Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.10, 2014

The Retention Rates of Students In Public Secondary Schools Using The Cox Proportional Hazard Model: A Case of Kisumu County, Kenya

Jacob Oketch Okungu, Dr. George Orwa, Dr. Joseph Mung'atu.
Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology,
P.O Box 62 000 00200, Nairobi, Kenya.
*E-mail: oketcho2000@gmail.com

Abstract

The study sought to propose a statistical model for the public Secondary School students' retention rates for Kisumu County. We used survival regression analysis in which students were grouped according to the their performance in KCSE, mean school fee payment, school category, sex, mean age and teacher - student ratio using a desirable survival function. The model was of interest because the study sought to address a life testing experience mostly restricted to survival models and most of the existing models on student retention addressed Universities and tertiary colleges' trend and were particularly developed outside the country. This was in spite of the fact that Kenya suffers from high dropout rates at the Secondary level. The annual Secondary School students data traced from a cohort in form one in the year 2010 to form four in the year 2013 were obtained from the Kisumu County Ministry of Education headquarters and analyzed using survival regression. The variables which were insignificant were dropped to get the desirable function. Survival rates of students in the Secondary Schools were traced at the end of each and every level. It was found out that dropping out of students was influenced by; the category of the school, average fee payment, performance in KCSE and sex, with more girls dropping out than boys and the dropping out mostly rampant at form two. The model will be of relevance to the concerned Secondary education stakeholders in improving the quality of education for it will inform the planning of necessary educational interventions to ensure enhanced retention rate and transition.
Keywords: Retention, Cox Regression, Drop-out rate, Probability Density function, Kaplan Meier.

1.0 Introduction

In the year 2003, the Kenyan government opened doors to all public schools with an aim to maximize access to basic primary education by all Kenyans. The student population implicitly increased in all the primary schools and consequently in all the public secondary schools. Of relevant concern is whether all those who were admitted at different levels successfully completed their primary education and were admitted into the secondary schools (Onyando and Omondi, 2008). It is this that motivated the project, whose purpose was to come up with a retention rate model for public secondary school students in Kisumu County. The secondary school segment in the education cycle of a Kenyan is important for three major reasons: It de-links one from elementary (primary) learning, it provides a chance for one to complete the cycle for basic education and anchors as the springboard to either tertiary or higher learning (Onyando and Omondi, 2008). However, pandemic secondary school dropout in Kenya is alarming.

As a nation, Kenya hopes to achieve Education for All (EFA) by the year 2015. This is an uphill task given the various challenges in the education sector. The year 2015 is also significant globally because it is the target year for the fulfillment of the eight-millennium goals. Kenya looks forward to have her people achieve the millennium goals together with other people worldwide (KIPPRA, 2010).The pivotal hinge for these important target goals is education levels of the people involved and look forward to benefit from the fruits of EFA, millennium goals and industrialization. For such matters therefore, Kenya is trying her best to have her people educated. This project will therefore inform the government on the nature of students' success and retention at different levels and therefore enable the government to make informed choices of various educational interventions to put in place either to improve or correct the drop-out rate.

1.1 Objectives of the study

The study was guided by the following objectives;

1. to propose a statistical model for retention rates of students in academic institutions,
2. to explore the properties of the proposed model,
3. to apply the proposed model to the case of Kisumu County

1.1 Assumptions of the study

The following assumptions were taken into consideration for the model to be appropriate;

- The study population (Kisumu County Students' enrolment) is assumed closed i.e. there is no immigration and out –migration of students with the neighboring sub-counties.
- Admissions take place only in form one.
- There is no class repetition and
- Drop–outs are assumed to be uniformly distributed.

2.0 Literature Review

Survival analysis techniques have been applied to longitudinal data in order to identify factors predictive of students ultimately experiencing a general event of interest. Chimka, Justin, R., Teri, R. and Kash (2007) used proportional hazards models to identify variables that showed significant differences in engineering students persisting to college graduation. Zwick and Jeffery (2005) constructed discrete-time survival models to estimate the conditional probability of a student graduating with bachelors degree based on students' science and mathematics scores.

