American Journal of Theoretical and Applied Statistics
2018; 7(1): 1-7
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20180701.11
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Application of Cox Regression in Modeling Survival Rate of Drug Abuse

Robert Kasisi, Joseph Koske, Mathew Kosgei

Department of Mathematics and Computer Science, Moi University, Eldoret, Kenya

Email address:
rkasisi@gmail.com (R. Kasisi), mkkosgei@yahoo.com (J. Koske), koske4@yahoo.co.uk (M. Kosgei)

To cite this article:
Robert Kasisi, Joseph Koske, Mathew Kosgei. Application of Cox Regression in Modeling Survival Rate of Drug Abuse. *American Journal of Theoretical and Applied Statistics*. Vol. 7, No. 1, 2018, pp. 1-7. doi: 10.11648/j.ajtas.20180701.11

Received: June 28, 2017; **Accepted**: July 10, 2017; **Published**: December 20, 2017

Abstract: Drug and substance abuse is a serious health problem in many countries. In Kenya drug abuse is one of the leading causes of mortality. Modeling the rate of survival of drug users involves determining time to relapse of drug users and the number of treatment episodes for full recovery. A study of the treatment programs that the subjects are enrolled was conducted. Those subjects who completed the treatment program and fully recovered from drug use were said to have survived while those who dropped out of the treatment program were said to have not survived. The objective of this study was to fit a cox repression model in determining a set of significant covariates for survival of drug users in Kenya. The dependent variable was survival time of the subject and the independent variables were age, gender, residence, marital status, job status, mode of drug abused and the type of drug abused. The study used data on drug use from Mathari National Hospital. Cox proportional hazards model was used to establish the hazard rate of a subject entering into drug use at different stages of life. Survival rate was 36.37% with the females having higher survival rates compared to male drug users. Age, gender, marital status and employment status were significant predictors of survival rate of drug users. The study recommended that subjects who were aged below 30 years, single and jobless required more intensive and specialized treatment. More intervention programs should be targeted to these subjects.

Keywords: Survival Rate, Cox Regression, Intervention Programs, Hazard Rate, Drug Abuse

1. Introduction

Several theories have emerged to explain reasons for drug abuse. Personality theory asserts that there exists certain traits for individuals who abuse drugs. Such traits include poor tolerance for frustration, poor impulse control, poor coping ability and low self-esteem. Individuals who possess such traits find it difficult to abstain from abuse of drugs. Peer group influence theory asserts that many young people indulge into drug use out of influence of their peers. As young people grow up they reduce their dependence on parents. They begin to depend on their friends whom most of them could be drug abusers and therefore get influenced to drug abuse. The theory of parental supervision asserts that lack of parental supervision leads young people to the abuse of drugs.

Most young people indulge into drug use as a result of peers. Some of them enter to drug abuse to break from parent's authority (Hempill, *et al*. 2011; Arteaga, *et al*. 2010). Some young people enter into drug abuse earlier than many adults suspect (Peterson, 2010; Fisher, *et al*. 2006). This habit of drug abuse continues in later life (Goldberg, 2011). Most of entry drugs into abuse by young people are cannabis, alcohol and cocaine. According to World Health Organization, (2007) alcohol is the primary and most dangerous abused drug. Guttannova, *et al,* (2011) studied the association between early onset of alcohol use and adult alcohol abuse and found that those who engage in regular drinking before age 21 years old had a greater rate of alcohol dependence.

In order to curb the problem of drug use it is important to study the factors that may improve the survival rate of the drug abuse subjects. Several researchers have studied the factors that influence survival rates based on of mode of drug taken, marital status, residence, education status, employment status, age and gender. Drug abuse subjects most commonly inject drugs such as heroin, Methamphetamine and Cocaine. IDUS are more likely than those using other routes to be

older (age 35+), unemployed, possess less than a high school education and reside in rural areas. IDUs also exhibited higher rates of abuse/ dependence, perceived need for substance abuse treatment and co-occurring physical and psychological problems.

2. Method

2.1. Cox Regression Model for Survival Data

The Cox Proportional Hazard (PH) Model is a multivariate regression method used to determine the effect of multiple covariates on the survival. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of covariates, while keeping the baseline hazards unspecified and can take only positive values. This model is defined as $h(t,x,\beta) = h_0(t)$ where $h(t,x,\beta)$ is the hazard function at time t with covariates $X = (X_1, X_2, \dots., X_p)'$.

