <right arrow> sign to move them to the categorical covariates box, and click *Continue*; Step 7 - Click the *Option* button, select c*lassification plots*, click display *At Last Step,* and click

Continue; and Step 8 - Click *OK.* Note that Step 8 – Click *Paste* to generate LOGISTIC REGRESSION syntax command lines as follows: LOGISTIC REGRESSION p_f_grp /METHOD = FSTEP(WALD) ung_bsa ung_gpa mcat_vr mcat_ps mcat_bs curr_grp gender ethnic hbcu course2f fresh_gp loan_amt /CONTRAST (curr_grp)=Indicator /CONTRAST (gender)=Indicator / CONTRAST (ethnic)=Indicator / CONTRAST (hbcu)=Indicator /CLASSPLOT /PRINT = SUMMARY / CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

### Major Findings for Logistic Regression Analysis

The sample size of 200 in this study was appropriate because the ratio (1 to 50) of the number of the explanatory variables (4) to the sample size (200) exceeded the minimum ratio of 1 to 10 recommended for the logistic regression study (Peng et al., 2002). It is crucial to assess the model collinearity, model assumption, model fitting, and model accuracy prior to interpretation of the research findings. The collinearity exists if one explanatory variable is a function of the other explanatory variables. There was no evidence of model collinearity because the tolerances ($TOL_i = 1 - R_i^2$, where $R_i$ is a multiple correlation coefficient between one explanatory variable ($X_i$) and the other explanatory variables (Xs) in the equation were large (a range of .571 to .971) based on a linear regression printouts (Norusis, 1985; and Menard, 1995). Moreover, the assumption of the logistic regression model was satisfied because the histogram of residuals appeared normally distributed with a mean of zero, and the residuals on the scatter diagram also appeared to be parallel with the X-axis, the indication of a constant variance.

As shown in the footnote of Table 1, the logistic regression model containing the four explanatory variables fits the data well based on the model chi-square test ($X^2$=83.40, df=4, and $p$<.001) and the small value of the – 2 log likelihood (159.23). Also, the improvement chi square value ($X^2$=38.01, df = 4, and $p$<.001) indicated that the model fits the data better than it had initially. The model-fitting statistic, namely the pseudo R-square, measured the success of the model by explaining the variations in the data. The pseudo R-square for Nagelkerke (0.49) was significantly different from zero. This indicated that forty-nine percent of variations in the binary outcome variable was accounted for by the explanatory variables. Finally, the prediction accuracy of 92% for the pass group and 84% for the combined pass and fail group were high. However, the prediction accuracy of 63% for the fail group was not high.

As illustrated in Table 1, the regression coefficients for the MCAT physical sciences score, MCAT biological sciences score, number of sophomore courses failed, and medical school freshman GPA were significantly different from zero at the .001 significance level using the Wald test. It was evident that these four explanatory variables significantly affected the USMLE Step 1 pass status. However, the research findings indicated that the undergraduate basic sciences average, undergraduate GPA, MCAT verbal reasoning score, gender, ethnicity, HBCU status medical school curriculum track, and financial aid support were not significantly associated with the licensure examination performances.

The logistic regression method yielded the following logistic regression equation to predict the USMLE Step 1 pass status: the estimated probability (Passing USMLE Step 1) = $P(X) = e^Z / (1 + e^Z)$, where e is the base of the natural logarithm, approximately 2.718; and $Z$ = - 12.21 + 0.62 * MCAT Physical Sciences Score + 0.51 * MCAT Biological Sciences Score - 2.53 * Number of Sophomore Courses Failed + 1.89 * Medical School Freshman GPA. Based on the contribution from each of the explanatory variables, the estimated probability can be obtained from this equation for a particular student. It can be said that if the estimated probability is greater than or equal to 0.5,

**Table 1**
**Logistic Regression Model for Predicting USMLE Step 1 Pass Status**

| Variables in the Equation | Logistic Regression Coefficient (ß) | Standard Error of ß SE (ß) | Odds Ratio ($e^{ß}$) |
|---|---|---|---|
| MCAT Physical Sciences Score | 0.62*** | 0.209 | 1.87 |
| MCAT Biological Sciences Score | 0.51*** | 0.144 | 1.66 |
| Number of Sophomore Courses Failed | -2.53*** | 0.639 | 0.08 |
| Medical School Freshman GPA | 1.89*** | 0.533 | 6.59 |
| Constant | -12.21*** | 2.317 | 0.00 |

*** p < .001 based on the Wald test [$X^2 = (ß/SE(ß))^2$ with df=1]
N=200
-2 Log Likelihood: -2LL=159.23
Model Chi Square Test: $X^2$=83.40, df=4, and p<.001
Improvement Chi Square Test: $X^2$=38.01, df=4, and p<.001
Pseudo R-square: Nagelkerke R-square = 0.49
Prediction Accuracy: Pass Group (92%), Fail Group (63%), and Combined Group (84%)

a student advances to the pass group of the USMLE Step 1. However, if the estimate probability is less than 0.5, a student falls into the fail group of the USMLE Step 1.

The effect (odds ratio) of each explanatory variable on the pass status of the USMLE Step 1 is shown in the last column of Table 1. When an average of the medical school freshman's GPA increased by one point, the odds of a student passing USMLE Step 1 increased by a factor of 6.59. This value indicates that the effect of the medical school freshman GPA on the USMLE Step 1 performance was very high. A change in the MCAT scores definitely led to a change in the USMLE Step 1 performances. If an average of the MCAT physical sciences score increased by one point, the odds of a student passing USMLE Step 1 increased by 87%. Meanwhile, when an average of the MCAT biological sciences score increased by one point, the odds of a student passing USMLE Step 1 increased by 66%. Furthermore, the odds of a student passing USMLE Step 1 increased by a factor of 0.08 when the number of courses failed in sophomore year increased by one. In other words, the odds of a student passing USMLE Step 1 decreased by a factor of 12.5 (an inverse of 0.08) because of one additional sophomore course failed in the basic science disciplines. This implied that the effect of sophomore course performances on the USMLE Step 1 pass rate was extremely high. A noteworthy finding is that the odds of a student passing the licensure examination depended heavily on sophomore course performances and medical school freshman GPAs as compared to those of pre-admission variables—MCAT physical sciences and biological sciences scores.

In summary, the medical school freshman GPA and sophomore course performance were significant explanatory variables for the USMLE Step 1 performance. The two MCAT scores (physical and biological sciences) were also significant explanatory variables, regardless of other pre-admission variables such as gender, ethnicity, and undergraduate BSA and GPA scores. As expected, the medical school freshman GPA was strongly correlated with the Step 1 performance while the number of courses failed during sophomore year was negatively associated with the Step 1 performance. It is evident that basic science disciplines have a predictive power for the medical licensure examination. This may imply that the medical school has implemented an adequate basic sciences curriculum, course instructions, and student assessment compared to other medical schools in the nation.

### Survival Analysis

The Cox regression model can be used to study the risk factors that are significantly associated with the timing of the critical event. The timeline known as survival time refers to the length of the time interval (month, semester, or year) between the onset time of the study and the timing of the event occurrence (Steinberg, 1999). The critical events cover not only negative and unpleasant experiences, such as dropout and academic difficulty, but also positive and pleasant results, such as passing certain examinations and graduation.

