PROJECT REPORT

AIDS CLINICAL TRIALS

USING COX PROPORTIONAL HAZARDS MODEL~~S~~

OPIM 5503 – Data Analytics Using R – SECB13

Team Name:

Rtists

Team members:

Sahil Sikka

Kunja Dutta

Sagar Swamy

Vinay Paladugu

Jyothsna Chivukula

Prasraban Mukhopadhyay

Contents

1. Executive Summary

2. Background and Techniques

2.1 Introduction to concepts in Survival Analysis

Survival Analysis – It is a branch of statistics to analyze the expected duration of time until one or more event of interest occurs. It attempts to answer questions such as what proportion of a population will survive past a certain time.

Event – Disease occurrence, death, recovery, or other experience of interest.

Time – The time from the beginning of an observation period to either an event or end of study or termination of the study.

Survival Function S(t) – The probability that the subject survives longer than the time t.

Time to Event – This is the main variable of interest in this subject. It is a positive random variable. It is the response variable which is also called as the failure time, event time or even the survival time. During most studies, this variable is subjected to truncation or even censoring as per needs. During this only limited information will be available.

Censor

In statistics, censored observations have incomplete information about their time to event or the survival time. There are mainly three types of censoring, left censored data, right censored data, or else interval censored data. In general, the right censored observations are common. During clinical tests or trials, the term censor refers to mathematically removing a subject (or patient) for the survival curve at the end of their follow-up time. This results in a reduction of the sample size for analysis after the time of censorship.

Right Censoring – This occurs when a subject leaves the study before the interested event occurs or if the study was ended before the event occurred.

Left Censoring – This occurs when the event of interest has already occurred before the subject was brought into the study. This situation is not encountered that frequent.

Interval Censoring – Here we mean that the random variable of interest is known only to lie within an interval instead of being observed exactly. This random variable is the time to some event such as death or failure.

Hazard

It is the probability when a subject under study at a time t has an event at that exact time. Hazard is mathematically denoted by h(t) or $\lambda(t)$. Survival functions generally focus on not having an event, whereas hazard function on the contrast focuses on the event occurring. As per the study, the hazard function is also called as the hazard rate or the failure rate. The Hazard function is a ratio of the probability density function to the survival function. The Hazard function contains the hazard rate, which is the rate of death for an item at a given age(x). In engineering systems, it is the frequency with which the component fails and is expressed in failures per unit of time. If the value of the hazard ratio is one, then it means that there is no difference in the survival time between two groups. But if the value is greater or less than one, then it means that the survival of one of the groups was better than the other.

2.2 Techniques

The different techniques used in survival analytics is broadly classified into three major categories, namely parametric methods, semi-parametric methods and non-parametric methods.

The method to be chosen depends on the problem scenario. Let us see what type of method to use under what situation, ~~by referring to the table shown below.~~

2.2.1 Parametric Methods

Whenever we have regression-based models, we can apply Parametric methods. There are three common methods for the same used for this case as mentioned below.

1. Exponential Model
2. Weibull Model
3. Lognormal Model

2.2.2 Non-Parametric Methods

We assume non-parametric models when our data might not meet the assumptions made in parametric models. In such cases, we rely totally on our sample dataset we have. The most popular models used is Kaplan-Meier Model.

2.2.2.1 Kaplan-Meier Model

1. Kaplan-Meier Model - This method computes the probability of dying at a certain point of time conditional to the survival up to that point. It utilizes the information of censored individuals till the point when the patient is censored. This way it maximizes utilization of available information on time to an event of the study sample. The equation used to derive survival probability at time ‘t’ is derived by the following equation:

$$S(t)= \quad \pi t\ (1-dt/nt),$$

where d_t/ n_t represents the probability of dying at time ‘t’ conditional to being at risk (alive) at ‘t -1’ time.

Advantages: It is one of the most commonly used models of survival analysis and its estimates are usually accurate to examine recovery rates, the probability of death, and the effectiveness of clinical treatments.

Limitation: This method is limited in its ability to estimate survival adjusted for covariates. Under such circumstances, the Cox Proportional Hazards Model is useful to estimate the covariate-adjusted survival analysis.

2.2.3 Semi-Parametric

Most commonly used semi-parametric model is Cox Proportional Hazard

2.2.3.1 Cox proportional Hazard

Cox proportional hazard model is a regression model essentially used for investigating the impact of several covariates over the time taken for the specified event to happen.

The Cox proportional Hazards model equation is:

$$\boxed{h(t) = h_o(t)\ \exp(b_1X_1 + b_2X_2 + \ldots . b_pX_p)}$$

Where

h(t) : The expected hazard at time t, or the rate of suffering the event in the next instant

$h_0(t)$: The baseline hazard and represents the hazard when all the predictors (or independent variables) X1, X2, XP are equal to zero.

$$\frac{h(t)}{ho(t)} = \exp\left(b_1X_1 + b_2X_2 + \cdots + b_pX_p\right)$$

$$Log(\frac{h(t)}{ho(t)}) = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

Where:

$$\frac{h(t)}{ho(t)} = \text{Hazard ratio}$$

The right-hand side of the equation is like that of a linear regression. The estimated coefficients in the cox proportional hazards model, for example, b_1 can be interpreted as the change in expected log of the hazard ratio relative to one unit change in X_1 holding all the other predictors constant. Thus, the predictors have a multiplicative or proportional effect on the predicted hazard.

Cox proportional hazards model is a semi-parametric model where there are no assumptions regarding the baseline hazard ($h_0(t)$) are made, however, there are some assumptions regarding the covariates which are as follows:

1. **Non-Informative Censoring**

   To satisfy this assumption, the data should be collected such that the censoring of the data is not related to the probability of an event occurring.

2. **Constant hazard ratio over time (Proportional Hazards)**

   Consider a simple model with one binary covariate (maintained / not maintained; treated/ not treated), it is assumed that the hazard function of both the classes is proportional over time.

3. **Linear Relationship between ‘log hazard’ and the ‘Continuous Covariates’**

   As with any linear regression, the continuous covariates are assumed to have a linear relationship with the log of hazard.

