# PROPORTIONAL HAZARDS MODEL FOR ASSESSING RISK FACTORS IN UNDERGRADUATE CALCULUS

V. J. DuRapau, Jr., Ph.D.
Xavier University of Louisiana
Department of Mathematics
New Orleans, LA 70125
vdurapau@xula.edu

**INTRODUCTION:** A major problem in analyzing factors contributing to students successfully completing a first semester Calculus course is the large number of students who withdraw from the course during the semester. Traditional statistical approaches (e.g., correlation and linear regression) may give biased results. Statistical techniques under the rubric of *survival analysis* offer ways of handling such censored data. The purpose of this paper is to suggest methods from survival analysis to address these challenges.

**METHODOLOGY:** This study involved observations of students over one-semester (fall 2003) at a small liberal arts university. One way to address the challenges associated with the relatively large number of withdrawals during the semester is to focus on the complementary event *unsuccessful in completing the course*. The methodology used in this study involves the notion of censored data.

Censoring occurs when the outcome or event of interest is not known for an individual during the period of observation. Consider four cases of students in *Calculus 1* during the fall 2003 semester.

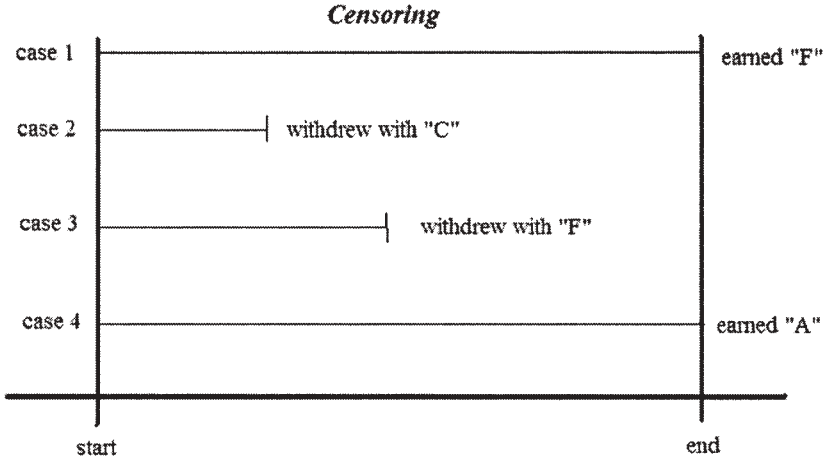


*Figure 1: Censoring (cases 2 & 4) & Event Occurrence (cases 1&3)*

Case 1 is a student who completed the semester with a failing grade *F*. Cases 2 and 3 represent students who withdrew during the semester, the first with a grade of *C* and the

second with a grade of *F*. Case 4 represents a student who completed the semester with a grade of *A*. The event of interest is failure to successfully complete the course (or simply "unsuccess"). In case 1, the event occurred during the period of observation (at the end of the semester) so no censoring occurred. Case 2 is an example of censoring. The student withdrew during the semester but had a *C* average to date in the course. This student is no longer at risk of failure. The student in case 3 withdrew during the semester and had an *F* average in the course at the time of withdrawal. The event "unsuccess" occurred for this student, so no censoring occurred. The student in case 4 completed the course with an *A*. The event "unsuccess" did not occur. This student is no longer at risk of failure. However, in the context of survival analysis terminology, this case is considered censored.

The event of interest for this study was

$$Evnt\_WDF = \begin{cases} 1, \text{ if "yes"} \\ 0, \text{ if "no"} \end{cases}$$

indicating whether or not the student (a) withdrew during the semester with a *D* or *F* average at the time of withdrawal or (b) earned a final grade of *D* or *F*. Many students, because of their major, need at least a *C* in *Calculus 1*.

Determining an exact time in days that a student withdraws from a class is difficult. Some students simply stop attending class but do not officially withdraw until near the end of the semester. A solution to this time problem is to use a proxy for time to event. There were five major tests (Modules 1 to 5) and a final exam. I used these six events to define the variable *Time to Event* (*Mods*), the proxy for time:

$$T\_Mods = \begin{cases} k, \text{ if event occurred } (Evnt\_WDF = 1) \text{ after Mod } k \text{ test, } k = 1,2,...,5 \\ 6, \text{ if event occurred } (Evnt\_WDF = 1) \text{ after final exam} \end{cases}$$

For example, if the last recorded test grade for a student was the Mod 4 test and the student had an *F* average to that point, then the time variable *T_Mods* = 4 for that student. Although *T_Mods* is discrete, I treated it as a continuous random variable.