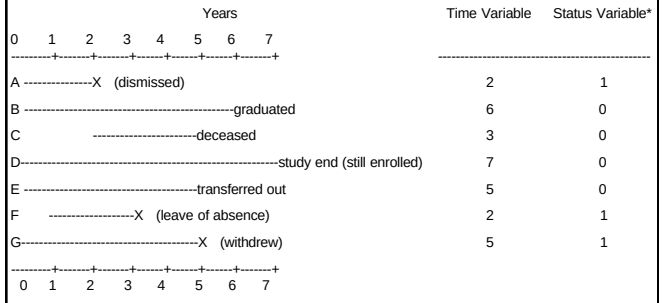
Survival analysis is a unique statistical approach for analyzing uncensored and censored data in a single study (Singer and Willett, 1991). With regard to uncensored data, they are often referred to as a complete set of timelines about the event occurring as opposed to some censored cases for which the event of interest does not occur during the study period (Kleinbaum, 1996; and Steinberg, 1999). For instance, if the graduation status within a two-year master's program is the event of interest, the timely graduation in two years can be considered as the uncensored data. Similarly, if the academic difficulty (dismissal, leave of absence, and withdrawal) from a seven-year doctorate program is the critical event of interest, the time-to-event for students A, F, and G illustrated in Table 2 belongs to the uncensored data. The censored data provide only partial information of timelines, which are collected under these circumstances: (1) students do not experience the event of interest (academic difficulty) before the study ends (student B - graduated, student D - still enrolled); (2) student withdraws from the study (student C - deceased); and (3) student is lost to follow-up during the study period (student E - transferred out). Because the occurrence of time-to-event for students (students B, C, D, and E) is not evident and hidden from view, the term 'censored' is applied to it. In essence, censoring occurs when investigators have some partial information about the time-to-event of the individual students, but they do not know the exact time variable. The survival time is measured in years from the time students enter the study. Students C and F entered the study at year two and one, respectively, which are different from the rest of the group who entered at the beginning of the study.

To understand survival analysis, one needs to begin with the survival function S(t). As indicated by researchers

(Braun and Zwick, 1993; Elandt-Johnson and Johnson, 1980; and Lee, 1992), the survival function can be expressed as S(t) = 1 – F(t) = P(T>t), where T represents a specific time of the event occurrence. In other words, the survival function cumulates the proportion of individuals surviving longer than the time variable, which is a complement of the cumulative distribution function. The survival function gradually decreases when more individuals experience the critical event. It depicts theoretically the survival experience of individuals from time zero to time infinity (Hosmer and Lemeshow, 1999; Kleinbaum, 1996; and Lee, 1992). The survival function has a probability value of one when the time variable equals zero and a probability value close to zero when the time variable approaches infinity. However, in practice, the estimated survival function is a step function rather than a smooth curve, which goes down at a step at each time interval to describe the survival events during the study period.

**Table 2**
**Example of Uncensored and Censored Data for the Doctorate Program**

The Outcome Variable is Academic Difficulty: Dismissed, Withdrew, and Leave of Absence because of Academic Reasons




| | Time Variable | Status Variable* |
|---|---|---|
| A | 2 | 1 |
| B | 6 | 0 |
| C | 3 | 0 |
| D | 7 | 0 |
| E | 5 | 0 |
| F | 2 | 1 |
| G | 5 | 1 |

* Status variable is coded as 1 for event occurring (uncensored data) and coded 0 for event not occurring (censored data).

If the survival function represents a positive or pleasant experience (e.g., surviving from dropout and academic difficulty), the hazard function h(t) is eventually considered as a negative or unhappy event (e.g., experiencing dropout and academic difficulty). The hazard function is a measure of the tendency of an individual to experience the critical event. Unlike the survival function, the hazard curve does not range from one to zero. Instead, it starts anywhere and goes up and down in any direction over time. The hazard function is designed to provide insight into the conditional failure rate (Kleinbaum, 1996). It explicitly refers to the instantaneous potential per unit time for the critical event to occur, given that individual has survived to a particular time (Hosmer and Lemeshow, 1999; Kleinbaum, 1996; and Lee, 1992). Investigators who have been exposed to calculus and mathematical statistics may desire to know the relationship of hazard and survival functions.

**Relationship between Hazard and Survival Functions**

The hazard function may be written as the conditional failure rate = h(t), where

$$h(t)=\lim_{\Delta t \to 0}\{P[t \le T < (t+\Delta t) \mid T > t] / [\Delta t]\}$$

$$=\lim_{\Delta t \to 0}\{P[t \le T < (t+\Delta t) \text{ and } T > t] / [\Delta t \; P(T>t)]\}$$

The definition of conditional probability, P(A|B) = P(A and B) / P(B) (Meyer, 1970), is applied to the hazard function above. That is P(A|B)= P [t ≤ T < (t+Δt) | T > t]; P(A and B) = P[t ≤ T< ( t+Δt) and T > t]; and P(B) = P(T > t), where A represents [t ≤ T < (t+Δt)], a specific survival time (T) of the event occurrence in a narrow time interval between t and t + Δt, and B refers to (T > t), a specific survival time (T) of the event occurrence greater than the time variable (t).

$$h(t)=\lim_{\Delta t \to 0}\{P[t \le T < (t+\Delta t) \text{ and } T > t] / [\Delta t \; P(T>t)]\}$$

$$= \lim_{\Delta t \to 0}\{P[t < T < (t+\Delta t)] / [\Delta t \; P(T > t)]\}$$

The method of combining sets (that is, events) to obtain a new set is applied to the numerators of the hazard function above. Given A and B are the two events, A∩B is the event which occurs if and only if A and B occur (Meyer, 1970). Therefore, the joint events of A and B, A∩B, or [t ≤ T< (t+Δt) and T > t] equals the overlap, intercept, or [t < T < (t+Δt)] area of the event A [i.e., t ≤ T< ( t+Δt)], and the event B (i.e. T > t).

$$h(t) = \lim_{\Delta t \to 0} \{P [t < T < (t+\Delta t)] / [\Delta t \; P(T>t)]\}= f(t) / S(t)$$

The hazard function h(t) above becomes a ratio of the two probabilities: the probability density function f(t) and the survival function S(t). The equality of the formula is based on two mathematical relations: (1) probability density function $f(t) = \lim_{\Delta t \to 0} \{P[t < T < (t+\Delta t)] / \Delta t\}$ as Δt ---> 0 (Lee, 1992); and (2) S(t) = P(T>t) (Braun and Zwick, 1993; Elandt-Johnson and Johnson, 1980; and Lee, 1992).

$$h(t) = f(t) / S(t) = \{d[1- S(t)]/dt\} / S(t) = - \{d [S(t)]/dt\} / S(t)$$

The equality of the hazard function above is derived from the substitutions of two mathematical expressions: (1) f(t) = d[F(t)]/dt; and (2) F(t) = 1 – S(t) into the equation h(t) = f(t) / S(t) (Braun and Zwick, 1993; Elandt-Johnson and Johnson, 1980; and Lee, 1992). Because the derivative (d/dt) with respect to the time variable is applied to F(t),

the hazard function can be reiterated as the change (slope) of an individual experiencing the hazard per unit time given that individual has survived longer than time (t).

There is a clearly defined relationship between the hazard and survival functions. Investigators can derive the hazard function from information regarding the survival function. As indicated by researchers, $h(t) = -[d\,S(t) / dt] / S(t) = -d[\log_e S(t) / dt]$ (Braun and Zwick, 1993; Elandt-Johnson and Johnson, 1980; and Lee, 1992). Note that the cumulative hazard function $H(t) = -\log_e S(t)$ or $H(t) = -\ln S(t)$ because the integral (or sum) is applied to both sides of this equation—$h(t) = -d[\log_e S(t)/dt]$. Investigators can also derive the survival function from information regarding the hazard function. Because $h(t) = -[d\,S(t) / dt] / S(t)$, the mathematical expression of the survival function can be written as $S(t)=e^{[-\int h(u)du]}$, where the integral ($\int$) is denoted by the area under the curve between u = zero and u = t (Elandt-Johnson and Johnson, 1980; Kleinbaum, 1996; and Lee, 1992). Therefore, it is clear that the survival function S(t) decreases as the hazard function h(t) increases, and vice versa.

### Kaplan-Meier Survival Analysis

Before proceeding to the Cox regression model, it is imperative to describe several terms that are frequently used in survival analysis: Kaplan-Meier estimator, survival function, conditional probability, and log-rank test. The Kaplan-Meier estimator allows investigators: (1) to use the definition of conditional probability to derive survival functions for distinct groups; and (2) to determine the program effectiveness by comparing the survival function among groups, respectively.

Because students' survival for the subsequent time interval depends on students' survival from the previous time interval, the Kaplan-Meier formula is merely the application of the conditional probability (Belle, et al. 2004; and Kleinbaum, 1996), which can be expressed as P(A and B) = P(B) P(A|B) as illustrated in Table 3. The formula can be interpreted as the probability of the occurrence of the joint events A and B equals the probability of the event B multiplied by the probability of the event A, given the occurrence of event B. Event A (After) refers to the occurrence of the student surviving for the subsequent time interval. Event B (Before) represents the occurrence of the student surviving for the previous time interval.