The model also enables us to compare two participants in the dataset based on their hazard ratios. For better understanding this let us assume that the model has only one predictor, i.e., X1, then the modified Cox proportional hazard equation would be

$$h(t) = h_0(t)\exp(b_1 X_1)$$

Now for comparing two subjects in terms of their expected hazards, assume that the first subject has X_1 = m and the second subject has X_2 = n. From the above equation, we get the expected hazards for the two subjects as h(t) = h_0(t)exp (b_{1a}) and h(t) = h_0(t)exp (b_{1b}), respectively.

$$h_m(t) = h_0(t)\exp(b_1 m)$$

$$h_n(t) = h_0(t)\exp(b_1 n)$$

The hazard ratio between observation m and n is as follows

$$\frac{h_m(t)}{h_n(t)} = \frac{\cancel{h_0(t)}\exp(b_1 m)}{\cancel{h_0(t)}\exp(b_1 n)}$$

$$\frac{h_m(t)}{h_n(t)} = \exp(b_1(m - n))$$

As we can clearly infer from the above equation that the hazard ratio is independent of time(t) which is the second assumption of cox proportional hazards model.

Advantages:

1. The model does not make any assumptions regarding the baseline hazard function. This is an advantage because in many cases, either the true form of the hazard function is unknown.
2. The inference procedures of the Cox model can easily handle the right-censored data.

2.2.4 Comparison between Kaplan-Meier and Cox Regression

Kaplan-Meier	**Cox Regression**
Used to compute the current probability of survival conditional upon survival until now	Used to analyze impact of multiple independent factors and collective impact on survival
Usually one categorical covariate	Multiple categorical and continuous covariates
Can be stratified based on two or more groups in the categorical covariate	Results can be stratified accounting other covariates
No assumptions – Non-parametric test	Assumptions made – semi-parametric test
Summary does not present statistical output	Statistical results available

3. AIDS Clinical Trial Case

3.1 Statement of Problem/Purpose

3.2 Dataset Description

Name: AIDS Clinical Trials Group Study 320 Data (actg320.dat)

Size: 1151 Observations, 16 Variables

Source: AIDS Clinical Trials Group

Descriptive Abstract:

The data come from a double-blind, placebo-controlled trial that compared the three-drug regimen of indinavir (IDV), open-label zidovudine (ZDV) or stavudine (d4T) and lamivudine (3TC) with the two-drug regimen of zidovudine or stavudine and lamivudine in HIV-infected patients (Hammer et al., 1997). Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measure was time to AIDS-defining event or death. Because efficacy results met a pre-specified level of significance at an interim analysis, the trial was stopped early.

Data Description:

Name	**Description**	**Codes/Values**
Id	Identification Code	1-1156
Time	Time to AIDS diagnosis or death	Days
Censor	Event indicator for AIDS defining diagnosis or death	1 = AIDS defining diagnosis or death, 0 = Otherwise
Tx	Treatment indicator	1 = Treatment includes IDV, 0 = Control group (treatment regime without IDV)
Txgrp	Treatment group indicator	1 = ZDV + 3TC 2 = ZDV + 3TC + IDV 3 = d4T + 3TC 4 = d4T + 3TC + IDV
Strat2	CD4 stratum at screening	0 = CD4 <= 50 1 = CD4 > 50
Sex	Sex	1 = Male, 2 = Female
Raceth	Race/Ethnicity	1 = White Non-Hispanic 2 = Black Non-Hispanic 3 = Hispanic (regardless of race) 4 = Asian, Pacific Islander 5 = American Indian, Alaskan Native 6 = Other/unknown
Ivdrug	IV drug use history	1 = Never 2 = Currently 3 = Previously
Hemophil	Hemophiliac	1 = Yes, 0 = No

Karnof	Karnofsky Performance Scale	100 = Normal; no complaint no evidence of disease 90 = Normal activity possible; minor signs/symptoms of disease 80 = Normal activity with effort; some signs/symptoms of disease 70 = Cares for self; normal activity/ active work not possible
cd4	Baseline CD4 count	Cells/milliliter (derived from multiple measurements)
Priorzdv	Months of prior ZDV use	Months
Age	Age at Enrollment	Years

3.3 Dataset Manipulation

Treatment group indicator 3 and 4 were removed from the dataset since very few values of it existed in the data set.

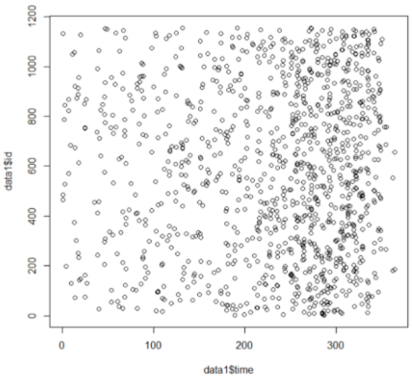
Name “Raceth” has been renamed to “Race” and values (1,2,3,4,5,6) have been changed to (“WNH”, “BNH”, “H”, “AI”, “A”, “U”).

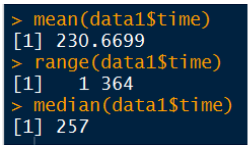
~~Values of sex (1,2) are changed to (“Male”, “Female”).~~

3.4 Data Exploration

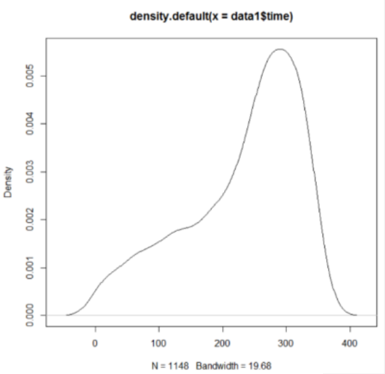
There is a total of 14~~15~~ variables in this dataset and 1151 observations. Exploring the important variables and checking a few interesting observations.

Time - Time to AIDS diagnosis or death – The mean time for this variable is found to be 230.67 days. The time variable spans between 1 day to 364 days in total and the median falls on the 257th day. To see how these patients time to diagnosis or death are distributed a plot is made for time against patient id as shown below.



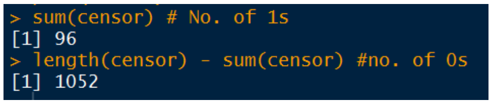


```
> mean(data1$time)
[1] 230.6699
> range(data1$time)
[1]   1 364
> median(data1$time)
[1] 257
```



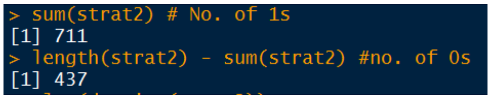
It is observed that most of the time to diagnosis or death falls towards the 200 – 360 days' bracket.