$h_0(t)$ is the arbitrary baseline hazard function that characterizes how the hazard function changes as a function of survival time. $\beta_0 = (\beta_1, \beta_2, \dots., \beta_p)'$ is a column vector of p regression parameters associated with explanatory variables. $e^{\beta' x}$ characterizes how the hazard function changes as a function of subject covariates. t is the failure time. Each individual has its own hazard function of survival time. Then, the above model becomes

$$h(t, X_i, \beta) = h_0(t) exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \quad (1)$$

Cox proportional hazard model is popular because it allows a flexible choice of covariates: time varying, time-independent, continuous and discrete. Two other issues that make it popular are that it does not make any assumption about the underlying survival distribution and also does not require estimation of the baseline hazard rate, $h_0(t)$ to estimate the regression parameters.

2.2. Estimation of Parameters in Proportional Hazard Model

Regression coefficients in the proportional hazards Cox model, which are the unknown parameters in the model, can be estimated using the method of maximum likelihood. In Cox proportional hazards model we can estimate the vector of parameters β without having any assumptions about the baseline hazard $h_0(t)$. Consider n independent subjects, the data that we need for the Cox proportional hazard model is represented by triplet

$$(t_i, \sigma_i, X_i), i = 1,2,3 \dots \dots . . n$$

where

t_i is the survival time for the ith subject

σ_i an indicator of censoring for the i^{th} subject given by 0 for censored and 1 for event/death

X_i a vector of covariates for individual $i(X_{i1}, X_{i2}, \dots, X_{ip})$. Then, the full maximum likelihood is defined as $L(\beta) = \prod_{i=1}^{n} h(t_i, X_i, \beta)^{\sigma_1} S(t_i, X_i, \beta)$ where $h(t_i, X_i, \beta) = h_0(t_i) e^{\beta' X_{i1}}$ is the hazard function for individual i. $S(t_i, X_i, \beta) = s_0(t_i) e^{\beta' X_{i1}}$ is the survival function for individual i. Then, the full maximum likelihood becomes

$$L(\beta) = \prod_{i=1}^{n} \left(h_0(t_i) e^{\beta' X_{i1}}\right)^{\sigma_1} s_0(t_i) e^{\beta' X_{i1}} \quad (2)$$

Full maximum likelihood requires that we maximize (3.1) with respect to the unknown parameter of interest, β, and unspecified baseline hazard and survival functions. This indicates that unless we explicitly specify the baseline hazard, $h_0(t)$ we cannot obtain the maximum likelihood estimators for the full likelihood.

2.3. The Breslow Approximation

This approximation is proposed by Breslow and Peto to modify the partial likelihood and has the form

$$L_p(\beta) = \prod_{i=1}^{n} \frac{exp(\beta' X_i)}{\left[\sum_{j \in R_{t(i)}} exp(\beta' X_i)\right]^{\sigma_i}} \quad (3)$$

Where

σ_i the number of relapses occurred at time *i*

s_i is the sum of covariates over σ_i subjects at time *i*

Then, the partial log is given as

$$L_p(\beta) = \sum_{i=1}^{m} \left[\beta' S_i - d_i ln \sum_{I \in R_{t(i)}} exp\,(\beta' X_i)\right] \quad (4)$$

Breslow maximum partial likelihood estimator, adjusted for tied observation is obtained, by differentiating equation (4) with respect to the components of β and setting the derivative equal to zero and solving for the unknown parameters.

2.4. Model Development

In any applied setting, performing a proportional hazard regression analysis of survival data requires a number of critical decisions. It is likely that we will have data on more covariates than we can reasonably expect to include in the model, so we must decide on a method to select a subset of the total number of covariates. When selecting a subset of the covariates, we must consider such issues as clinical importance and statistical significance, (Hosmer and Lemeshow, 1999).

3. Results

3.1. Survival Difference

The pairwise comparison tests for the covariates showed that the survival of patients based on age, gender, marital status and job status were statistically significantly different ($p<0.05$) while there were no significant differences between type of drug used, mode of taking the drug and residence of the subjects ($p>0.05$). This was shown in table 2.

Table 1. *Survival difference.*

Covariate	Median time	χ^2	Log rank
age	18	38.44	52.25
gender	19	7.06	7.18
Marital status	18	8.55	8.87
Job status	19	17.46	18.54

From the entire follow up period of two years the study obtained a survival rate of 36.37% based on the total time. Relapse subjects constituted 30.9% (63 subjects) of the study, while subjects without relapse comprised of 69.1% (99 subjects). The results of the cox proportional model were presented graphically (figure 1).

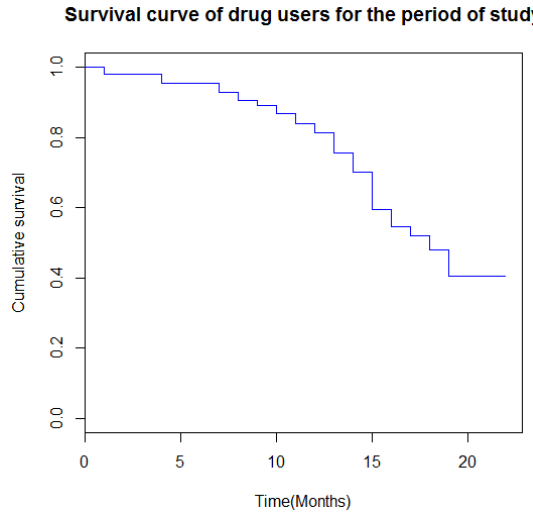


Figure 1. *Survival function for drug abuse.*

3.2. Analysis of Maximum Likelihood Estimates

Those predictors that were significant were selected using the maximum log partial likelihood of the model $(-2LL)$. The results showed larger reduction in $-2LL(\hat{\beta})$ for residence that reduced the value of the null model to 34.75 with a p value of 0.0300 followed by job status with a likelihood ratio of 15.31, p value of 0.0000, marital status with a likelihood ratio of 9.11, p value of 0.0036, age with a likelihood ratio of 6.04, p value of 0.011, gender with likelihood of 1.06, p value of 0.027, drug type with a likelihood ratio of 0.64, p value of 0.2351 and finally mode of taking the drug with a likelihood ratio of 0.92 with a p value of 0.3371. Using this procedure covariates were eliminated in accordance to their magnitude in which they reduced the $-2LL(\hat{\beta})$. Those predictors that were significant were considered for the next multivariable analysis at p-value of 0.25. These predictors included age, residence, job status, Type of drug, gender and marital status. Age, gender, job status and marital status had strong associations with survival time of drug users at P-value less than 0.05. The Covariate that was not significant was mode taken and was therefore removed from the model. The study then fitted initial multiple Cox proportional model by considering the six covariates that were significant. This was followed by another Cox proportional regression model fitted by eliminating covariates which were not significant at p value of 0.05. From the total of seven covariates, residence, type of drug and mode of taking the drug (p-value >0.05) were eliminated from the model. The final Cox model comprised of four covariates age, gender, job status and marital status. The importance of the variables which were not significant in the univariate analysis as predictors or useful confounder of survival experience of patients and their effects was then assessed. The effect of those variables not significant in the analysis was also examined. These variables were added one sequentially into the cox model containing the four variables significant at 5% significance level.

Then the improvement on $-2LL(\hat{\beta})$ was determined for significance. The results showed that none of those variables were significant and therefore they were removed from the model. Then Wald test was used to assess the significance of reasonable and possible interactions. The null hypothesis tested was that the model with only main effect fitted the model equally well as the model having the main effects and their interactions as predictors. The decision for rejection of null hypothesis was reached if $-2LL_2 - (-2LL_1) > \chi^2(\alpha = 0.05) = 3.84$. Thus, the interaction of each variable was assessed. Accordingly, none of the variables had significant interaction with the other variables. Therefore, the final model contained only the main effects [Table 2]. The study also found out that the model containing the four significant covariates age, gender, marital status and job status had the smallest value of Akaike Information criteria of 532.3751 compared to the model containing all the variables which had Akaike information criteria of 535.3489. This suggested that that model with the smallest AIC and BIC values was the best for the study, [table 3].

Table 2. *Analysis of maximum likelihood.*

Variable	DF	$se(\beta)$	Z	$p > chi$	HR	LR	CI	Sig
residence	1	0.3128	5.231	0.0300	5.136	34.75	2.782-9.482	0.0300
Age	1	0.1334	-2.541	0.0111	0.7126	6.04	0.5487-0.9254	0.011
Job status	1	0.1618	-4.023	0.0000	0.5215	15.31	0.3797-0.7161	0.0000
Marital status	1	0.1746	2.909	0.0036	1.662	9.11	1.18-2.34	0.0036
Drug type	1	0.06662	0.809	0.235	1.055	0.64	0.9262-1.203	0.2351
Mode taken	1	0.1619	-.972	0.331	0.8544	0.92	0.622-1.173	0.3314
gender	1	0.1564	0.873	0.027	0.7628	1.06	1.185-1.9774	0.0269

Table 3. *AIC and BIC Values of Cox model covariates.*

Model 1 (original model)		Model 2 (optimal model)	
AIC	BIC	AIC	BIC
535.3489	548.2077	532.3751	540.9477

Model 1

$$h(t,x) = h_0(t)\exp(-0.345x_{age} + 0.0901x_{gender} + 0.5663x_{marital.status} + 0.07907x_{job.status} + 0.0665x_{mode.taken} + 0.061x_{type.of.drug}) \quad (5)$$

Model 2

$$h(t,x) = h_0(t)\exp(0.03534x_{age} + 0.02363x_{gender} + 0.56934x_{marital.status} + 0.077103x_{job.status}) \quad (6)$$

After removing the variables that did not deserve to be in the model, table 4 shows that the age of a patient, gender, marital status and job status of the subject significantly affect the survival rate of drug abuse. Therefore the fitted model was

$$h(t,x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \quad (7)$$

$h_0(t)$ =baseline hazard rate at time t and x is the observation

β=estimated coefficient for observation x, the independent variable.

$$h(t,x) = h_0(t)\exp(0.03534x_{age} + 0.02363x_{gender} + 0.56934x_{marital.status} + 0.077103x_{job.status}) \quad (8)$$

Table 4. *Analysis of maximum Likelihood Estimates for Model 1.*

Variable	DF	Coef	$se(\beta)$	Z	HR	$p > chi$	Sig	CI
gender	1	0.02363	0.125	-2.51	0.7316	0.01206	0.048	0.5732-0.9338
Marital status	2	0.56934	0.1704	-4.17	0.4915	0.000	0.0057	0.3520-0.6864
Job status	3	0.77103	0.1817	3.121	1.7629	0.002	0.0016	1.2348-2.5171
age	3	0.3534	0.1326	2.472	1.2830	0.0010	0.0010	1.0452-2.3592

Analysis of deviance was carried out to test goodness of fit of the proposed model. It was found that the model was a good fit with p values less than the standard p value of 0.05 upon adding the covariates sequentially. More precisely adding covariate gender to the model had a significant impact of 0.0267, adding covariate age to the model with age had a significant impact of 0.006, marital status had a significant impact of 0.003 to the model with both gender and age covariates while adding covariate job status to the model containing gender, age and marital status as covariates had a significant impact of 0.0001306, (table 5).

Table 5. *Analysis of Deviance table.*

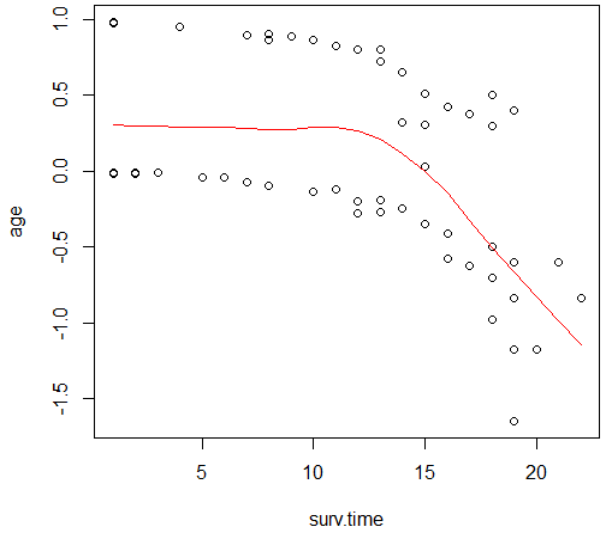
	loglik	Chisq	Df	Pr (>|Chi|)
NULL	-280.10			
Gender	-277.64	4.9101	1	0.0267000 *
age	-273.87	7.5493	1	0.0060033 **
Marital status	-269.50	8.7332	1	0.0031246 **
Job status	-262.0	19 14. 6328	1	0.0001306 ***

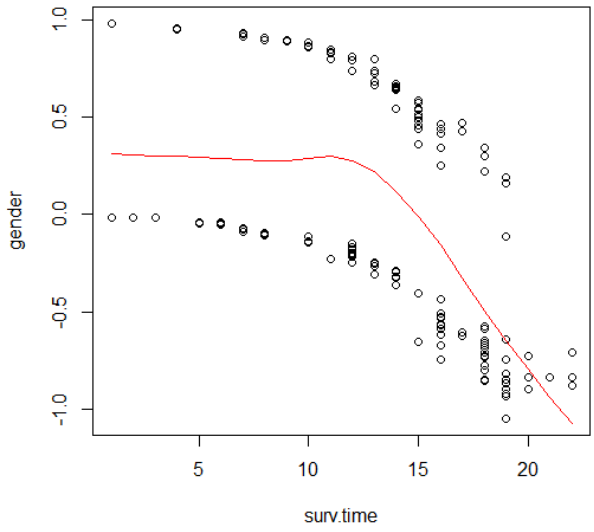
3.3. Model Diagnostics

After identifying the final preliminary model the next step was to diagnose the fit of the model. As in the case for a linear or generalized model, it was desirable to determine whether a fitted Cox regression model adequately described the data. The three kinds of diagnostics that were considered in the study were, violation of the assumption of proportional hazards, effect of influential observations and nonlinearity in the relationship between the log hazard and the covariates.

3.3.1. Assessment of Linearity of Covariates in the Model

The study sought to check whether the correct functional form of the continuous covariate held in the model proposed to describe the data. The hypothesis of interest was that the effect of the covariate was linear in the log hazard. Graphical technique of the plots of the martingale residuals was used to assess the linearity of relation of continuous covariate in which the correct functional form was understood. The study obtained plots shown in Figure 2.





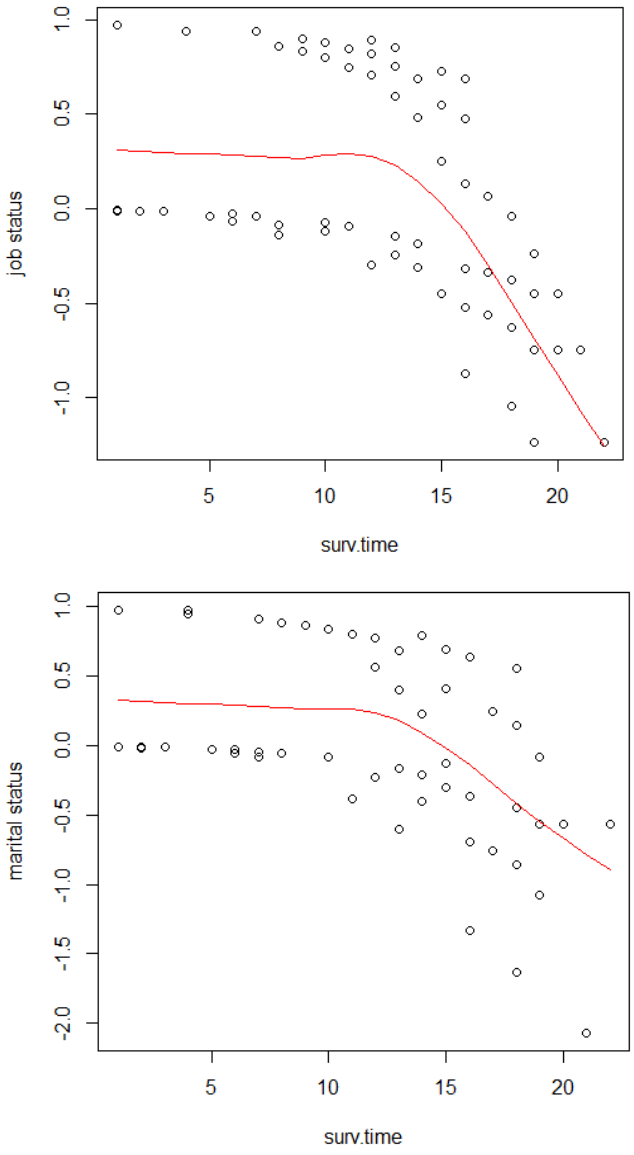


Figure 2. *Assessment of linearity of covariates in the model.*

Figures for age, gender, marital status and job status show the plot of martingale residuals versus each covariate. For the covariates, age, gender, job status and marital status, the plots did not show systematic patterns or trend and the resulting smoothed plots (LOESS) were approximately between -1 and +1.

Therefore the plots of martingale residual confirmed that age, gender, job status and marital status of a subject had an approximate linear relationship with the survival time. Therefore the study concluded that the model containing covariates age, gender, job status and marital status was an appropriate model to describe the data, since it had passed all tests of fitting a cox proportional hazards model.

3.3.2. Assessment of the Proportional Hazards Assumption

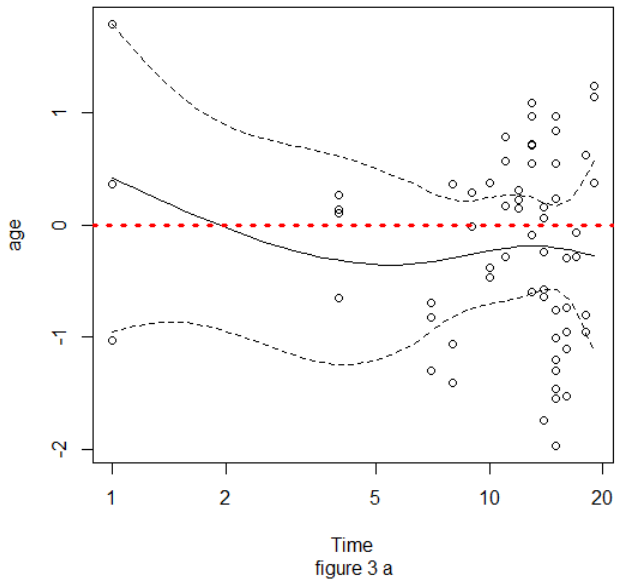
The assumption of proportional hazards states that the hazard ratios are constant overtime. That is, the risk of failure must be the same no matter how long subjects had been followed. In order to test this assumption, Cox model was employed and a graphical display used to substantiate the same. Thus, in this study, using a test based on the interaction of the covariates with the log of time and also using the plot of the scaled Schoenfeld residuals the assumption was used to see if the assumption of proportionality was violated or not. Therefore, one of the statistical tests for proportional hazards assumption was to generate time varying covariates by creating interactions of the predictors and a function of survival times, usually covariate times the log of time, and including them in the model. If any of the time dependent covariates were significant then those predictors did not exhibit a proportional effect over the study period. That is the proportional hazard assumption failed to hold. The result of the test is given in table 6. The table shows the Wald chi-square value and corresponding P-values for each covariate. Since the P-value of the Wald test was greater than 0.05 for all covariates, there was no evidence against the proportionality of hazard assumption. The global test also gave a p value that was not significant suggesting that the assumption had not been violated (p=0.230).

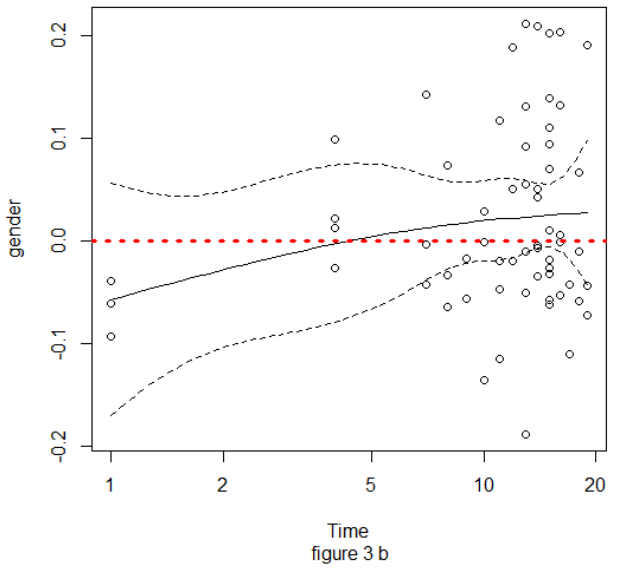
Table 6. *Assessment of proportional hazards assumption.*

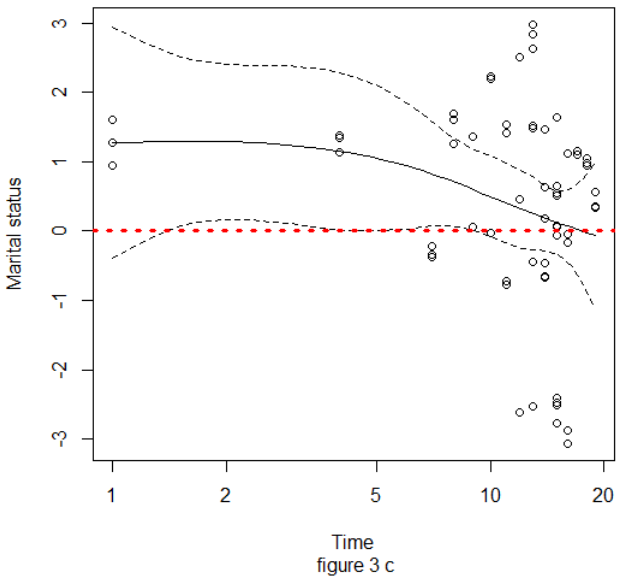
Covariate	Rho	Chisq	Probability value
Age	0.206	1.508	0.22
Job status	0.113	0.718	0.397
Marital status	-0.176	1.705	0.192
gender	0.214	2.783	0.221
Global	NA	4.313	0.230

In addition, the assumption of proportionality was also assessed graphically by plotting the scaled Schoenfeld residuals of each covariate against log time [figures 3 a, b, c, d].

All interactions of covariates with the logarithm of survival times were modeled together with the main effects and Wald statistic used to test the significance of the interaction terms at 5% level of significance. The result of the test indicated that none of the coefficients of interaction terms were significant at 5% level (i.e. higher p-value). The results revealed the non-significance of time-dependent covariates. On the other hand, there were no covariates which showed a trend/pattern with the time that indicated the hazard ratios would be constant over the study time. This showed that there was no sufficient evidence to reject the null hypothesis that the coefficients of the time varying covariates (interaction terms) were zero. Thus there was no enough evidence against proportionality assumption to hold. Furthermore, plotting the scaled Schoenfeld residuals of each covariate against log time was used to check whether the assumption of proportional hazards was violated or not. This plot indicated that the residuals were random and LOESS curve was smooth and almost had zero slope. This also suggested that the plots supported proportionality assumption to hold.







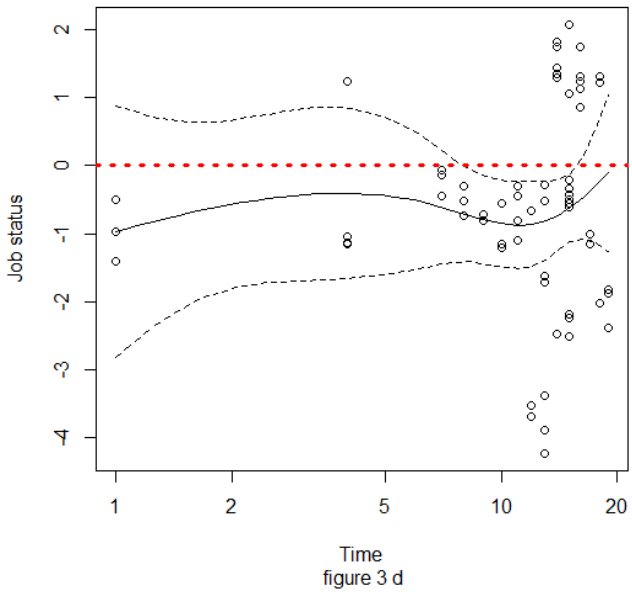


Figure 3. *Assessment of proportional hazards assumption.*

4. Discussion

The study conducted a follow up study on a cohort of drug users who had enrolled in the beginning of July 2013 to the end of our study period, that is, June 2015. Information on these subjects was obtained from referring to the subjects' medical records for the entire period that they were in or attended the hospital. Factors such as age, marital status, employment status, residence, type of drug abused, mode of taking the drug and gender of the subjects were studied. From the results of the study it was found that on average subjects attending treatments were aged 33 years. The median survival time on the basis of marital status was 18 months, 19 months for employment status, 19 months for gender and 18 months for age. This meant higher survival rates were reported for married and employed subjects attending treatment. In an overall test, equality of survival times for all the subjects involved in the follow up study, based upon the differences in group mean, was significant (Wilcoxon statistic=103, df=27, $p = 7.44 \times 10^{-11} < 0.05$). This means therefore that survival curves of at least two covariates were significantly different.

The pairwise comparison tests for the covariates showed that the survival of patients based on age, gender, marital status and job status were statistically significantly different ($p<0.05$). This implied that age, gender, marital status and job status were significant predictors of drug abuse. Therefore treatment programs should tailor their treatment programs on the basis of these covariates. Such that subjects who divorced, unemployed, single, young and mostly male subjects should be given more treatment attention compared to the other subjects for ease of recovery. From the entire follow up period of two years the study obtained a survival rate of 36.37% based on the total time. Relapse subjects constituted 30.9% (63 subjects) of the study, while subjects without relapse comprised of 69.1% (99 subjects).

5. Conclusion

Survival rate of drug users was examined by analyzing the data on the 162 drug use patients. Analysis was done with respect to age, employment, gender, and marital status. Marital status was identified as a predictor associated with higher observed survival time for drug abuse in multivariate analysis, where married patients were more likely to recover compared to unmarried patients. In addition, job status was also significant where employed subjects were found to have a higher survival rate compared to unemployed subjects. Gender and age were also shown in multivariate analysis to be independent prognostic factors for observed survival. The study results for our case were not very different from those of other countries. The positive association observed in this study was consistent with similar studies done in other countries.

However, there is insufficient evidence to largely support the association between marital status and drug use survival probability in Kenya. Additional follow up and larger scale studies need to be implemented in Kenya in order to depict the association of marital status with the drug use survival. Such studies present a priority in light of the escalating prevalence of drug abuse. It is expected that Kenya's undiagnosed drug use population is as numerous as the diagnosed group.

It was also found that during enrollment the hazard rate of drug use was higher and experienced gradual decline towards end of the study period. The study reveals that youth are at higher risk of indulging into drug abuse compared to the older population. Those subjects who had historical backgrounds of drug use were more exposed to drug use. It was also observed that married people had a higher survival rates compared to the unmarried and divorced subjects and observed that injecting drug users were more likely to quit drug use compared to oral drug users. The study revealed that age, gender, marital status and employment status were significant factors for survival rate of drug users.

Therefore, these results provide a foundation of evidence and an essential element for raising public awareness, advocacy and improving health care service delivery with regards to drug abuse. Continued monitoring and evaluation of the drug use survival estimates is a vital component to developing future targeted and effective programs and policies in Kenya.

The study also recommends that data collection on drug and substance abuse should not be restricted to the primary substances of abuse but should consider as well cases of multiple drug abuse of substances. More intervention programs should target the youth. It is also recommended rehabilitation facilities should be set up within reach for the subjects to enable early detection and treatment of drug abuse and anti-drug abuse campaigns should also be targeted to fight multiple drug use. Other key areas that could help solve the problem of drug abuse include conducting regular educational campaigns, distributing specialized treatment centers in order to cater for multiple drug use, early reporting of addiction problems and the need to offer cost effective and efficient treatment services to drug addicts.

References

[1] Arteaga, I., Chen, C. C. & Reynolds, A. J. (2010) Childhood predictors of adult substance abuse "*Children and Youth Services" Review*, 32, pp. 1108-1120.

[2] Cox, R. G., Zhang, L., Jonhson, W. D. & Bender, D. R. (2007). Academic performance and use: Findings from a state survey of public high school students. *Journal of School Health, 77* (3), pp. 105-155.

[3] Breslow N & Crowley, (1974). A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. *Annals of Statistics. Volume 2, Number 3, 437-453.*

[4] Collett, D. (2003). Modeling Binary Data. *Chapman and Hall, London, 2nd Edition.*

[5] Efron, B, (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of American Statistical Association, 72, 557-565.*

[6] Fisher & Roget, N, (2006). The Drug Abuse Treatment Outcomes. *Journal of psychoactive drugs.*

[7] Gomberg, E. S. (1994). Risk factors for drinking over a woman's life span. *Alcohol Health & Research World, 18,* 220-227.

[8] Guttannova P. (2011). Adolescent transitioning and substance misuse. *Journal of Substance Abuse, 5,* 1-14.

[9] Hemphill (2011). The role of psychology in the prevention of youth violence. *Australian psychological association*.

[10] Hosmer, D. W. and Lemeshow, S. (1999). Applied Survival Analysis: Regression Modeling of Time to Event Data. *Wiley, New York.*

[11] Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241 (1982).