Wickens, John and Singer (1991) stated that educational researchers should employ survival analysis techniques in order to study topics such as student persistence and teacher attrition, because one of the best reasons to apply survival analysis is that standard statistical techniques require knowledge of when the event occurred (the outcome) for each sample member. The prior education research indicates that the use of survival analysis techniques can be quite powerful in modeling educational event occurrences. The ability to test time - varying predictors as well as time invariant predictors is particularly valuable benefit of applying survival analysis techniques. Most of the survival analyses have been carried out at the universities and colleges with a lot of gap in the secondary schools, this is what motivated the study.

3.0 Survival Models Used In The Study

3.1 Survival Function

The survival function is used to represent the probability that a student survives from the start of secondary education to sometime beyond t.

$$s(x) = \Pr(X > x) \qquad (1)$$

Also the integral of the probability density function $f(x)$:

$$s(x) = \Pr(X > x) = \int_0^x f(t)dt \qquad (2)$$

Thus given a survival function, we can calculate the probability density function

$$f(x) = -\frac{ds(x)}{dx} \qquad (3)$$

3.2 The Hazard Function

Hazard function is widely used to express the risk or hazard of drop out (failure) at some time *x*, and is obtained from the probability that a student drops out at time *t*, conditional on he/she having survived to time *x*. This is sometimes called instantaneous failure rate.

It is defined as

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \frac{\Delta x}{X} \ge x]}{\Delta x} \qquad (4)$$

$$= \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \frac{\Delta x}{X}]}{\Delta x Pr(X \ge x)} \qquad (5)$$

$$= \lim_{\Delta x \to 0} \frac{\mathrm{F(x+x\Delta)-F(x)}}{\Delta x s(x)} \qquad (6)$$

$$= \frac{1}{s(x)}\left\{\lim_{\Delta x \to 0} \frac{F(x+x\Delta)-F(x)}{\Delta x}\right\} \quad (7)$$

Since *x* is a continuous random variable,

$$h(x) = \frac{f(x)}{s(x)} = -\frac{d}{dx}\log[s(x)] \quad (8)$$

The cumulative hazard is

$$H(x) = \int_0^x h(u)du = -\log[s(x)] \quad (9)$$

Thus for a continuous lifetime of the students

$$s(x) = e^{-H(x)} = e^{-\int_0^x h(u)du} \quad (10)$$

3.3 The exponential Distribution

The probability density function (pdf) of an exponential distribution is

$$f(x;\beta) = \begin{cases} \beta e^{-\beta x}, & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (11)$$

Here $\beta > 0$ is the parameter of the distribution, often called the *rate parameter.*

We also note that if the drop-out rate is constant, that is if the hazard rate is constant in a given student population, then it follows an exponential distribution. The Cox model therefore follows the exponential distribution form.

3.4 The Cox Proportional Hazard Model.

Is a statistical technique for exploring the relationship between the survival of a subject and several explanatory variables. It is based on a modeling approach to the analysis of survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival (Walters, 2009).

4.0 Methodology

4.1 Model Building

A statistical model is basically an assumption relating effects of different levels of factors involved in an experiment alongside one or more terms representing the error effects. The study proposes to model the retention rates of secondary School students in Kisumu County. The model follows the Cox Proportional Hazard model with adjustments. Survival analysis examines and models the time it takes for events to occur. The prototypical such event is death, which in this project was 'drop out' from which the name ‘survival analysis’ and much of its terminology derived, though the ambit of application of survival analysis is much broader. Essentially the same methods are employed in a variety of disciplines under various rubrics like ‘event-history analysis’ in sociology. In this project, terms such as survival are to be understood generically. Survival analysis focuses on the distribution of survival times. Although there are well known methods for estimating unconditional survival distributions, most interesting survival modeling examines the relationship between survival and one or more predictors, usually termed covariates (Terry and Patricia, 2000).

As opposed to the common use of linear regression, this study adopted survival regression analysis to come up

with survival model which was derived from a desirable survival function (Wickens, 2004). The model was based on the simple Cox Proportional Hazard model which depends on the hazard rate. The number of students dropping out (deaths) of the academic institutions at each stage from the year 2010 to 2013 was used to derive the hazard function, *h(t,x)*. Since the survival function uses the hazard function as the primary theoretical concept, the suggested model followed the Cox Proportional Hazard Model;

$$h(x,t) = h_0(t)e^{(\beta_1 x_1 + \cdots + \beta_6 x_6)} \quad (12)$$

$$= h_0(t)e^{\beta^\tau X} \quad (13)$$

where X is a set of measurements

$$X = [X_1, X_2, X_3, X_4, X_5, X_6] \quad (14)$$

and $h_{\{0\}}(t)$ is the baseline hazard function that depends only on time but not the covariates.