Censor - Event indicator for AIDS-defining diagnosis or death – This variable can take only binary values. There are 96 ones and 1052 zeros present.



```
> sum(censor) # No. of 1s
[1] 96
> length(censor) - sum(censor) #no. of 0s
[1] 1052
```

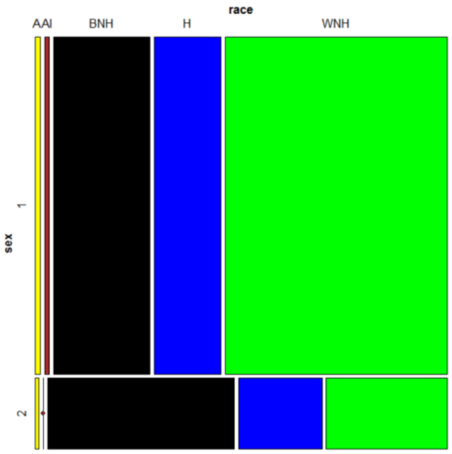
Tx - Treatment indicator – this is used to differentiate between the group whose treatment includes IDV from the other group which doesn't use it. There is a total of 572 patients treated with IDV and 576 without it. The distribution is not biased and uniformly distributed as shown below.



```
> sum(strat2) # No. of 1s
[1] 711
> length(strat2) - sum(strat2) #no. of 0s
[1] 437
```

as 0, else 1. There is a total of 711 cases where the CD4 count of patients was more than 50, with 437 being in a critical state of less than 50.

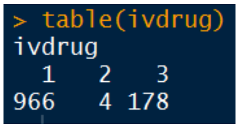
Race and **sex** – To understand the different race of people involved in this drug testing and the population gender, we can look at the visual below.



Sex 1= Males and 0 = Females. A total 949 males are part of this trial. Most the patients belong to the White Non-Hispanic population, with the Black Non-Hispanic being the next major race in the dataset. Among the females, the Black Non-Hispanic (BNH) race is the most dominant and among the males, White Non-Hispanic (WNH) race is the most dominant.

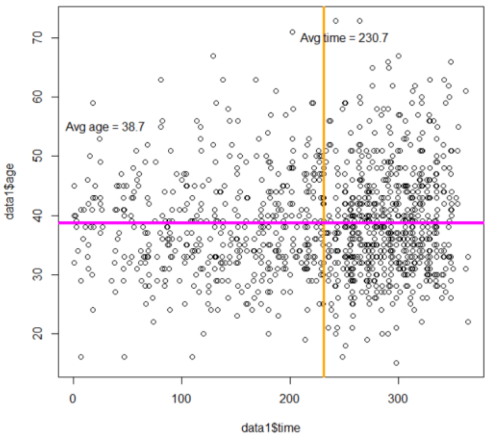
The below code displays a detail of the trial population with respect to race and gender.

Ivdrug – This column speaks of the historical usage of IV drug on them. There are 178 patients who were previously exposed to the IV drug, 4 of them are currently being exposed to it and remaining 966 patients were never exposed to it.



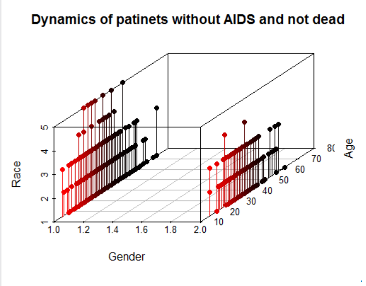
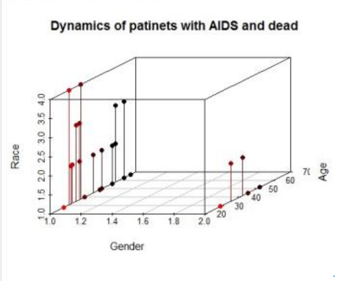
```
> table(ivdrug)
ivdrug
  1   2   3
966   4 178
```

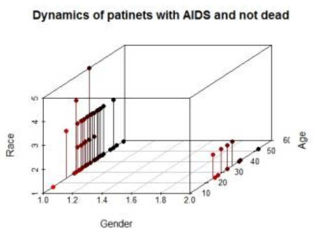
Age – The range of age of the population under trial is between 15 years to 73 years.



The time to diagnosis or death was more towards 250 days to 360 days, the average time to detect is found to be roughly 231 days. Even the population distribution as per age is denser between 30 years to 50 years with the mean age being about 39 years. The visual towards the left shows the age and population distribution.

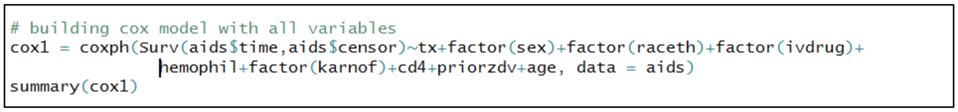
Interpretation of data regarding patients with and without AIDS based on race, gender, and age.

~~Plot~~	~~Observation~~
	~~Out of the patients without AIDS and who are not dead, Males are more than females. Most of them being white, Black, and Hispanic.~~
	~~Out of patients with AIDS and dead, we can see that there are more Males than females who are mostly White and Black.~~

	~~Out of patients who are diagnosed with AIDs and are still surviving, there are more White, Black and Hispanic.~~

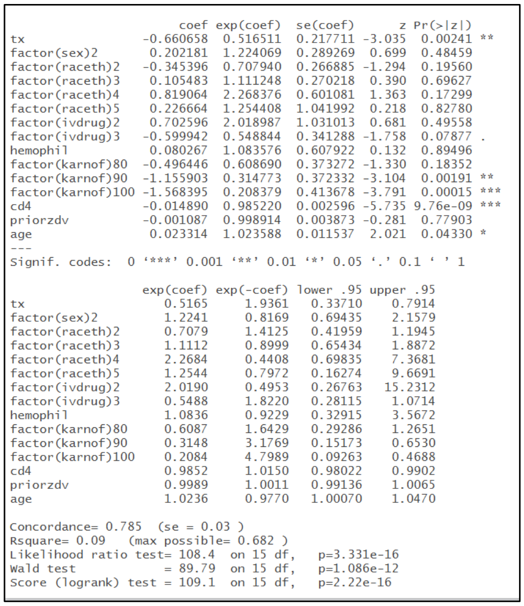
3.5 Modeling

The model fitting ~~are~~ Cox proportional hazards model with all the available covariates is given



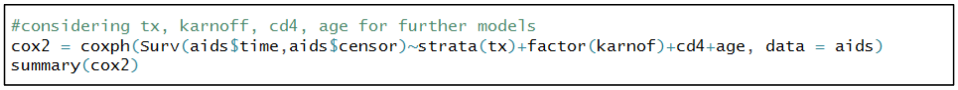
```
# building cox model with all variables
cox1 = coxph(Surv(aids$time,aids$censor)~tx+factor(sex)+factor(raceth)+factor(ivdrug)+
               hemophil+factor(karnof)+cd4+priorzdv+age, data = aids)
summary(cox1)
```

The summary statistics obtained are as follows:



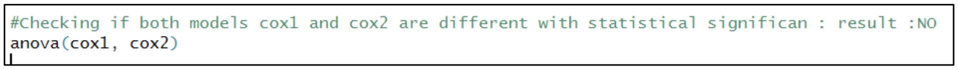
```
                       coef exp(coef)  se(coef)      z Pr(>|z|)    
tx                -0.660658  0.516511  0.217711 -3.035  0.00241 ** 
factor(sex)2       0.202181  1.224069  0.289269  0.699  0.48459    
factor(raceth)2   -0.345396  0.707940  0.266885 -1.294  0.19560    
factor(raceth)3    0.105483  1.111248  0.270218  0.390  0.69627    
factor(raceth)4    0.819064  2.268376  0.601081  1.363  0.17299    
factor(raceth)5    0.226664  1.254408  1.041992  0.218  0.82780    
factor(ivdrug)2    0.702596  2.018987  1.031013  0.681  0.49558    
factor(ivdrug)3   -0.599942  0.548844  0.341288 -1.758  0.07877 .  
hemophil           0.080267  1.083576  0.607922  0.132  0.89496    
factor(karnof)80  -0.496446  0.608690  0.373272 -1.330  0.18352    
factor(karnof)90  -1.155903  0.314773  0.372332 -3.104  0.00191 ** 
factor(karnof)100 -1.568395  0.208379  0.413678 -3.791  0.00015 ***
cd4               -0.014890  0.985220  0.002596 -5.735 9.76e-09 ***
priorzdv          -0.001087  0.998914  0.003873 -0.281  0.77903    
age                0.023314  1.023588  0.011537  2.021  0.04330 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
tx                   0.5165     1.9361   0.33710    0.7914
factor(sex)2         1.2241     0.8169   0.69435    2.1579
factor(raceth)2      0.7079     1.4125   0.41959    1.1945
factor(raceth)3      1.1112     0.8999   0.65434    1.8872
factor(raceth)4      2.2684     0.4408   0.69835    7.3681
factor(raceth)5      1.2544     0.7972   0.16274    9.6691
factor(ivdrug)2      2.0190     0.4953   0.26763   15.2312
factor(ivdrug)3      0.5488     1.8220   0.28115    1.0714
hemophil             1.0836     0.9229   0.32915    3.5672
factor(karnof)80     0.6087     1.6429   0.29286    1.2651
factor(karnof)90     0.3148     3.1769   0.15173    0.6530
factor(karnof)100    0.2084     4.7989   0.09263    0.4688
cd4                  0.9852     1.0150   0.98022    0.9902
priorzdv             0.9989     1.0011   0.99136    1.0065
age                  1.0236     0.9770   1.00070    1.0470

Concordance= 0.785  (se = 0.03 )
Rsquare= 0.09   (max possible= 0.682 )
Likelihood ratio test= 108.4  on 15 df,   p=3.331e-16
Wald test            = 89.79  on 15 df,   p=1.086e-12
Score (logrank) test = 109.1  on 15 df,   p=2.22e-16
```

By analyzing the summary statistics, it is observed that only a few variables contribute to the model (p-value < 0.05) which are Tx, Karnof, and CD4. A subsequent model is built considering only these variables. To compare the impact of treatments with IDV and without IDV the model is stratified based on Tx. Stratification subsets the dataset with respect to treatment and fits the hazard function for those individual receiving a three-drug treatment and individuals receiving a two-drug treatment(Placebo).



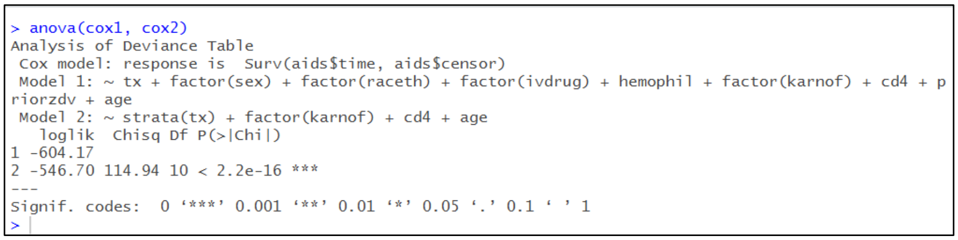
```
#considering tx, karnoff, cd4, age for further models
cox2 = coxph(Surv(aids$time,aids$censor)~strata(tx)+factor(karnof)+cd4+age, data = aids)
summary(cox2)
```

ANOVA is used to check if the first model with all the covariates and the model with limited covariates are statistically same,



```
#Checking if both models cox1 and cox2 are different with statistical significan : result :NO
anova(cox1, cox2)
```

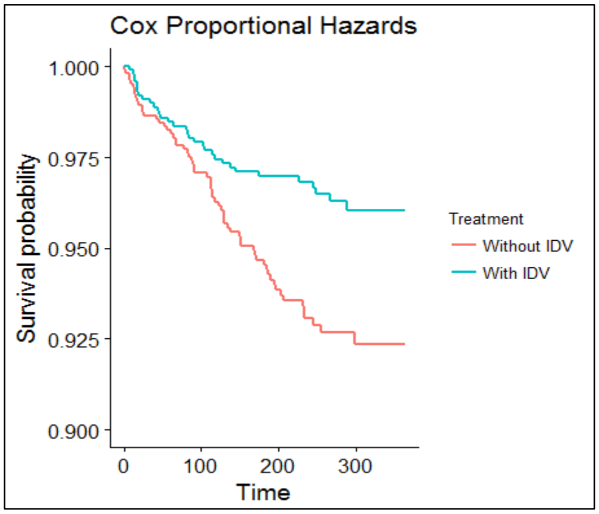
The results can be seen below



```
> anova(cox1, cox2)
Analysis of Deviance Table
 Cox model: response is  Surv(aids$time, aids$censor)
 Model 1: ~ tx + factor(sex) + factor(raceth) + factor(ivdrug) + hemophil + factor(karnof) + cd4 + p
riorzdv + age
 Model 2: ~ strata(tx) + factor(karnof) + cd4 + age
   loglik  Chisq Df P(>|Chi|)
1 -604.17
2 -546.70 114.94 10 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The p-value is insignificant which proves that both the models are not different statistically and hence the model with a limited number of variables is chosen for further analysis for simplicity of the model.

Using ggsurvplot function available is the survminer package the survival curves shown below are plotted



It can be clearly inferred from the graph that the population which has IDV in their treatment regime have a higher probability of survival, i.e., they had a higher chance of escaping the event than compared to the population that did not have IDV.

Since the Cox proportional Hazards model is based on three main assumptions it is very important to validate the assumptions from the results of the model thus far built.

1. **Non-Informative Censoring**:

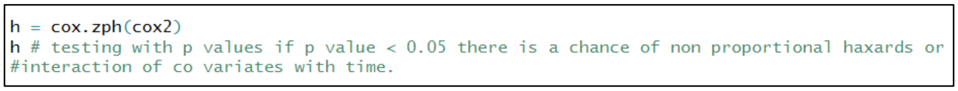
   The description given with the dataset says that the number of patients lost to follow-up is independent of the patient being diagnosed with aids. Hence the data collected during the study satisfies this assumption.

2. **Proportional Hazards Assumption**:

The assumption is satisfied if there the hazards ratio remains constant and does not vary with time for each of the covariates. This can be verified by analyzing the residual plots of the model. The survminer package provides both graphical and statistical methods to check for non-proportional hazards.

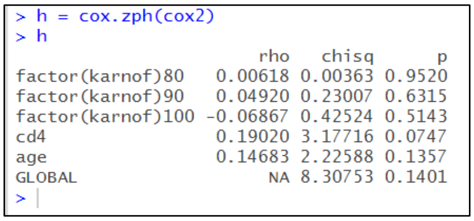
Cox.zph function can be used to check if any covariate contributes a non-proportional hazard

The results obtained were as follows:



```
h = cox.zph(cox2)
h # testing with p values if p value < 0.05 there is a chance of non proportional haxards or
#interaction of co variates with time.
```

Results of this test are shown below



```
> h = cox.zph(cox2)
> h
                        rho   chisq      p
factor(karnof)80    0.00618 0.00363 0.9520
factor(karnof)90    0.04920 0.23007 0.6315
factor(karnof)100 -0.06867 0.42524 0.5143
cd4                 0.19020 3.17716 0.0747
age                 0.14683 2.22588 0.1357
GLOBAL                   NA 8.30753 0.1401
> 
```

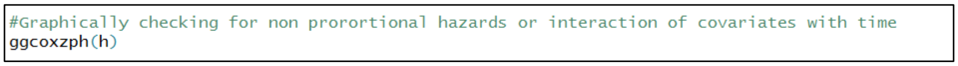
P-value of any of the covariates less than 0.05 suggests that there is an interaction between that particular covariate with time and the hazards are non-proportional (also, the ratio of the hazard is not constant). If this is the case a stratified cox proportional hazards model can be used to model such data where the variable having an interaction with time is stratified and a time interaction is accounted. If the variable is continuous then binning the variable and converting it to a categorical variable is suggested.

In the dataset considered above there are no covariates with a p-value less than 0.05 and hence the proportional hazards assumption is satisfied.

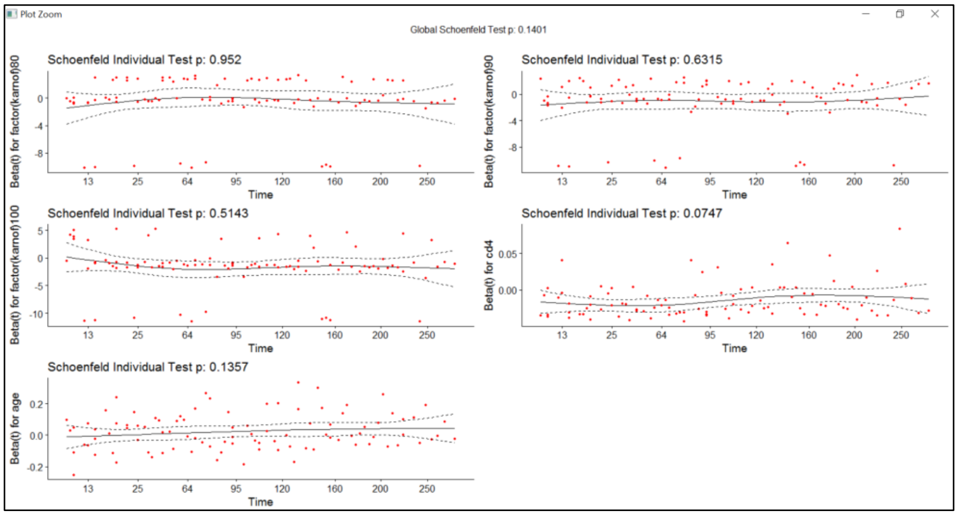
Graphical Approach:

The assumption can also be validated by analyzing the Schoenfeld residuals of each covariate.

The Schoenfeld residuals for each covariate can be plotted by using the ggcoxzph function.



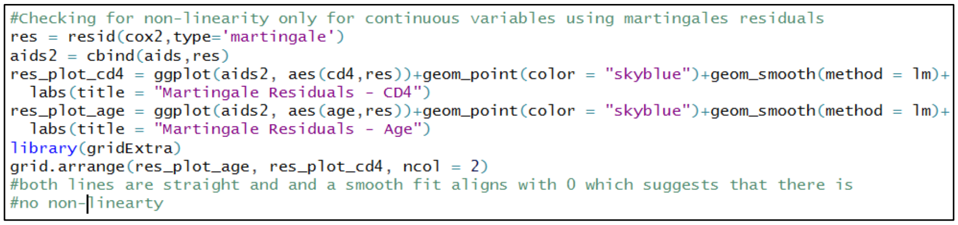
```
#Graphically checking for non prorortional hazards or interaction of covariates with time
ggcoxzph(h)
```



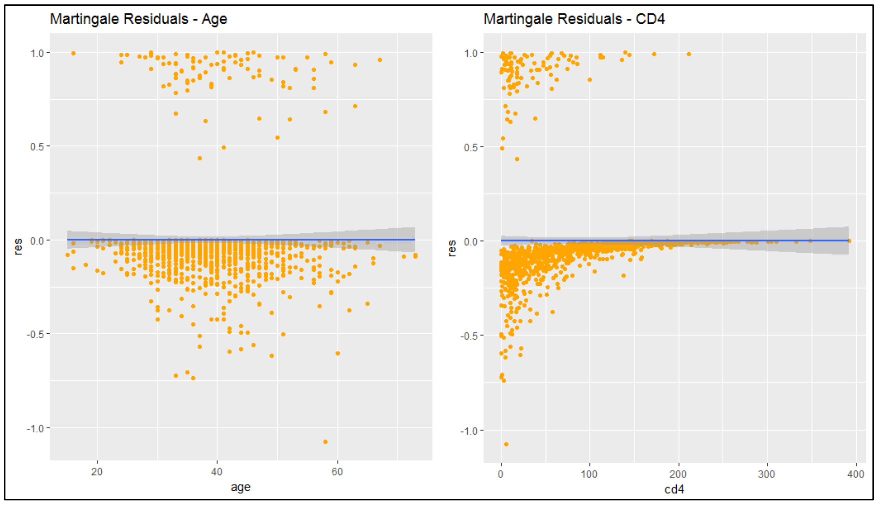
The hazards seem to be fairly linear with time, suggesting proportional hazards.

3. **Linear Relationship between 'log hazard' and the 'Continuous Covariates'**:

To satisfy this assumption, there should be a linear relationship between the residuals and the continuous covariates used in the model. The 'Martingale' residuals are plotted against the continuous covariates 'cd4' and 'age'.



```
#Checking for non-linearity only for continuous variables using martingales residuals
res = resid(cox2,type='martingale')
aids2 = cbind(aids,res)
res_plot_cd4 = ggplot(aids2, aes(cd4,res))+geom_point(color = "skyblue")+geom_smooth(method = lm)+
  labs(title = "Martingale Residuals - CD4")
res_plot_age = ggplot(aids2, aes(age,res))+geom_point(color = "skyblue")+geom_smooth(method = lm)+
  labs(title = "Martingale Residuals - Age")
library(gridExtra)
grid.arrange(res_plot_age, res_plot_cd4, ncol = 2)
#both lines are straight and and a smooth fit aligns with 0 which suggests that there is
#no non-linearty
```

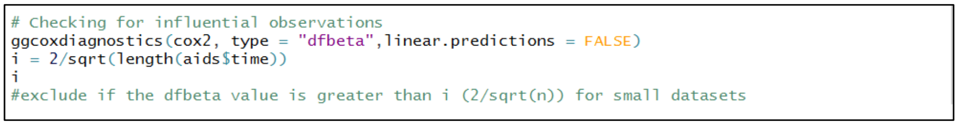


The blue line shows the lowess fit of the martingale residuals, the non-linearity can be detected if the lowess fit is parallel with the 0 in the above graphs. In the dataset considered above the lowess fit line is a straight line perfectly aligning with 0, thus it can be concluded that the covariates CD4 and age have a linear relationship with log of hazard.

3.6 Influential Observations

The model built is also checked to verify if there are any influential observations that might impact the goodness of the fit. The function 'ggcoxdiagnostics' is used to detect the influential observations. The argument "dfbeta", plots the estimated changes in the regression coefficients upon deleting each observation iteratively.

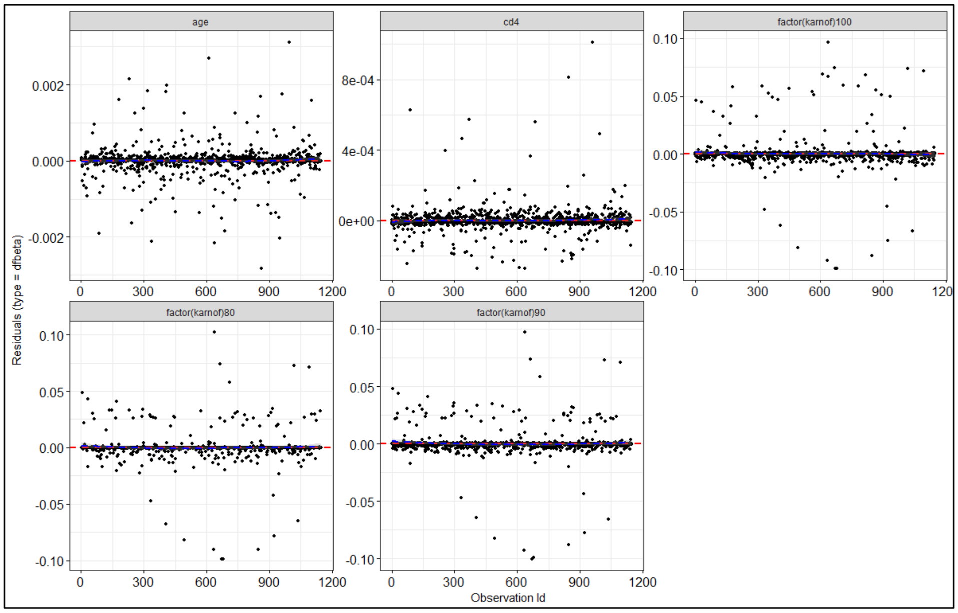
An observation is considered influential if the "dfbeta" value is greater than $\frac{2}{\sqrt{n}}$ where n is the number of observations in the dataset. For small datasets the criteria is less than 1.



```
# Checking for influential observations
ggcoxdiagnostics(cox2, type = "dfbeta",linear.predictions = FALSE)
i = 2/sqrt(length(aids$time))
i
#exclude if the dfbeta value is greater than i (2/sqrt(n)) for small datasets
```

Note: Influential observations are only considered for continuous variables and thus the emphasis of this test is of “age” and “CD4” variables only.

The plots if “dfbeta” values are shown below

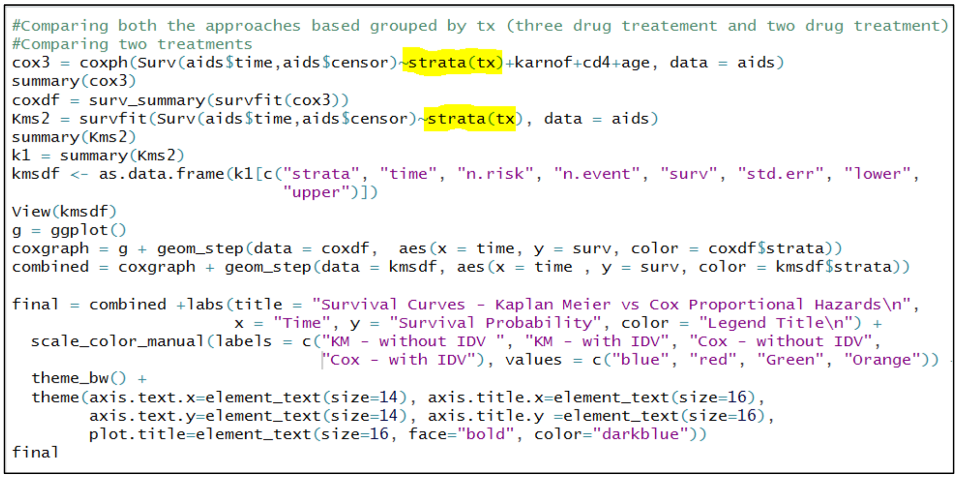


It can be inferred from the plots that none of the ‘dfbeta’ value is greater than ‘0.007876397’, it can be concluded that the dataset does not have any influential observation.

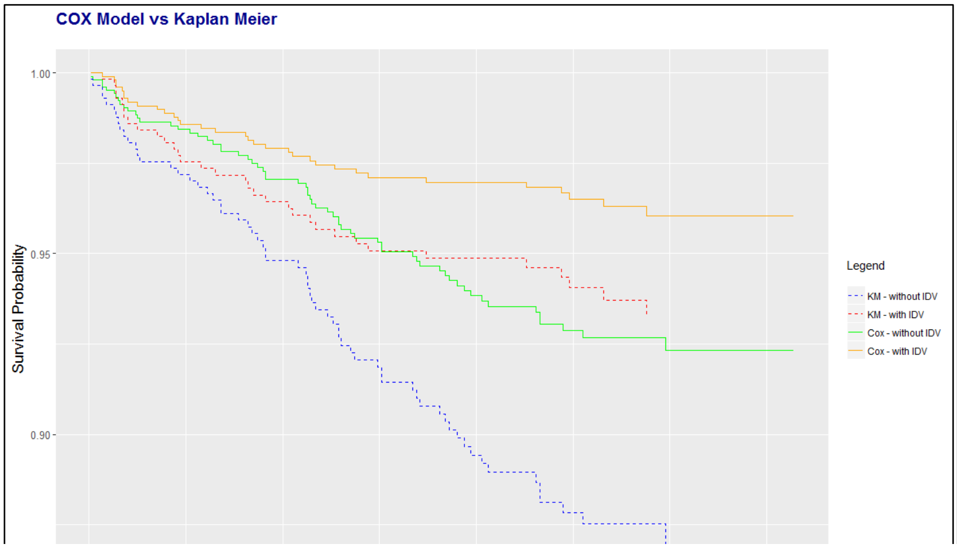
3.7 Model comparison

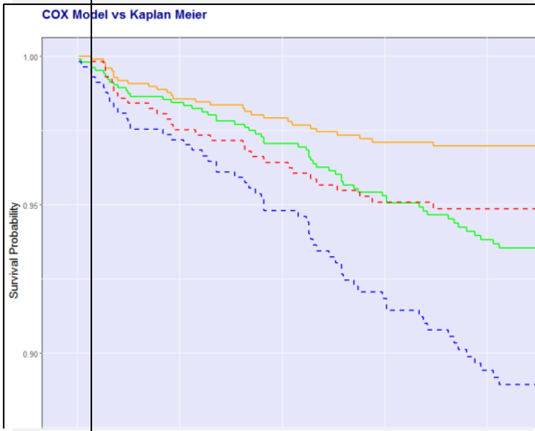
To compare the survival curves of ‘Cox Proportional’ and ‘Kaplan-Meier’ model grouping the individuals receiving three drug treatment and individuals receiving two-drug treatmen t(stratification on Tx). which in our study was divided into two sub-groups, i.e., one where the treatment regime included IDV (tx=1) and the other where the treatment regime was without IDV (tx=0).

A ggplot containing all the four survival curves is plotted to graphically see the difference between each of these curves.



```
#Comparing both the approaches based grouped by tx (three drug treatement and two drug treatment)
#Comparing two treatments
cox3 = coxph(Surv(aids$time,aids$censor)~strata(tx)+karnof+cd4+age, data = aids)
summary(cox3)
coxdf = surv_summary(survfit(cox3))
Kms2 = survfit(Surv(aids$time,aids$censor)~strata(tx), data = aids)
summary(Kms2)
k1 = summary(Kms2)
kmsdf <- as.data.frame(k1[c("strata", "time", "n.risk", "n.event", "surv", "std.err", "lower",
                            "upper")])
View(kmsdf)
g = ggplot()
coxgraph = g + geom_step(data = coxdf,  aes(x = time, y = surv, color = coxdf$strata))
combined = coxgraph + geom_step(data = kmsdf, aes(x = time , y = surv, color = kmsdf$strata))

final = combined +labs(title = "Survival Curves - Kaplan Meier vs Cox Proportional Hazards\n",
                       x = "Time", y = "Survival Probability", color = "Legend Title\n") +
  scale_color_manual(labels = c("KM - without IDV ", "KM - with IDV", "Cox - without IDV",
                                "Cox - with IDV"), values = c("blue", "red", "Green", "Orange"))
  theme_bw() +
  theme(axis.text.x=element_text(size=14), axis.title.x=element_text(size=16),
        axis.text.y=element_text(size=14), axis.title.y =element_text(size=16),
        plot.title=element_text(size=16, face="bold", color="darkblue"))
final
```



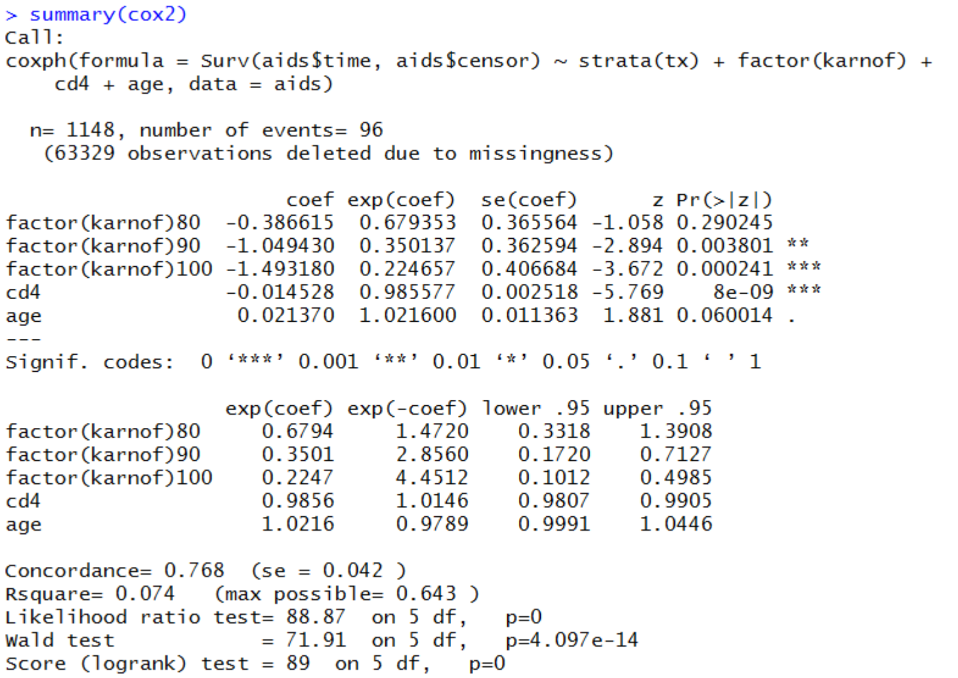


As we can interpret from the plot above, the fitting is better for COX Model as compared to Kaplan-Meier Model. Hence, we select COX Model in such scenarios.

4. Result and Conclusion

After thorough analysis of various models, the Cox Proportional Hazard model, with the predictor variables ‘Treatment Indicator (tx)’, ‘Karnofsky Performance Scale (Karnof)’, ‘Baseline CD4 count (cd4)’ and ‘Age at Enrollment (age)’ came out to be the best model.

The summary statistics of the final model are as follows:



```
> summary(cox2)
Call:
coxph(formula = Surv(aids$time, aids$censor) ~ strata(tx) + factor(karnof) +
    cd4 + age, data = aids)

  n= 1148, number of events= 96
   (63329 observations deleted due to missingness)

                       coef exp(coef)  se(coef)      z Pr(>|z|)
factor(karnof)80  -0.386615  0.679353  0.365564 -1.058 0.290245
factor(karnof)90  -1.049430  0.350137  0.362594 -2.894 0.003801 **
factor(karnof)100 -1.493180  0.224657  0.406684 -3.672 0.000241 ***
cd4               -0.014528  0.985577  0.002518 -5.769    8e-09 ***
age                0.021370  1.021600  0.011363  1.881 0.060014 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
factor(karnof)80     0.6794     1.4720    0.3318    1.3908
factor(karnof)90     0.3501     2.8560    0.1720    0.7127
factor(karnof)100    0.2247     4.4512    0.1012    0.4985
cd4                  0.9856     1.0146    0.9807    0.9905
age                  1.0216     0.9789    0.9991    1.0446

Concordance= 0.768  (se = 0.042 )
Rsquare= 0.074   (max possible= 0.643 )
Likelihood ratio test= 88.87  on 5 df,   p=0
Wald test            = 71.91  on 5 df,   p=4.097e-14
Score (logrank) test = 89  on 5 df,   p=0
```

Analyzing the summary statistics enabled us to make the following interpretations regarding the model:

Categorical Variables:

1. By using the three drug regime an individual's risk of being diagnosed with Aids reduces by 48.18 % when compared to an individual on two drug regime.
2. An individual having karnofsky performance scale index of 100 is at 0.22 times the risk of being diagnosed with aids than an individual having karnofsky index of 70.
3. For every 1 year increase in age the risk of being diagnosed with aids increases by 2.2%.

~~The hazard coefficients are obtained for ‘Karnofsky Performance Scales 80, 90 and 100’, while the coefficient results of ‘Karnofsky Performance Scale 70’ are omitted.~~

1. Since the hazard coefficient for ‘cd4’ is less than 1, it implies that for every one unit increase in the value of ‘cd4’, the probability of being diagnosed with aids would decrease by ‘1.4%’.
2. Since the hazard coefficient for ‘age’ is greater than 1, it implies that a year increase in ‘age’, would increase the risk ~~result in the probability~~ of being diagnosed with aids by ‘2.1%’.

Concordance Rate:

The ‘Concordance Rate’ of ‘0.768’ mean that the for a pair of subjects, the model would be correctly able to predict 76.8 % of the times the subject for which the event of interest would happen sooner. In our case, it means that around 76 out of 100 times, the model would be able to correctly predict which among the subjects would be ~~diagonised~~diagnosed with aids sooner.

Likelihood Ratio Test:

The likelihood ratio test is statistically comparing the model fits of the final model (the model with the selected predictors) with the baseline model (the model with no predictor variables), to check if the difference made by the predictor variables is statistically significant.

Since, the Likelihood Ratio test value is 88.87, which very high, we can conclude that the predictor variables are indeed enabling the model to fit better and thus, improve the predictive accuracy.

Wald Test:

The Wald Test is done to check the hypothesis that all the variable coefficients are simultaneously equal to zero.

5. Reference

1. ~~Reference:~~ http://data.princeton.edu/pop509/ParametricSurvival.pdf
2. https://cran.r-project.org/web/packages/survminer/survminer.pdf
3. https://cran.r-project.org/web/packages/survival/survival.pdf
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332/
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2940174/
6. http://www.diva-portal.org/smash/get/diva2:161225/FULLTEXT01.pdf
7. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival_print.html

6. APPENDIX

Statistical Modeling of Parametric Methods – Exponential Model, Weibull Model and Lognormal Model

Exponential: The exponential distribution has constant hazard $\lambda(t) = \lambda$. Thus, the survivor function is $S(t) = \exp\{-\lambda t\}$ and the density is $f(t) = \lambda \exp\{-\lambda t\}$. It 1 can be shown that $E(T) = 1/\lambda$ and $\mathrm{var}(T) = 1/\lambda 2$. Thus, the coefficient of variation is 1. The exponential distribution is related to the extreme-value distribution. Specifically, T has an exponential distribution with parameter λ, denoted $T \sim E(\lambda)$, iff $Y = \log T = \alpha + W$ where $\alpha = -\log \lambda$ and W has a standard extreme value (min) distribution, with density $fW\ (w) = e\ w-e\ w$. This is a unimodal density with $E(W) = -\gamma$, where $\gamma = 0.5722$ is Euler's constant, and $\mathrm{var}(W) = \pi\ 2/6$. The skewness is -1.14. The proof follows immediately from a change of variables.

Weibull: T is Weibull with parameters λ and p, denoted $T \sim W(\lambda, p)$, if $T\ p \sim E(\lambda)$. The cumulative hazard is $\Lambda(t) = (\lambda t)\ p$, the survivor function is $S(t) = \exp\{-(\lambda t)\ p\}$, and the hazard is $\lambda(t) = \lambda\ p\ ptp-1$. The log of the Weibull hazard is a linear function of log time with constant p $\log \lambda + \log p$ and slope $p - 1$. Thus, the hazard is rising if $p > 1$, constant if $p = 1$, and declining if $p < 1$. The Weibull is also related to the extreme-value distribution: $T \sim W(\lambda, p)$ iff $Y = \log T$

= $\alpha + \sigma W$, where W has the extreme value distribution, $\alpha = -\log \lambda$ and $p = 1/\sigma$. The proof follows again from a change of variables; start from W and change variables to $Y = \alpha + \sigma W$, and then change to T = e Y.

Lognormal: T has a lognormal distribution iff $Y = \log T = \alpha + \sigma W$, where W has a standard normal distribution. The hazard function of the log-normal distribution increases from 0 to reach a maximum and then decreases monotonically, approaching 0 as $t \to \infty$. As $k \to \infty$ the generalized extreme value distribution approaches a standard normal, and thus the generalized gamma approaches a log-normal.