International Scholars Journals

International Journal of Education Research and Reviews ISSN 2329-9843 Vol. 7 (8), pp. 001-004, August, 2019. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Interaction and survival analysis of graduation data

Justin R. Chimka[1*] and Lauren Holloway Lowe[2]

Justin Chimka, 4207 Bell Engineering Center, Fayetteville, AR 72701, USA, (479) 575-7392, USA.

Accepted 16 May, 2019

This is the second in a series of articles describing ongoing research that involves studying engineering college student graduation using Cox proportional hazards models. The first article, called "Proportional hazards models of graduation," was based on main effects models of graduation controlling for descriptors such as in -state residence, hometown population, and student major. This second article attends to first-order interaction terms between pairs of previously considered main effects. Survival analysis of graduation data here suggests significance of standardized math scores, and English and Science ACT scores, under certain circumstances that might not have been discovered without the examination of interaction.

Key words: Graduation, proportional hazards, retention, survival analysis.

INTRODUCTION

This is the second in a series of articles describing ongoing research that involves studying engineering college student graduation using Cox proportional hazards models. The first article reported significance of standardized math test scores, gender and science ACT scores in explaining variation in student graduation based on main effects models of graduation controlling for descriptors such as in-state residence, hometown population, and student major (Chimka et al., 2007 - 2008). This second article, "Interaction and survival analysis of graduation data," attends to first-order (or two-factor) interaction terms between pairs of previously considered main effects. Interaction between factors occurs when "the difference in response between the levels of one factor is not the same at all levels of the other factors" (Montgomery et al., 2001).

Recently there have been other journal articles added to the statistical models literature of college student retention. Berkovitz and O'Quin (2006-2007) estimated a logistic regression model of graduation for readmitted students controlling for GPA upon readmission. The study found younger students and those having partici-pated in pre freshman orientation more likely to graduate. Academic dismissal prior to readmission decreased probability of eventual graduation. Davidson and Beck (2006- 2007) used logistic regression to model probability of re enrollment as a function of responses to the Survey of Academic Orientations (Davidson et al., 1999).

Gansemer-Topf and Schuh (2006) have examined the relationship between institutional expenditures and selectivity, and graduation and retention rates. Results of multiple linear regression analysis suggested new resource allocation strategies to enhance graduation and retention. In addition, multiple regressions have been used to analyze graduation rates at public community colleges as functions of most notably ratio of part time faculty (Jacoby, 2006). Relatively controversial logistic regres-sion models of Scott et al. (2006) "suggest that with equivalent resources and student populations, public schools would graduate a slightly larger percentage of students than privates."

Perhaps the only other recent research into the impact of time varying risk factors on student retention has been by Randolph et al. (2006). They use non proportional hazards to estimate the influence of demographic factors, early school experiences, school involvement, family income and maternal employment status on rate of high school dropout.

*Corresponding author. E-mail: jchimka@uark.edu.

Table 1. Explanatory variables.

Variable	Value
fem	1 if the student is female, 0 if the student is male
ok	1 if the student attended high school in Oklahoma, 0 otherwise
oohu	Percent owner-occupied housing units in the 3-digit ZIP Code tabulation area of the student's high school according to Census 2000
tp	Total population in the 3-digit ZIP Code tabulation area of the student's high school according to Census 2000
engact	English ACT score
mathact	Math ACT score
readact	Reading ACT score
sciact	Science ACT score
mathsat	SAT Math score
verbsat	SAT Verbal score
math	Maximum (*mathact*, SAT math score translated to ACT)
test	Maximum (composite ACT score, SAT total score translated to ACT)

Table 2. Descriptive statistics.

	SAT Scores Only	Both SAT and ACT Scores	ACT Scores Only	Entire Sample
Students	62	186	181	429
Grad. Rate	0.581	0.559	0.425	0.506
Proportion *fem*	0.226	0.188	0.249	0.219
Proportion *ok*	0.097	0.495	0.845	0.585

MATERIALS AND METHODS

Observation of student graduation has been explained as survival data where Cox (1972) proportional hazards models of time from enrollment to graduation are of interest. For a thorough development of our methods the reader should consult Chimka et al. (2007-2008) along with Walpole et al. (2002) for a good communication on the philosophy of maximum likelihood, and a mathematical statistics book for reference.

The cohort followed for six and one half years is composed entirely of 429 first-time full -time students declaring an engineering major after being admitted in fall 1995 to the University of Oklahoma directly from high school. Table 1 provides the variables examined here, along with their possible values, and Table 2 presents descriptive statistics including graduation rates. It should be noted that while students of the cohort have declared an engineering major at some times, their majors are not necessarily engineering at all times. Thus there is the time dependent covariate *engmaj*.

We begin by attempting to analyze the three specific datasets originally presented in Chimka et al. (2007 - 2008) and consider additionally interactions between the main effects. The datasets are 1) SAT scores only, 2) ACT scores and male students only, 3) female students only. Results appearing in Tables 3 through 5 require the following interpretation. The hazard ratio or relative risk appears in column 2. Hazard ratios greater than one indicate increases in likelihood of graduation. Standard errors appearing in column 3 show variability of hazard ratio point estimates. Column 4 shows probabilities associated with appropriate values from the standard normal distribution.