Therefore

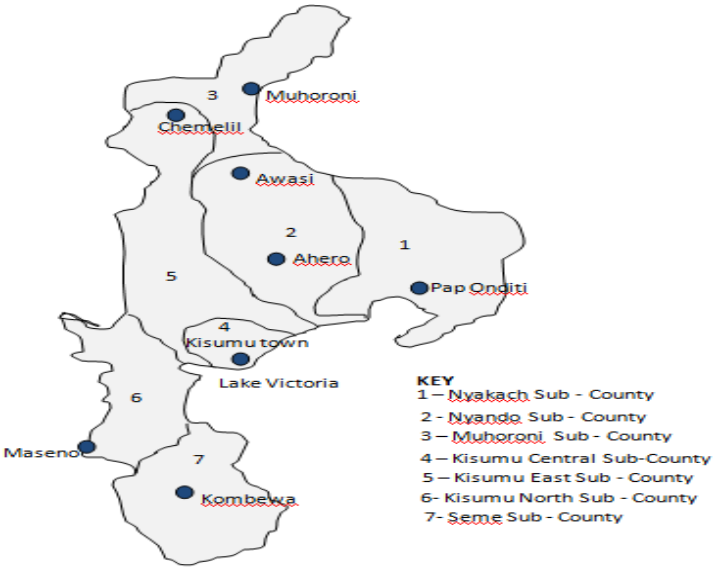
$$h(x,t) = h_o(t,\alpha)exp(\beta^\tau X) \quad (15)$$

where α are some parameters influencing the baseline hazard function.

It is worth noting that the hazard was decomposed into a product of two terms; $h_o(t,\alpha)$, a term that depends on time but not the covariates and $exp(\beta^\tau X)$, a term that depend on the covariates but not time (Cox, 1997). The covariates; X_1 was performance in KCSE, X_2 mean fee payment, X_3 mean age, X_4 the School category, X_5 sex of the students in the Schools and X_6 the teacher - student ratio. The exponential model is preferred because it has a constant hazard function.

4.2 Research Study Area

The study site considered in this research was the Kisumu County in the Lake region the former Nyanza province. The area was chosen because it had the highest drop-out rate and lowest retention rate in primary Schools. Kisumu County boarders Lake Victoria to the North, Nandi county to the south, Siaya County to the East, Homa–Bay County to the South – West and Kericho County to the west.



Geography of Kisumu – County, Source Google maps

5.0 Results And Data Analysis

Empirical results

Table 1: The number of drop outs by Category of Schools

Category	Dropped		Total
	No	Yes	
County	7082	1487	8569
Extra – County	2309	164	2473
National	494	13	507
Total	9885	1664	11549

Table 2: The number of drop outs by Sex/Gender

Sex/Gender	Dropped		Total
	No	Yes	
Female	4022	869	4891
Male	5863	795	6658
Total	9885	1664	11549

Table 3: The analysis of the key variables

Variable		Mean	Standard Error	[95% confidence interval]	
KCSE	No	5.864682	0.0070512	5.850861	5.878502
	Yes	5.479471	0.0289097	5.422807	5.536134
Student teacher ratio	No	39.19727	0.0184018	39.16121	39.23334
	Yes	40.182	0.0793248	40.02653	40.33748
Average age	No	16.57691	0.0056799	16.56577	16.58804
	Yes	17.0056	0.0205379	16.96535	17.04586
Average fee	No	27063.89	68.83727	26928.97	27198.81
	Yes	22495.57	235.5191	22033.95	22957.2

Table 4: The log rank test for equality of survivor functions by gender/sex

$\chi^2 = 78.77$

$p = 0.000$

Sex	Events observed	Events expected
Female	869	695.33
Male	795	166.00
Total	1664	

Table 5: Log-rank test for equality of survivor functions by school category.