Applying the visual examination method on the survival curves, investigators can detect the difference between two survival functions. For example, if two survival functions are separated in the first half of the study period, but thereafter, are somewhat closer to each other, then a large gap forms. This suggests that the intervention strategy is more effective earlier during the study period. In Kaplan-Meier survival analysis, the log-rank test allows investigators to test the significant differences of survival functions at different follow-up times among the study groups (Kleinbaum, 1996). By comparing the survival curves for the experimental group (with the intervention strategy) and the control group (without the intervention strategy), investigators can determine the effectiveness of the intervention strategy if the experimental group appears to be superior.

**Table 3**
**Example of Calculating the Estimated Survival Function S(t) for the Doctorate Program***

The Outcome (or Event) Variable of Interest is Academic Difficulty: Dismissed, Withdrew, and Leave of Absence because of Academic Reasons

| Time (in years) $t_i$ | Risk Set*** $R(t_i)$ | Number of Events | Number of Censored | Estimated S(t)** or % of Surviving |
|---|---|---|---|---|
| 0 | 7 students' survival time $\geq$ 0 year | 0 | 0 | 1 |
| 2 | 7 students' survival time $\geq$ 2 years | 2 | 1 | 1 x 5/7 = .7143 |
| 5 | 4 students' survival time $\geq$ 5 years | 1 | 3 | .7143 x 3/4 = .5357 |

* Survival times (in years): $2^+$, $2^+$, 3, $5^+$, 5, 6, and 7 come from Table 2, where + stands for the occurrence of academic difficulty
** Kaplan Meier formula: P(B) P(A|B) = P(A and B), e.g., S(t=5)=(.7143) (3/4) =.5357
*** Each student in $R(t_i)$ has a survival time = $t_i$

### Cox Regression Equation

Unlike the Kaplan-Meier estimator, the Cox regression method allows investigators to generate the hazard function as a function of the time variable, risk factors, and baseline hazard. Investigators can calculate the measure (relative risk or relative hazard) of the risk factors and interpret the hazard ratio. The hazard ratio is the ratio of two hazard functions that allows investigators to measure the association between the risk factors and the effects of such factors on the risk functions.

The Cox regression model can be expressed as $h(t, X) = h_o(t)e^{\beta 1X1 + \beta 2X2 + ... + \beta pXp}$ (Hosmer and Lemeshow, 1999; Kleinbaum, 1996; and Lee, 1992). The hazard function, h(t, X), is a function of the baseline hazard $h_o(t)$ and the risk factors (X). The baseline hazard function is similar to the constant of the linear regression model, which represents the value of the hazard function before the risk factors are taken into account. The base of the natural logarithm is denoted by e that equals approximately 2.718. The risk factors may include continuous and categorical variables. The continuous variables, e.g., grade-point-averages and test scores are measured by an interval or ratio scale. The categorical variables, e.g., gender and course grades, are measured by a nominal and ordinal scale, respectively. The regression coefficients ($\beta$) are unknown parameters to be estimated by the partial maximum likelihood estimation approach (Hosemer and Lemeshow, 1989; and Kleinbaum, 1996). The term 'partial' likelihood is applied because the likelihood equation calculates probabilities only for cases of the event occurrence rather than all cases. For a specific value of the time variable, the hazard function depends on the quantities of $e^{\beta 1X1 + \beta 2X2 + ... + \beta pXp}$ and $h_o(t)$. It exhibits graphically the variation in the time-to-event occurrence (Kleinbaum, 1996). Because the hazard function is a non-linear curve,

it is necessary to use the iterative process to find better approximations of the regression coefficients that satisfy the partial likelihood equation.

To assess the model goodness of fit, the Cox regression model allows investigators to address the research question, "Does the model containing the variables in question tell us more about the outcome variable than a model that does not include these variables?" (Willett and Singer, 1991). In particular, it allows investigators to perform the likelihood ratio test, which compares the likelihood for the intercept only model to the likelihood for the model containing the risk factors within each model. The logic of hypothesis testing for the likelihood ratio test is the same for Cox regression and logistic regression models. Based on the significance of the likelihood ratio test, investigators can claim that at least one of the risk factors contributed to the relative hazard. In addition, Cox regression analysis uses the Wald test to examine whether or not each individual regression coefficient is significantly different from zero. The logic of hypothesis testing for the Wald test is also similar for both the Cox and logistic regression models. If an individual regression coefficient is significantly different from zero, the corresponding risk factor significantly contributes to the hazard function of an event occurring.

Cox regression analysis is known as the proportional hazards model because the model assumes that two hazard functions are proportional to each other over time (Kleinbaum, 1996). For example, given two students, if one initially has twice as much relative risk than the second student, then the relative risk for the first student is two times that of the second student at all time points. Also, given two student groups, if one group initially has three times the relative risk compared to the second group, the relative risk of the first group is three times more than that of the second group during the study period. In other words, the model assumes that the relative hazard is constant across time for two different students and student groups, respectively.

### Interpretations of Relative Hazard and Hazard Ratio

The exponential expression of each regression coefficient in the Cox regression model is called relative risk or relative hazard. The magnitude of the relative hazard indicates the direction of the association between the outcome variable and the corresponding risk factor. If the relative hazard is greater than one (i.e. positive regression coefficient), the hazard of a student experiencing the critical event increases as the value of the risk factor increases. This implies that the relative hazard is positively associated with the risk factor. Moreover, if the relative hazard becomes one (i.e. regression coefficient is zero), the risk factor has no effect. On the other hand, if the relative hazard is a positive fraction and less than one (i.e. negative regression coefficient), the relative hazard of a student experiencing the critical event decreases as the value of the risk factor increases. This implies that the relative hazard is negatively associated with the risk factor.

Investigators should focus on the hazard ratio (HR) when they study the effects of the risk factors on the critical event in Cox regression analysis. The value of the hazard ratio reflects the strength of the relationship between a specific risk factor and the effect of that factor. The hazard ratio is a ratio of two hazard functions, which can be expressed as $HR = [h_M(t)] / [h_N(t)] = [h_o(t).e^{\beta X}] / [h_o(t).e^{\beta X^*}] = e^{\beta(X - X^*)}$ (Kleinbaum, 1996). The hazard ratio is the ratio of the relative hazard to the risk factor (X) changes, i.e., the relative hazard of X = 0 compared to the relative hazard of X = 1. The hazard ratio decreasing or increasing is based on the positive or negative regression coefficient. The hazard ratio can also be interpreted as the multiplicative change, rather than additive change, in the hazard of a student experiencing the critical event based on every unit of the change in a specific factor, holding other factors as constant. If the hazard ratio of the $M^{th}$ group (without the intervention strategy) versus the $N^{th}$ group (with the intervention strategy) equals three, investigators may conclude that students in the $M^{th}$ group experience three times greater hazard compared to those in the $N^{th}$ group. Moreover, if the risk factor (X) with a value of four is for student M and the same risk factor with a different value (X*) of three is for student N, the hazard ratio equals $e^{\beta (X - X^*)} = e^{\beta (4 - 3)} = e^{\beta}$. Assuming the hazard ratio is two ($e^{\beta} = 2$, where $\beta = \log_e 2$), the hazard of a student experiencing academic difficulty is two times greater for student M compared to student N. Note that the baseline hazard function, $h_o(t)$, appears in both the numerator and denominator terms of the hazard ratio, which can be cancelled out if the proportional assumption of the hazard function is held true.

### Example of Cox Regression Analysis

A Cox regression model was used to analyze a sample of 200 matriculated students randomly selected from the population in 1993-1997. The study attempted to answer the following research questions: "How well can the following risk factors explain students encountering academic difficulty: demographics, undergraduate GPAs, medical college admission test scores, medical school academic performances, medical school curriculum tracks, and financial aid support amounts?" and "In which months do students experience the highest risk of academic difficulty based on the risk factors mentioned above?" Thus, the aims of the study were: (a) to identify risk factors that are significantly associated with the hazard ratio of students experiencing academic difficulty; (b) to provide insight into the measure (relative risk, relative hazard) of the effects of the risk factors mentioned above; and (c) to detect at what month students are most likely to experience academic difficulty.