The proportional hazards model used in this study is

$$h_i(t) = [h_0(t)]e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}}, i = 1,2,...,n$$

where

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

and the $p = 4$ covariates were

$x_{i1} = DVMath_i$ Took Developmental Math (0 = no; 1 = yes)
$x_{i2} = Atmp1070_i$ Number of previous attempts of Calculus 1 (0, 1, ...)
$x_{i3} = preTstN_i$ Pre-Test score (max = 25)
$x_{i4} = PCbypass$ By-passed Pre-Calculus (0 = no; 1 = yes)

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

I used *Statistical Package for the Social Sciences (SPSS) 12.0 for Windows* to perform a stepwise Cox regression with *T_Mods* as the time to event, *Evnt_WDF* as the event indicator, and with the entry criteria for the first three covariates set to forward likelihood ratio test. Covariates that are not statistically significant will not appear in the final model. Since I wanted the fourth covariate (*PCbypass*) to appear in the final model, I entered it in a separate block with entry method "Enter" specified. Both *DVMath* (took *Developmental Math*) and *PCbypass* (by-passed Pre-Calculus?) are categorical variables (1 for "yes" and 0 for "no"). Separate survival and hazard functions were generated for each of the two values of *PCbypass* at the mean of any other covariates remaining in the model at the end of the stepwise procedure. Also, the baseline hazard function was estimated.

**RESULTS:** Descriptive statistics were computed but are not shown here. Table 1 shows the parameter estimates for the variables in the final proportional hazards model.

**Variables in the Equation**

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| DVMath | -.258 | .314 | .677 | 1 | .411 | .772 | .417 | 1.429 |
| PreTstN | -.109 | .031 | 12.193 | 1 | .000 | .897 | .844 | .954 |
| PCbypass | .471 | .227 | 4.320 | 1 | .038 | 1.602 | 1.027 | 2.499 |

*Table 1: Variables in the Final Model*

The final model predicted hazard for the Cox regression model is

$$h_i(t) = [h_0(t)]e^{\beta_0 + (-.258)DVMath_i + (-.109)preTstN_i + (.471)PCbypass_i}, i = 1,2,...,n$$

$$= [h_0(t)]e^{\beta_0} \cdot (0.772)^{DVMath_i} \cdot (0.897)^{preTstN_i} \cdot (1.602)^{PCbypass_i}, i = 1,2,...,n$$

The "unsuccess" hazard for a student who did not take *Developmental Math* is 0.772 times that of a student who did take *Developmental Math*. (In the categorical variable recoding during analysis, *DVMath* = 1 for a student who did not take *Developmental Math*.) The "unsuccess" hazard is multiplied by 0.897 for each unit increase in *PreTstN*. In other words, as expected, the higher the Pretest score, the lower the hazard and the higher the survival rate, that is, the more likely the student successfully completes *Calculus 1*. The "unsuccess" hazard for a student who did not bypass *Pre-Calculus* is 1.602 times that of a student who did bypass the *Pre-Calculus*.

Hazard and survival curves were generated. Only the latter are shown here. Figure 2 shows separate model-predicted survival curves for values of *PCbypass*. Students who by-passed *Pre-Calculus* and went directly into *Calculus 1* (*PCbypass* = 1) have a survival curve that is higher than those who did not by-pass *Pre-Calculus* (*PCbypass* = 0).