We also attend to tests of the proportional hazards assumption; different individual explanatory variables' hazard curves must be proportional to an arbitrary, unspecified baseline hazard function. The assumption was rejected using Schoenfeld (1982) residuals with "$\alpha = 0.05$" (Grambsch and Therneau, 1994). There exist tests of model significance and proportional hazards, sometimes referred to as global tests. Relevant statistics have the χ^2 distribution. All tests were performed using Stata Statistical Software (StataCorp, 2005).

RESULTS

For students submitting only SAT scores upon application, there are not enough observations to support the interaction model. There is nothing to amend these results of the first article (Chimka et al., 2007-2008): Students with better SAT Math scores, and female students, were more likely to graduate among students submitting only SAT scores upon application.

For the interaction model using male students with ACT scores only, the proportional hazards assumption is violated due to interaction between *engmaj* and *ok*, so four separate models were next estimated. a) *engmaj* = 0, *ok* = 0 b) *engmaj* = 0, *ok* = 1 c) *engmaj* = 1, *ok* = 0 d) *engmaj* = 1, *ok* = 1.

For male students with ACT scores only attending high school outside of Oklahoma (separate models a. and b. above), there are not enough observations to estimate

Table 3. ACT scores and male students only (engmaj = 0, ok = 1; n = 109 obs. for 62 students).

	Hazard ratio	Standard error	*P* > |*z*|
oohu	0.70	0.13	0.049
tp	1.00	1.41E-5	0.070
engact	0.04	0.04	0.001
readact	0.57	0.41	0.443
sciact	2.66	2.22	0.239
oohu × *tp*	1.00	1.97E-7	0.326
oohu × *engact*	1.02	0.01	0.050
oohu × *readact*	1.02	0.01	0.086
oohu × *sciact*	0.98	0.01	0.083
tp × *engact*	1.00	3.59E-7	0.006
tp × *readact*	1.00	5.02E-7	0.001
tp × *sciact*	1.00	4.37E-7	0.900
engact × *readact*	0.98	0.01	0.227
engact × *sciact*	1.08	0.03	0.004
readact× *sciact*	0.98	0.01	0.056

Table 4. ACT scores and male students only (engmaj = 1, ok = 1; n = 133 obs. for 105 students).

	Hazard ratio	Standard error	*P* > |*z*|
oohu	0.49	0.38	0.362
tp	1.00	6.44E-5	0.318
engact	0.01	0.04	0.236
readact	59.47	217.57	0.264
sciact	48.17	80.99	0.021
oohu × *tp*	1.00	1.04E-6	0.145
oohu × *engact*	1.03	0.02	0.261
oohu × *readact*	0.98	0.02	0.503
oohu × *sciact*	0.98	0.02	0.462
tp × *engact*	1.00	3.30E-6	0.228
tp × *readact*	1.00	3.10E-6	0.303
tp × *sciact*	1.00	6.10E-7	0.002
engact × *readact*	1.01	0.02	0.528
engact × *sciact*	1.00	0.03	0.908
readact × *sciact*	0.94	0.03	0.095

the models. For models where *ok* = 1, see "Tables 3 "student's major is not engineering) and 4 (student's major is engineering), and keep in mind results share common students, as *engmaj* is time-varying.

Among students from Oklahoma high schools, while non-engineering majors, those with better English ACT scores were less likely to graduate (*P* > |*z*| = 0.001). This result would seem strange, and it is not true of the asso-

Table 5. Female students only (n = 214 observations for 94 students).

	Hazard ratio	Standard error	*P* > |*z*|
engmaj	0.01	0.03	0.224
ok	0.04	0.14	0.405
oohu	0.93	0.09	0.449
tp	1.00	2.49E-6	0.004
math	0.37	0.26	0.160
test	2.15	1.06	0.121
engmaj × *ok*	1.69	1.14	0.433
engmaj × *oohu*	1.07	0.04	0.092
engmaj× *tp*	1.00	6.73E-7	0.135
engmaj× *math*	0.74	0.05	0.000
engmaj× *test*	1.31	0.12	0.002
ok × *oohu*	1.00	0.05	0.932
ok × *tp*	1.00	1.23E-6	0.773
ok × *math*	0.81	0.09	0.052
ok × *test*	1.34	0.17	0.017
oohu × *tp*	1.00	5.45E-8	0.041
oohu × *math*	1.02	0.01	0.020
oohu × *test*	0.98	0.01	0.025
tp × *math*	1.00	1.56E-7	0.105
tp × *test*	1.00	1.37E-7	0.105
math × *test*	1.00	0.01	0.645

ciated group of students while engineering majors (*P* > |*z*| = 0.236). There may be ground for belief here that students with greater English ACT scores should for whatever reason try even harder than other students to persist as engineering majors once they have declared themselves as such. This model as a whole is statistically significant (global *P-value* > $\chi^2 = 0.0000$), and there is no evidence that the proportional hazards assumption for the entire model has been violated (global *P-value* > $\chi^2 = 0.9910$).

Among the associated students from Oklahoma high schools, while engineering majors, there is the same significant result as found in the main effects model for ACT scores and male students only: Students with greater Science ACT scores are more likely to graduate (Chimka et al., 2007-2008). The new model as a whole is statistically significant (global *P-value* > $\chi^2 = 0.0000$), and there is no evidence that the proportional hazards assumption has been violated (global *P-value* > $\chi^2 = 0.9947$).

The same result is not evident for associated non-engineering students, but there is significant interaction among them between English ACT and Science ACT (*P* > |*z*| = 0.004). This means that, while Science ACT is not generally important to describing variation in graduation of non-engineering students, it very well may be for certain ranges of English ACT. Let us continue to examine this interaction with respect to medians based on all relevant observations of English ACT (25) and Science ACT (24). For relevant students with English ACT scores less than the median, just 1/6 students (about 17%) with Science ACT scores less than 24 graduated, while 9/22 students (about 41%) with Science ACT scores greater than or equal to 24 did graduate.

Finally we consider interactions in addition to main effects for female students only. This model as a whole is statistically significant (global *P-value* > $\chi^2 = 0.0000$), and there is no evidence that the proportional hazards assumption for the entire model has been violated (global *P-value* > $\chi^2 = 0.9774$). Controlling for other descriptors and all available first-order interaction, we find significant interaction between *engmaj*, and both *math* and *test*.

Note that *math* is the maximum of student's math ACT score and SAT math score translated to ACT; *test* is the maximum of student's composite ACT score and SAT total score translated to ACT. Let us examine in more detail the interaction between *engmaj* and *math*, since it is more specific than *test*. Since *engmaj* is dichotomous we next stratify appropriately and fit separate main effects models for females while they are (not) engi-neering majors. We find that, while engineering majors, female students with greater *math* scores are more likely to graduate; while non-engineering majors, math scores among females may not really matter. Remember these are results of main effects models controlling for *ok*, *oohu*, *tp* and *test*.

DISCUSSION

Though variables were dropped due to predictors lacking independence in our first study, effects of interaction have now been seen to illuminate results otherwise less obvious. In summary, results reported here are the following.

Male students from Oklahoma high schools with better English ACT scores were less likely to graduate while non-engineering majors, suggesting it may be more important for students (having declared engineering as a major) with greater English ACT scores to persist as engineering majors. Otherwise the greater English ACT score seems to become a liability in relation to graduation. In this same group, but for students while they are engineering majors, students with greater Science ACT scores are more likely to graduate. This result holds true for students while non-engineering majors, if those students have lesser English ACT scores. Finally we have observed among female students that greater math scores increase the probability of graduation while major is engineering. Otherwise there is no such significant result for female students.

If interest in proportional hazards models of graduation continues to persist, we might find results of the following future work. The notion of time dependent effects seems interesting and worthy of more attention. While our survival analysis has been to determine student attributes affecting college graduation, future work should perhaps turn to predicting probability of response or mean lifetime. Finally, more should be done to illustrate proportional hazards models of graduation with log-log plots of survival, and Kaplan-Meier and predicted survival plots.

REFERENCES

Berkovitz RA, O'Quin K (2006-2007). Predictors of graduation of readmitted "at risk" college students. J. Coll. Stud. Retent. 8(2):199-214.

Chimka JR, Rhoads TR, Barker K (2007-2008). Proportional hazards models of graduation. J. Coll. Stud. Retent. 9(2): 221-232

Cox DR (1972). Regression models and life-tables. J.R. Stat. Soc. Series B 34(2): 187-220.

Davidson WB, Beck HP (2006-2007). Survey of academic orientations scores and persistence in college freshmen. J. Coll. Stud. Retent. 8(3): 297-305.

Davidson WB, Beck HP, Silver NC (1999). Development and validation of scores on a measure of six academic orientations in college students. Educ. Psychol. Measure. 59(4): 678-693.

Gansemer-Topf AM, Schuh JH (2006). Institutional selectivity and institutional expenditures: examining organizational factors that contribute to retention and graduation. Res. High. Educ. 47(6): 613-642.

Grambsch PM, Therneau TM (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81(3): 515-526.

Jacoby D (2006). Effects of part-time faculty employment on community college graduation rates. J. High. Educ. 77(6):1081-1103.

Montgomery DC, Runger GC, Hubele NF (2001). Engineering statistics. New York: John Wiley and Sons, Inc.

Randolph KA, Fraser MW, Orthner DK (2006). A strategy for assessing the impact of time-varying family risk factors on high school dropout. J. Fam. Issues 27(7):933-950.

Scott M, Bailey T, Kienzl G (2006). Relative success? Determinants of college graduation rates in public and private colleges in the US. Res. High. Educ. 47(3): 249-279.

Schoenfeld D (1982). Partial residuals of the proportional hazards regression model. Biometrika, 69(1): 239-241.

StataCorp (2005). Stata Statistical Software: Release 8.0, College Station, TX: Stata Corporation.

Walpole RE, Myers RH, Meyers SL, Ye K (2002). Probability and statistics for engineers and scientists (7th ed.). Upper Saddle River, NJ: Prentice Hall.