		$\chi^2 = 234.39$ $p = 0.000$
Category	Events observed	Events expected
County	1487	1221.58
Extra-County	164	366.25
National	13	79.17
Total	1664	1664.00

Figure 1: Kaplan-Meier survival estimate

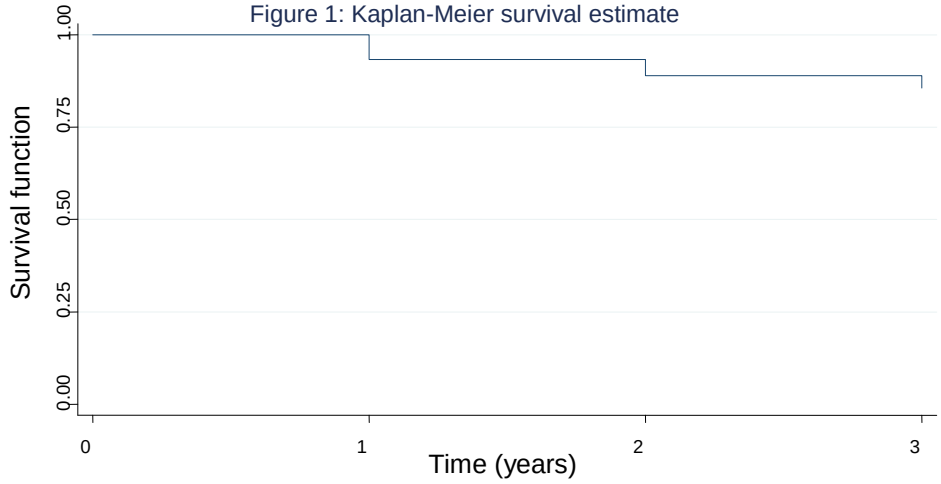
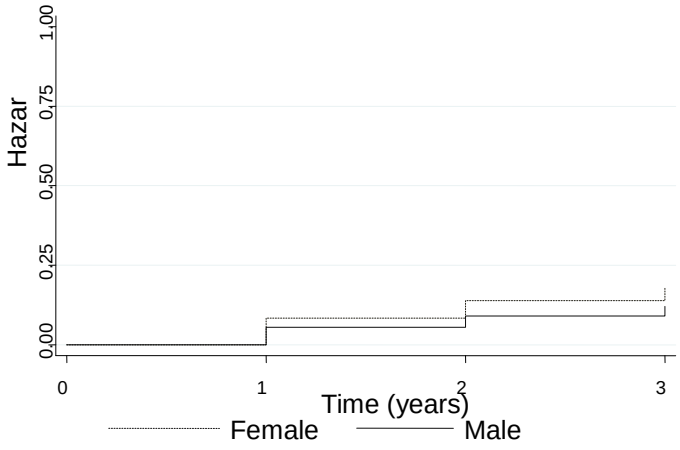


Figure 2: Kaplan-Meier failure estimates by sex



The Kaplan Meier curve in Figure 2 shows lower hazards of dropping among males compared to females.

Table 6: The failure time analysis

Category	Total	Mean	Minimum	Median	Maximum
No. of Subjects	11649		1221.58		
Number of records	32532	2.816867	1	3	3
Entry time		0	0	0	0
Exit time		2.819638	1	3	3
Subjects with gap	0				
Time on gap if gap	0				
Time at risk	32564	2.819638	1	3	3
Failures	1664	0.1440817	0	0	1

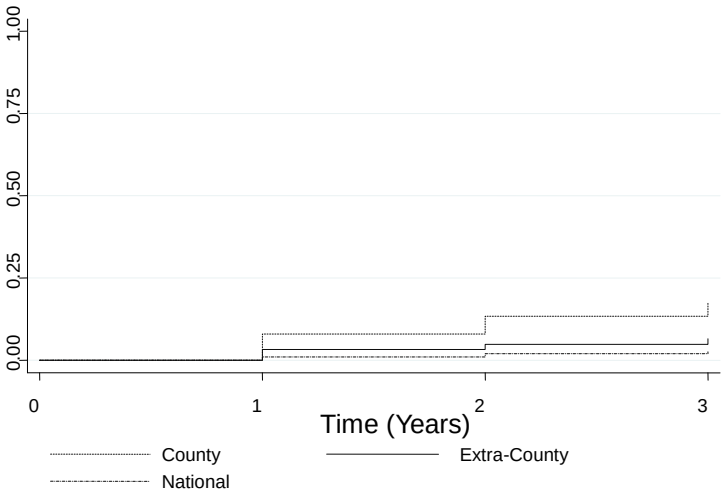


Figure3:Kaplan-Meier failure estimates by school category

The Kaplan Meier curve in figure 3 shows that national category has lower hazards of dropping compared to the rest

