UNIVERSITY OF NAIROBI

SCHOOL OF MATHEMATICS

ANALYSIS OF FACTORS RELATED TO RECIDIVISM.

A PROJECT REPORT SUBMITTED TO THE SCHOOL OF MATHEMATICS IN
PARTIAL FULFILLMENT FOR
A MASTER OF SCIENCE DEGREE IN SOCIAL STATISTICS

Isaac Masese Onsando

I the undersigned declare that this project report is my original work and to the best of my knowledge has not been presented for the award of a degree in any other university

Isaac Masese Onsando

.... 22/11/2012

Signature *Date*

Declaration by Supervisor

This project report has been submitted for examination with my approval

as a university supervisor

Prof.Jamm Ottieno

.... 22/11/2012

Signature *Date*

DEDICATION

This project is dedicated to my father Zachariah who selflessly stood by me during the entire study period.

EXECUTIVE SUMMARY

It is generally agreed that amongst the objectives of a punitive or correctional system is the prevention of the re-occurrence of criminal behavior on the part of those who have deviated from the socially accepted and legally prescribed rules of conduct. The aim of many programs of treatment of criminal offenders is the prevention of recidivism, and the choice of any treatment plan is based upon the conviction that of what is known about the characteristics of an offender, a particular method of treatment will best realize the objectives of the correctional system Offenders are placed on probation, incarcerated, released on parole, etc., when the accumulated evidence suggests that they possess that pattern of attributes which tends to insure the favorable response of offenders not recidivating. The accuracy of this choices determines not only the effectiveness of any treatment program in achieving the desired results, but also the manner in which the community is safeguarded against predatory individuals. The procedures involved in making such decisions, the kind of information utilized in arriving at decisions are of the utmost importance in the administration of criminal justice.

The relationship between various factors and demographic characteristics of released offenders and recidivism is investigated in a sample of 402 adult male offenders in this study. Age, work experiences, marital status, as well as the education level contributed significantly to the relapse to crime. A unit increase in age reduces the risk of relapse by 4%, released convicts from slum dwellings are three times more likely to recidivate than their counterparts from all other dwellings, those with previous work experience have 60% less chance of recidivating than those without, released felons that were married before arrest have 50% less chances of recidivating than their un-married counterparts and a unit increase in education level reduces the risk of recidivating by 44%. However, those released on parole did not show any differences with those released after serving their entire sentences. Interactions between some of this variables also showed significance in predicting recidivistic behaviour.

ACKNOWLEDGEMENT

I am first and foremost deeply indebted to my supervisor Prof.Jam Ottieno without whose guidance this work would not have taken this shape. My appreciation also goes to all the lecturers of the school of mathematics who inspired me in one way or another. I acknowledge the support I got from my class colleagues whose discussions in and out of class helped me get this far. Finally to my family for the encouragement during the entire period

To all, I thank you most sincerely.

Table of Contents

Chapter One

General Introduction

1.0 Background

Gresham Sykes (1958) acknowledged five pains of incarceration: isolation from the larger community; lack of material possessions; blocked access to heterosexual relationships; reduced personal autonomy; and reduced personal security. Sykes argues that these deprivations comprise what is referred to as prisonization, that is, alienation from the larger society. Additionally, criminologists argue that many inmates bring to prison a commitment to criminal subcultures and criminal norms (Irwin and Cressey 1962). Both the deprivations of imprisonment and the imported criminogenic norms, criminologists argue, facilitate the growth of inmate subcultures favoring a normative orientation hostile to prison management and supporting a continuation of criminal behavior after release from prison (Thomas and Petersen 1977; Kassebaum et al.)

Risk assessment has become a key activity in our criminal justice system, with profound consequences for public safety and particularly for the offender. Many decisions throughout an offender's progression in the criminal justice system involve risk assessment, including sentencing, security classification, parole decisions, treatment needs, and supervision intensity. Risk assessment is also inherent in two of the three principles of effective correctional treatment (risk/need/responsivity; Andrews & Bonta, 2006; Andrews et al., 1990). Treatment intensity should be directly proportional to the offender's risk (the risk principle), and treatment should target the criminogenic needs of the offender (those needs related to criminal behaviour; the need principle). Adherence to either principle requires a risk assessment. Given this reliance on risk assessment and the implications for public safety and for the offender, it is important for risk evaluators to be cognizant of the history of risk assessment, its various methods, their empirical support, and the limitations of current research.

An offender's demographic characteristics play a significant part in the risks they pose to reoffend when they are released from prison. May et al. (2008) state that recidivism carries an inverse relationship with age. This relationship indicates that recidivism decreases as age increases. May et al. found that "reoffending...was the highest for those aged from 18 to 20, and lowest for those aged 40 and over". Langan and Levin (2002) also found that older prisoners are much less likely to recidivate than younger prisoners.

Francis J. Carney (1967) supports the findings that the most crucial variables in terms of predicting recidivism or non-recidivism were found to be the combination of age at present commitment and prior penal record. It is clear that those subjects who are relatively old and who have had no previous commitments are quite likely to be non-recidivists. On the other hand, those who are relatively young and who have been previously committed to a correctional institution are likely to be recidivists.

Race is another significant factor captured in a lot of research. Langan and Levin (2002) found in their study that blacks are more likely to recidivate than whites and those of nonhispanic origin are more likely to recidivate than those of Hispanic origin. This is attributable mostly to their social- economic status than to any other differences that may exist between these races. In Kenya, and specifically Nairobi, the distinct social classes manifest themselves in the residential backgrounds of offenders, which determine largely their education levels and work experience and consequently their susceptibility to a life of crime. Often the majority of the prison population is comprised of offenders from slum dwellings characterized by high levels of unemployment, drug use, and illiteracy. These offenders can therefore be considered as being at high risks of re-offending because once released, they go right back to these circumstances and slowly relapse back to crime simply to survive.

Coley and Barton (2006) estimated that over half of all African American males who drop out of high school are incarcerated at some point in their lives. While in 1980, 10% of African American high school dropouts were imprisoned, that percentage grew to 37% by 2008 (Western and Pettit 2010). This enormous increase in imprisonment among people with low levels of education illustrates “an astonishing level of institutionalization and a great growth in incarceration rates among the least educated reflects increasing class inequality in incarceration" (Western and Pettit 2010). The crisis of mass incarceration in our society largely reflects the high rates of imprisonment among school dropouts.

The differences between the offenses committed by each gender, as well as the overwhelming disparity of criminals between genders, has a lot to do with the social roles instilled in the human culture. Women are far more likely to be invested in the family unit, and therefore have less time to be involved in other, more dangerous or illegal activities outside the home as males (Messner & Rosenfeld, 2007). Langan and Levin (2002) agree that men are more likely to recidivate than women, which has a lot to do with society’s gender role socialization. This finding may have changed recently with the current shift in the definition of today's gender roles. Because of the clearly enormous differences between the genders, this factor has been found to be insignificant in most studies and most researchers have opted to narrow down on male offenders, bearing in mind that they constitute a clear majority of the prison population worldwide.

Solomon et al. (2004) found in their study that gender (male), race (particularly black and then white), age (specifically ages 20-39, average age being 35.7 years old), sentence length are significant in terms of recidivism. Soloman et al. therefore state that "ex-prisoners returning to communities with high unemployment rates, limited affordable housing options, active drug markets, and few services may be more likely to relapse and recidivate" (Soloman et al., 2004). It can be seen, therefore, that many offenders who are released from prison choose criminal behavior on a rational basis as the more conventional means of attaining social capital.

The number of previous convictions between an individual's first arrest and the present is strongly believed to be predictive of potential reoffense rates. The intensity of the record is also a factor (Beck & Shipley, 1989). Thus, "the number of times a prisoner has been arrested in the past is a good predictor of whether that prisoner will continue to commit crimes after being released" (Langan & Levin, 2002). In addition to the number of times, incarcerated, previous criminal activity can include deviant behavior that does not result in an individual's incarceration. This behavior can include activities an individual engages in on a regular basis, such as having ties to a gang or living in a crime-ridden area as these factors may pressure individuals into engaging in criminal behavior. In Kenya, this is compounded by the high levels of poverty that lead some offenders to prefer residing in prison because they are assured of a meal and a place to sleep unlike when they are free and uncertain of their next meal.

Data is collected on a daily basis in all the 93 prisons in Kenya. For each offender admitted into prison, a file is opened. Prison's case files contain information on caseloads and case characteristics. Prison's data is summarized for national level statistics by using case files to fill data collection forms that are then sent to the Research and Statistics Unit at the prisons headquarters on a weekly basis for eventual computation into national statistics. The data is mainly analyzed manually using the variables in the forms (annex 9). The analysis is in the form of simple descriptive statistics that give frequencies, percentages increases, and decreases and rates of change. This analysis provides information on type of offence, total number of offences and conviction rates among other variables

The development of effective methods for predicting whether an individual released from prison eventually returns or not is a major concern in rehabilitation. In the present climate of high inmate populations and shrinking resources it is more critical than ever to gather and report valid data on factors that may have an effect on the perpetuation of criminal behavior and to present that information in such a way that it is useful to public safety professionals in making security, classification, programming and release decisions that will improve each offender's potential for successful reintegration into society and ultimately enhance public safety.

High rates of recidivism result in tremendous costs both in terms of public safety and in money spent to arrest, prosecute, and incarcerate re-offenders. For many men aged 20–40, the prison door is a revolving one. Commit serious crime; get arrested and incarcerated; spend some time in prison; get out; commit more crimes; get arrested and incarcerated; and so on. Any effort to reduce recidivism must recognize that the diversity of the prison population requires solutions that can address a myriad of inmate needs. No single program can reduce recidivism significantly because many different factors affect it. Released inmates encounter a range of common problems that contribute to returning to criminal behaviors. Careful evaluation of rehabilitation programs is necessary to identify those that merit replication.

This study is looking to answer questions such as, what is the likelihood that an inmate who is released today will come back to prison. What factors influence recidivism rates? Do age, gender, and different social economic groups show differences in recidivism rates? So as to identify groups most likely to fail when they are released and consequently determine where to devote scarce correctional and community resources. The research questions in this study concern the risk associated with various individual characteristics on the probability of reoffending. Specifically, they address the disparity between this characteristics and their cumulative effect on hazards of re-offence. The research questions are as follows;

1. Which individual characteristics contribute to the propensity for recidivism of an offender?

2. What are the relative risks associated with each attribute?

3. What are the implications of these results for offender management and interventions?

This project report is separated into several distinct chapters. The next chapter discusses the literature review that exists about recidivism studies and the application of Cox's proportional hazard model in many other related studies. Chapter three discusses the research design and methodology of this project. This chapter presents the research questions, discusses the data set, and presents the research design and the methods of analyses. Chapter four presents the results. Finally, chapter five presents a discussion of the results presented in chapter four, discusses limitations of the research, provides suggestions for future research, and presents a conclusion.

Chapter Two

Literature review

2.1 Introduction

This chapter looks at studies that have been conducted on both the subject of recidivism and the use of Cox's proportional hazard. It categorizes these studies by first looking at those touching on the various factors and demographic characteristics considered to be related to recidivism and later on the various models used to study recidivism. The review will look at where available the objectives of the study, the methods of data collection and the form of data collected, methods of data analysis as well as the findings. It ends with a summary on how each literature contributes to shaping this research project.

2.2. Factors related to Recidivism

2.2.1. Sentence length

Sentence length describes the entirety of time or the percentage of time served. The total length of a given sentence (a sentence of twenty years as opposed to a sentence of two years) may influence the rate of recidivism once an offender is released. In Kruttschnitt, Uggen, and Shelton's (2000), they state that incarceration can have a criminogenic effect as it reduces job stability, weakens social bonds, and limits the ability to accumulate social capital. Longer sentences and time served in prison may be particularly damaging in that respect.

In contrast, May et al. (2008) indicates that longer terms are statistically significant in predicting recidivism. Their study found that offenders with a sentence of four years or more are less likely to commit another crime than offenders sentenced to a term of incarceration of one year or less. In addition, "the odds of reoffending were reduced for prisoners who were in custody for the first time" (May et al., 2008).

In "Recidivism of Prisoners Released in 1994", Langan and Levin (2002) state that "no evidence was found that spending more time in prison raises the recidivism rate" and that the "results were mixed regarding whether serving more time reduces recidivism" (Langan & Levin, 2002). There are many reasons to look at total time served as an indicator of potential recidivism. Previous criminal history and specific demographic characteristics such as age, sex, and race are continually associated with the risk of continued offending as well as previous drug use and sentence length. Langan and Levin (2002), however, did not find a relationship between sentence length and recidivism.

Song and Leib's (1993) study analyzes "the effect of prison or jail sentences on recidivism." Advocates of longer sentences do so with the argument that longer sentences are a benefit to public safety while those who advocate shorter sentences do so with the argument shorter sentences are a benefit to cost effectiveness. Both sentence types are tied to their ability to reduce recidivism rates. Longer periods of incarceration are argued to reduce crime in three ways. First, the offender is prevented from committing additional crimes against the public while in prison. This type of crime prevention is known as incapacitation and is known as a form of deterrence. Advocates for shorter sentences argue that "certainty of punishment is more important than duration of punishment in deterring offenders from reoffending" because many offenders continue to commit crime due to a variety of reasons: physical addiction, limited life choices, illiteracy, poor job training and the idea that prison is a school for criminals which emphasizes the use of criminal efforts in everyday life (Song & Leib, 1993).

Incapacitation prevents recidivism through "longer sentences, mandatory minimums, and reduced parole" (Leipold, 2006). A lot of the justification for longer sentences lies in the knowledge that half of all inmates released from prison will recidivate by being convicted for a new crime. The recidivism rate increases when you include offenders who are simply arrested for a new crime after their release. Leipold (2006) points out that almost 70% of the cohort in a 1994 study was arrested for a new crime after release.

According to Piliavin et al. (1986), "prior research has failed to unearth a consistent deterrent influence of perceived severity of formal sanctions". This lack of deterrent effect can lead to recidivism. Recidivism constitutes a failure of the criminal justice system to do its job. This fact, in part, is especially true when one looks at Langan and Levin's (2002) report on recidivism which states the average length of prison sentence was 5 years but offenders were typically " released after serving 35% of their sentence, or about 20 months" on average. Leipold (2006) points out that many individuals will likely commit additional crimes when released from prison.

Kevin L. Nunes, et al.(2007),studied Incarceration and Recidivism among Sexual Offenders and in particular examined it as a dichotomous variable (incarceration vs. community sentence) and as a continuous variable (length of incarceration). The primary purpose of the study was; to examine the association between incarceration and sexual and violent recidivism while controlling for risk. consider whether incarceration interacts with risk, and to address the possibility that there is a non- linear relationship between incarceration and recidivism in sex offender.

Follow-up data consisted of 627 male offenders who were assessed at the Royal Ottawa Hospital, Sexual Behaviours Clinic, between 1983 and 1995. To examine the magnitude of the differences between recidivists and non-recidivists, Cohen's d's were calculated. By convention, d's of around 0.20, 0.50, and 0.80 are respectively considered small, medium, and large effect sizes (Cohen, 1992). The results showed that sexual recidivism was not significantly associated with incarceration for the index offense and the effect size was very small. A series of logistic regressions were performed to examine the association between incarceration or length of incarceration and recidivism while controlling for risk. The odds ratios were reported. There was no evidence that incarceration or length of incarceration was associated with recidivism differently depending on risk, as measured by the RRASOR. To address the possibility of a non-linear relationship between length of incarceration and recidivism, the logistic and Cox regressions above were re-run with the quadratic (length of incarceration squared) and cubic (length of incarceration cubed) terms for length of incarceration added to the equation. In all cases, neither squared nor cubed length of incarceration was significantly associated with sexual or violent recidivism.

According to Kohl et al. (2008), incarceration is found to be negatively associated with rates of recidivism. As time served in prison increased, recidivism rates decreased. Inmates who served six months in prison or less had a recidivism rate of almost half while those offenders who served over six months in prison had a rate of just fewer than 0.45. Inmates who served five years or more had a rate of recidivism of 0.30. Kohl et al. (2008) also found that "recidivism rates did not differ significantly among those released after serving 6 months or less with those released after 7-12 months. Kohl et al. (2008) also found that offenders who "returned to prison were young, single, and more likely to commit non-violent crimes". They also found that offenders released from prison had dense, length criminal histories.

2.2.2. Paroled vs. Un-conditional release.

Kohl et al. (2008) found that paroled inmates were 45 percent of the recidivism reported while those offenders who left prison via expiration of sentence recidivated at a rate of 36 percent.