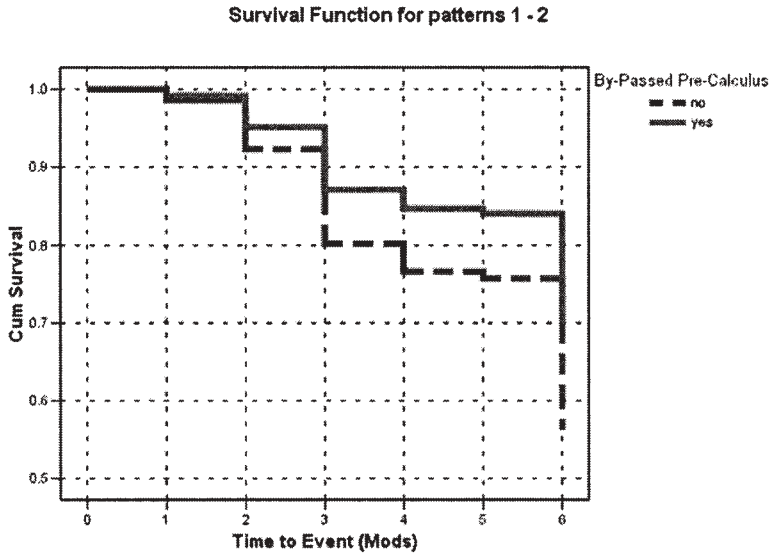


*Figure 2: Cumulative Survival Functions*

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

Cox regression was used to estimate parameters (regression coefficients) of the model. A forward stepwise procedure was used to include or exclude covariates in block 1 (*DVMath*, *Atmp1070*, *PreTstN*). Based on criteria set in the stepwise procedure, only *DVMath* and *PreTstN* were included in the model at the end of block 1. Since I wanted the covariate *by-passed Pre-Calculus* (*PCbypass*) to appear in the final model, I included this variable in block 2 of the model building process. The coefficient for *DVMath* in the final model was not statistically significant after the variable *PCbypass* was included.

Although the time variable *T_Mods* as defined in this study is discrete, I treated it as a continuous random variable. For one of the 10 sections of the course, I did have a measure in days of time to event (*T_Days*). Using the same proportional hazards model

described in this paper with *T_Days* as the time variable, only *PreTstN* was significant in the final model. Of the 22 students in that one section, the event *Evnt_WDF* occurred for 7 while the other 15 students were censored. The value of $e^{\beta}$ for *PreTstN* in this model was 0.791, somewhat lower than the 0.897 value for the model with *T_Mods* for time. The status variable *Evnt_WDF* was used to indicate whether or not the event of interest (failure to successfully complete the course) occurred.

The 303 subjects included in this study were a convenience sample, namely, the students enrolled in Calculus 1 for fall 2003 at the private liberal arts university. Inferences to other populations may not be appropriate. However, the methodology proposed in this paper for investigating risk factors can be applied to other courses in which there is a relatively large number of students withdrawing during the semester. If a time to event variable is not defined, or one wishes to analyze the data independent of time, binary logistic regression with *Evnt_WDF* as dependent variable may be used.

## References


1. T. A. Cesa (April 13, 1993), "Using logistic regression to model whether admits will register." UC Berkeley.
2. D. R. Cox (1972), "Regression Models and Life Tables" *Journal of the Royal Statistical Society B*, 34, 187-220.
3. S. L. DesJardins (May 1993), "Using hazard models to study student careers." Presented at the *33rd Annual Forum of the Association for Institutional Research.*
4. V. J. DuRapau, Jr. (February 25, 2004) "Results of an Ongoing Assessment of Calculus 1." Presented to Mathematics faculty, Xavier University of Louisiana.
5. Hsi-Wen Liao (1998), "A simulation study of estimators in stratified proportional hazards models." http://www.ats.ucla.edu/stat/sas/library/nesug98/p118.pdf .
6. B. W. Lindgren (1976), *Statistical Inference* (3rd ed). New York: Macmillan Publishing Co.
7. D. E. Matthews, V. T. Farewell (1988), *Using and Understanding Medical Statistics*, 2nd revised ed. S. Karger.
8. V. K. Rohatgi (1984), *Statistical Inference.* New York: John Wiley & Sons.
9. P. B. Seetharaman (Washington University, St. Louis, MO) & Pradeep K. Chintagunta (University of Chicago) (March 12, 2002), "The Proportional Hazard Model for Purchase Timing: A Comparison of Alternative Specifications."
10. *Statistical Package for the Social Sciences (SPSS) for Windows 12.0* (2004). Chicago: http://www.spss.com .
11. D. D. Wackerly, W. Mendenhall III, R. L. Scheaffer (2002), *Mathematical Statistics with Applications*, 6th ed. Pacific Grove, CA: Duxbury.