Table 7: The life tables for the students' transition

Interval		Beginning Total	Deaths	Lost	Survival	Std. Error	95%Conf. Interval	
0	1	11817	0	2	1.0000	0.0000	.	.
1	2	11815	775	4	0.9344	0.0023	0.9298	0.9387
2	3	11036	526	27	0.8898	0.0029	0.8840	0.8953
3	4	10483	406	10077	0.8235	0.0041	0.8152	0.8341

Table 8: The cohort characteristics

Cohort	Person-time	Failures	Rate	[95% Conf. Interval]	
Total	32564	1664	0.05109937	0.0487022	0.0536145

Table 8 shows that out of the 32564 students in the chosen cohort, 1664 dropped out which is an incidence rate of 0.05109937. The number of students who dropped out is a significant percentage.

Table 9: The Survival Analysis By Sex.

Sex	Person time	Failures	Rate	[95% confidence interval]	
Female	13563	869	0.06407137	0.05995	0.0684761
Male	19001	795	0.0418399	0.0390303	0.448518
Total	32564	1664	0.050109937	0.0487022	0.0536145

Table 10: Survival analysis by School Category

Category	Person time	Failures	Rate	[95% confidence interval]	
County	23843	1487	0.06236631	0.0592756	0.0656181
Extra-County	7216	164	0.02272727	0.019502	0.0264859
National	1505	13	0.00863787	0.00501156	0.0148761
Total	32564	1664	0.05109937	0.0487022	0.0536145

Table 11: Cox proportional hazard model

Variable	Coefficient	Standard Error	Z	*p*	[95% confidence interval]	
Category	-0.6683251	0.0969717	-6.89	0.0000	-0.8583861	-0.4782641
Gender	-0.4320194	0.0506912	-8.62	0.0000	-0.5301963	-0.3338425
KCSE	-0.0506482	0.0144943	3.49	0.0000	0.0222399	0.0790566
Average age	0.0409454	0.0624711	0.66	0.512	-0.0814958	0.1633865
Average fee	-0.0000113	0.00000217	-5.18	0.000	-0.0000155	0.000007
Student teacher ratio	0.0051654	0.0048975	1.05	0.292	-0.0044335	0.0147642

Table 12: Adjusted Cox Hazard Model with Hazard ratios

Variable		Hazard ratio	Standard Error	Z	*P*	[95% Confidence interval]	
						Log likelihood =-15289.088	
						χ^2 (16)=389.83	
						p = 0.0000	
Time independent	Gender	0.6486026	0.032625	-8.61	0.000	0.5877098	0.7158045
	School category	0.3188847	0.102212	-3.57	0.000	0.1701368	0.5976805
Time dependent	KCSE	1.049758	0.0167621	3.04	0.002	1.017414	1.083131
	Average age (aa)	1.054203	0.0689548	0.88	0.377	0.9323204	9.203352
	Average fee (af)	0.999989	2.25e-0.6	-4.92	0.000	0.999846	0.999994
	Student –teacher ratio (str)	1.00185	0.0053488	0.35	0.729	0.9914214	1.012389

Discussion

The log rank test tests whether there are differences in risk in dropping out of school within gender and also school category. From table 4, the p = 0.0000 within gender shows that they varied significantly. Also from Table 5, dropping out also within school category was significantly different. Figure 2 shows the Kaplan Meier curve which gave a visible indication of risk of survival and in this case boys were associated with lower hazards of dropping compared to girls. Also from Figure 3, the hazards of dropping out among those in national schools was lowest followed by Extra county schools then county schools with highest hazards.

Overall survival curve is shown in figure 1 which shows that by the end of four years over 75% would not have dropped from school. This shows that though the students dropped out from forms one to four, most of the students are retained to completion or graduation with about 25% dropping out, a significant number though.