The outcome variable of interest was a continuous

measure—the hazard function of students experiencing academic difficulty. Academic difficulty refers to dismissal, withdrawal, and leave of absence due to academic reasons. The time variable was the survival time in months. The time origin of this study was the matriculation year, which was not at the same calendar year for each matriculating class. A combined data set for a five-year period was adopted for this study based on the justification of class comparability in academic difficulty and risk factors. The study ended in the 36th month because no single student experienced academic difficulty beyond the 36th month within the study population. The status variable was labeled as difficult (*difficult* for the SPSS command) and coded as 1 for students who experienced academic difficulty and 0 for students who did not experience academic difficulty during the study period. The risk factors with a continuous measure were undergraduate basic-sciences average (BSA), undergraduate GPA, medical college admission test scores (MCAT physical sciences, biological sciences, and verbal reasoning scores), medical school freshman GPA, number of sophomore courses failed, and the financial aid loan amount. The risk factors with a discrete measure were gender (1 for male and 0 for female), ethnicity (1 for African American and 0 for Non-African American), historical black colleges and universities status (1 for HBCU graduate and 0 for Non-HBCU graduate), medical school curriculum track (labeled as curr_grp for the SPSS command, 1 for four-year curriculum track and 0 for five-year curriculum track).

This Cox regression model was constructed by means of the forward selection procedure. At each step, the risk factor with the smallest observed significance level of the Wald statistic was entered into the model. The default p value of .05 was the entry criterion for the risk factors. Next, all risk factors in the model were examined to see if they met the default removal criterion (p=.10). The Wald statistics for all risk factors in the model were examined, and the risk factor with the largest observed significance level for the Wald statistic was removed from the model. If there were no risk factors that met removal criterion, the next eligible risk factor was entered into the model. This process continued until no risk factors met entry or removal criterion.

**SPSS PC Commands for Cox Regression Analysis**

The 12 steps of the SPSS PC Version 12.0 commands required to perform Cox regression analysis are as follows: Step 1 - Click *Analyze*, click *Survival*, and click *Cox Regression*; Step 2 - Click on time variable (*month*), and click <right arrow> sign to move it to the time box; Step 3 – Click on status variable (*difficult*) and click <right arrow> sign to move it to the status box; Step 4 - Click the *Define Event* button, key in 1 in the single value box, and click *Continue*; Step 5 - Click all risk factors (*ung_bsa, ung_gpa, mcat_vr, mcat_ps, mcat_bs, gender, ethnic, hbcu, course2f, fresh_gp,* and *loan_amt*) and click <right arrow> sign to move it to the covariates box; Step 6 - Click the method options and select *Forward-Wald*; Step 7 – Click on stratifying variable (*curr_grp*) and click <right arrow> sign to move it to the strata box; Step 8 - Click the category option, click <right arrow> sign to move the covariates (*gender, ethnic, hbcu*) to the categorical covariates box, and click Continue; Step 9 - Click the *Plots* button, select *Hazard* plots, and click *Continue*; Step 10 - Click the *Option* button, select display model information; Step 11 – select *Display Baseline Function,* and click *Continue*; and Step 12 - Click *OK.* Note that Step 12 – Click *Paste* to generate COXREG syntax command lines as follows: COXREG month /STATUS=difficul(1)/ STRATA=curr_grp/CONTRAST(gender)=Indicator/ CONTRAST(ethnic)=Indicator/CONTRAST(hbcu)=Indicator /METHOD=FSTEP (WALD) ung_bsa ung_gpa mcat_vr mcat_ps mcat_bs gender ethnic hbcu course2f fresh_gp loan_amt /PLOT HAZARD /PRINT=SUMMARY BASELINE /CRITERIA= PIN (.05) POUT(.10) ITERATE(20).

**Major Findings for Cox Regression Analysis**

Data were analyzed to examine the relationship between the risk factors and the hazard of a student experiencing academic difficulty. The curriculum track demonstrated its significant contribution to the hazard function ($\beta$=-1.33, $p<.001$, and $e^{\beta}$=0.265), and the log-minus log (LML) plot of the survival functions appeared not to be parallel in the first run of Cox regression analysis. These findings support evidence of the violation of the proportional hazards assumption. Therefore, the stratified Cox regression model, stratifying on curriculum track, was implemented. No strong evidence of the violation of proportionality was found based on the parallel pattern of the LML plot of the survival curves.

As illustrated in the footnote of Table 4, the model fits the data quite well based on the model chi-square test ($\chi^2$=39.09, df=3, and $p<.001$) and the small value of the –2 log likelihood (280.24). Also, the improvement chi square ($\chi^2$=44.60, df = 3, and $p<.001$) indicated that the model fits the data better than it had initially. Note that the initial –2 log likelihood was 324.84 in which the model contained only the constant term. Because the three explanatory variables were added in the model, the –2 log likelihood became 280.24, a decrease (improvement) of 44.60 units.

In this study, the following eight risk factors were not significantly associated with the hazard of a student experiencing academic difficulty: undergraduate basic sciences average, undergraduate GPA, MCAT physical science, MCAT biological science, ethnicity, HBCU status, Medical school freshman GPA, and financial aid loan amount. However, the risk factor, MCAT verbal reasoning score, had the highest Wald statistic (32.5) and entered the model equation in the first step. This was followed by

the inclusion of two more risk factors—gender and number of sophomore courses failed. As shown in Table 4, the regression coefficients for MCAT verbal reasoning score and gender were significantly different from zero at the .001 significance level. Also, the regression coefficient for the number of courses failed during sophomore year was significantly different from zero at the .05 significance level.

It was evident that these three risk factors—MCAT verbal reasoning score, gender, and number of courses failed in sophomore year—were significantly associated with academic difficulty. The hazard ratio was used to express the measure of the effects of these risk factors. For instance, the number of courses failed in the sophomore curriculum exhibited the hazard ratio of 2.11. It indicated that a student who failed an additional course had about a two times greater hazard of experiencing academic difficulty as opposed to a student who did not. On one hand, the hazard ratio (0.35) for gender demonstrated that male students had 0.35 times greater hazard of experiencing academic difficulty compared to female students. Because the hazard ratio of 0.35 was a positive fraction and less than one, it is more meaningful to interpret that male students had almost three times (inverse of 0.35) less of a hazard than female students to experience academic difficulty. On the other hand, the hazard ratio (0.51) concerning MCAT verbal reasoning indicated that a one-unit increase in the MCAT verbal reasoning score was associated with a decrease in relative risk of academic difficulty by a factor of two (inverse of 0.51).

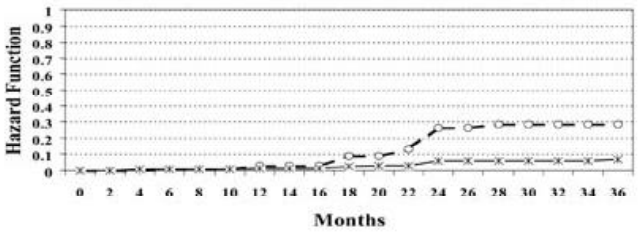
**Table 4**
**Cox Regression Model for Students Experiencing Academic Difficulty**

| Variables in the Equation | Logistic Regression Coefficient (ß) | Standard Error of ß SE (ß) | Odds Ratio ($e^{ß}$) |
|---|---|---|---|
| MCAT Physical Reasoning Score | -0.68*** | 0.119 | 0.51 |
| Gender (1 for male; 0 for female) | -1.05*** | 0.377 | 0.35 |
| Number of Sophomore Courses Failed | 0.75* | 0.309 | 2.11 |

*p < .05 and *** p < .001 based on the Wald test [$X^2 = (ß/SE(ß))^2$ with df=1]
N=200
-2 Log Likelihood: $X^2$=280.24
Model Chi Square Test: $X^2$=39.09, df=3, and p<.001
Improvement Chi Square Test: $X^2$=44.60, df=3, and p<.001

The hazard function h(t, X) against time (t) was plotted graphically based on the stratified Cox regression model. Figure 1 displays the hazard curves showing the four-year curriculum track is lower than the five-year curriculum track. This suggested that students in the four-year curriculum track were less likely to experience academic difficulty. According to the pattern of the hazard function, students in the five-year curriculum track experienced academic difficulty starting at the 20th month (first semester of the sophomore year), peaked at the 30th month (second semester of the sophomore year), and maintained the same level of hazard through the rest of the study period. The implication of this research finding was that the medical school should focus on academic support for students in the five-year curriculum track to address this extended period of academic difficulty.

**Figure 1**
**Hazard Curves for Students Experiencing Academic Difficulty**




### Similarities between Logistic and Cox Regression Models

As illustrated in Table 5, logistic and Cox regression models share several common characteristics. The model similarities are mainly reflected in the principle of modeling strategies, the objective of regression models, the method of parameter estimations, and the procedure of variable selections.

In the principle of modeling strategies, both models include all relevant explanatory variables at the initial stage of model fitting and achieve parsimony and consistency at the completion stage of model fitting. All relevant explanatory variables can be included in a single model with both methods as well because multiple effects can be simultaneously studied, and the effect of individual explanatory variables can be examined while others are held constant. Moreover, when fewer explanatory variables are sufficient to explain the occurrence of the event, investigators do not need elaborate explanations and unnecessary variables in models. Furthermore, investigators need to demonstrate the consistency of the model structure. It is important that significant explanatory variables and the effects of these variables are as identical as possible when the models are constructed over time.

Another characteristic that the two models have in common is the study objective. Both logistic and Cox regression models are applicable to the occurrence of the binary outcome or critical event. They are designed to study the relationship between certain student learning outcomes and their relevant explanatory variables. The logistic regression model allows investigators to identify explanatory variables that significantly contribute to the probability of a student obtaining a binary outcome while

the Cox regression model allows investigators to identify risk factors that significantly contribute to the hazard function of a student experiencing the critical event.

Both methods use the maximum likelihood estimation technique to estimate the regression coefficients and to construct the non-linear model equations. The principle of maximum likelihood is basically the same, although the Cox regression model uses the partial likelihood estimation (considering probabilities for the event occurrence rather than censored cases). The maximum likelihood estimation technique is applied to estimate the regression coefficients such that the likelihood of observing data is maximized. This technique has been discussed in many statistical books (Eliason, 1993; Hosemer and Lemeshow, 1989; Kleinbaum, 1996; and Pampel, 2000). First, the probability distribution (i.e., model equation) of students obtaining the outcome is prepared. Secondly, the joint probability distribution is derived based on the product of probability distributions under the assumption of independence. Thirdly, logarithm transformation is applied to the joint probability distribution to yield the log likelihood functions. Lastly, the iterative procedure is undertaken to estimate the regression coefficients. The iterative procedure begins with the value for individual parameters in the log likelihood functions, followed by a cycle of adjusting initial values to improve the model fitting. This procedure ultimately achieves the maximum likelihood estimates of the regression coefficients.

Logistic and Cox regression analyses use the same procedures such as enter, forward, and backward elimination to select significant explanatory variables (SPSS Inc, 2002). These procedures are briefly described as follows: (a) Enter procedure—All explanatory variables are forced to be included in the model in one step; (b) Forward stepwise procedure—Explanatory variables are included in the model one at a time based on the highest Wald or the likelihood-ratio statistics with the entry criterion (p=.05). As each new variable is added to the model, all of the existing explanatory variables in the model are evaluated for removal based on the Wald or likelihood-ratio test with the removal criterion (p=.10) When no more variables meet the entry or the removal criteria, the algorithm of selecting significant variables stops; and (c) Backward procedure—All of the explanatory variables are entered into the model during the first step. The explanatory variables that meet removal criterion (p =.10) are removed sequentially. When no more variables meet the removal criteria, the algorithm of selecting significant variables stops.

In both models, investigators utilize the –2 log likelihood value as criterion to make a judgment concerning the significance of the explanatory variables. The –2 log likelihood (-2LL), a chi square statistic, is important because it tells the probability of obtaining the binary outcome given the established parameter estimates. It is also a measure of how well the estimated parameters fit the data; a small value of the –2 log likelihood means the model fits the data well. (Norusis, 1985). Both models allow investigators to perform the likelihood ratio test that compares the likelihood for the intercept only model to the likelihood for the model containing the explanatory variables within each analysis. If the p value is less than the predetermined significance level (á=.05, .01, or .001), investigators may claim that at least one of the explanatory variables or risk factors in the model significantly contribute to the outcome variable. An additional similarity of the two models includes the use of the Wald statistic as criterion to test the association between individual explanatory variable and the outcome variable. If the p value for the Wald test is less than the predetermined significance level, investigators may conclude that a specific explanatory variable significantly contributes to the probability or the hazard function of experiencing a critical event.

Logistic and Cox regression models allow investigators to interpret the effect (odds ratio and hazard ratio) of specific explanatory variables on the outcome variable. The interpretations of the odds ratio and hazard ratio are the same, although the calculations are quite different. Logistic regression analysis uses the odds ratio ($e^{\beta}$) to indicate that an average one-unit of change in the explanatory variable leads to a change in the odds of a student obtaining a binary outcome by a factor of $e^{\beta}$. Cox

**Table 5**
**Summary of Similarities between Logistic and Cox Regression Models**

| Similarities of Two Models | Logistic and Cox Regression Analyses |
|---|---|
| Principle of Modeling Strategies | Inclusion of all relevant explanatory variables or risk factors at the initial stage of the model fitting; and achieving parsimony and consistency upon the completion stage of the model fitting |
| Objective of Regression Models | Logistic Regression:<br>To identify explanatory variables (X) that significantly contribute to the probability, P(X), of student obtaining a binary outcome<br><br>Cox Regression:<br>To identify risk factors (X) that significantly contribute to the hazard function h(t,X) for the duration and timeline of a student experiencing the critical event |
| Method of Parameter Estimations | The principle of the maximum likelihood estimation is applied to estimate the regression coefficients |
| Procedure of Variable Selections | Enter, forward, or backward procedures are used to select significant explanatory variables or risk factors to form the regression model |
| Test of Significant Predictors | Similar to the F test in linear regression, both models use minus two log likelihood (-2LL) test for the significance model fitting (All explanatory variables do not contribute to the outcome occurrence vs. At least one of the explanatory variables contribute to the outcome occurrence)<br>Similar to the t test in linear regression, both models use the Wald test for the significance of the individual regression coefficients. |
| Interpretation of Magnitude Effects | Logistic Regression:<br>To provide insight into the measure (odds, odds ratio, $e^{(\beta 0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$) of the effects of the explanatory variables<br><br>Cox Regression:<br>To provide insight into the measure (relative hazard, hazard ratio, or $e^{(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}$) of the effects of risk factors |

regression analysis utilizes the hazard ratio ($e^{\beta}$) to indicate that an average one-unit of change in the risk factor contributes to a change in the hazard of a student experiencing the critical event.

### Differences between Logistic and Cox Regression Models

As described in Table 6, the distinct characteristics of logistic and Cox regression models can be distinguished by the formation of research questions, the expression of model equations, the assessment of model fittings, and the verification of model assumptions.

Logistic regression analysis focuses on a critical event occurring or not occurring. Examples of research questions formed include "Will the student experience the occurrence of the critical event, 'yes' or 'no'?" and "What explanatory variables contribute to the occurrence of the critical event, 'success' or 'failure'?" However, Cox regression analysis is concerned with the duration and timing of the occurrence of the event. The research questions may be framed as "At what time (month, semester, year) does the student experience the critical event at the highest risk?", and "What risk factors contribute to the occurrence of the critical event at different time of the study periods?"

A key difference between logistic and Cox regression analyses is to use distinct model equations to study the relationship between the binary outcome and the explanatory variables. The logistic regression model is written as $P(X) = e^Z / (1+e^Z)$, where $Z = \beta_o + \beta_1X_1 + \beta_2X_2 +...+ \beta_pX_p$, and P(X) is the probability of obtaining a binary outcome. The Cox regression model may be expressed as $h(t,X)=h_o(t)e^Z$, where $Z = \beta_1X_1+ \beta_2X_2 +...+ \beta_pX_p$. The hazard function, h(t,X), is a function of time (t), the risk factors (X), and the baseline hazard $h_o(t)$. The baseline hazard is dependent on the time variable, acting as a constant term and contributing to the hazard function in a multiplicative manner.

A noteworthy difference between the two models is the pattern of non-linear curves. In the logistic regression model, the main focus is to estimate the probability of obtaining the binary outcome. The estimated probability is a continuous measure that begins with zero and increases as a smooth S-shaped curve. The value of the probability is between zero and one depending on the explanatory variables and the regression coefficients. However, for the Cox regression model, the hazard function is a continuous variable dependent on the duration and timeline of experiencing the critical event. The hazard function is a rate, which is ranged from zero to positive infinity. It begins with any positive value and goes up and down depending on the product of the baseline hazard and the risk factors.

Logistic regression, unlike the Cox regression model, has the capability of assessing the prediction power of the model and performing future predictions. In the logistic regression model, the prediction results can be used as criteria to make a judgment regarding accuracy of the classifications. The cross-tabulating method is used to categorize the predicted and the actual responses into a 2 by 2 table that indicates the accuracy of the classification results. However, in the Cox regression model, the classification results are not readily available to assess the model accuracy. Although the logistic regression model can be used to perform future predictions based on the known explanatory variables for prospective students, the Cox regression model is not capable of performing such predictions. On the right side of the Cox regression equation, all risk factors (X) for prospective students are the observed values that are ready to be placed into the equation. However, the time variable (t) in the baseline hazard $h_o(t)$ for prospective students is unknown, therefore it cannot be placed into the equation to perform future predictions.

The availability of pseudo R-squares also differs between the two models. The pseudo R-square measures the success of the model in explaining the variations in the

**Table 6**
**Summary of Differences between Logistic and Cox Regression Models**

| Differences of Two Models | Logistic Regression Analysis | Cox Regression Analysis |
|---|---|---|
| Formulation of Research Questions | Will the critical event occur (yes or no-the binary outcome)? What explanatory variables contribute to the occurrence of the critical event? | When (at what time) will the critical event occur (yes or no-the binary outcome)? What risk factors contribute the timeline of the occurrence of the critical event? |
| Expression of Model Equations | $P(X)=e^Z/(1+ e^Z)$, where P(X) is the probability of event occurrence; e is the base of the natural logarithm; $Z = ß_0 + ß_1X_1+ ß_2X_2+...+ ß_pX_p$; ßs are regression coefficients; and Xs are explanatory variables | $h(t,X) = h_0(t)\ e^Z$, where h(t,X) is hazard function of the event occurrence given time variable (t); $h_0(t)$ is baseline hazard; e is the base of the natural logarithm; $Z = ß_1X_1+ ß_2X_2+...+ ß_pX_p$; ßs are regression coefficients; and Xs are risk factors |
| Pattern of Non-Linear Curves | P(X) is a continuous measure (probability of a student experiencing the critical event such as pass/fail, achiever/nonachiver)<br><br>P(X) begins with zero and increases as a smooth S-shaped curve. P(X) is the probability value between zero and one depending on the explanatory variables and the regression coefficients<br><br>P(X) cannot be used to detect the timeline of the occurrence of the critical events. | h(t, X) is a continuous measure (hazard function for the timeline of a student experiencing the critical event such as dropout, dismissal, and withdrawal)<br><br>The hazard function h(t,X) begins with any positive value and goes up and down. It is a rate between zero and positive infinity depending on the product of baseline hazard and the function of risk factors.<br><br>h(t, X) can be used to detect the timeline of the occurrence of the critical events. |
| Assessment of Accurate Predictions | Explanatory variables are the observed values readily to be plugged into the equation to perform predictions<br><br>The prediction results allow investigators to identify potential at-risk students to participate in the mandatory intervention program.<br><br>The classification results are available to assess prediction accuracy. | Risk factors are the observed values readily available for the predictions. However, the time variable (t) in the baseline hazard function ho(t) is unknown value that cannot be plugged into the equation to perform predictions.<br><br>The classification results are not available to assess prediction accuracy. |
| Assessment of Model Fittings | Pseudo R-square is readily available to assess the model fittings. | Pseudo R-square is not available to assess the model fitting. |
| Verification of Model Assumptions | Residuals are normally distributed with a mean of zero and a constant variance. The histogram and scattergram of residuals are plotted to check for the normality and the homogeneity of variance. | Hazard ratio are constant across time. If the hazards are proportional, the survival curves generated by log minus log (LML) should be parallel. |

data, which means it indicates that the proportion of variations in the outcome variable is accounted for by the explanatory variables. In the logistic regression model, the pseudo R-squares are used as criteria to assess the model goodness of fit. However, in the Cox regression model, the pseudo R-squares are not readily available to assess the model fitting in SPSS PC Version 12.0 commands.

Finally, another key difference between the two models lies in the model assumptions. For the logistic regression model, residuals are assumed to have a mean of zero and a variance of P(X) [1-P(X)]. Investigators must check for violation of the assumption by plotting the histogram and scatter diagram for residuals. The model assumption is satisfied if (1) the histogram of the residuals is normally distributed with a mean of zero and (2) the residuals on the scatter diagram appear to be parallel with the X-axis, (i.e., the indication of a constant variance). The Cox regression model states that there is a multiplicative relationship between the baseline hazard function $h_0(t)$ and the function of the risk factors. As a result, the ratio of the hazard functions for an event with different values for the risk factor does not depend on time (t). Therefore, investigators need to check for violation of this assumption by using the log-minus log (LML) plot of the survival function. If the hazards are proportional, the survival curves generated by LML should be parallel (Steinberg, 1999). Under the assumption of proportional hazards, the resulting curves should be parallel and separated only by a constant vertical difference.

## Summary

In this study, the main objectives for constructing logistic and Cox regression models were accomplished. For logistic regression analysis, the explanatory variables contributing to the probability of a student passing the USMLE Step 1 were identified. It was evident that the MCAT (physical and biological sciences) scores, number of sophomore courses failed, and medical school freshman GPAs were significantly associated with the USMLE Step 1 performances. The study results confirmed that MCAT scores and medical school course performances were significant predictors of the USMLE Step 1 (Chen et al., 2001; and Haught and Walls, 2002). The implication of the study results was that the medical school should continue its effort to recruit and admit qualifying students with high MCAT scores, and to strengthen teaching and learning to ensure student success on the licensure examination.

With regard to Cox regression analysis, the method indicated that academic difficulty was significantly accounted for by risk factors such as MCAT verbal reasoning score, gender, and number of sophomore courses failed. Moreover, students in the five-year curriculum track experienced academic difficulty during the first semester of their sophomore year, peaked at the second semester of the sophomore year and maintained the same level of risk through the rest of the study period. The research results were consistent with the literature stating that an increase in the relative risk for a student experiencing academic difficulty was significantly associated with a low MCAT score (Huff and Fang, 1999), and students at risk for academic difficulty remained at risk throughout the first three years of medical school (Fang, 2000). The implication of this study was that the medical school addressed academic difficulty issues through academic development and support services.

## Strengths

Both logistic and Cox regression models are typically used for data analysis concerning binary outcomes such as admitted/not admitted, enrolled/not enrolled, and graduated/not graduated. In particular, the logistic regression method is capable of allowing investigators to answer some important questions linked to learning outcomes"Is student performance likely to improve or not improve after the implementation of tutorial and remedial programs?" "Are student graduation and attrition status significantly associated with the explanatory variables concerning student characteristics and college learning environment?" and "Can the probability of students expressing overall college satisfaction be estimated by certain explanatory variables concerning academic programs and services?"

In Cox regression analysis, the hazard function is a function of the time-to-event and risk factors. This function provides investigators with the valuable information to answer certain learning outcome questions such as "How many semesters elapse before students experience academic difficulty?" "What risk factors are significantly associated with the occurrence of academic difficulty?" "At what month are students likely to pass the licensure examination?" "What explanatory variables significantly contribute to the success of licensure examination?" "How many years pass before students graduate from the college?" and "To what extent student's timely and delayed graduation are accounted for by the explanatory variables concerning the overall quality of educational program and student support?"

Clearly, these two models are proven to be useful tools in studying explanatory variables that are significantly associated with binary outcomes. Moreover, the effect of a specific explanatory variable or risk factor on the event occurrence can be investigated holding other explanatory variables or risk factors constant. Both methods also allow investigators to assess the model fittings by means of the likelihood ratio test. Furthermore, using the logistic regression model, investigators can perform classifications, and subsequently evaluate the predictive power.

### Limitations

This study is not completely conclusive because of some methodological limitations. For example, given both models were built to study the effects of the explanatory variables on the binary outcomes—passing licensure examination (yes or no) or experiencing academic difficulty (yes or no)—the investigator is unable to use the same techniques to study multiple outcome categories. These outcomes categories include three pass- or fail-groups of licensure examination (e.g., first-time pass, second-time pass, and fail at least two times) and three categories of academic difficulty (e.g., dismissal, withdrawal, and leave of absence), respectively.

The Cox regression model is also known as the Cox proportional hazards model and has to satisfy the model assumption of proportional hazards. If the model assumption is seriously violated, the analysis results could be inaccurate and misleading. However, it is difficult to make judgments about the proportional hazards on the LML plot of survival curves partly because the functions are non-linear instead of straight lines. In other words, the validity of the model assumption may not be properly evaluated because of the limitations of the LML assessment tool.

### Major Alternatives

To study the outcome variable with multiple categories as a function of the explanatory variables, polytomous logistic regression analysis is a possible alternative (Peng, et al., 2002). Polytomous logistic regression analysis includes ordered and multinomial logistic regression models. The ordered logistic regression model is applicable to three or more ordinal outcome categories (e.g., first-time pass, second-time pass, and at least two-time fail groups for licensure examination). The model is called the cumulative logit model because it is based on the cumulative response probabilities of being in a category or lower (Walters, et al., 2001). The model is also called the proportional odds model because it assumes that the corresponding regression coefficients in the link function are equal for each explanatory variable. Therefore, the model assumption has to be verified carefully by the parallel lines test (SPSS, Inc., 2002). If the model assumption is satisfied, investigators can proceed to interpret the effects of the explanatory variables.

If the model assumption is violated in the ordered logistic regression analysis, the multinomial logistic regression model should be considered as another alternative. The outcome variable of multiple nominal groups includes the reference group and the target groups. For instance, in passing licensure examination, the first-time pass group is labeled as the reference group, the second-time pass group is coded as target group 1, and at least two-time fail group is considered as target group 2. Two model equations are generated for the nominal outcome with the three groups. In addition, the two sets of relative risk rates are calculated when the probability of a student falling into a specific target group is compared to that of a student in the reference group (Walters, Campbell, and Lall, 2001). The multinomial logistic regression analysis relaxes the model assumption of the proportional odds. It does not require that investigators verify the assumption of parallel lines because the relationship between the explanatory variables and the effects of these variables depends on the outcome category (Plank and Jordan, 1997).

An implicit feature of the Cox regression model is reflected in the model assumption of proportional hazards for all time intervals. If the LML plot of survival curves for assessing the validity of the model assumption is not successful, the smoothed plots of the scaled Schoenfeld residuals proposed by Therneau and Grambsch (Belle, et al, 2004) may be a better approach. The Schoenfeld residual method not only provides the investigator with an easier visual interpretation, it also offers a statistical test for the proportional hazards assumption (Belle, et al., 2004). An additional alternative for detecting the violation of the model assumption is the likelihood ratio test (Palmer, et al., 2003). Departure from the assumption of proportional hazards can be analyzed by the likelihood ratio test comparing models with and without the stratifying variable by the covariate interaction terms (e.g., the curriculum track by the three explanatory variables—MCAT verbal reasoning score, gender, and the number of sophomore courses failed, respectively, in the present study). If no statistical significance is found in these interaction terms, then the model assumption of proportional hazards is not violated.

In the Cox regression model, if the effects of the explanatory variables change during time in a practical situation (e.g., age and financial aid amount fluctuate over years), it may lead to a violation of the model assumption. For this reason, the extended Cox regression model that utilizes the time-dependent variables should be considered as an alternative to analyze data for that do not require the model assumption (Klein and Moeschberger, 1997; and Kleinbaum, 1996). The extended Cox regression model allows investigators to study the effect of the explanatory variables along with the time-dependent variables on the hazard function.

### Editor's Notes

Recently there has been an increased interest in using various methodologies that better fit the situations we encounter. While Multiple Linear Regression with its Ordinary Least Squares has been part of our methodology for many years, it has always had certain limitations. For example in predicting probabilities it has the unfortunate characteristic of going both negative and also going greater than one. Logistic Regression deals with this by creating

a dependent variable, the log odds, that can range from negative to positive and that can exceed 1.0. While Logistic Regression has been discussed in the statistical literature for numerous years, it has become more prevalent in institutional research during the last several years.

A second use of Linear Regression has been to explain the time between a process starting and completing. Here again the linearity of relationships tends to be suspect. In addition the regression model is not able to handle data where specific completion date is not available. Cox Regression is capable of dealing with both of these issues. As with logistic regression, it uses a function of the exponential and converts the relationship into one that is linear in logarithms.

While both the Logistic and Cox procedures are part of many standard statistical software, such as SPSS, correctly using them is frequently not intuitive. This is where the article by Dr. Chen provides an extremely valuable service. First it provides a good summary of when and why a researcher would want to use these types of models. In addition it provides an example of their use. In fact it gives you the step-by-step procedures needed to run your own analysis in SPSS. Next it gives a brief interpretation of the results.

What I think you will find most exciting however is Table 6 where there is a head to head comparison of the two methodologies. As an exercise to consider when linear regression should be used, the reader may well want to add an additional column to Table 6, title the column Linear Regression, and fill in the cells. For example under Pattern of Non-Linear Curves, one might want to list the use of various nonlinear transformations such as quadratic and cubic terms. The reader might also want to add rows to Table 6 such as Interpretation of Regression Weights. The interpretation of these weights for the Logistic and Cox procedures can be found in the text of the article.

This suggestion also applies to Table 5 where you may want to put in the similarities that Multiple Linear Regression has with Logistic and Cox Regression.

Also found in this article is the discussion of the relationships between Hazard and Survival functions. As such this paper represents an excellent first step or primer and explains techniques that greatly extend our analytical methodologies. Readers do need to be warned, however, that if they are interested in using these techniques they must focus on learning more about them. For example when using Chi Square and testing within stepwise or nested models, it is extremely important to select appropriate models that are theoretically relevant. The selection of the specific stratification of the model, where stratification is desirable, is also an extremely important element in the methodology. In terms of building the models, Dr. Chen selected Forward-Wald to build his survival model. There are other options such as simultaneously entering the variables or backwards elimination of variables The decision to use a specific strategy should be selected based on the situation and the intent of the researcher.

Another option the researcher has in Logistic Regression is to adjust the cut-point for classifying and observation into the two categories. Dr. Chen used the default of .5 but you may want to use a proportion that's closer to the actual split in the sample.

This article will give you an extremely good start in appropriately using two alternatives to the traditional regression methodology. As you begin to use one or both of the alternatives, the references he provides contain the information that will be essential for you to use the methodologies appropriately.

**References**

Belle, Gerald Van., Fisher, Lloyd D., Heagerty, Patrick J., and Lumley, Thomas. (2004). Biostatistics: A Methodology for the Health Sciences. Wiley-Interscience (2nd edition)

Braun, Henry I. and Zwick, Rebecca (1993). Empirical Bayes Analysis of Families of Survival Curves: Applications to the Analysis of Degree Attainment. Journal of Educational Statistics. Vol. 18., No. 4, pp. 285 - 303

Case, S.M., Swanson, D.B., Ripkey, D.R., Bowles, L.T., and Melnick, D.E. (1996). Performance of the Class of 1994 in the New Era of USMLE. Academic Medicine 72:31S-33S

Chen, Chau-Kuang, Campbell, Vickie C., and Suleiman, Ahmad, (2001). Predicting Student Performances at a Minority Professional School, Association for Institutional Research 2001 Annual Forum Paper, ERIC No. ED457714

Cox, D. R. (1972). Regression Models and Life Tables. Journal of the Royal Statistical Society, Series B, 34, 187-202

Cox, D. R., and Oakes, D. (1984). Analysis of Survival Data. London: Chapman & Hall.

DesJardins, Stephen L. and Moye, Melinda J., (2000). Studying the Timing of Student Departure from College, Air 2000 Annual Forum Paper, ERIC No. ED445650

DesJardins, S. L., Ahlburg, D. A., & McCall, B.P. (1997). Using Event History Methods to Model the Different Modes of Student Departure from College. Paper presented at the Association for Institutional Research 37th Annual Forum, Lake Buena Vista, Florida

Dey, Eric L. and Astin, Alexander W. (1993). Statistical Alternatives for Studying College Student Retention: A Comparison Analysis of Logit, Probit, and Linear Regression. Research in Higher Education Vol 34, No. 5

Elandt-Johnson, Regina C. and Johnson, Norman L. (1980). Survival Models and Data Analysis, John Wiley and Sons, New York.

Eliason, Scott R. (1993). Maximum Likelihood Estimation: Logit and Practice, a Sage University paper, 07-096, Newbury Park, CA: Sage

Fang D. (2000). 1998 AAMC CAMCAM: Fact Sheet, Vol. 2, No. 12, P.3

Gill, Jeff (2000). Generalized Linear Model: A Unified Approach. Sage Publication, Thousand Oaks, California.

Hachen. David S., Jr. (1988). The competing risks model: A method for analyzing processes with multiple types of events. Sociological Methods & Research, 17 21-54.

Han, Tianqi and Ganges, Tendaji W. (1995). A Discrete-Time Survival Analysis of the Education Path of Specially Admitted Students. ERIC No. ED387033

Haught, Patricia A. and Walls, Richard T. (2002). Adult Learners: Relationships of Reading, MCAT, and USMLE Step 1 Test Results for Medical Students. ERIC No. ED464943

Hojat, Mohammadreza, and others (1995). Primary Care and Non-Primary Care Physicians: A Longitudinal Study of Their Similarities Differences, and Correlates before, during, and after Medical School. Academic Medicine. Vol 70, n1 suppl pS17-S28

Hosemer, David W., and Lemeshow, Stanley, (1989). Applied Logistic Regression. New York: John Wiley & Sons, Inc.

Hosemer, David W., and Lemeshow, Stanley, (1999). Applied Survival Analysis. New York: John Wiley & Sons, Inc.

Huff, K.L. and Fang, D. (1999). When Are Students Most at Risk of Encountering Academic Difficulty? A Study of the 1992 Matriculants to U.S. Medical Schools. Academic Medicine 74:454-460

Klein, John P., and Moeschberger, Melvin L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag, New York

Kleinbaum, David G. (1996). Survival Analysis: A Self-Learning Text, Springer-Verlag, Inc., New York

Lee, Elisa T. (1992). Statistical Methods for Survival Data Analysis, John Wiley & Sons, Inc.

Matthews, D.E. and Farewell, V.T., (1996). Using and Understanding Medical Statistics. 3rd Revised Edition. Basel:S Karger, 245 pp.

McDaniel, Cleve and Graham, Steven (1999). Student Retention in an Historically Black Institution. Paper presented at the annual Forum of the Association for Institutional Research

Menard, Scott, (1995). Applied Logistic Regression Analysis. Thousand Oaks, California: Sage Publications.

Meyer, Paul L. (1970). Introductory Probability and Statistical Applications, Addison-Wesley Publishing Company, 2nd edition

Miller, R. G. (1981). Survival Analysis. New York: Wiley

Norusis, Marija J. (1985). Advanced Statistics Guide, SPSS, Inc.

Palmer, Julie; Wise, Lauren A.; Horton, Nicholas J.; Adams-Campbell, Lucile L.; and Rosenberg, Lynn (2003). Dual Effect of Parity on Breast Cancer Risk in African-AmericanWomen. Journal of the National Cancer Institute, Vol. 95, No. 6, March 19, 2003

Pampel, Fred C., (2000). Logistic Regression: A Primer. Thousand Oaks, California: Sage Publications.

Plank, Stephen B. and Jordan, Will J. (1997). Reducing Talent Loss. The Impact of Information, Guidance, and Actions on Postsecondary Enrollment, Report No. 9 Eric No: ED405429

Peng, Chao-Ying Joanne, So, Tak-Shing Harry, Stage, Frances K., and St. John, Edward P., (2002). Use and Interpretation of Logistic Regression, Research in Higher Education v43, (3).

Peterson, T., (1984). A Comment on Presenting Results of Logit and Probit Models. American Sociological Review, 50(1), 130-131.

Ronco, Sharon L. (1994). Meandering ways: Studying student dropout with survival analysis. Paper presented at the annual Forum of the Association for Institutional Research, New Orleans, LA.

Sadler, William E., Cohen, Frederic L, and Kockesen, Levent, (1997). Factors Affecting Retention Behavior: A Model To Predict At-Risk Students, AIR 1997 Annual Forum Paper, ERIC No. ED410885

Singer, J.D. & Willett, J.B. (1993). It's about time: Using discrete time survival analysis to study the duration and the timing of events. The Journal of Educational Statistics, 18, 115-l95.

Singer, J.D. & Willett, J.B. (1991). Modeling the Days of Our Lives: Using Survival Analysis When Designing and Analyzing Longitudinal Studies of Duration and the Timing of Events. Psychological Bulletin, 110(2), 268-290.

SPSS, Inc. (2002). SPSS Advanced Models 10.0., Chicago, IL.

Steinberg, Milton (1999). Cox Regression Example, SPSS Advanced Models 10.0 SPSS Inc. Chicago, IL

Strayhorn, Gregory (2000). Pre-Admission Program for Underrepresented Minority and Disadvantaged Students: Application, Acceptance, Graduation Rates, and Timeliness of Graduating from Medical School, Academic Medicine, V75 n4 p355-61

Therneau, T. M., and Grambsch, P. (2000). Modelling Survival Data: Extending the Cox Model. Springer-Verlag, New York

Walters, S.J., Campbell, M.J., and Lall, R (2001). Design and Analysis of Trials with Quality of Life as an Outcome: A Practical Guide. Journal of Biopharmaceutical Statistics 11(3), 155-176.

Willett, J.B. & Singer, J.D. (1995). It's Deja Vu all over again: Using multiple-spell discrete-time survival analysis. The Journal of Educational and Behavioral Statistics, 20, 41-67.

Willett, J.B. & Singer, J.D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. Review of Educational Research, 61, 407-450.

*IR Applications* is an AIR refereed publication that publishes articles focused on the application of advanced and specialized methodologies. The articles address applying qualitative and quantitative techniques to the processes used to support higher education management.

Editor:
Gerald W. McLaughlin
Director of Planning and Institutional Research
DePaul University
1 East Jackson, Suite 1501
Chicago, IL 60604-2216
Phone: 312/362-8403
Fax: 312/362-5918
gmclaugh@depaul.edu

Managing Editor:
Dr. Terrence R. Russell
Executive Director
Association for Institutional Research
222 Stone Building
Florida State University
Tallahassee, FL 32306-4462
Phone: 850/644-4470
Fax: 850/644-8824
air@mailer.fsu.edu

## AIR *IR Applications* Editorial Board



Authors can submit contributions from various sources such as a Forum presentation or an individual article. The articles should be 10-15 double-spaced pages, and include an abstract and references. Reviewers will rate the quality of an article as well as indicate the appropriateness for the alternatives. For articles accepted for *IR Applications*, the author and reviewers may be asked for comments and considerations on the application of the methodologies the articles discuss.

Articles accepted for *IR Applications* will be published on the AIR Web site and will be available for download by AIR members as a PDF document. Because of the characteristics of Web-publishing, articles will be published upon availability providing members timely access to the material.

Please send manuscripts and/or inquiries regarding *IR Applications* to Dr. Gerald McLaughlin.