2.2.3. Number of Prior Convictions

According to Kohl et al. (2008), the odds of reoffending swell with the number of prior convictions. Thus, "the number of times a prisoner has been arrested in the past is a good predictor of whether that prisoner will continue to commit crimes after being released" (Langan & Levin, 2002, p.10). In addition to the number of times one has been incarcerated, previous criminal activity can include deviant behavior that does not result in an individual's incarceration. This behavior can include activities an individual engages in on a regular basis, such as having ties to a gang or living in a crime-ridden area as these factors may pressure individuals into engaging in criminal behavior.

2.3. Demographic Characteristics

2.3.1. Age

Francis J. Carney (1967) found that age was the most powerful variable in terms of discriminating between recidivists and non-recidivists. It was discovered that slightly over half of the recidivists (51.0%) were twenty-five or younger at the time of their present commitment, while only about one-third of the non-recidivists (33.0%) fell into this age range. In addition, it was found that about one out of five non-recidivists (20.7%) were forty-five or older, while only one out of fifty of the recidivists (2.0%) were in this category. It was further shown that the mean age of recidivists (26.9) was significantly lower than that of non-recidivists (33.6). Such a difference is so striking that the probability of it occurring by chance is less than one in a thousand.

May et al. (2008) state that recidivism carries an inverse relationship with age. This relationship indicates that recidivism decreases as age increases. May et al. found that "reoffending...was the highest for those aged from 18 to 20, and lowest for those aged 40 and over". Langan and Levin (2002) also found that older prisoners are much less likely to recidivate.than younger prisoners.

2.2.4. Drug Use

Drug use is an important factor related to recidivism due to the very nature of the drug trade itself. Illegal drugs are characterized by violence because individuals often have to resort to violence to sell, receive payment, and keep individuals from alerting law enforcement. When the end itself is illegal, the means to attain are often illegal as well. May et al. (2008) re-affirms that those who report a difficulty with drugs before custody were much more likely to reoffend. This study found that three-quarters of the participants who recidivated within a year of their release had reported a problem with drugs before their incarceration.

2.2.5. Previous Recidivism Studies.

Anna Ferrante, Nini Loh & Max Maller (1999) were contracted by the Ministry of Justice, Australia to calculate the recidivism estimates of offenders attending the KOP program (consists of a series of three-hour sessions with groups of approximately ten Aboriginal offenders who are either serving community based orders, serving prison sentences, or on parole). Their aim was to;

i. To estimate the probability of re-offending of the KOP offenders

ii. Calculate the median time to re-offending and

iii. To estimate recidivism probabilities and time-to-fail measure of offenders who had not undertaken the KOP program.

A quasi-experimental design was used in this study. Information about the offending and re-offending patterns of offenders was obtained from the Recidivism database maintained by the Crime Research Centre. For this study, probabilities of re-arrest have been estimated from a parametric statistical model fitted to the observed follow-up times of the specific offender group(s) under review. The Weibull mixture model is described by various parameters including p, a parameter representing the probability of ultimate or long-term failure, λ (lambda) which is related to the rate of failure, and α (alpha) which describes the "shape" of the Weibull curve. The Weibull model used here also incorporates covariates so that differences between sub-groups of the population under analysis can be tested. The covariates sex, race, prior arrests, age, offence-type and completion-status are adopted. Offenders who completed the KOP program successfully were found to have the same ultimate probability of re-arrest (0.93) than those who did not complete the program. However, the rate of re-arrest was significantly higher for the unsuccessful program attendees than for the rest: median time to re-arrest for unsuccessful completions was 0.9 years, compared with 1.5 years for those successfully completing the program.

2.3 Cox's Proportional Hazard Model

Jiayi Ni (2008) applied the Cox's proportional hazard model to the stock exchange market by fitting the model to stock data in the Shanghai Security Market. By September of that year, the Shanghai stock exchange's benchmark index had plunged 64% since the start of the year, reaching a 52-week low and crashing past the 2000 points barrier. Though it was an overall crash of a stock market, differences still existed among individual stocks. Some experienced wild ups and downs in price, while others rapidly fell down, almost straight down to half of their highest prices. Her paper therefore aimed at finding out what the main factors were that influenced price performances of quoted companies, and what kind of companies were more likely to survive this meltdown. Dismissing the macro factors such as a change of the stamp tax on stock trading and macro economy regulation and control, it focused on the financial data of each individual stock. The time origin was defined as the date when a stock's price reached its highest point in that year. The end-point is the date its price dropped to below 40% of that price for the first time. The number of days between these two dates was then the survival time of a stock. The length of this study was 8 months, from Jan 1 to Aug 31 in 2008. Data were collected from stocks in the SSE 50 Index. The index selected the 50 largest stocks of good liquidity and representativeness in the Shanghai security market. Referring to the semi-annual report of each company, 6 factors were considered at first, including earning per share (EPS), net asset per share (NAPS), cash flow per share (CFPS), return on equity (ROE), growth rate of operating profit (GROP), and the percentage of released non- floating shares (RNF) by the end of study. Also stocks are divided into 14 sectors by industry. The Communication Device sector was selected as the reference for the sectors and one dummy variable was created for each of the other 13 sectors. The Pearson Correlation Coefficients matrix was then calculated to ascertain which two factors are highly correlated, and then only one of them was to be included in the model to reduce the multicollinearity among the factors and led to the drop of EPS. The regression model showed that release of more and more non-floating shares is not a main cause of the nose-diving in stock market, and return on equity also does not affect the stock price a lot, while NAPS, CFPS and GROP are positively related to stock survival times, that is, it is the companies' good financial condition, high liquidity and growth rate of earning capacity that make their stocks survive longer in the market.

Mohamad Amin Pourhoseingholi et al (2007) compared two survival regression methods – Cox regression and parametric models - in patients with gastric adenocarcinomas who registered at Taleghani hospital, Tehran. They studied 746 cases from February 2003 through January 2007. Gender, age at diagnosis, family history of cancer, tumor size and pathologic distant of metastasis were selected as potential prognostic factors and entered into the parametric and semi parametric models. Weibull, exponential and lognormal regression were performed as parametric models with the Akaike Information Criterion (AIC) to compare the efficiency of models. They begin by describing the Cox's semi-parametric model as one where the baseline hazard takes no particular form and then they link it to parametric survival models through alternative functions for the baseline hazard, that is, by letting the baseline hazard be a parametric form such as Weibull, Gompertz, Exponential, and Lognormal. The aim of this study was to investigate the comparative performance of Cox and parametric models in a survival analysis of patients with gastric carcinoma using the Akaike Information Criterion (AIC). The survival results from both Cox and Parametric models showed that patients who were older than 45 years at diagnosis had an increased risk for death, followed by greater tumor size and presence of pathologic distant metastasis. In multivariate analysis Cox and Exponential were found to be similar. Although it seemed like there may not have been a single model that is substantially better than others, in univariate analysis the data strongly supported the log normal regression among parametric models which could be an alternative to Cox.

Brian D. Bunday and Victor A. Kiri (1992) in their study titled 'Analysis of Censored Recidivism Data Using a Proportional Hazard-Type Model'; set to find out the effects of individual characteristics and correctional factors on the future criminal behaviour of offenders. They looked at the possibility that some of these effects may diminish significantly with time and described how the simple proportional hazards model can be adapted to account for such transient effects. Their aim was to determine the extent to which the propensity for recidivism of an offender is attributable to his/her individual characteristics. They included a transience parameter in the model that recognized that all the explanatory factors will cease to be effective on the hazard, irrespective of the strength of their individual levels of significance, once the threshold period has elapsed. The data for this study concern 307 males, born in 1953 in England and Wales, who were followed up during the period 1962-1981. Censoring of the time to failure arises for those individuals not reconvicted by 1981, at which time the study ended.

They considered the variables age of the individual at his first conviction and the type of sentence passed for this conviction-fine, probation or some other non-custodial correction, or custodial. The number of crimes taken into account when passing this first sentence was also recorded and they dichotomized this variable to take the values 0 or 1 depending on whether the number of such crimes was 1 or more than 1. For the dichotomous variables they divided all individuals into two groups, they thus had a two-sample problem and the log-rank test allowed them to test whether the survivor functions for the two groups differ. This analysis was carried out before the estimation of the β parameter for that variable in the proportional hazards model. The first analysis involved the use of the simple hazard proportion model and the validation of the model was checked using a residual plot and found to be good. Unfortunately, the data lack detail on some of the individual characteristics found elsewhere to have a significant influence on the risk of recidivism; namely, race, employment and marital status, as well as records on alcohol and hard drugs.

The 2009 Florida Prison Recidivism Study report produced by the department of corrections of the Florida state looks at releases from 2001 to 2008 to answer questions such as, what is the likelihood that an inmate who is released today will come back to prison?, what factors influence recidivism rates? Do age, gender, and racial groups show differences in recidivism rates? So as to identify groups most likely to fail when they are released and consequently determine where to devote scarce correctional and community resources. For this report, recidivism is defined as a return to prison, either because of a new conviction or because of a violation of post prison supervision. The follow-up periods (typically reported as three years) are calculated from prison release date to the date of readmission to prison. The basic rates for tables and graphs are computed from Kaplan-Meier estimates of the survival curve using right-censored data. The analyses of factor significance are conducted using Cox models (proportional hazards regression) of the same data. The analysis used a 5% level of significance and, to determine the factors in order of importance, a stepwise selection routine for determining which factors to include. Various groups are compared in terms of their recidivism rates (i.e. male vs. female, various violent crimes, non-violent crimes, and age's e.t.c). Female inmates recidivate at a much lower rate than male inmates, among inmates who were in prison for violent offenses, those in for murder or manslaughter have the lowest recidivism rates, among inmates who were in prison for non-violent offenses, those in for weapons offenses have the lowest recidivism rates whereas burglars released during this period have the highest recidivism rates and the older an inmate is at time of release, the less likely he is to return to prison. When these factors are put in a Cox regression model, it is found that a male inmate is 50.6% more likely to fail than a female inmate with all other factors held constant, each additional disciplinary report that an inmate incurs while incarcerated, increases his likelihood of recidivating by 1.0% .

Summary

Previous criminal history and specific demographic characteristics such as age, sex, and race are continually associated with the risk of continued offending as well as previous drug use and the sentence length. Most offenders who go on to commit additional crime after serving time in prison have significant criminal histories, are of a specific age, race, education level and sex, and have a noted substance abuse problem. These variables I will adopt in this research project as well because of their proven significance to ascertain if their influence will be consistent with the Kenyan situation.

The methodological developments of survival analysis that have had the most profound impact in the study of recidivism are; the Kaplan-Meier method for estimating the survival function, the log-rank test for comparing the equality of two or more survival distributions for the different categories of the same variable and the Cox proportional hazards (PH) model for examining the covariate effects on the hazard function for survival data with the presence of censoring.

Unlike the other studies reviewed above, I use univariate analysis to check all the risk factors before proceeding to the full multivariate Cox PH model including all the potential risk factors and treatment arms. The Wald test and the Likelihood test are considered in each univariate Cox PH model.

After a Cox PH model is fitted, the adequacy of this model, including the PH assumption and the goodness of fit, need to be assessed. The PH assumption checking with graphical method and a statistical test method based on the Schoenfeld residuals has been described. Finally, I assess goodness of fit by residual plots.

Chapter Three

Cox regression model

3.1 Introduction

In survival models, the hazard function for a given individual describes the instantaneous risk of experiencing an event of interest within an infinitesimal interval of time, given that the individual has not yet experienced that event. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of explanatory variables, or covariates, but keeps the baseline hazard as an arbitrary, unspecified, nonnegative functional of time.

Definition 3.1.1; The Cox Proportional Hazards model is given by

$$h(t|x) = h_0(t)exp\ (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = h_0(t)exp\ (\boldsymbol{\beta}'\boldsymbol{x})$$

where $h_0(t)$ is called the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero, $\boldsymbol{x} = (x_1, x_2, x_3, \dots x_p)$ is the values of the vector of explanatory variables for a particular individual, and $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3, \dots, \beta_p)$ is a vector of regression coefficients.

The corresponding survival functions are related as follows:

$$S(t|x) = S_0(t)e^{\sum_{i=1}^{p}(\beta_i x_i)}$$

This model, also known as the Cox regression model, makes no assumptions about the form of $h_0(t)$ (non-parametric part of model) but assumes parametric form for the effect of the predictors on the hazard (parametric part of model). The model is therefore referred to as a semi-parametric model.

The beauty of the Cox approach is that this vagueness creates no problems for estimation. Even though the baseline hazard is not specified, we can still get a good estimate for regression coefficients $\boldsymbol{\beta}$, and hazard ratio.

The measure of effect is called hazard ratio. The hazard ratio of two individuals with different covariates X and X^* is

$$\widehat{HR} = \frac{h_0(t)exp\ (\beta' x)}{h_0(t)exp\ (\beta' x^*)} = \exp\ (\sum \beta'\ (x - x^*)$$

This hazard ratio is time-independent, which is why this is called the proportional hazards model.

3.2 Partial likelihood estimate for unique failure times

Fitting the Cox proportional hazards model, we wish to estimate $\boldsymbol{\beta}$. One approach is to attempt to maximize the likelihood function for the observed data with respect to $\boldsymbol{\beta}$. A more popular approach is proposed by Cox [3] in which a partial likelihood function that does not depend on $h_0(t)$ is obtained for $\boldsymbol{\beta}$. Partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameters ($h_0(t)$ in the Cox PH model). In this section, we will construct the partial likelihood function based on the proportional hazards model.

Let $t_1, t_2, \dots, t_n$ be the observed survival time for n individuals. Let the ordered failure time of r individuals be $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ and let $R\,(t_{(j)})$ be the risk set just before $t_{(j)}$. So that $R\,(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The conditional probability that the i^{th} individual fails at $t_{(j)}$ given that one individual from the risk set on $R\,(t_{(j)})$ fails at $t_{(j)}$ is;

P (individual i fails at $t_{(j)}$| one failure from the risk set $R\left(t_{(j)}\right)$ at $t_{(j)}$)

$$= \frac{h_i(t_{(j)})}{\sum_{k\in R(t_{(j)})} h_k(t_{(j)})}$$

$$= \frac{h_0(t_{(j)})\exp\left(\beta' x_i(t_{(j)})\right)}{\sum_{k\in R(t_{(j)})} h_0(t_{(j)})\exp\left(\beta' x_k(t_{(j)})\right)}$$

$$= \frac{\exp\left(\beta' x_i(t_{(j)})\right)}{\sum_{k\in R(t_{(j)})} \exp\left(\beta' x_k(t_{(j)})\right)}$$

Then the partial likelihood function for the Cox PH model is given by;

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp\left(\beta' x_i(t_{(j)})\right)}{\sum_{k\in R(t_{(j)})} \exp\left(\beta' x_k(t_{(j)})\right)}$$

in which $x_i(t_{(j)})$ is the vector of covariate values for individual i who dies at $t_{(J)}$. Cox (1972) discussed the general method of partial likelihood. Note that this likelihood function is only for the uncensored individuals. Let $t_1, t_2, \dots, t_n$ be the observed survival time for n individuals and δ_i be the event indicator, which is zero if the i^{th} survival time is censored, and unity otherwise. The likelihood function in equation (3.1) can be expressed by

$$L(\beta) = \left[\prod_{i=1}^{r} \frac{exp\,(\beta' x_i(t_i))}{\sum_{k \in R(t_i)} exp\,(\beta' x_k(t_i)}\right|^{\delta_i}$$

where $R\,(t_i)$ is the risk set at time . The partial likelihood is valid when there are no ties in the dataset. That means there are no two subjects who have the same event time.

3.3 Partial likelihood for repeated failure times

The case when two or more individuals are recorded as failing at the same time is more complex. Here I discuss three of the methods commonly referred to in the reviewed literatures. In all I will adopt the following notation;

- $t_{(i)}$ is the i^{th} ordered unique failure time (so if four failures occur at times 1, 1, 3, 3, t(1) = 1 and t(2) = 3);
- I is the total number of unique failure times;
- D (t) is the set of individuals who fail at time t.

Breslow's Method

$$l_p(\beta, x) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\left(\sum_{j \in R(t_{(i)})} \phi_j\right)^{|D(t_{(i)})|}}$$

Note that $|D(t_{(i)})|$ is the number of individuals that fail at time $t_{(i)}$.

Breslow's method is the default for many statistical packages. But it is not the default for R. R uses Efron's partial likelihood, as it is considered a closer approximation to the exact partial likelihood.

Efron's Method

For the case when two or more individuals are recorded as failing at the same time then Efron's method [4] gives:

$$L(\beta) = \prod_{j=1}^{r} \frac{\prod_{j\in D(t_j)} \exp(\beta' X_j)}{\prod_{k=1}^{D(t_j)} \left[\sum_{j\in R(t_j)} \exp(\beta' X_j) - \frac{k-1}{D(t_j)} \sum_{j\in D(t_j)} \exp\left(\beta' X_j\right)\right]}$$

Which is the default for R and $\boldsymbol{D(t_j)}$ is the set of individuals who fail at time $\boldsymbol{t_j}$.

Exact Method

The exact method computes the "probability" that 'A' evented first followed by 'B' OR 'B' evented first followed by 'A.' Since both 'A' and 'B' had the same observed survival times, we assume that each sequence (permutation) is equally likely:

$$P(A,B) = \left[\frac{h(A)}{h(A)+h(B)+h(C)}\right] \cdot \left[\frac{h(B)}{h(B)+h(C)}\right]$$

$$P(B,A) = \left[\frac{h(B)}{h(A)+h(B)+h(C)}\right] \cdot \left[\frac{h(A)}{h(A)+h(C)}\right]$$

$$L_1 = \left[\frac{h(A)}{h(A)+h(B)+h(C)}\right] \cdot \left[\frac{h(A)}{h(B)+h(C)}\right] + \left[\frac{h(B)}{h(A)+h(B)+h(C)}\right] \cdot \left[\frac{h(A)}{h(A)+h(C)}\right]$$

The covariates are related to the hazard function in the usual way, e.g. $\boldsymbol{h(A)} = \boldsymbol{exp\{\beta_1 * x_A\}}$ where $\boldsymbol{x_A}$ is same variable of interest such as age. The likelihood function would relate the outcomes to the unknown parameters such as$\boldsymbol{\beta_1}$. If there are a total of $\boldsymbol{n_{T_l}}!$ tied event times for a particular time,$\boldsymbol{t_i}$, then the exact method must evaluate $\boldsymbol{n_{T_l}}!$ Separate permutations of possible outcomes. Therefore, the likelihood function can become extremely complicated (very quickly) as $\boldsymbol{n_{T_l}}!$ increases.

You can see that the exact method quickly becomes computationally intensive when there are large numbers of ties.

That is if we let $\phi_i = exp(\beta' X)$, then suppose individuals labeled 1–5 are at risk at time t_i i.e. in $R(t_i)$and that of these, individuals 1–3 fail at time t_i.Then Breslow's method gives as the contribution from time·t(i) to the partial likelihood

$$\frac{\phi_1\phi_2\phi_3}{(\phi_1+\phi_2+\phi_3+\phi_4+\phi_5)^3}$$

The exact method gives;

$$\frac{\phi_1\phi_2\phi_3}{\phi_1\phi_2\phi_3 + \phi_1\phi_2\phi_4 + \phi_1\phi_2\phi_5 \ldots + \phi_3\phi_4\phi_5}$$

Then Efron's method gives;

$$\frac{\phi_1\phi_2\phi_3}{(\phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5)\left(\frac{2}{3}\phi_1 + \frac{2}{3}\phi_2 + \frac{2}{3}\phi_3 + \phi_4 + \phi_5\right)\left(\frac{1}{3}\phi_1 + \frac{1}{3}\phi_2 + \frac{1}{3}\phi_3 + \phi_4 + \phi_5\right)}$$

Note that in the absence of ties, all three reduce to the no-ties partial likelihood.

3.4 Assumptions

There are a few assumptions about Cox proportional hazard model

1. The proportional hazards models assume that the hazard ratio of two people is independent of time and it is valid only for time independent covariates. This means that the hazard functions for any two individuals at any point in time are proportional. In other words, if an individual has a risk of death at some initial time point that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.
2. Each study group has a hazard function that is a positive multiple of the baseline hazard, r × $h_0(t)$.
3. Explanatory variables act only on the hazard ratio r. They do not affect the baseline hazard.
4. Independence of observations
5. Sufficient data for inference
6. Censoring is independent of the event of interest.

3.5 Proportional hazard assumption checking

The main assumption of the Cox proportional hazards model is proportional hazards. Proportional hazards means that the hazard function of one individual is proportional to the hazard function of the other individual, i.e., the hazard ratio is constant over time. There are several methods for verifying that a model satisfies the assumption of proportionality.

3.5.1 Graphical method

We can obtain Cox PH survival function by the relationship between hazard function and survival function

$$S(t,x) = S_0(t)\exp\left(\sum_{i=1}^{p} \beta_i x_i\right)$$

Where $\boldsymbol{x} = (x_1, x_2, \dots, x_p)'$ are the values of the vector of explanatory variables for a particular individual? When taking the logarithm twice, we can easily get

$$ln[-lnS(t,x)] = \sum_{i=1}^{p} \beta_i x_i + ln[-lnS_0(t)]$$

Then the difference in log-log curves corresponding to two different individuals with variables $x_1 = (x_{11}, x_{12}, \dots, x_{1p})$ and $x_2 = (x_{21}, x_{22}, \dots, x_{2p})$ is given by

$$ln[-lnS(t,x_1)] - ln[-lnS(t,x_2)] = \sum_{i=1}^{p} \beta_i (x_{1i} - x_{2i})$$

This does not depend on t. This relationship helps us to identify situations where we may have proportional hazards. By plotting estimated log (-log (survival)) versus survival time for two groups we would see parallel curves if the hazards are proportional.

This method does not work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Furthermore, the curves are sparse when there are few time points and it may be difficult to tell how close to parallel is close enough.

3.5.2 Tests based on the Schoenfeld residuals.

The other statistical test of the proportional hazards assumption is based on the Schoenfeld residual [4]. The Schoenfeld residuals are defined for each subject who is observed to fail. If the PH assumption holds for a particular covariate then the Schoenfeld residual for that covariate will not be related to survival time.

Schoenfeld residuals give us the difference between the covariate value observed at a failure time, and the weighted expected value of the covariate, for those observations still in the risk set. If a covariate's effect is unaffected by t, then this difference should be 0 at all failure times.

Assume there are p covariates and n independent observations of time, covariates, and censoring, which are represented as (t_i, x_i, c_i), where i = 1, 2... n, and c_i = 1 for uncensored observations and zero otherwise.

From Hosmer and Lemeshow (1999), Schoenfeld residuals are:

$$x_{ik} - \sum_{j \in R(t_i)} x_{jk} p_j$$

This gives the difference in the covariate value at t minus a weighted average of the covariate (weighted by the probability of failure, which is derived from the partial likelihood estimator). For example;

At time=7, suppose we have 5 cases and two variables, x_1, x_2: x= (55, 0); (45, 1); (67, 0); (58, 1) ;(70, 1).

Suppose x_{1j} = (55, 0) fails at time=7. Imagine the probability of failure at time=7 (derived from partial likelihood) for each case is: 0.10, 0.05, 0.30, 0.20, 0.30.Schoenfeld residual for x_1:55 − {55(.10) + 45(.05) + 67(.30) + 58(.20) + 70(.30)} =55 − 60 = −5

Schoenfeld residual for x_2:0− {0(.10) +1(.05) +0(.30) +1(.20) +1(.30)} =0−.55 =−.55

So this test is accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual survival times. The null hypothesis is that the correlation between the Schoenfeld residuals and the ranked survival time is zero. Rejection of null hypothesis concludes that PH assumption is violated.

In R, tests for the proportional-hazards assumption are obtained from **cox.zph**, which computes a test for each covariate, along with a global test for the model as a whole:

Plotting the object returned by cox.zph produces graphs of the scaled Schoenfeld residuals against transformed time. Systematic departures from a horizontal line are indicative of non-proportional hazards

3.6 Cox proportional hazards model diagnostics

After a model has been fitted, the adequacy of the fitted model needs to be assessed. The model checking procedure below is based on residuals. In linear regression methods, residuals are defined as the difference between the observed and predicted values of the dependent variable. However, when censored observations are present and partial likelihood function is used in the Cox PH model, the usual concept of residual is not applicable. A number of residuals have been proposed for use in connection with the Cox PH model. I will only describe the Schoenfeld residual.

3.6.1 Schoenfeld residuals

The Schoenfeld residuals were originally called partial residuals because the Schoenfeld residual for i^{th} individual on the j^{th} explanatory variable X_j is an estimate of the i^{th} component of the first derivative of the logarithm of the partial likelihood function with respect to β_j. From equation (3.2), this logarithm of the partial likelihood function is given by;

$$\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{p} \delta_i \left(x_{ij} - a_{ij}\right)$$

where x_{ij} is the value of the j^{th} explanatory variable $j = 1, 2, \dots, p$ for the i^{th} individual and

$$a_{ij} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\beta' x_l\right)}{\sum_{l \in R(t_i)} \exp\left(\beta' x_l\right)}$$

The Schoenfeld residual for i^{th} individual on X_j is given by $r_{p_{ji}}$=$\delta_i\left(x_{ij} - a_{ij}\right)$. The Schoenfeld residuals sum to zero.

3.7 Strategies for analysis of non-proportional data

3.7.1 Stratification

A situation that sometimes occurs is that hazards are not proportional on an overall basis, but that they are proportional in different subgroups of the data. In problems of this kind, it may be assumed offenders in each of the subgroups, or strata, have a different baseline hazard function, but that all other explanatory variables satisfy the proportional hazards assumption within each stratum. Suppose that the offenders in the j^{th} stratum have a baseline hazard function h_{0j}(t), for j= 1, 2... g, where g is the number of strata. The effect of explanatory variables on the hazard function can then be represented by a proportional hazard model for h_{ij}(t), the hazard function for the i^{th} individual in the j^{th} stratum, where $i = 1, 2, \dots, n_j$, say, and n_j is the number of individuals in the j^{th} stratum. We then have the stratified proportional hazard model, according to which

$$h_{ij}(t) = exp(\beta' x_{ij})h_{0j}(t)$$

Where x_{ij} is the vector of values of p explanatory variables, $X_1, X_2, \dots, X_p$, recorded on the ith individual in the jth stratum.

3.5.2 Interaction

Another way of accommodating non-proportional hazards is to build interactions between covariates into the Cox regression model; such interactions are themselves considered as covariates.

3.5.3 Hypothesis testing

There are three tests that are commonly used to test the hypothesis that a covariate has no effect. These are the Wald test, the score test and the likelihood ratio test. All test

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

for the model

$$h(t|x) = h_0(t)e^{X\beta'}$$

A likelihood ratio comparison: contrasting two log-likelihood values effectively identifies systematic differences between two survival time models. First, a log-likelihood value is estimated under the condition that $\beta = 0$ (h_0(t) = h_1(t)) and then under the condition that$\beta \neq 0$. The comparison of the log-likelihood values likely reflects any important difference in survival time between the compared groups (significance).

The likelihood ratio test-statistic,

X_C^2= −2 [log ($L_{\beta=0}$) −log ($L_{\beta\neq0}$)]

Where$L_{\beta=0}$ *is the log partial likelihood,*

has an approximate chi-square distribution with one degree of freedom when $\beta = 0$. The p-value is P ($X^2 \geq X_C^2 | \ \beta = 0$).

Alternatively, a Wald test statistic defined as;

$$X_{Wald}^2 = \left|\frac{\beta}{S_\beta}\right|^2$$

Has an approximate chi-square distribution with 1 degree of freedom when $h_i(t) = h_0(t) or\ \beta = 0$. The associated p value is $P(X^2 \geq X^2_{wald} | \beta = 0)$.

Fitting the Cox's PH Model using the R statistical package

The Cox proportional-hazards regression model is fit in R with the *coxph* function (located in the survival

Library in R) and takes the form

Coxph (Surv (time, event) ~age+sex, method=c ("efron","breslow","exact"), data=)

The right-hand side of the model formula for coxph is the same as for a linear model. The left-hand side is a survival object, created by the Surv function. In the simple case of right-censored data, the call to Surv takes the form Surv (time, event), where time is either the event time or the censoring time, and event is a dummy variable coded 1 if the event is observed or 0 if the observation is censored. Among the remaining arguments to coxph: method indicates how to handle observations that have tied (i.e., identical) survival times. The default "efron" method is generally preferred to the once-popular "breslow" method; the "exact" method is much more computationally intensive and data, a data frame in which to interpret the variables named in the formula.

Chapter Four

Data Analysis and Interpretation.

4.1 Introduction

In this chapter, the description of the sample understudy is presented first, followed by the description of the variables understudy, the descriptive analysis, and then the fitting of the model. The survival function estimate is reported first, then the differences within the variables understudy are examined using the log-rank test and the results are accompanied by graphs showing how the estimates of the survival functions of the various categories of the same variable behave. Next I examine the univariate Cox's proportional hazard model and then a summary of the multivariate model results concludes the chapter.

4.2 Description of data set

4.2.1 Study population and objective

The following data is derived from Industrial Area prison in Nairobi. A sample of 402 male offenders released in January 2003 has been observed. For each the attributes of interest have been recorded and tabulated as shown below;

Week	*arrest*	*fin*	*age*	*residence*	*wexp*	*mar*	*paro*	*prio*	*educ*
20	*1*	*0*	*38*	*1*	*0*	*0*	*1*	*3*	*3*
17	*1*	*0*	*22*	*1*	*0*	*0*	*1*	*8*	*4*
25	*1*	*0*	*19*	*0*	*1*	*0*	*1*	*13*	*3*
52	*0*	*1*	*51*	*1*	*1*	*1*	*1*	*1*	*5*
52	*0*	*0*	*19*	*0*	*1*	*0*	*1*	*3*	*3*

Thus, for example, the first individual was arrested in week 20 of the study, while the fourth individual was never rearrested, and hence has a censoring time of 52.

4.2.3 Description of variables

The variables used in this study come from the variable names used by Allison (1995), from whom these variable descriptions are adapted. These variables are used as independent and dependent variables in this research and represent concepts outlined in the literature review. The variable week is the dependent variable representing time to re-arrest from the beginning of the study, arrest is an indicator variable for those arrested and those who were not and the variables age, residential dwelling, work experience, marital status, education level attained, number of prior offences and format of release are the independent variables

The variables and codes for this data are as in the following table:

Variables	Description	Codes/Values
Week	week of first arrest after release	
Arrest	the event indicator	1= arrested, 0=not arrested
Age	age at the time of release	in years
Res	a categorical variable	1= slum, 0=others
Wexp	work experience prior to incarceration	1= yes, 0=no
Mar	marriage status at time of release	1= yes, 0=no
Paro	individual was released on parole	1= yes, 0=no
Prio	number of prior convictions	
Educ	education level	2=pri,3=sec,4=skill,5=college,6=univ

4.3 Statistical analysis and results

4.3.1 Descriptive and non-parametric analysis

Four hundred and two offenders released in the months of December 2002 and January 2003 are considered and the period of study ran up to December of 2003. Among these offenders the mean age is 24.74 years (SD = 6.043), with the youngest being 17 years and the oldest 44 years. The mean number of prior offences is 1.642 (SD=0.827). 105 arrests are made during the period of study and so the recidivism rate for the entire sample is 0.26 (105 out of 402). Of the sample considered here, 295 come from slum dwellings and represent 73 percent of all offenders under study. 221 out of the 402 have some form of work experience before release and 117 are married. The number of those who did not serve their full sentences or were released on parole is 250(62% of those released). Average time of opportunity to reoffend is 36.41 weeks (SD = 15.15).

I first examined the univariate relationship between recidivism, age, work experience, the number of prior offences, education level, and the release formula (paroled, released). For the dichotomous recidivism variable, arrest was coded as one and non-arrest as zero. These coding formats I used in all analyses. To examine the magnitude of the differences between recidivists and non-recidivists based on each variable, log-rank test are considered. But first I plot the K-M estimate of the survival function; in R the function survfit () is used to find the Kaplan-Meier estimate of the survival function. There are three arguments of particular interest: formula, conf.int, and conf.type. Formula will be a survival object and it is the only required input:

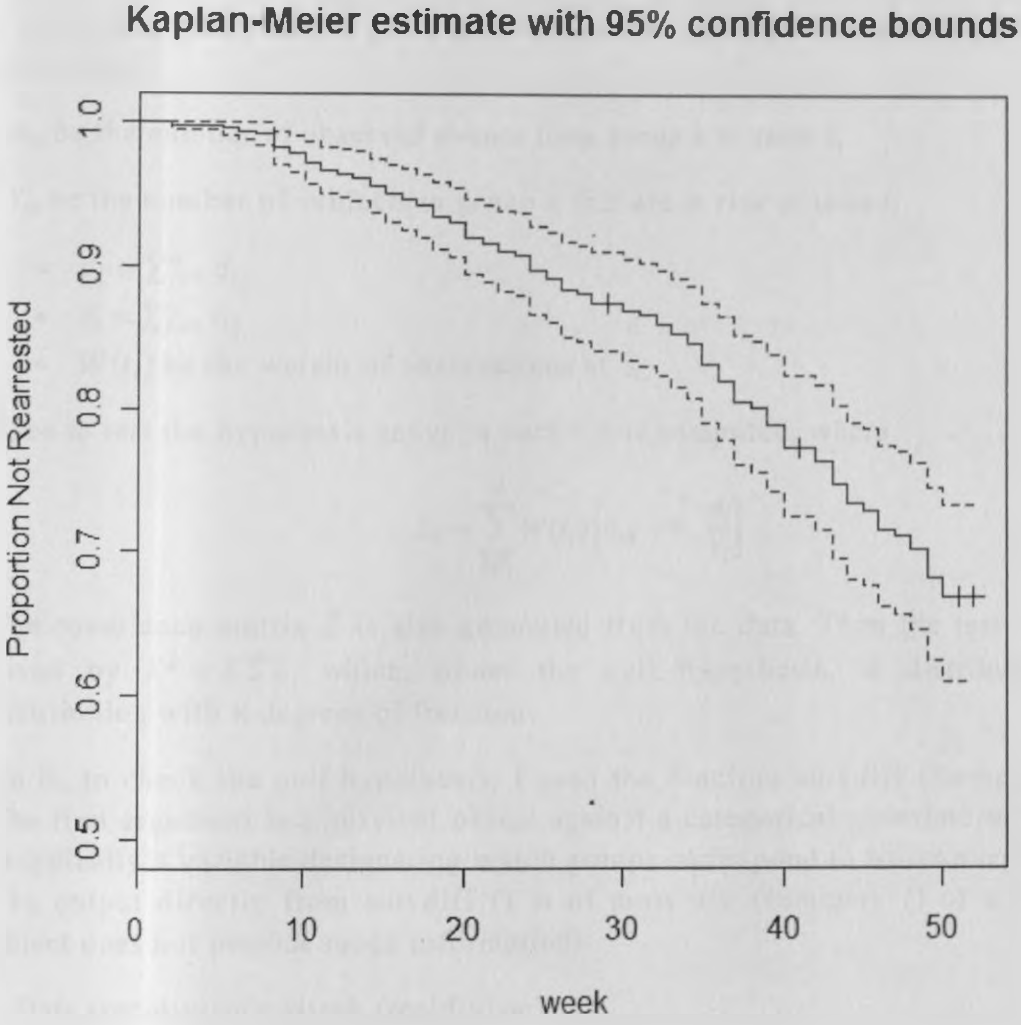


Figure 1: Estimated survival function S(t) for the Cox regression of time to rearrest on several predictors. The broken lines show a point-wise 95-percent confidence envelope around the survival function.

Next, given two or more samples, is there a difference between the survival times? Setting up hypotheses for this problem,

Let

- t_i be times where events are observed (assume these are ordered and there are D such times),
- d_{ik} be the number of observed events from group k at time t_i
- Y_{ik} be the number of subjects in group k that are at risk at time t_i
    - $d_i = \sum_{j=1}^{n} d_{ij}$
    - $Y_i = \sum_{j=1}^{n} Y_{ij}$
    - $W(t_i)$ be the weight of observations at t_i

Then to test the hypothesis above, a vector Z is computed, where

$$Z_k = \sum_{i=1}^{D} W(t_i)\left[d_{ik} - Y_{ik}\frac{d_i}{Y_i}\right]$$

The covariance matrix $\hat{\Sigma}$ is also computed from the data. Then the test statistic is given by $X^2 = Z'\hat{\Sigma} Z$, which, under the null hypothesis, is distributed as χ^2 distribution with n degrees of freedom.

In R, to check the null hypothesis, I used the function survdiff (formula, rho=0). The first argument is a survival object against a categorical covariate variable that is typically a variable designating which groups correspond to which survival times. The output directly from survdiff () is of most use (summary () of a survdiff () object does not provide much information).

```
> Data (recidivism); attach (recidivism)
```

```
> Survdiff (Surv (week, arrest) ~ age) # output omitted
```

The second argument shown, rho, the default is rho=0, which corresponds to the log-rank test.

Comparing survival times for different Residential areas

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Res=0	100	16	35.5	10.70	16.7
Res=1	301	88	68.5	5.54	16.7

Chisq= 16.7 on 1 degrees of freedom, p= 0.029

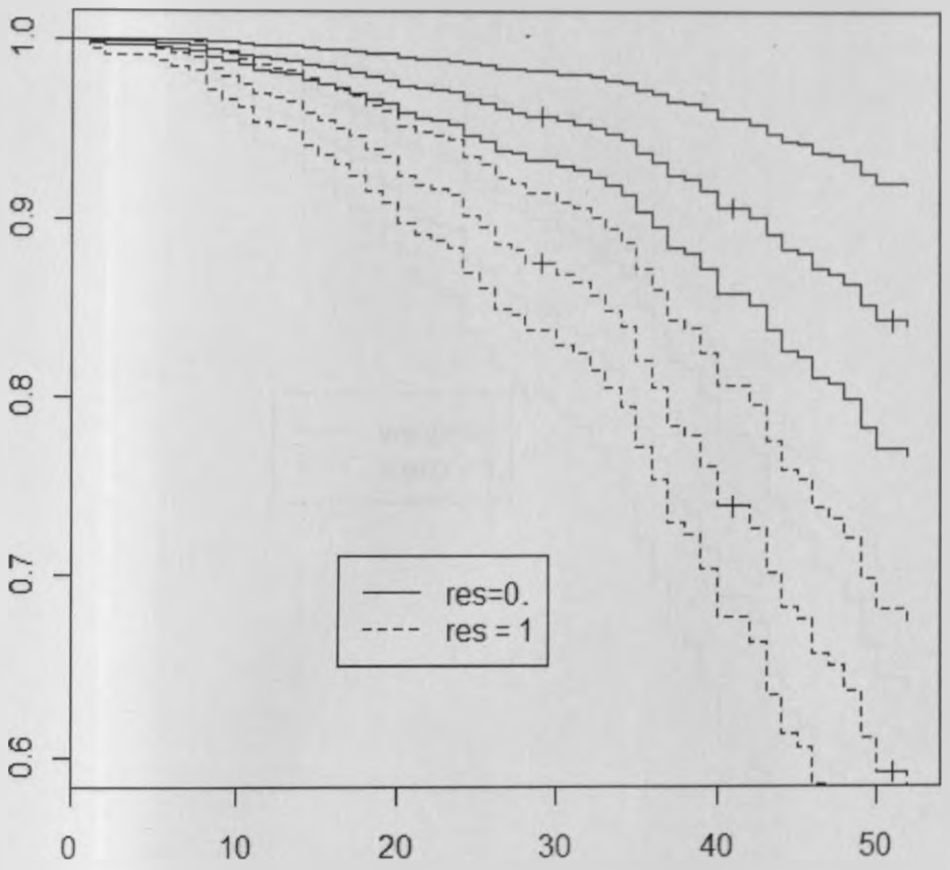


Figure 2: Estimated survival functions for those residing in slums (res = 1) and others (res = 0). Other covariates are fixed at their average values. Each estimate is accompanied by a point-wise 95-percent confidence envelope.

Work Experience

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Wexp=0	186	65	36.4	22.4	35.4
Wexp=1	215	39	67.6	12.1	35.4

Chisq= 35.4 on 1 degrees of freedom, p= 2.74e-09

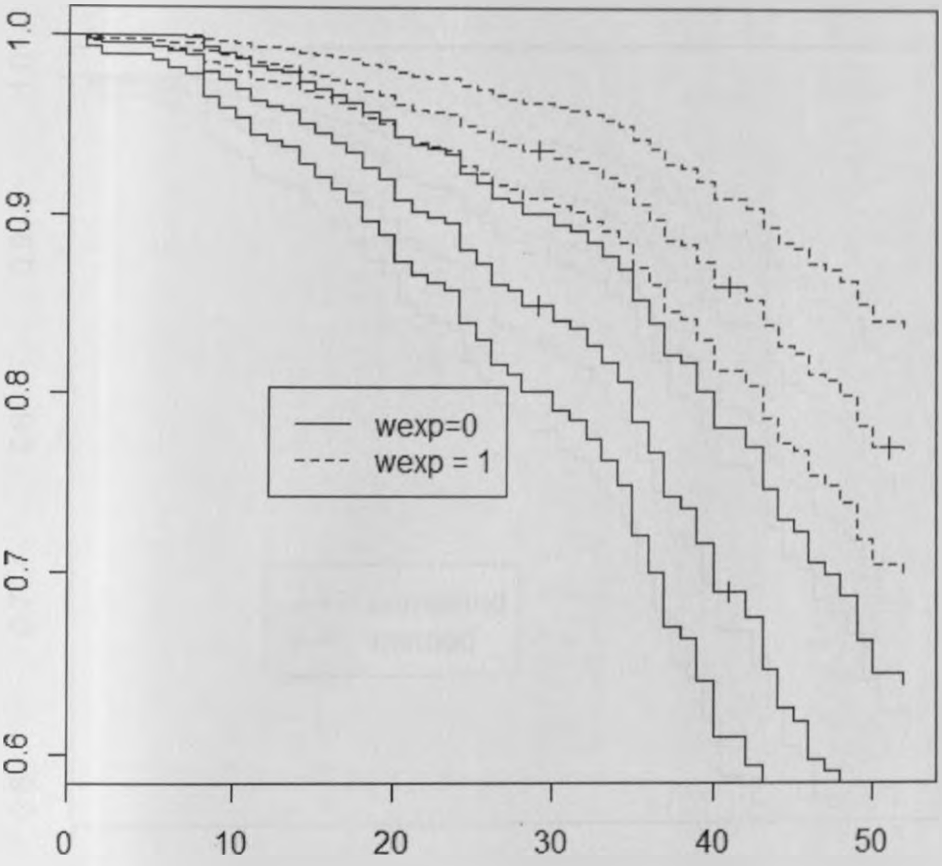


Figure 3: Estimated survival functions for those with work experience (wexp = 1) and those without (wexp= 0). Other covariates are fixed at their average values. Each estimate is accompanied by a point-wise 95-percent confidence envelope.

Marriage Status

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Mar=0	331	92	81.9	1.23	5.88
Mar=1	70	12	22.1	4.58	5.88

Chisq= 5.9 on 1 degrees of freedom, p= 0.0153

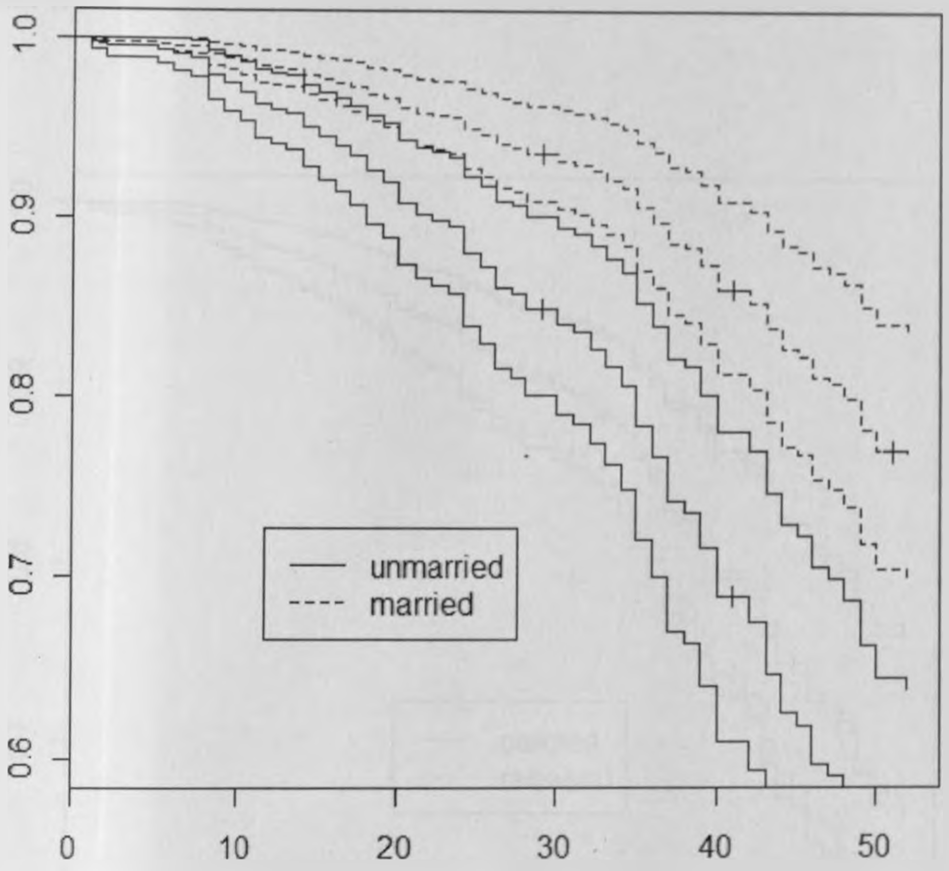


Figure 4: Estimated survival functions for those who are married and those who are not. Other covariates are fixed at their average values. Each estimate is accompanied by a point-wise 95-percent confidence envelope.

The graph returned supports the p-value of the test that indicates significant differences in survival functions of these two groups of offenders.

Paroled versus Released

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Paro=0	152	41	41.2	0.001387	0.00232

```
Paro=1       249     63             62.8          0.000911            0.00232
 Chisq= 0 on 1 degrees of freedom, p= 0.962
```

Potting the resulting graph supports the no-difference conclusion reached by the log-rank test

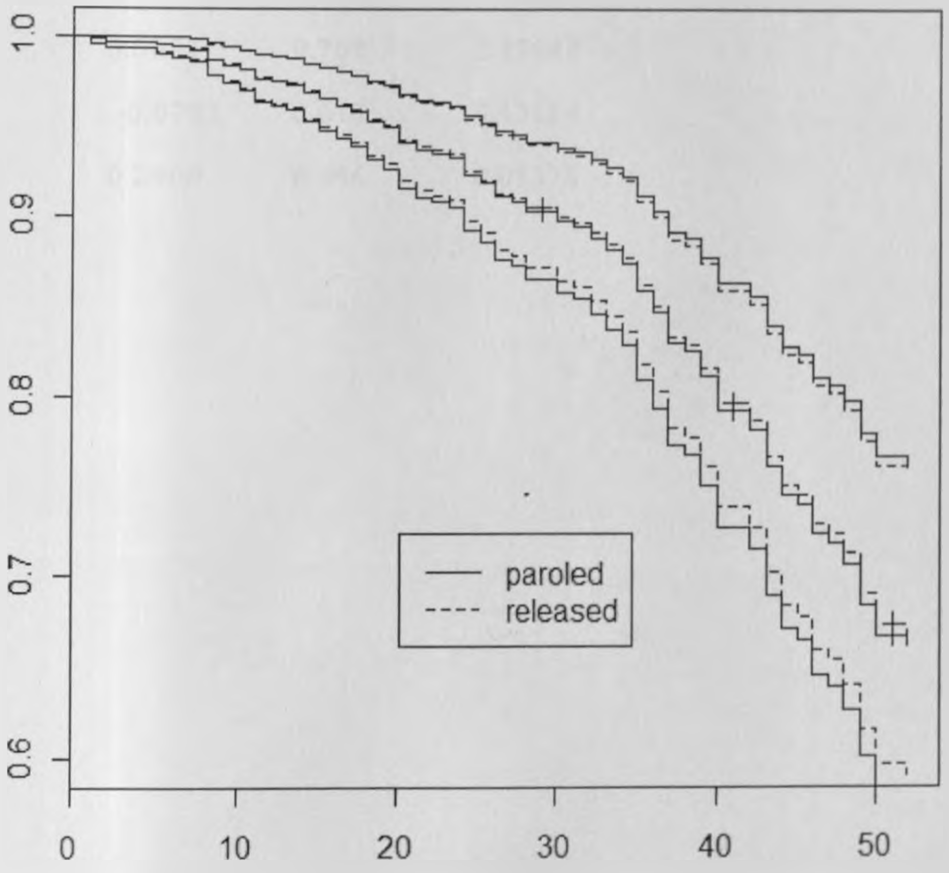
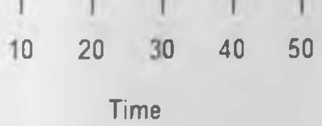
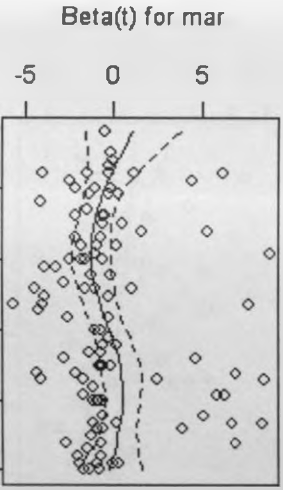


Figure 5: Estimated survival functions for those who were paroled and those who are released without condition. Other covariates are fixed at their average values

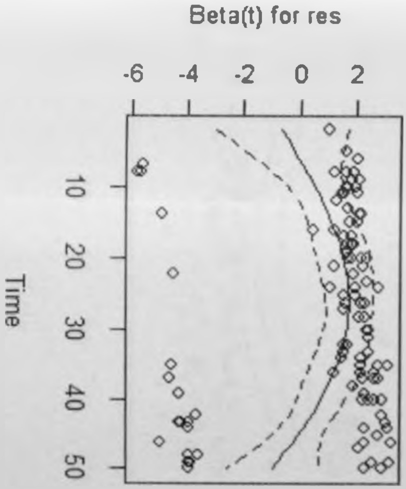
The variables were then tested for conformity with the proportionality assumption by applying the following command;

```
temp1 <- cox.zph (fit1, transform="identity", global=F) ; print (temp1)

                 rho          chisq            p
age           -0.2915         8.996         0.00271
res           -0.1018         1.023         0.31193
wexp           0.2190         4.898         0.02689
mar            0.1090          1.258         0.26211
paro           0.0851         0.782         0.37642
prio            -0.0781       0.619         0.43154
educ           0.2900         6.066         0.01378
```

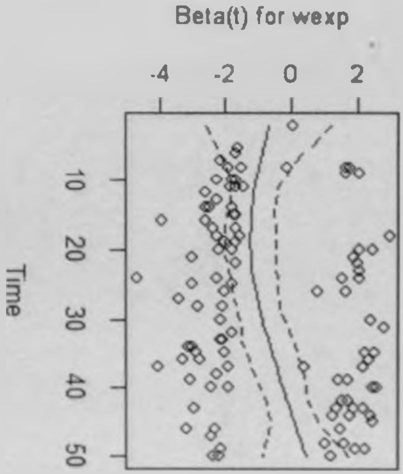

10
20
30
40
50
Time



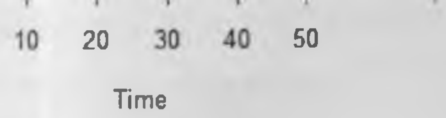
Beta(t) for mar

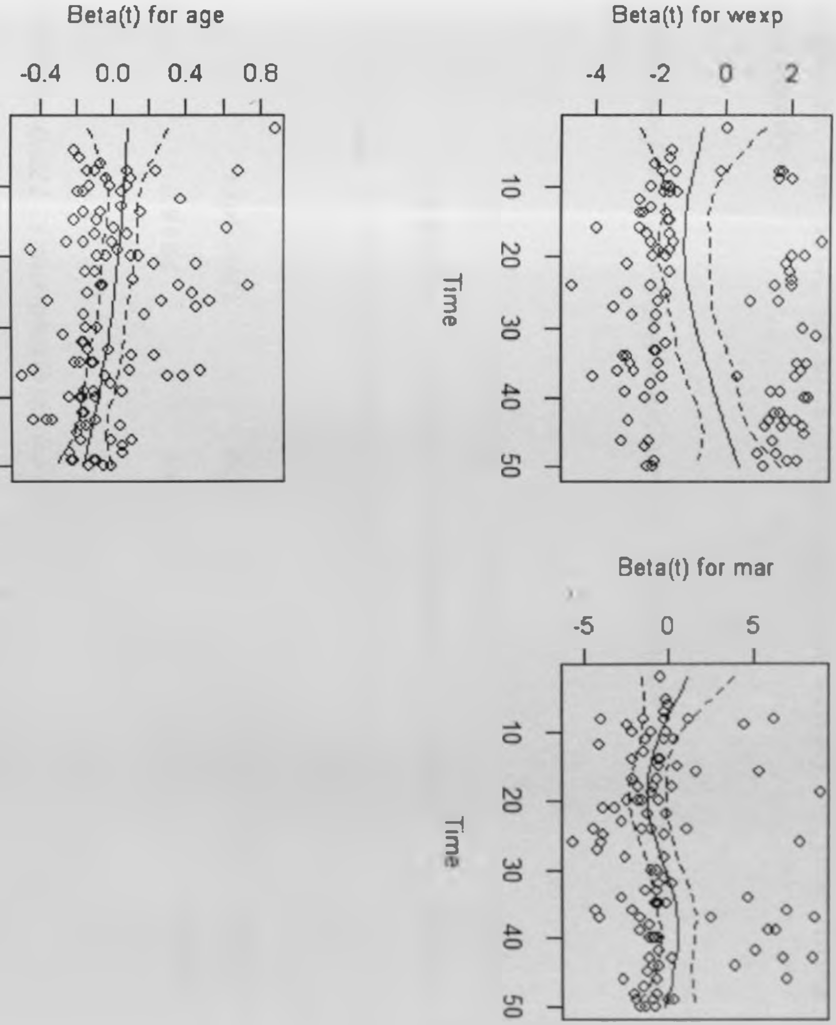


Beta(t) for res



Beta(t) for wexp


10
20
30
40
50
Time


Beta(t) for age
-0.4
0.0
0.4
0.8
Beta(t) for wexp
-4
-2
0
2
10
20
30
40
50
Time
Beta(t) for mar
-5
0
5
10
20
30
40
50
Time

4.4 The fitted Cox's PH model.

Here we consider the best way to develop a model with multiple potential predictors. For now, let us approach the data naively by first fitting each term individually.

Age

```
Call:coxph(formula = Surv(week, arrest) ~ age, data = recid); n= 402, number of events=
105

          coef        exp(coef)    se(coef)     z          p-value

age       -0.05506    0.94642      0.02058      -2.675     0.00746 **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef)          exp(-coef)         lower .95    upper .95

age       0.9464             1.057              0.909        0.9854

Rsquare= 0.021   (max possible= 0.945)

Likelihood ratio test= 8.52 on 1 df,   p=0.003512

Wald test = 7.16 on 1 df,   p=0.007462

Score (logrank) test = 7.33 on 1 df,   p=0.006794
```

>Age affects survival, with younger offenders at more risk of re-offence than older ones. The coefficient of determination(R^2=0.021) indicates that these two variables share a considerably small amount of variance.

Residence

```
Call:coxph(formula = Surv(week, arrest) ~ res, data = recid); n= 402, number of events=
105

          coef       exp(coef)        se(coef)      z          p-vaue

res       0.9077     2.4787           0.2551        3.558      0.000374 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef)      exp(-coef)      lower .95      upper .95
res       2.479          0.4034          1.503          4.087

Rsquare= 0.037   (max possible= 0.945)
Likelihood ratio test= 15.03 on 1 df,   p=0.0001058
Wald test = 12.66 on 1 df,   p=0.0003739
Score (logrank) test = 13.52 on 1 df,   p=0.0002364
```

> The residence of a released offender affects his chances of re-offending, with those hailing from slam dwellings being more than twice at risk than their counter-parts from all other forms of dwellings. The coefficient of determination (r^2=0.037) indicates that the two variables share a small amount of variance.

Work Experience

```
Call: coxph (formula = Surv (week, arrest) ~ wexp, data = recid); n= 402, number of
events= 105

          coef         exp(coef)       se(coef)      z          p-value
wexp      -0.9160      0.4001          0.1995        -4.592     4.38e-06 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef)      exp(-coef)      lower .95      upper .95
wexp      0.4001         2.499           0.2706         0.5915

Rsquare= 0.052   (max possible= 0.945)
Likelihood ratio test= 21.53 on 1 df,   p=3.482e-06
Wald test  = 21.09  on 1 df,   p=4.38e-06
```

```
Score (logrank) test = 22.56 on 1 df,   p=2.037e-06
```

> Having work experience reduces the chances of re-offence by nearly half. The coefficient of determination (*Rsquare= 0.052*) indicates that a weak relationship between work experience ad time to re-arrest though it has statistical significance as they share a very small amount of variance.

Marital status

```
Call:coxph(formula = Surv(week, arrest) ~ mar, data = recid); n= 402, number of events=
105


            coef          exp(coef)     se(coef)        z           Pr(>|z|)

mar          -0.6364       0.5292        0.2486        -2.559        0.0105 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


            exp(coef)           exp(-coef)          lower .95     upper .95

mar          0.5292              1.890               0.3251        0.8615


Rsquare= 0.018   (max possible= 0.945)

Likelihood ratio test= 7.41 on 1 df,   p=0.00649

Wald test = 6.55  on 1 df,   p=0.01048

Score (logrank) test = 6.77 on 1 df,   p=0.009247
```

Though not as influential as the covariates that we have seen so far, marriage is still highly significant (married men are less likely to re-offend than their un-married counterparts by half). The coefficient of determination (R^2=0.018) indicates that the dependent and independent variable in this case share an extremely small amount of variance.

Education

```
Call:coxph(formula = Surv(week, arrest) ~ educ, data = recid);n= 402, number of
events= 105


            coef          exp(coef)           se(coef)      z              Pr(>|z|)
```

```
educ     -0.27501     0.75957          0.07234   -3.802         0.000144 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


            exp(coef)        exp(-coef)        lower .95         upper .95
educ        0.7596           1.317             0.6592            0.8753


Rsquare= 0.036   (max possible= 0.945 )
Likelihood ratio test= 14.73 on 1 df,   p=0.0001241
Wald test = 14.45 on 1 df,   p=0.0001437
Score (logrank) test = 14.73 on 1 df,   p=0.0001242
```

> Education level of a released offender remains to be one of the most significant predictors of recidivism with those with higher education level having 25 percent less chances of re-offence. The coefficient of determination (Rsquare=0.036) indicates that less than 4% of the variance in hazard of re-arrest can be accounted for by its linear relationship with the education level of the offender.

Paroled/released

```
Call: coxph(formula = Surv(week, arrest) ~ paro, data = recid); n= 402, number of
events= 105
            coef          exp(coef)          se(coef)      z          p-value
paro        0.02173       1.02197            0.20018       0.109      0.914


            exp (coef)    exp(-coef)         lower .95           upper .95
paro        1.022         0.9785             0.6903              1.513


Rsquare= 0   (max possible= 0.945)
Likelihood ratio test= 0.01 on 1 df,   p=0.9135
```

```
Wald test = 0.01 on 1 df,   p=0.9136

Score (logrank) test = 0.01 on 1 df,   p=0.9136
```

Whether an offender is released on parole or unrestricted release has no bearing on his chances of re-arrest. The hypothesis of equal hazards between the two groups is soundly accepted and the coefficient of determination (Rsquare=0) neither indicates that this variable is correlated to time to rearrest nor helps to explain any variance observed.

Number of prior offences

Call: coxph (formula = Surv (week, arrest) ~ prio, data = recid); n= 402, number of events= 105

```
          coef        exp(coef)        se(coef)     z          Pr(>|z|)

prio      -0.0580     0.9436           0.1213       -0.478     0.633


          exp(coef)        exp(-coef)        lower .95        upper .95

prio      0.9437           1.060             0.744            1.197


Rsquare= 0.001   (max possible= 0.945)

Likelihood ratio test= 0.23 on 1 df,   p=0.6297

Wald test = 0.23 on 1 DF,   p=0.6326

Score (logrank) test = 0.23 on 1 df,   p=0.6326
```

The number of prior offences has no bearing in the chances of re-offending.the hypothesis that those who have committed more crimes than others are more susceptible to failure is rejected and the coefficient of determination indicates that none of the variation observed in the chances of re-arrest is attributable to this variable.

Since age, marital status, work experience and education level all seem to influence survival; a natural second step would be to fit a regression model incorporating all of them:

```
Coxph (formula = Surv(week, arrest) ~ age + res + wexp + mar +paro + prio + educ, data
= recid)

           coef      exp(coef)        se(coef)     z          p
age       -0.0252    0.975            0.0247      -1.019     0.31000
res        0.8374     2.310           0.2586       3.238     0.00120
wexp     -0.6988     0.497            0.2100      -3.328     0.00087
mar       -0.2872    0.750            0.2971       -0.967    0.33000
paro      -0.1215    0.886            0.2026       -0.600    0.55000
prio     -0.0258     0.975            0.1226      -0.211     0.83000
educ     -0.2418     0.785            0.0736      -3.288     0.00100

Likelihood ratio test=50.2  on 7 df, p=0  n= 402, number of events= 105
```

```
Call: coxph(formula = Surv(week, arrest) ~ age + res + wexp + mar +paro + prio + educ,
data = recid);n= 402, number of events= 105

          coef         exp(coef)        se(coef)      z          p-value
age        -0.02517     0.97514          0.02469     -1.019      0.308005
res       0.83738      2.31030          0.25862      3.238       0.001204 **
wexp     -0.69879     0.49719          0.20996     -3.328       0.000874 ***
mar      -0.28722     0.75035          0.29712      -0.967      0.333709
paro      -0.12155     0.88555          0.20258      -0.600      0.548513
prio      -0.02581     0.97452          0.12259      -0.211      0.833261
educ      -0.24184     0.78518          0.07355      -3.288      0.001009 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
        exp(coef)    exp(-coef)    lower .95    upper .95
age      0.9751       1.0255        0.9291       1.0235
res      2.3103       0.4328        1.3917       3.8353
wexp     0.4972       2.0113        0.3295       0.7503
mar      0.7503       1.3327        0.4191       1.3433
paro     0.8855       1.1292        0.5954       1.3172
prio     0.9745       1.0261        0.7664       1.2392
educ     0.7852       1.2736        0.6798       0.9069
Rsquare= 0.117   (max possible= 0.945 )
Likelihood ratio test= 50.19  on 7 df,   p=1.326e-08
Wald test   = 45.77  on 7 df,   p=9.674e-08
Score (logrank) test = 48.31  on 7 df,   p=3.092e-08
```

From the output of R program, it is observed that the test for the regression parameters equal to zero is rejected with a Likelihood ratio test= 49.78 value for 5 degrees of freedom and p-value of 0.

In the output presented, the beta coefficients show just how much each independent variable predicts the dependent variable and therefore how significant each independent variable is. The beta weights varied significantly. The residential background has a beta weight of 0.84 and therefore had the greatest effect in predicting the hazards of re-arrest followed by work experience.

Of the attributes considered age, work experience, marital status as well as the education level continue to contribute significantly to the relapse to crime i.e. if all other attributes are held constant,

i. A unit increase in age reduces the risk of relapse by 4%.

ii. Released convicts from slum dwellings are three times more likely to recidivate than their counterparts from all other dwellings.

iii. Those with previous work experience have 60% less chance of recidivating than those without.

iv. Released felons that were married before arrest have 50% less chances of recidivating than their un-married counterparts.

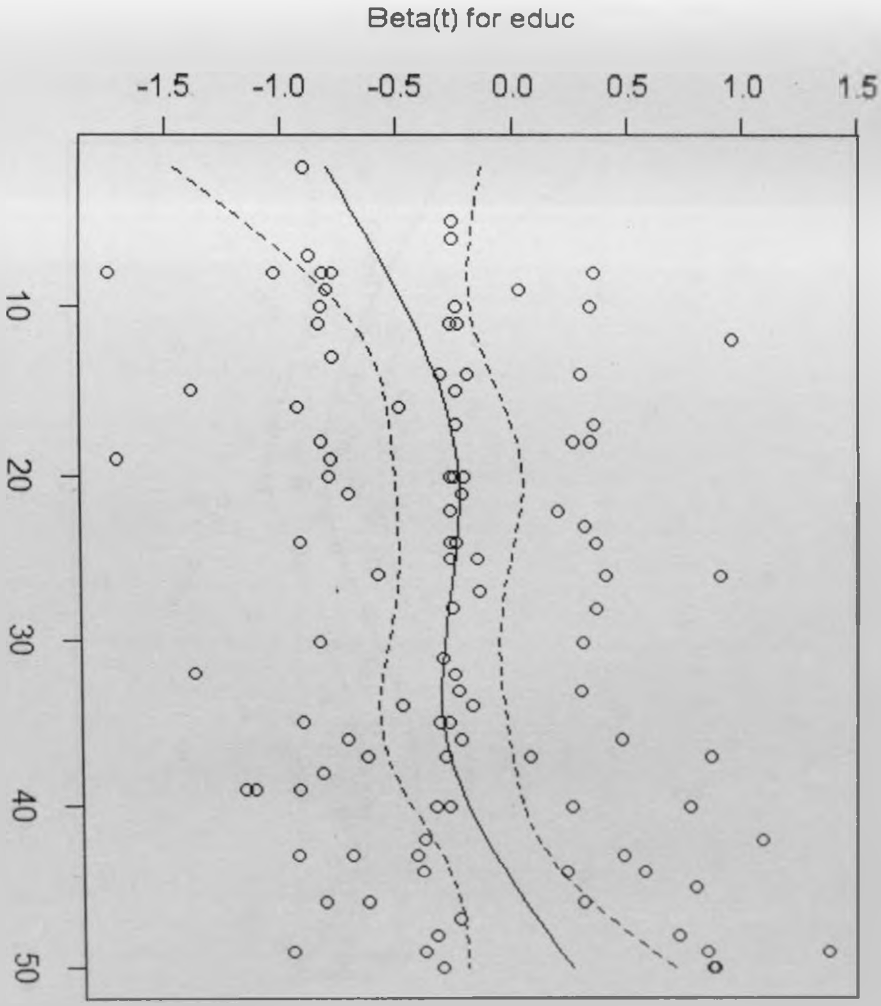
v. A unit increase in education level reduces the risk of recidivating by 44%.

Tests and graphical diagnostics for proportional hazards give;

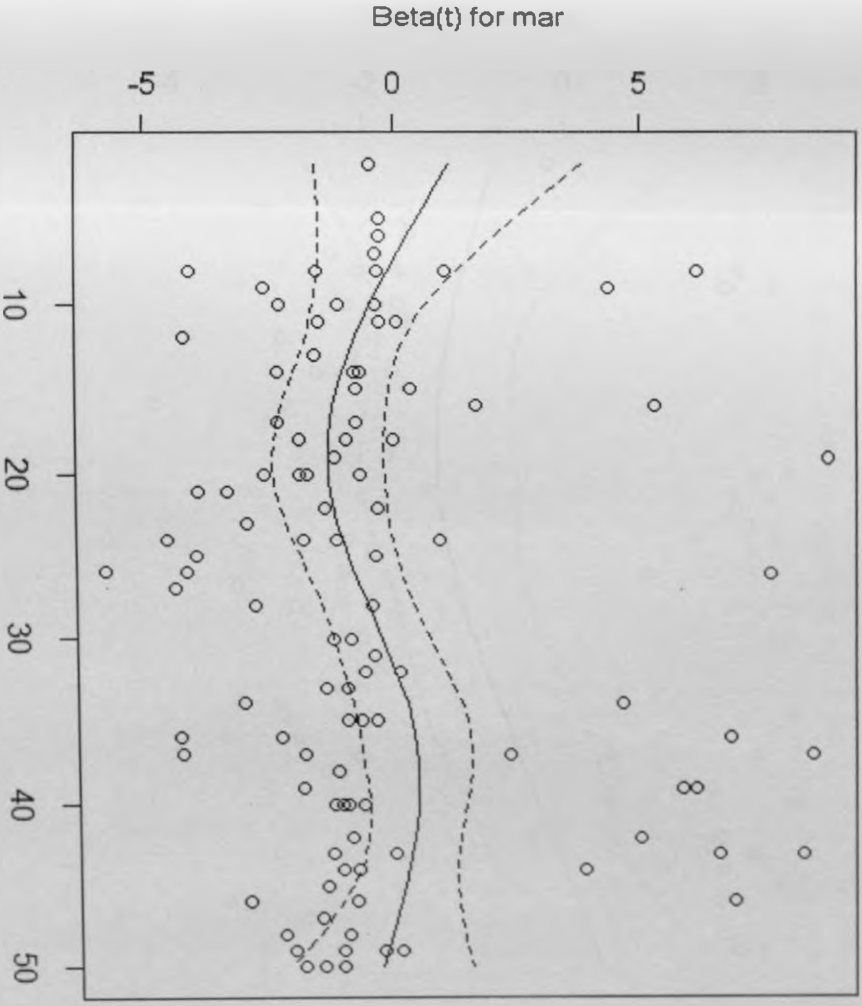
```
> temp1 <- cox.zph(fit1 ,transform="identity", global=TRUE)

> print(temp1)

            rho        chisq          p

age        -0.2915     8.996      0.00271

res        -0.1018     1.023      0.31193

wexp        0.2190     4.898      0.02689

mar         0.1090     1.258      0.26211

paro        0.0851     0.782      0.37642

prio       -0.0781     0.619      0.43154

educ        0.2900     6.066      0.01378

GLOBAL        NA      21.893      0.00265
```

Here, the covariates age, work experience and education show violation of the proportional hazard assumption with p-values that are greater than 0.05. Plotting the object returned by this test produces graphs of the scaled Schoenfeld residuals against time which can also be used in the verification process.

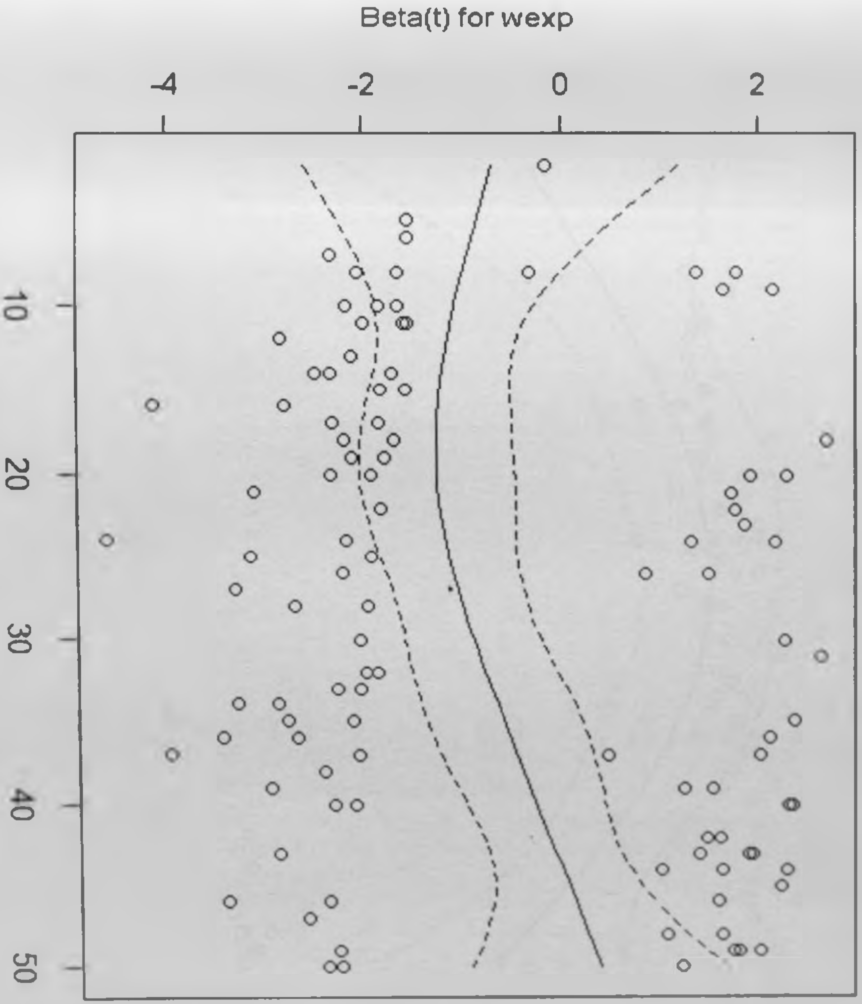
Time

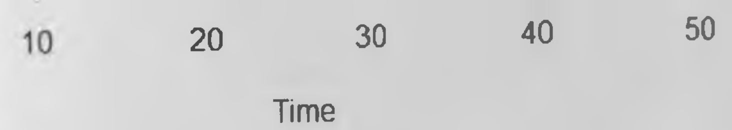

Beta(t) for educ
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
10
20
30
40
50

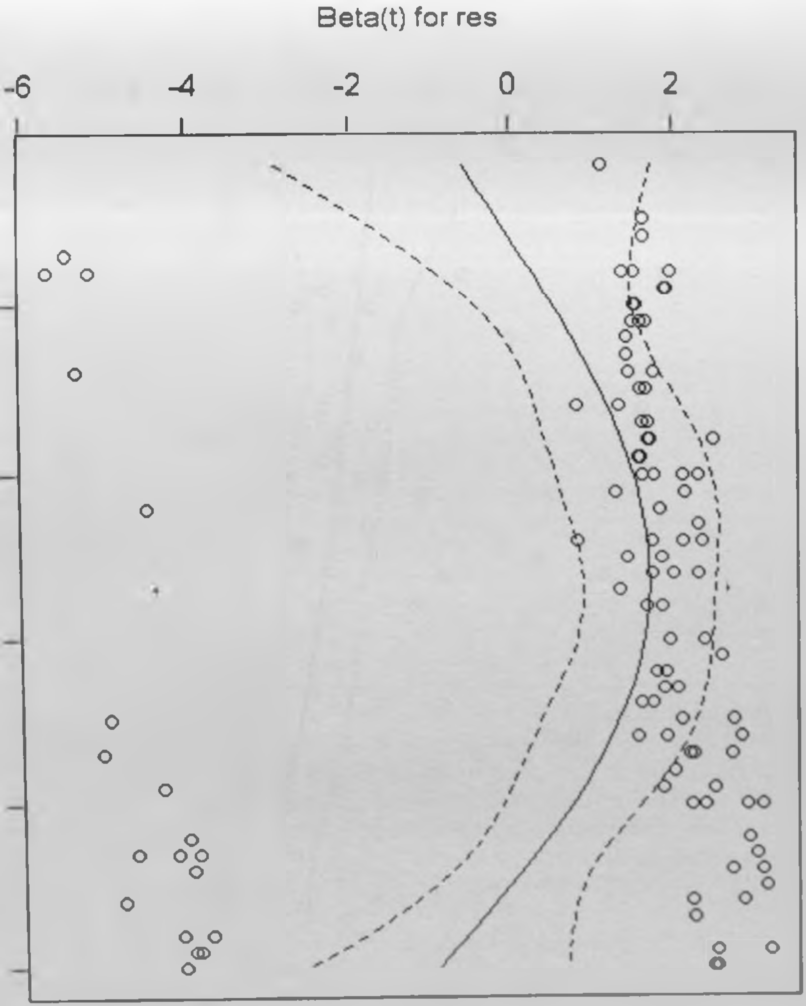
Time

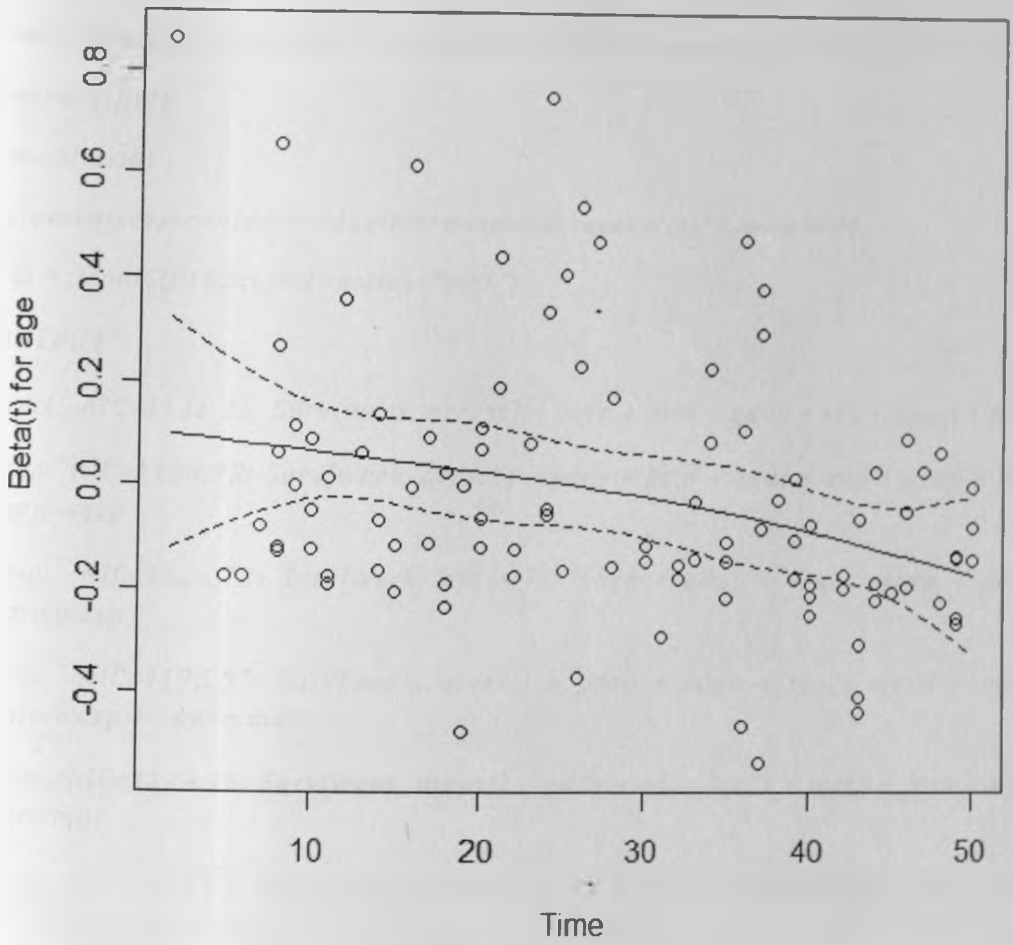

Beta(t) for mar
-5
0
5
10
20
30
40
50

Time


Beta(t) for wexp
-4
-2
0
2
10
20
30
40
50


10
20
30
40
50
Time


Beta(t) for res
-6
-4
-2
0
2



Systematic departures from a horizontal line are indicative of non-proportional hazards. The assumption of proportional hazards appears to be supported for the covariates age and prior, but there appears to be a trend in the plot for residence, work experience, marital status, parole and education level.

One way of accommodating non-proportional hazards is to build interactions between covariates into the Cox regression model; such interactions are themselves considered as covariates. There are two main ways to define which interactions to include or drop. One approach is the use of the AIC to determine the parameter which best improves the model. This is easy to implement in R, I simply run the command stepAIC, which at each step, and tries to add in each excluded covariate and remove each included covariate. It then chooses the model that minimizes the AIC.

```
library(survival,MASS)

fit4<-coxph(Surv(week,arrest)~paro+prio+educ+res+wexp+mar+age,data=recid);fit4

summary(fit4)

attach(recid)

Scope=list(upper=~(paro+educ+res+wexp+mar+age+prio)^2,lower=~1)

fitA = stepAIC(fit4,Scope,direction="both")
```

OUTPUT

```
Start: AIC=1131.22; Surv(week, arrest) ~ paro + prio + educ + res + wexp + mar + age

Step: AIC=1128.73; Surv(week, arrest) ~ paro + prio + educ + res + wexp + mar + age +
paro:wexp

Step: AIC=1126.73; Surv(week, arrest) ~ paro + educ + res + wexp + mar + age +
paro:wexp

Step: AIC=1125.57; Surv(week, arrest) ~ paro + educ + res + wexp + mar + age +
paro:wexp + paro:mar

Step: AIC=1124.16; Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar

Step: AIC=1123.42; Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar + res:wexp

Step: AIC=1122.84; Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar + res:wexp + educ:wexp

Step: AIC=1120.17; Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar + res:wexp + educ:wexp + educ:res

Step: AIC=1119.07; Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar + res:wexp + educ:wexp + educ:res + educ:mar
```

At the last stage, it seeks either to remove one of the two covariates already selected or add a single other covariate or add an interaction between the two existing covariates, but none of these models improves the AIC, so it stops there. Running the selected model and its interactions gives me;

```
fit44<-coxph(Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp +
paro:mar + res:wexp + educ:wexp + educ:res + educ:mar)

> summary(fit44)

Call:coxph(formula = Surv(week, arrest) ~ paro + educ + res + wexp + mar + paro:wexp
+ paro:mar + res:wexp + educ:wexp + educ:res +  educ:mar)

n= 402, number of events= 105


                coef       exp(coef)   se(coef)    z        Pr(>|z|)
paro          -0.69840     0.49738     0.27924  -2.501     0.01238 *
educ          -0.77309     0.46158     0.24280  -3.184     0.00145 **
res            0.03567     1.03632     0.74531   0.048     0.96182
wexp          -1.81476     0.16288     0.72683  -2.497     0.01253 *
mar           -0.04411     0.95684     0.76083  -0.058     0.95376
paro:wexp      0.87771     2.40539     0.43176   2.033     0.04206 *
paro:mar       0.93027     2.53518     0.57159   1.628     0.10363
res:wexp      -1.04687     0.35103     0.56572  -1.851     0.06424 .
educ:wexp      0.41633     1.51638     0.16915   2.461     0.01384 *
educ:res       0.45968     1.58357     0.22644   2.030     0.04235 *
educ:mar      -0.32965     0.71918     0.18931  -1.741     0.08164 .
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


          exp(coef)    exp(-coef)    lower .95    upper .95
paro      0.4974       2.0105        0.28774      0.8598
educ      0.4616       2.1665        0.28680      0.7429
res       1.0363       0.9650        0.24048      4.4659
wexp      0.1629       6.1396        0.03919      0.6769
mar       0.9568       1.0451        0.21539      4.2507
```

```
paro:wexp   2.4054     0.4157     1.03198     5.6066
paro:mar    2.5352     0.3944     0.82694     7.7722
res:wexp    0.3510     2.8487     0.11583     1.0639
educ:wexp   1.5164     0.6595     1.08850     2.1125
educ:res    1.5836     0.6315     1.01600     2.4682
educ:mar    0.7192     1.3905     0.49624     1.0423

Rsquare= 0.161   (max possible= 0.945 )

Likelihood ratio test= 70.34  on 11 df,   p=1.052e-10

Wald test = 63.25  on 11 df,   p=2.294e-09

Score (logrank) test = 74.77  on 11 df,   p=1.500e-11
```

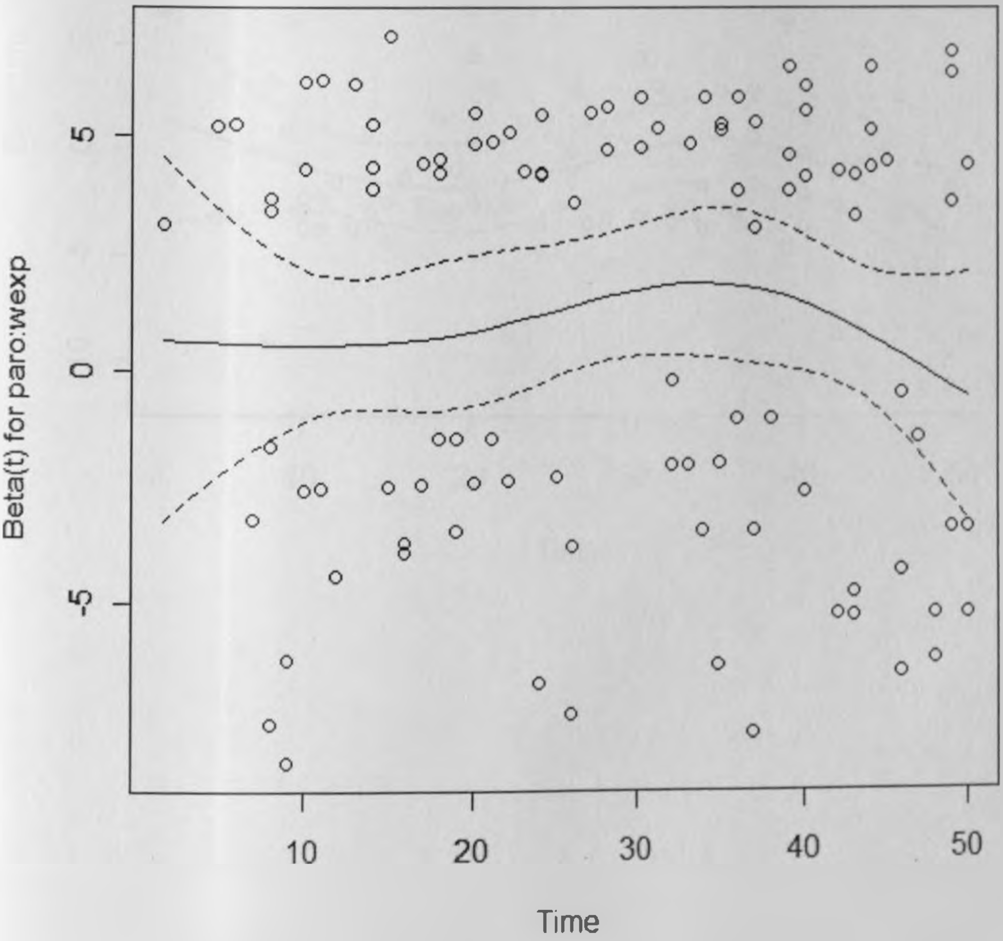
The interactions between education, work experience, marital status and residence show significance in predicting chances of re-offence. The proportion of variation accounted for by the model also declines significantly when compared to the model without interactions.

Now, checking this fit for conformity with proportional hazard assumption and general goodness of fit(indicated as global in the following table) indicates that non of the remaining variables show violations of the assumption and that the model in general is a good fit.

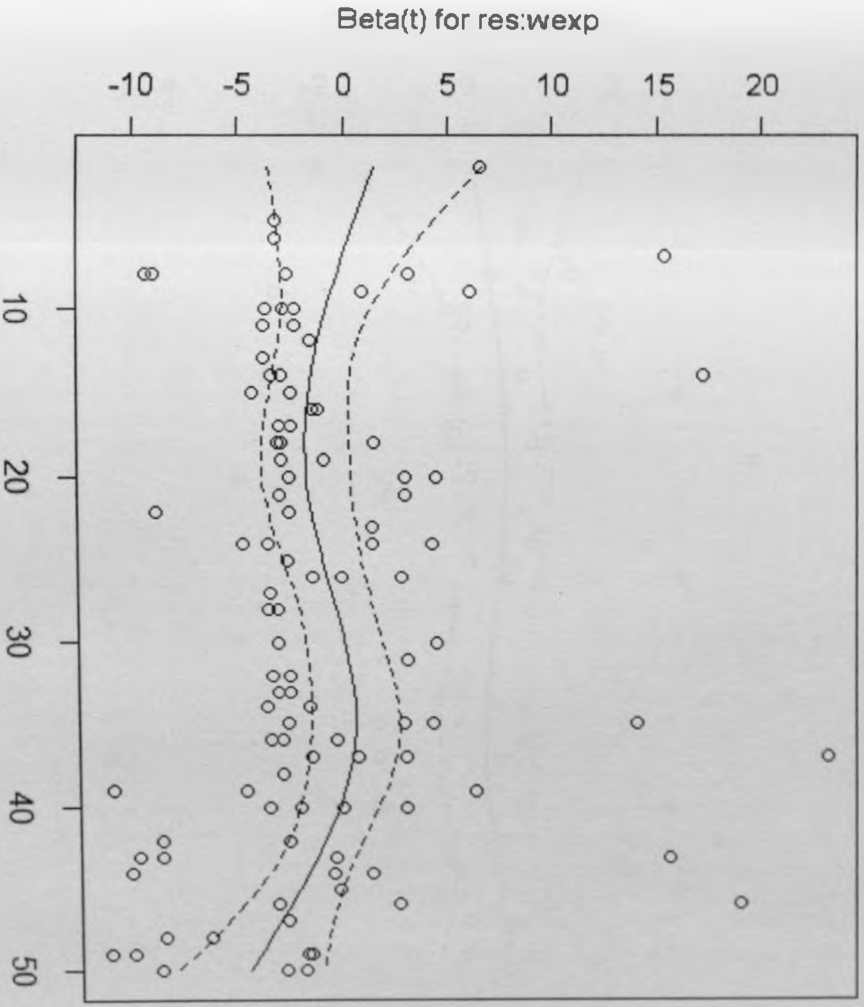
```
temp1 <- cox.zph(fit44 ,transform="identity", global=T)

> print(temp1)

                rho       chisq      p
paro          0.0506     0.27543    0.600
educ          0.0466     0.19741    0.657
res          -0.0407     0.12356    0.725
wexp          0.0615     0.34118    0.559
mar          -0.0463     0.22231    0.637
paro:wexp     0.0158     0.02769    0.868
paro:mar     -0.1033     1.14295    0.285
```

res:wexp	*-0.0755*	*0.62279*	*0.430*
educ:wexp	*0.0081*	*0.00493*	*0.944*
educ:res	*0.0162*	*0.02274*	*0.880*
educ:mar	*0.1316*	*1.58624*	*0.208*
GLOBAL	*NA*	*12.45899*	*0.330*

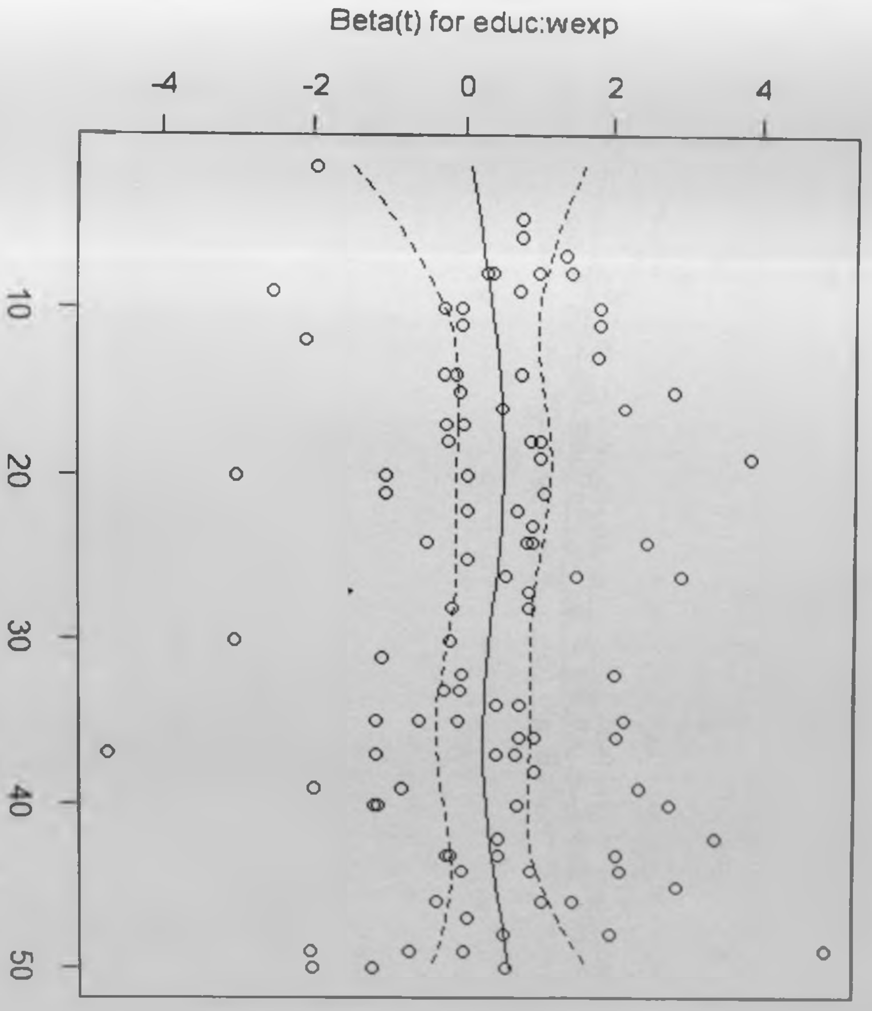
Plotting the results for interactions, I observe;

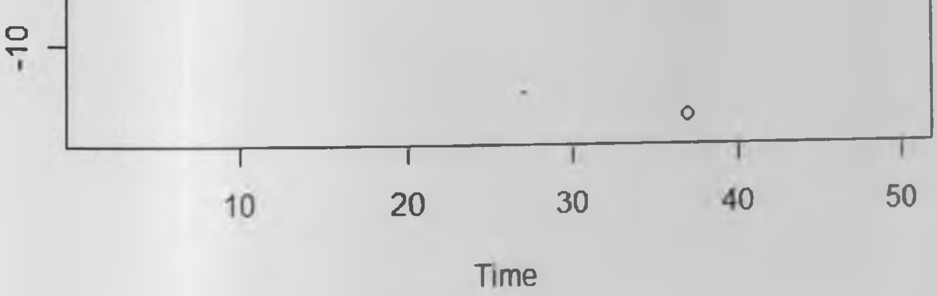


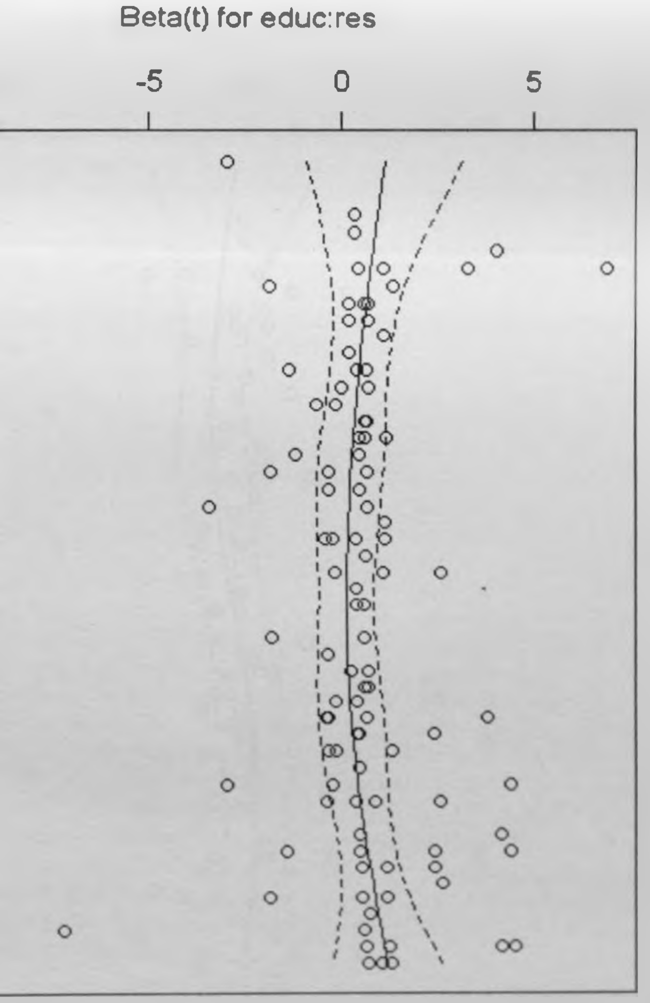
Time

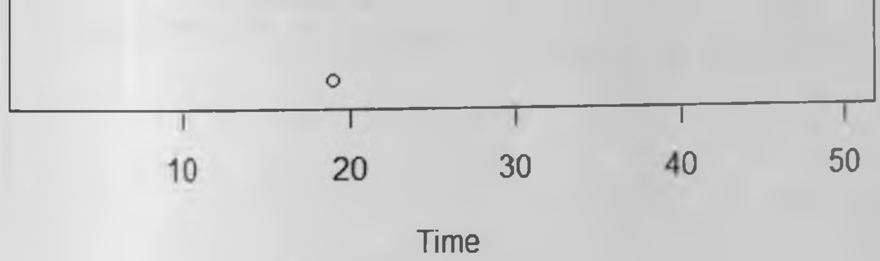

Beta(t) for res:wexp
-10
-5
0
5
10
15
20
10
20
30
40
50

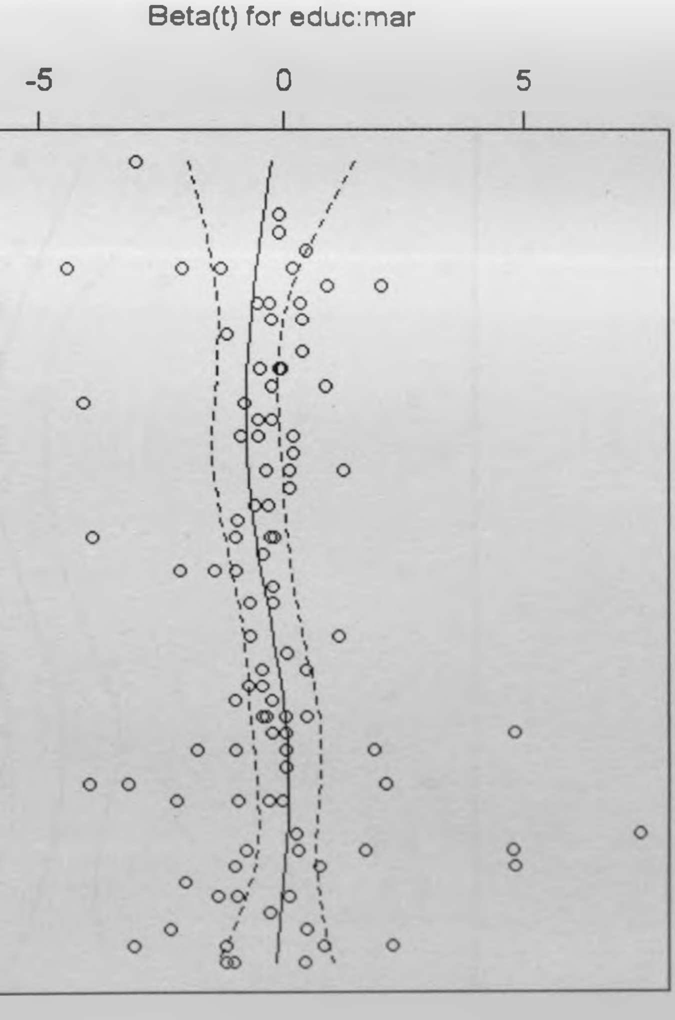
Time

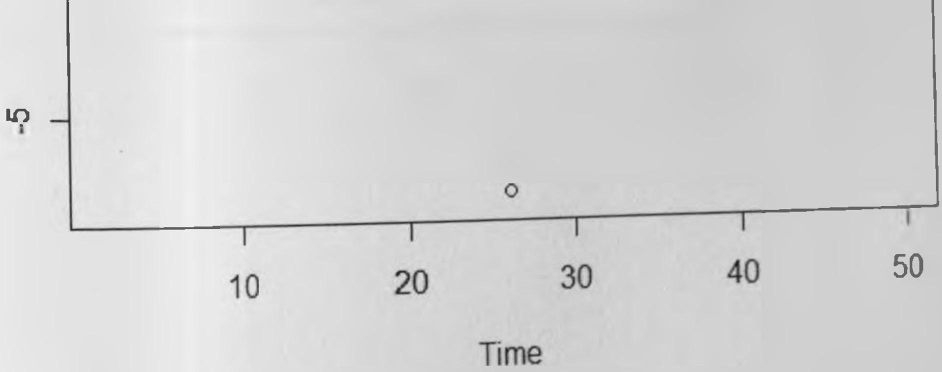

Beta(t) for educ:wexp
-4
-2
0
2
4
10
20
30
40
50

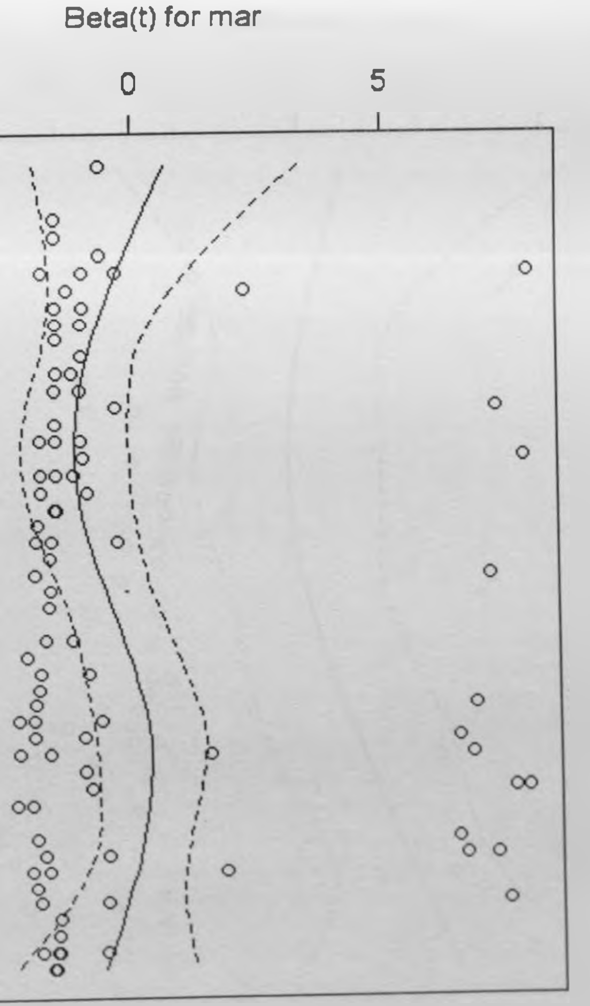

-10
10
20
30
40
50
Time


Beta(t) for educ:res
-5
0
5

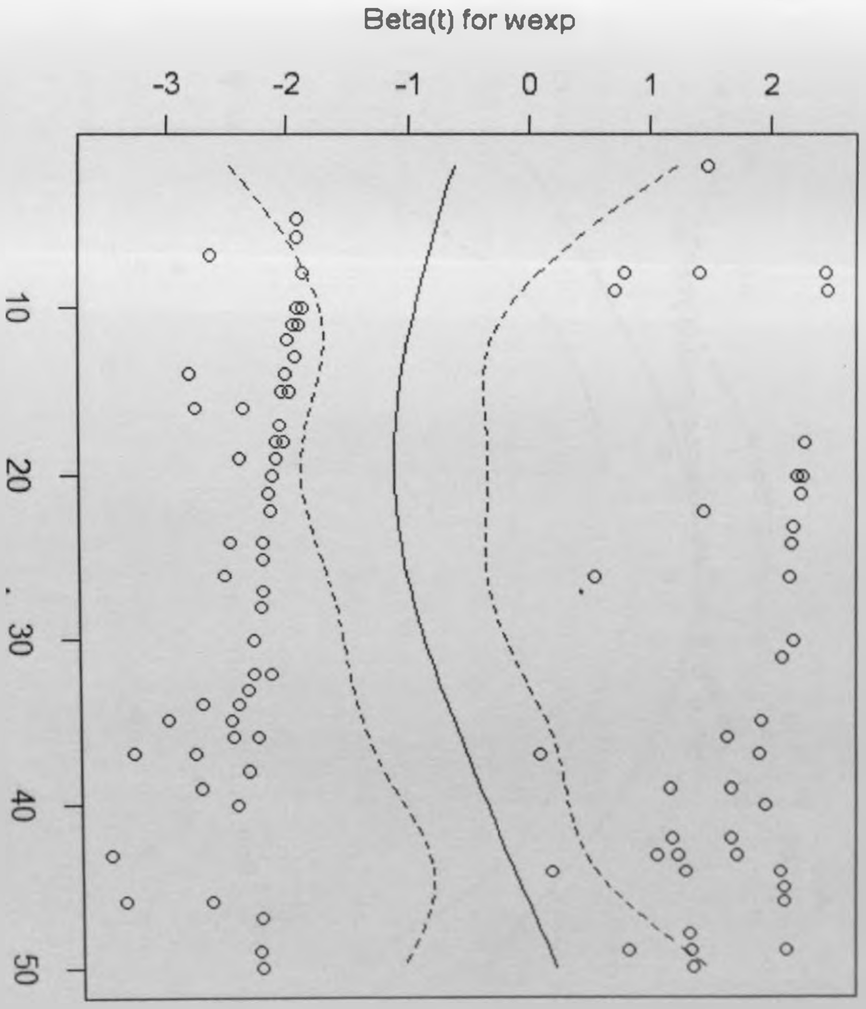

10
20
30
40
50
Time

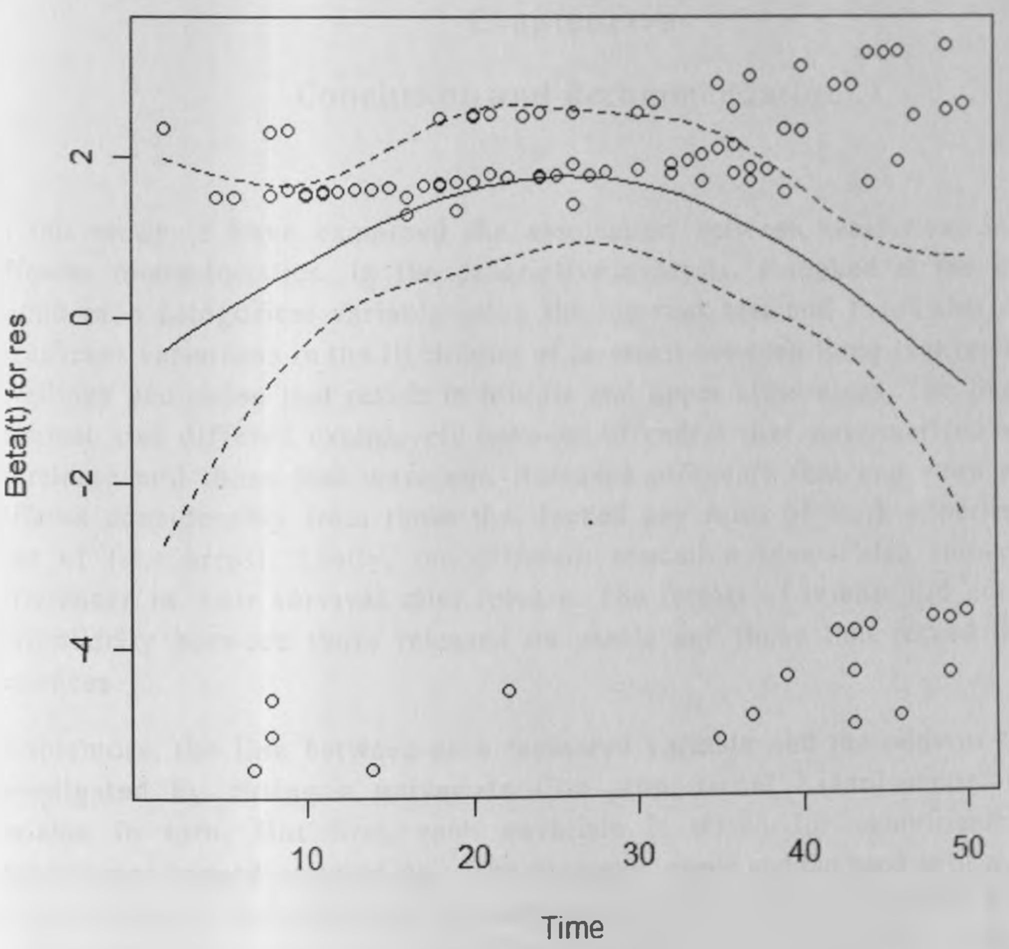

Beta(t) for educ:mar
-5
0
5


-5
10
20
30
40
50
Time


Beta(t) for mar
0
5

Time


Beta(t) for wexp
-3
-2
-1
0
1
2
10
20
30
40
50



The interaction of variables achieves a slight improvement as can be seen in the above figures and hence enables the model as a whole to comply with this assumption.

Chapter Five

Conclusion and Recommendation

In this study, I have examined the association between recidivism and various offender characteristics. In the descriptive analysis, I looked at the differences within each categorical variable using the log-rank test and found that there were significant variations in the likelihood of re-arrest between those that reside in slum dwellings and those that reside in middle and upper class areas. The likelihood of re-arrest also differed extensively between offenders that were married at the time of release and those that were not. Released offenders that had work experience differed considerably from those that lacked any form of work experience at the time of first arrest. Lastly, the different education levels also showed notable differences in their survival after release. The format of release did not offer any dissimilarity between those released on parole and those that served their entire sentences.

Furthermore, the link between each measured variable and the odds of re-arrest is investigated by fitting a univariate Cox proportional hazard model with each variable in turn. But first, each covariate is tested for compliance with the proportional hazard assumption. The covariates, parole and the number of prior offences show violation of the assumption and are consequently dropped from the multivariate analysis but not before their individual contributions are assessed. The univariate analysis yields noteworthy contributions to the hazard of recidivism for the variables age, residential area, work experience, education level and marriage. The covariates, number of prior offences and the format of release continue to show insignificant contributions to the chances of re-arrest .

A natural second step then becomes incorporating all the variables in the model. Of the attributes considered age, work experiences, marital status as well as the education level continue to contribute significantly to the relapse to crime i.e. if all other attributes are held constant, a unit increase in age reduces the risk of relapse by 4%, released convicts from slum dwellings are three times more likely to recidivate than their counterparts from all other dwellings, those with previous work experience have 60% less chance of recidivating than those without, released felons that were married before arrest have 50% less chances of recidivating than their un-married counterparts and a unit increase in education level reduces the risk of recidivating by 44%. The clear violation of the PH assumption is mitigated by introducing a step-wise procedure of determining which covariates to interact and this achieves a tremendous improvement in the compliance of the resulting model though in general the model reduces the proportion of observed differences it is able to explain.

In addition to this, the coefficient of determination is weak for each of the attribute considered both in the univariate analysis and in the overall model. Having such low coefficient of determination has a lot to do with the small sample size as well as incomplete variable selection and I will therefore recommend the inclusion of more variables such as sentence length, type of offence and to incorporate a larger region to better understand the role played by demographic characteristics in the propensity of recidivism. Similarly, due to the extensive number of offenders that come from the same residential backgrounds and therefore share the same social-economic challenges, I will recommend the use of frailty models to ascertain their susceptibility to crime.

Bibliography

May, C., Sharma, N., & Stewart, D. (2008). Factors linked to reoffending: A one-year follow-up of prisoners who took part in the resettlement surveys 2001, 2003 and 2004. Ministry of Justice.

Anna Ferrante, N. L. (September 1999). Measurement of the Recidivism of Offenders attending the Kimberley Offender Program. *Crime Research Centre* , 1-17.

Langan, P. A., & Levin, D. J. (2002). Recidivism of prisoners released in 1994. U.S. Department of Justice, Office of Justice Programs. Washington, DC: Bureau of Justice Statistics.

Diez, D. *Survival Analysis in R.*

Efrat Aharonovich, P. X. (2005, August). Postdischarge Cannabis Use and Its Relationship to Cocaine, Alcohol, and Heroin Use: A Prospective Study. *Am J Psychiatry* , pp. 1506-1514.

Florida Department of Corrections. (2009). *2009 Florida Prison Recidivism Study Releases From 2001 to 2008.* Florida : Florida Department of Corrections.

Fox, J. (Februrary 2002). Cox Proportional-Hazards Regression for Survival Data. *journal of royal statistics* , 1-18.

Kruttschnitt, C., Uggen, C., & Shelton, K. (2000). Predictors of desistance among sex offenders: The interaction of formal and informal social controls. Justice Quarterly, 17 (1), 61-87.

Frank S. Pearson, D. S. (3 july 2012). The Effects of Behavioral/Cognitive-Behaviora lPrograms on Recidivism. *National Development and Research Institutes, Inc* , 1-21.

Hanagal, David D. (2011). *Modeling survival data using frailty models.* Danver: Taylor and Francis Group.

Hill, S. J. (May 9, 2011). *A Quantitative Analysis of Factors Related to Recidivism.* Auburn, Alabama: Graduate Faculty of Auburn University.

Jason P. Fine, D. V. (2001). Simple estimator for a shared frailty regression model. *Biometrics* , 1-24.

Martinussen, C. B. (2004). An Estimating Equation for Parametric Shared Frailty Models with Marginal Additive Hazards. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 66, No.1* , 207-220.

Ni, J. (2009). *Application of Cox Proportional Hazard Model to the Stock Exchange Market.* Beijing: Ball State University.

Per Kragh Andersen, J. P. (1997). Estimation of Variance in Cox's Regression Model with Shared Gamma Frailties. *International Biometric Society* , 1476-1477.

Petersen, J. H. (1998). An Additive Frailty Model for Correlated Life Times. *International Biometric Society* , 647-649.

Qi, J. (2009). *Comparison of Proportional Hazards and Accelerated Failure Time Models.* Saskatchewan: Department of Mathematics and Statistics University of Saskatchewan.

RUTERE SALOME KAGENDO. (2005). *FACTORS PRECIPITATING RECIDIVISTIC BEHAVIOURS AMONG.* Nairobi: Adventure works press.

Tommi Härkänen, H. H. (Sep,2003). A Non-Parametric Frailty Model for Temporally Clustered Multivariate Failure Times. *Scandinavian Journal of Statistics, Vol. 30, No. 3* , 523-533.

Christine Achieng' Okoth Obondi*. (2001). *EFFECTIVE RESETTLEMENT OF OFFENDERS BY STRENGTHENING 'COMMUNITY REINTEGRATION FACTORS': KENYA'S EXPERIENCE.* NAIROBI: 145TH INTERNATIONAL TRAINING COURSE PARTICIPANTS AND OBSERVERS' PAPERS.