Table 11 shows the adjusted Cox model with coefficient values (β).The column p value shows whether the variable in the model is significant. Variable with $p < 0.05$ was included in the model and their coefficients substituted in the equation. For our case; category, sex, performance in KCSE and average fee payment were significant in the model and so the model was reduced to contain only these variables. The model therefore took the form in equation 16.

$$h(x,t) = h_o(t)\exp(0.0506482X_1 - 0.0000113X_2 - 0.668351X_4 - 0.4320194X_5) \quad (16)$$

The Table 12 shows the same adjusted model but reporting hazards. From the table, we can say that boys in reference to girls are associated with reduced hazards of dropping out from schools as depicted by Hazard ratio (HR): 0.6486, 95% confidence interval (0.587 to 0.7156) p = 0.000). Also in reference to county schools, national schools and extra county schools were associated with lower hazards of dropping; HR: 0.0.32, 95% confidence interval (0.0.17 to 0.60) , p = 0.000) and HR: 0.51, 95% confidence interval (.41 to 0.62) p = 0.000) respectively. Since the confidence intervals for county and extra county schools overlaps then we can conclude that the hazards of dropping out from the two categories do not vary.

Table 9 shows incidence of 0.0510 which shows about 5 students dropping for every 100 person years schooling which is equivalent to 1 student somewhere between classes dropping out of school for every 20

enrolling in form one compared to girls which is at most one or no incidence for every 20 boys joining form one compared to 6 incidence for every 20 girls joining form one .The school category section shows incidence of dropping out of 6,2,and 1 for county, extra county and national schools respectively for every 20 students joining form one.

From Table 9 out of the 11549 students who enrolled in form one, 1664 dropped out of school in the course of study and the total time at risk contributed by all students was 32564 years

From Table 12, a unit increase in KCSE results was associated with higher hazards of dropping HR: 1.12, 95% confidence interval (1.02 to 1.08) p = 0.000) while a unit increase in average school fees was associated with reduced hazards of dropping HR: 0.999, 95% confidence interval (0.991 to 0.9999), P=0.000). While Table 7 shows the life tables, where majority of those dropping out happened at form two.

Conclusion

According to the findings of this research, drop-out rate in Kisumu County is influenced by performance in KCSE, School category, average fee payment and gender/sex. Therefore, the Cox proportional hazard model is most suitable for it shows the interaction between the covariates and the dependent variable. It has been found out that drop- out rate; increases with decrease in KCSE performance, increases with increase in fee payment, decreases with the category of the School and such that more girls drop out of the schools than boys. This study was meant to provide future survivorship for the students in this region in order to help improve the retention and graduation rates to 100%.

Recommended further area of research

The Cox Proportional hazard model should be applied in a larger area especially the whole country in order to ascertain the academic future of the nation based on retention and graduation rate.

Acknowledgement

Much appreciation goes to all who in a way or the other guided us throughout the duration of this study. We are greatly indebted to the financial support from Higher Education Loans Board (HELB) and many thanks to the Ministry of Education- Kisumu County for providing us with the data for the study as well as the staff of Jomo Kenyatta University of Agriculture and Technology (JKUAT) at Kisumu CBD campus.

References

Chimka, Justin, R. T. R. and Kash, B. (2007). *Proportional hazards models of graduation. Journal of American Educational Research,* Washington D.C.

Cox, D. R. (1997). *Regression Models and Life Tables*. Royal Statistical Society, New York.

Onyando, R. M. and Omondi, M. (2008). *Counting The Costs of Teenage Pregnancy and School's Drop Out In Kenya*. Centre for the Study of Adolescence, Nairobi.

Singer, D. and Willet, J. (2006). *Using Discrete - Time Survival Analysis to study duration and the time of event*. Springer-Verlag, Chicago.

Tyler, S., S. B. and Ryan, M.(2000). *Survival Analysis Using Cox Proportional Hazards Modelling For Single And Multiple Event Time Data*. Naval Health Research centre, San Diego, CA.

Terry, M. T. and Patricia, M. G. (2000). *Modelling Survivall Data: Extending the Cox Model*. Springer Science + Business Media, New York.

Walters, J. S. (2009). *What is A Cox Model?* Hayward Medical Communications, New York.

Wickens, John, B. and Singer (1991). *From whether to when: new methods for studying student drop out and teacher attrition*. University of California, Los Angeles.

Wickens, D. T. (2004). *The General Linear Model.* University of California, Los Angeles.

Zwick, R. and Jeffery, G. (2005). *Predicting college grades and degree completion using high school grades and Science and Technology Scores: the role of student ethnicity and first language*. The journal of American Education Research, Washinton D.C.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar