Survival Analysis Project: HIV Clinical Trial

Juste Simanauskaite & Patricia Rivera

Contents

Introduction

HIV (Human Immunodeficiency Virus) is a disease known as an immune system disorder, which causes severe destruction of white blood cells that are responsible for fighting infection. The presence of this disorder is a lead-in for a human to be more prone to infections and cancer diseases. AIDS is the final stage of HIV, which is not always developed in HIV patients. Zidovudine (AZT) is known as antiretroviral medication for prevention of HIV/AIDS, whereas lamuvidine (3TC) is an inhibitor medication that works in decreasing HIV and hepatitis B. Previously, it has been founded that three-drug combinations, in particular, with a previous exposure to AZT, have shown the most significant resulted in reducing HIV-1 RNA concentrations. Therefore, this study used indinavir sulfate (a synthetic antiviral agent that inhibits HIV protease activity) in combination with AZT and 3TC as well as variation of placebo treatments to determine the potency of triple drug therapy in the cases of advanced HIV-1 patients. The study hypothesized that a three-drug combination, including a HIV-protease inhibitor and two nucleoside analogues (AZT and 3TC) would alter the progression of the HIV-1 disease. The study was successful in reaching significant data of the clinical superiority of a three-drug approach with inidavor over a treatment containing only a two-drug combination.

The current analysis of the data from a study conducted by Hammer et al. in 1997 considers the response variable to be *time*, which here describes the amount of time in days for the time of death, AIDS diagnosis, or the termination of the study. Another important variable used for the analysis is *censor*, which indicates the participants of the study that survived till the termination of the study without dying or being diagnosed AIDS. The study explored the influence of the explanatory variable *tx*. referring to the treatment group that was differentiated into: a control (placebo group) and a treatment group that included IDV (indinavir) # Methods

The study was a randomized, double-blind, and a placebo-controlled trial that compared a three-drug treatment of indinavir (Crixivan), zidovudine (AZT) and lamivudine (3TC) with a two-drug treatment. Patients were selected based on the factor that they had no more than 200 CD4 cells per cubic millimeter at least 3 months prior to AZT

therapy. The patients had to be more than 16 years old, with a diagnostic documentation of HIV-1 infection, having no more than 1 week of prior lamuvidine treatment, and a Karnofsky score of at least 70.

The approved patients received 200mg of open-label zidovudine three times daily and 150mg of lamuvidine two times daily and were randomly assigned to a placebo or a treatment of 800mg of indinavir every eight hours.

Some modifications were made to the protocol. In October of 1996 prior exposure to AZT was reduced to at least 3 months and permitted patients with no tolerance for this drug to enter the study with stavudine as a substitute.

Patients diagnosed with AIDS-defining events were offered an open-label assignment of the indinavir treatment with nor reveal of their initial treatment assignments. All of these cases had to be reviewed via a blind procedure by the study chair.

Follow ups were made at weeks 4,8, and 16 and every eight weeks afterwards. CD4 cell counts and Plasma HIV-1 RNA concentrations were measured twice at baseline and at weeks 4,8,24, and 40.

The statistical analysis methods used to interpret results were Kaplan-Meier estimates, log-rank tests, and proportional hazards models. The p-values, estimates of treatment differences and 95% confidence intervals were not adjusted for repeated analysis.

```
## [1] 851  16
```

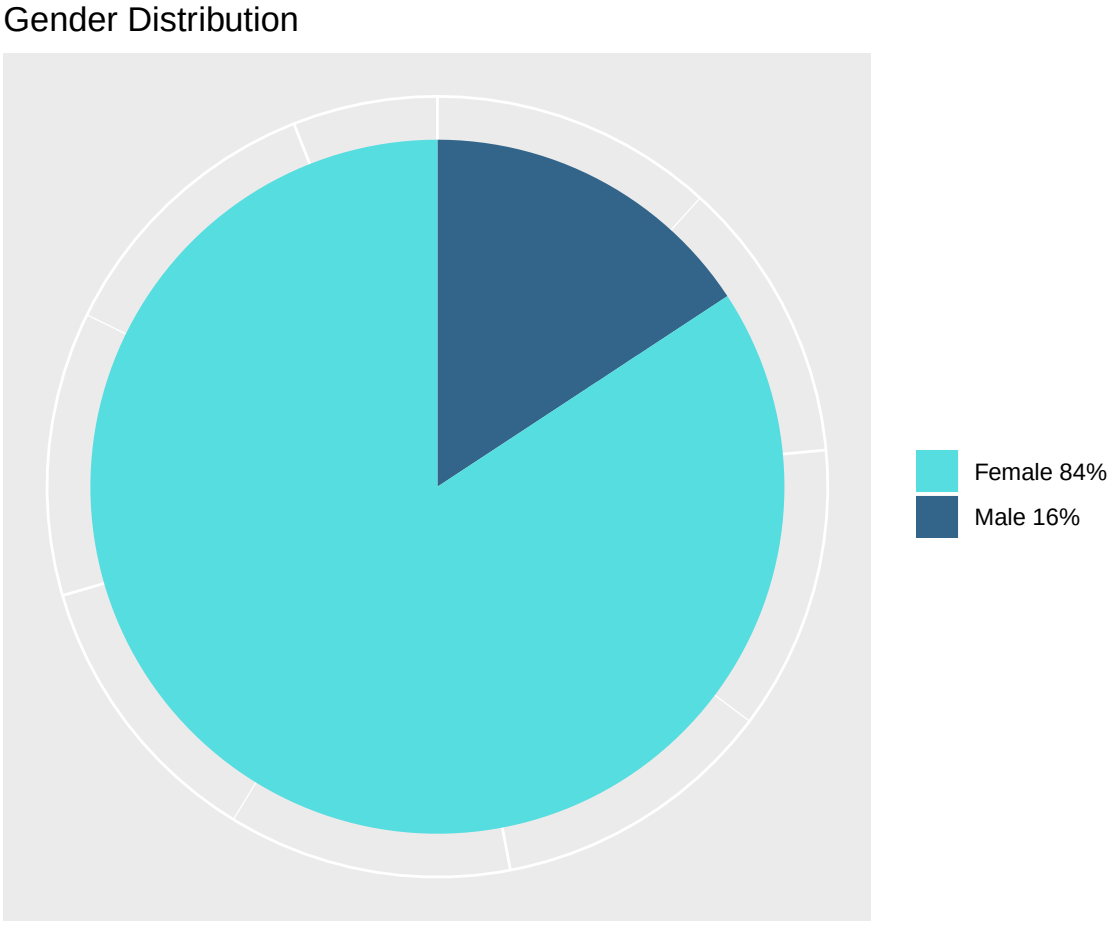
Data-Set Analysis

The data set contains a sample size equal to 851 participants and 16 different variables. Out of these participants 782 were considered as uncensored data point, which indicates that these patients survived through the course of the study without diagnosis of AIDS and/or death. 69 were found to be censored meaning that either there was an occurrence of death or AIDS diagnosis, out of which it is known that 20 patients died throughout the course of the study.

```
#Survival Analysis
#mutation of age
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")
aids <- aids %>%
  mutate(age = ifelse(age <= 20, "under20",
                              ifelse(age <=30, "20-30",
                                     ifelse(age <= 40, "30-40",
                                            ifelse(age <=50, "40-50",
                                                   ifelse(age <=60, "50-60",
                                                          ifelse(age <=70, "60-70",
                                                                 "over70")))))) %>%
  mutate(age = factor(age,
                      levels = c("under20", "20-30", "30-40","40-50", "50-60","60-70","over70")),
         sex = ifelse(sex == 2, "male","female"))

aids <- aids %>%
  mutate(cd4 = ifelse(cd4 <=50, "0-50",
                              ifelse(cd4 <=100, "50-100",
                                     ifelse(cd4 <= 150, "100-150",
                                            ifelse(cd4 <=200, "150-200",
                                                   ifelse(cd4 <=250, "200-250",
                                                          ifelse(cd4 <=300, "300-350", "350+")))))))
```

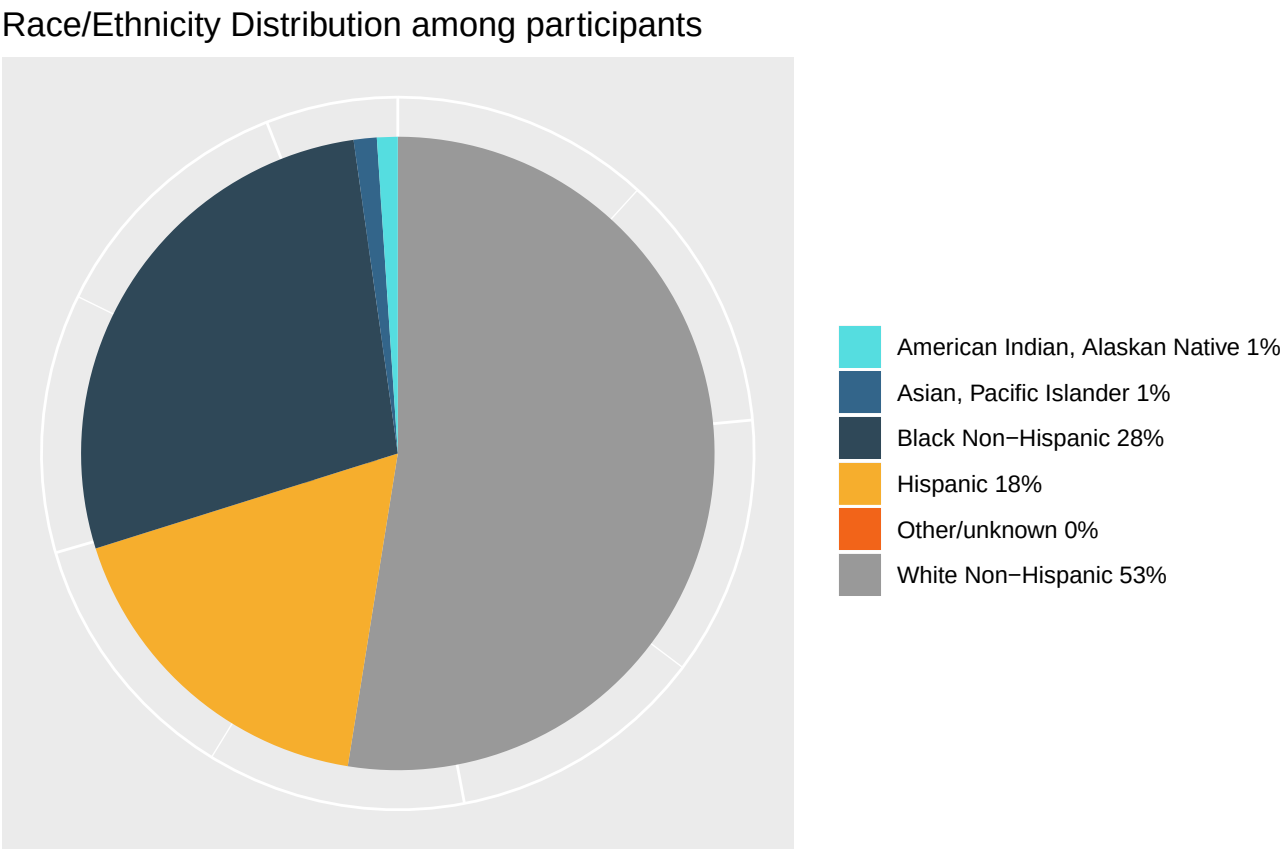
Since there are many values of the explanatory variable *age* in the original data, we've decided to mutate the variable into age categories from under 20 to over 70 in increments of 10 years. Similar modifications have been made to the baseline *CD4* count, just in increments of 50 up until 350+. Furthermore, we changed the labeling and representation of *sex* into "male" and "female" instead of "1" and "2" in the data.

```r
male<-sum(aids$sex=="male")
female<-sum(aids$sex=="female")
slices <- c(male, female)
lbls <- c("Male","Female")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
df = data.frame(slices = slices,labels =  lbls)
sexplot<- ggplot(df,aes(x = factor(1),y=slices,  fill = labels)) +
         geom_bar(stat="identity", width = 1)+
        coord_polar(theta = "y")+
        theme(axis.line = element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank())
print(sexplot + ggtitle("Gender Distribution")+
        scale_fill_manual(values=c("#55DDE0", "#33658A",
                                   "#2F4858"))+
        labs(x = NULL, y = NULL, fill = NULL))
```



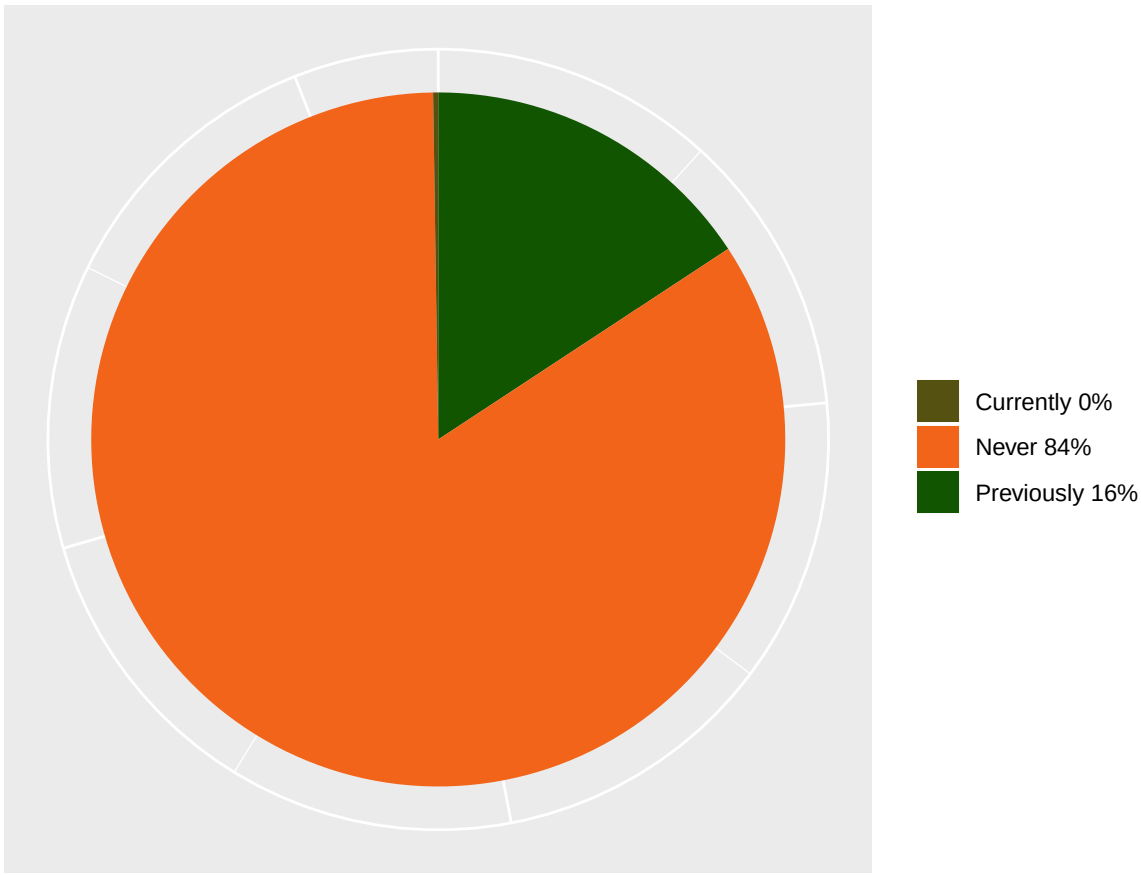
The Pie Chart represents the gender distribution in the sample, with 84% male and 16% female. This shows the potential for the data to not be able to correctly represent the difference of the data variance by gender, if there were to be one. Therefore, gender is something to look into in future data analysis.

```r
wnh<-sum(aids$raceth==1)
bnh<-sum(aids$raceth==2)
h<-sum(aids$raceth==3)
api<-sum(aids$raceth==4)
aian<-sum(aids$raceth==5)
oth<-sum(aids$raceth==6)
slices <- c(wnh,bnh,h,api,aian,oth)
lbls <- c("White Non-Hispanic", "Black Non-Hispanic", "Hispanic","Asian, Pacific Islander",
          "American Indian, Alaskan Native", "Other/unknown")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
df = data.frame(slices = slices,labels =  lbls)

ethplot<- ggplot(df,aes(x = factor(1),y=slices,  fill = labels)) +
         geom_bar(stat="identity", width = 1)+
        coord_polar(theta = "y")+
        theme(axis.line = element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank())
print(ethplot + ggtitle("Race/Ethnicity Distribution among participants")+
        scale_fill_manual(values=c("#55DDE0", "#33658A",
                                   "#2F4858", "#F6AE2D", "#F26419",
                                   "#999999"))+ labs(x = NULL, y = NULL,
                                                     fill = NULL))
```



The distribution of race/ethnicity shows that the greatest number of participants consists of white non-Hispanic identifying individuals, with black non-Hispanic following and Hispanic as the 3rd largest represented group.

```
never<-sum(aids$ivdrug==1)
cur<-sum(aids$ivdrug==2)
prev<-sum(aids$ivdrug==3)
slices3 <- c(never,cur,prev)
lbls3 <- c("Never", "Currently", "Previously")
pct3 <- round(slices3/sum(slices3)*100)
lbls3 <- paste(lbls3, pct3)
lbls3 <- paste(lbls3,"%",sep="")
df3 = data.frame(slices = slices3,labels =  lbls3)
ivplot<- ggplot(df3,aes(x = factor(1),y=slices3,  fill = labels)) +
         geom_bar(stat="identity", width = 1)+
        coord_polar(theta = "y")+
        theme(axis.line = element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank())
print(ivplot + ggtitle("IV Drug Use History")+
        scale_fill_manual(values=c( "#555111",
                                    "#F26419",
                                   "#125500")) +
        labs(x = NULL, y = NULL, fill = NULL))
```



From this chart we see that most of the participants (84%) have never used IV drugs, whereas 16% of participants have some type of history of usage and none of the participants reported to be currently using the drugs.

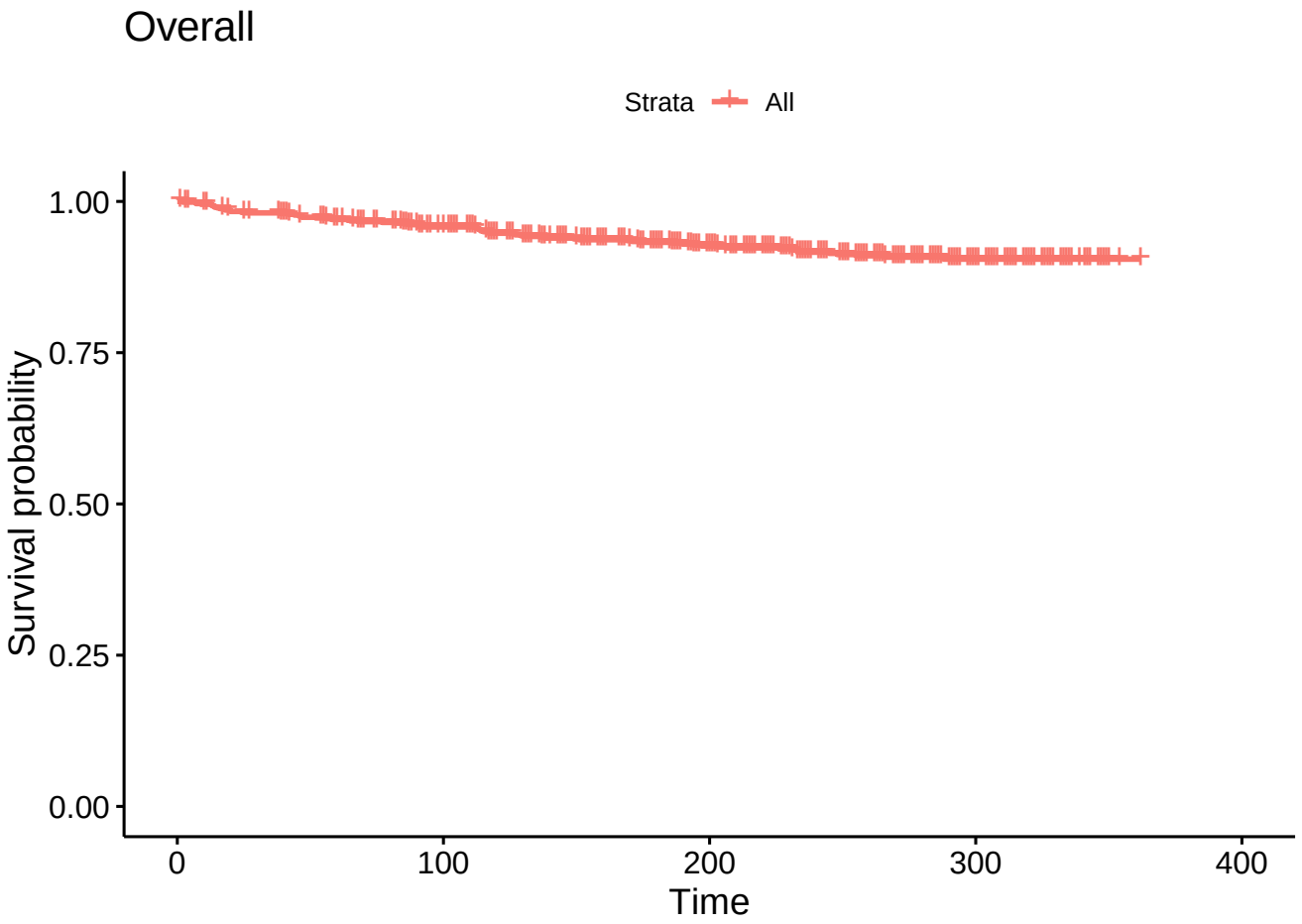
Results

Survival Analysis

Kaplan-Meier Curves

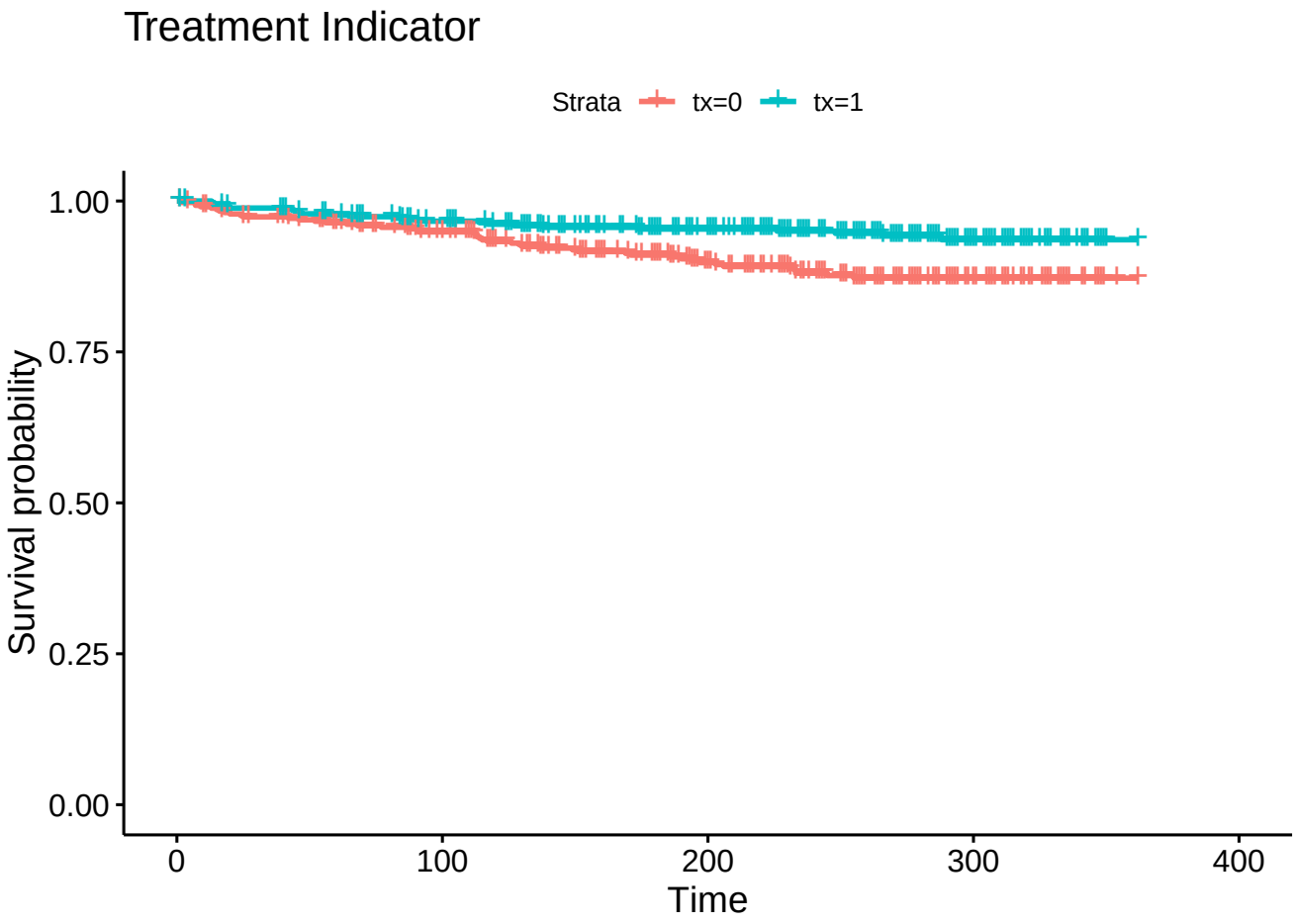
The following graph is a representation of a Kaplan Meier Curve for all participants in the study, we can see that only a few participants dies or were diagnosed with AIDS during the study as the slope of the curve is not experiencing a high decrease.

```
fit <- survfit(Surv(time,censor)~1, data = aids)
ggsurvplot(fit,data = aids,conf.int = FALSE) + ggtitle("Overall")
```



The following graph is a representation of the Kaplan-Meier survival probability based on the treatment indicator. In this case *tx=0* was the control group and *tx=1* the treatment group that was given IDV. Already, we can see a trend in the graph that the control group shows a lower survival probability with time. According to the log-rank test, we see that the p-value for the test statistic is equal to 0.002 (<0.05), thus we can reject the null hypothesis that the two population survival functions are the same, and the alternative is accepted, which says that the survival curves are different. The Wilcoxon test also provides us with a small p-value of 0.06 , which again rejects the null and goes in agreement with our primary conclusion.

```
fit1 <- survfit(Surv(time,censor)~tx, data = aids)
ggsurvplot(fit1,data = aids,conf.int = FALSE) + ggtitle("Treatment Indicator")
```




```
'Log-Rank'
```

```
## [1] "Log-Rank"
```

```
survdiff(Surv(time,censor)~tx, data = aids, rho=0)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ tx, data = aids, rho = 0)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## tx=0 422       46     33.3      4.83      9.35
## tx=1 429       23     35.7      4.51      9.35
##
##  Chisq= 9.3  on 1 degrees of freedom, p= 0.002
```

```
'Wilcoxon'
```

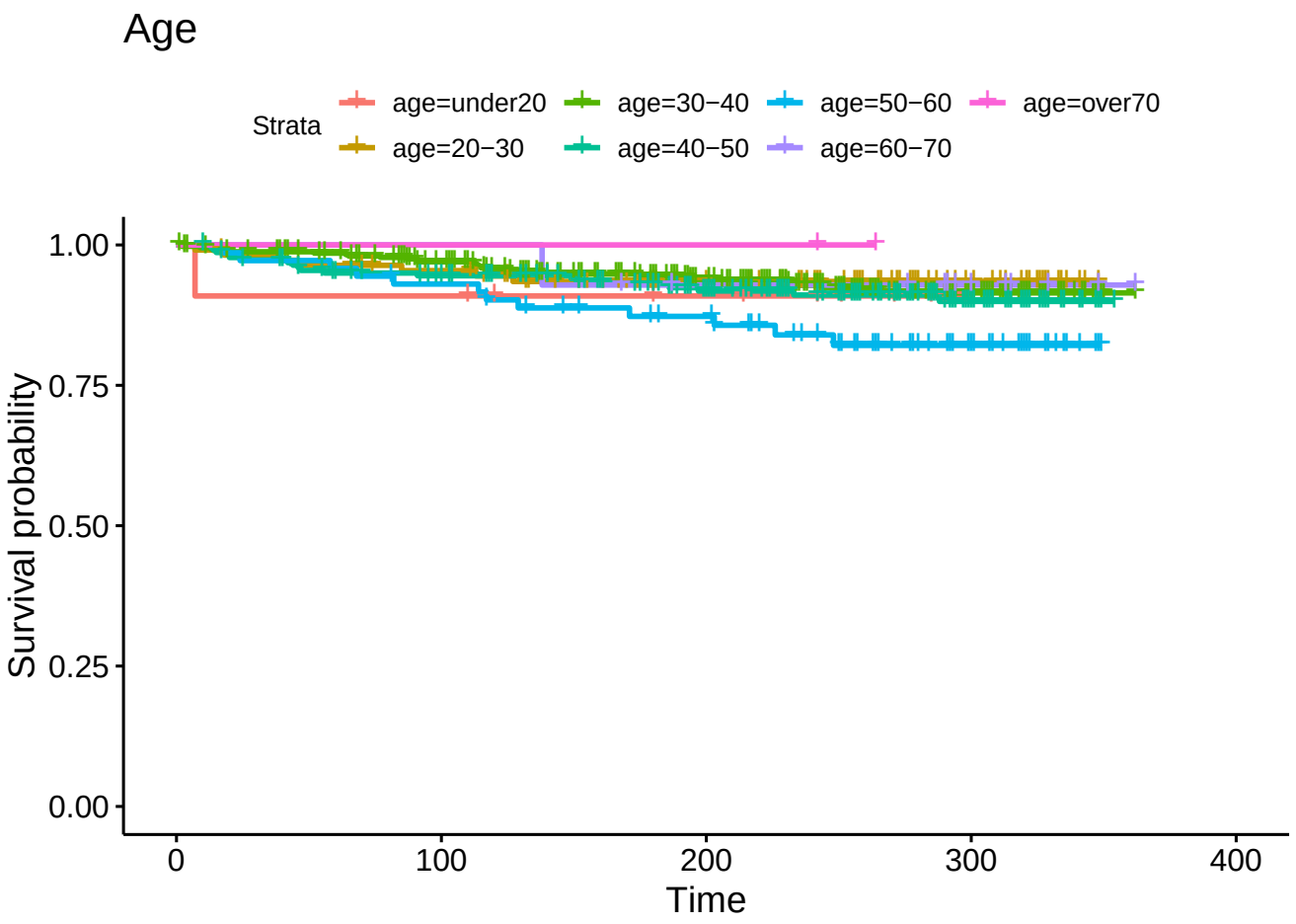
```
## [1] "Wilcoxon"
```

```
survdiff(Surv(time,censor)~tx, data = aids, rho=1)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ tx, data = aids, rho = 1)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## tx=0 422     44.0     31.9      4.57      9.24
## tx=1 429     22.1     34.2      4.28      9.24
##
##  Chisq= 9.2  on 1 degrees of freedom, p= 0.002
```

The Kaplan-Meier curves for survival probability based on *age* as an explanatory variable, with ether a grouped or ungrouped methods. If we leave the variable unchanged (ungrouped), according to log-rank and Wilcoxon tests, we get p-values of 0.008 and 0.009 respectively, (<0.05), thus rejecting the null of no difference, therefore it is considered a signifcant variable in the data set. (*This graph and resulting values are not shown due to the extensiveness of the data output*) However, with grouping of the age variable, we achieve a more consise, yet a nonsignificant Kaplan-Meier distribution, where the p-values in log-rank and Wilcoxon tests are equal to 0.3, therefore, failing to reject the null of no difference, indicating thhat it has no significant effect on the overall survival probability.

```
fit2 <- survfit(Surv(time,censor)~age, data = aids)
ggsurvplot(fit2,data = aids,conf.int = FALSE) + ggtitle("Age")
```




```
'Log-Rank'
```

```
## [1] "Log-Rank"
```

```
survdiff(Surv(time,censor)~age, data = aids, rho=0)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ age, data = aids, rho = 0)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=under20   11        1    0.824    0.0378    0.0383
## age=20-30    111        7    9.065    0.4703    0.5418
## age=30-40    416       29   33.543    0.6152    1.1980
## age=40-50    225       19   17.995    0.0561    0.0759
## age=50-60     72       12    6.098    5.7134    6.2712
## age=60-70     14        1    1.294    0.0669    0.0682
```

```
## age=over70    2        0    0.182    0.1817    0.1823
##
##  Chisq= 7.1  on 6 degrees of freedom, p= 0.3
```

```
'Wilcoxon'
```

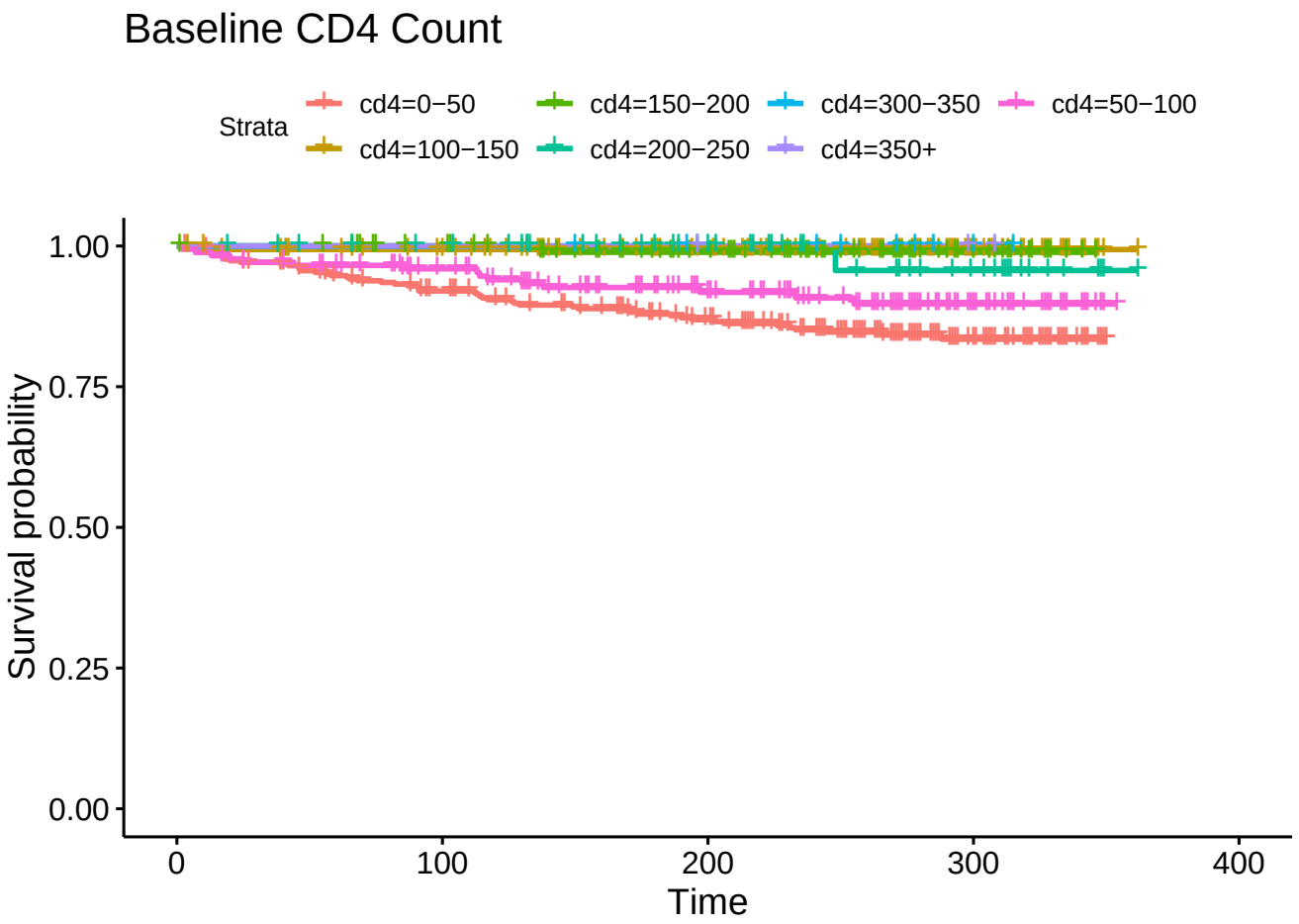
```
## [1] "Wilcoxon"
```

```
survdiff(Surv(time,censor)~age, data = aids, rho=1)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ age, data = aids, rho = 1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=under20   11    0.999    0.790    0.0553    0.0583
## age=20-30    111    6.812    8.684    0.4037    0.4850
## age=30-40    416   27.640   32.137    0.6292    1.2782
## age=40-50    225   18.280   17.238    0.0629    0.0889
## age=50-60     72   11.423    5.836    5.3495    6.1293
## age=60-70     14    0.941    1.236    0.0706    0.0753
## age=over70     2    0.000    0.174    0.1738    0.1821
##
##  Chisq= 7  on 6 degrees of freedom, p= 0.3
```

The following graph is a representation of the Kaplan-Meier survival probability based on the Baseline CD4. Due to the close proximity of the curves it's harder to see the significance of the differences. However, according to the log-rank test, we see that the p-value for the test statistic is equal to 5e-07 (<0.05), thus we can reject the null hypothesis, The Wilcoxon test again provides us with a small p-value of 5e-07, which again rejects the null and goes in agreement with our primary conclusion that the curves are significantly different.

```
fit3 <- survfit(Surv(time,censor)~cd4, data = aids)
ggsurvplot(fit3,data = aids,conf.int = FALSE) + ggtitle("Baseline CD4 Count")
```




```
'Log-Rank'
```

```
## [1] "Log-Rank"
```

```
survdiff(Surv(time,censor)~cd4, data = aids, rho=0)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ cd4, data = aids, rho = 0)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## cd4=0-50     346       51   28.279    18.256    30.974
## cd4=100-150 162        1   13.444    11.519    14.315
## cd4=150-200 104        1    8.467     6.585     7.511
## cd4=200-250  51        1    4.047     2.294     2.439
## cd4=300-350  10        0    0.882     0.882     0.894
## cd4=350+      3        0    0.275     0.275     0.276
## cd4=50-100  175       15   13.606     0.143     0.178
##
##  Chisq= 40  on 6 degrees of freedom, p= 5e-07
```

```
'Wilcoxon'
```

```
## [1] "Wilcoxon"
```

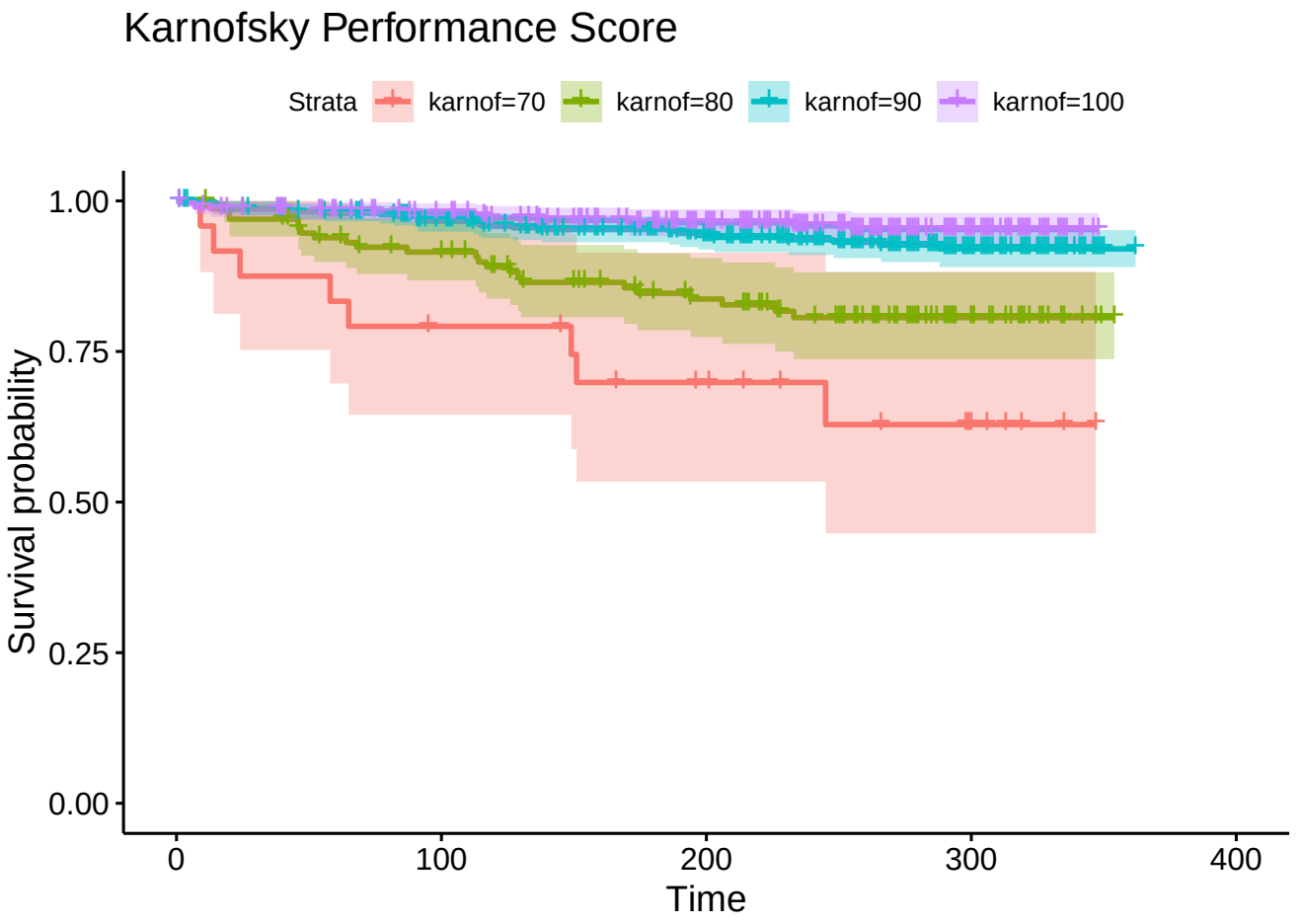
```
survdiff(Surv(time,censor)~cd4, data = aids, rho=1)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ cd4, data = aids, rho = 1)
##
```

```
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## cd4=0-50     346   48.834   27.072    17.494    30.954
## cd4=100-150  162    0.995   12.872    10.959    14.210
## cd4=150-200  104    0.942    8.114     6.339     7.540
## cd4=200-250   51    0.914    3.881     2.268     2.513
## cd4=300-350   10    0.000    0.844     0.844     0.893
## cd4=350+       3    0.000    0.263     0.263     0.276
## cd4=50-100   175   14.409   13.049     0.142     0.184
##
##  Chisq= 40  on 6 degrees of freedom, p= 5e-07
```

The final explanatory variable we're investigating as a part of our model is the Karnofsky Performance Scale. The Kaplan-Meier curves for this variable present a higher amplitude of distributions across the survival scale. The p-values of both log-rank and the Wilcoxon Test again present with the same significant p-value of 5e-10 (<0.05), thus we can reject the null hypothesis of no difference between the curves and consider this a significant variable in the construction of our model.

```
aids_fit_time_k <- survfit(Surv(time, censor) ~karnof   , data=aids)
ggsurvplot(aids_fit_time_k, data=aids,  conf.int = TRUE) +
  ggtitle("Karnofsky Performance Score")
```




```
'Log-Rank'
```

```
## [1] "Log-Rank"
```

```
survdiff(Surv(time,censor)~karnof, data = aids, rho=0)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ karnof, data = aids,
##     rho = 0)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## karnof=70   24        8      1.7     23.43     24.04
## karnof=80  133       23     10.4     15.26     17.98
## karnof=90  399       26     32.8      1.41      2.68
## karnof=100 295       12     24.1      6.08      9.36
##
##  Chisq= 46.2  on 3 degrees of freedom, p= 5e-10
```

```
'Wilcoxon'
```

```
## [1] "Wilcoxon"
```

```
survdiff(Surv(time,censor)~karnof, data = aids, rho=1)
```

```
## Call:
## survdiff(formula = Surv(time, censor) ~ karnof, data = aids,
##     rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## karnof=70   24     7.71     1.63     22.69     24.24
## karnof=80  133    22.02     9.97     14.57     17.90
## karnof=90  399    24.82    31.40      1.38      2.75
## karnof=100 295    11.55    23.09      5.77      9.26
##
##  Chisq= 46.3  on 3 degrees of freedom, p= 5e-10
```

To decide what variables to use in our Cox proportional hazards (PH) model, we can use backwards selection to determine what explanatory variables are most important to the model, along with the likelihood ratio test to compare differences in our models. Using *time_d* and *censor_d* as the response variables, we begin with the full model that includes the rest of the variables (minus time and censor) as the explanatory variables. We remove the variable with the highest p value first and create a new model without this variable. Using the likelihood ratio test between the two models, we then determine whether or not the variable added significance to the model. We continued this process until we arrived at the best model which uses *tx*, *karnof*, *cd4* and *age* as the explanatory variables.

```
### COX PH MODEL USING BACKWARDS SELECTION ####
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")

#full model
cp_full<- coxph(Surv(time,censor)~.-time_d -censor_d, data = aids)
cp_full$loglik
```

```
## [1] -452.6325 -410.5565
```

```
cp_full
```

```
## Call:
## coxph(formula = Surv(time, censor) ~ . - time_d - censor_d, data = aids)
##
##                coef  exp(coef)   se(coef)      z        p
## id        0.0005002  1.0005003  0.0003626  1.380 0.167711
## tx       -0.7164183  0.4884988  0.2583962 -2.773 0.005562
## txgrp            NA         NA  0.0000000     NA       NA
## strat2    0.2334775  1.2629844  0.4034960  0.579 0.562834
## sex       0.3963112  1.4863318  0.3218751  1.231 0.218226
## raceth   -0.0133058  0.9867823  0.1415732 -0.094 0.925121
```

```
## ivdrug   -0.2748505  0.7596857  0.1855538 -1.481 0.138541
## hemophil  0.5493026  1.7320447  0.6099731  0.901 0.367835
## karnof   -0.0595861  0.9421544  0.0142836 -4.172 3.02e-05
## cd4      -0.0174343  0.9827168  0.0048476 -3.596 0.000323
## priorzdv -0.0013102  0.9986907  0.0047307 -0.277 0.781815
## age       0.0242606  1.0245573  0.0135542  1.790 0.073471
##
## Likelihood ratio test=84.15  on 11 df, p=2.311e-13
## n= 851, number of events= 69
```

```
#reduced model 1
cp_red1<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug, data=aids)
cp_red1$loglik
```

```
## [1] -452.6325 -411.7891
```

```
#likelihood ratio test and p-value
s1 <- 2*(cp_full$loglik[2]-cp_red1$loglik[2])
1-pchisq(s1,1)
```

```
## [1] 0.116385
```

```
#reduced model 2
cp_red2<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv,
                data=aids)
cp_red2$loglik
```

```
## [1] -452.6325 -411.8488
```

```
#likelihood ratio test and p-value
s2 <- 2*(cp_red1$loglik[2]-cp_red2$loglik[2])
1-pchisq(s2,1)
```

```
## [1] 0.7297236
```

```
#reduced model 3
cp_red3<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth, data=aids)
cp_red3$loglik
```

```
## [1] -452.6325 -411.9199
```

```
#likelihood ratio test and p-value
s3 <- 2*(cp_red2$loglik[2]-cp_red3$loglik[2])
1-pchisq(s3,1)
```

```
## [1] 0.7061007
```

```
#reduced model 4
cp_red4<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth -strat2, data=aids)
cp_red4$loglik
```

```
## [1] -452.6325 -412.1975
```

```
#likelihood ratio test and p-value
s4 <- 2*(cp_red3$loglik[2]-cp_red4$loglik[2])
1-pchisq(s4,1)
```

```
## [1] 0.4562002
```

```
#reduced model 5
cp_red5<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
```

```
                -raceth -strat2 -hemophil, data=aids)
cp_red5$loglik
```

```
## [1] -452.6325 -412.5191
```

```
#likelihood ratio test and p-value
s5 <- 2*(cp_red4$loglik[2]-cp_red5$loglik[2])
1-pchisq(s5,1)
```

```
## [1] 0.4225607
```

```
#reduced model 6
cp_red6<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth -strat2 -hemophil -sex, data=aids)
cp_red6$loglik
```

```
## [1] -452.6325 -413.1175
```

```
#likelihood ratio test and p-value
s6 <- 2*(cp_red5$loglik[2]-cp_red6$loglik[2])
1-pchisq(s6,1)
```

```
## [1] 0.2739592
```

```
#reduced model 7
cp_red7<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth -strat2 -hemophil -sex -id, data=aids)
cp_red7$loglik
```

```
## [1] -452.6325 -413.8789
```

```
#likelihood ratio
s7 <- 2*(cp_red6$loglik[2]-cp_red7$loglik[2])
1-pchisq(s7,1)
```

```
## [1] 0.2171892
```

```
#reduced model 8
cp_red8<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth -strat2 -hemophil -sex -id, data=aids)
cp_red8$loglik
```

```
## [1] -452.6325 -413.8789
```

```
#likelihood ratio
s8 <- 2*(cp_red7$loglik[2]-cp_red8$loglik[2])
1-pchisq(s8,1)
```

```
## [1] 1
```

```
#reduced model 9
cp_red9<- coxph(Surv(time,censor)~.-time_d -censor_d -txgrp -ivdrug -priorzdv
                -raceth -strat2 -hemophil -sex -id -age, data=aids)
cp_red9$loglik
```

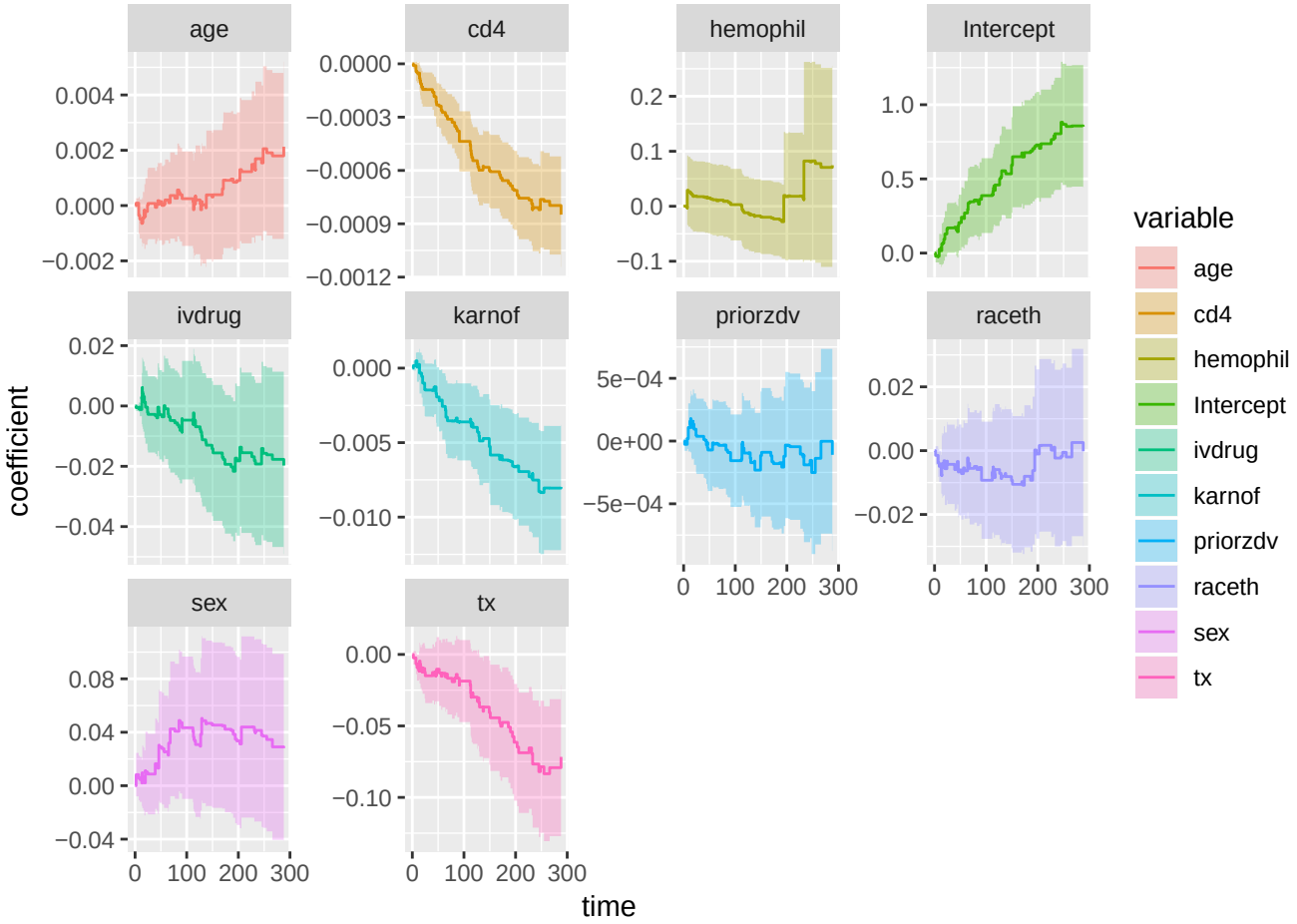
```
## [1] -452.6325 -415.0276
```

```
#likelihood ratio
s9 <- 2*(cp_red8$loglik[2]-cp_red9$loglik[2])
1-pchisq(s9,1)
```

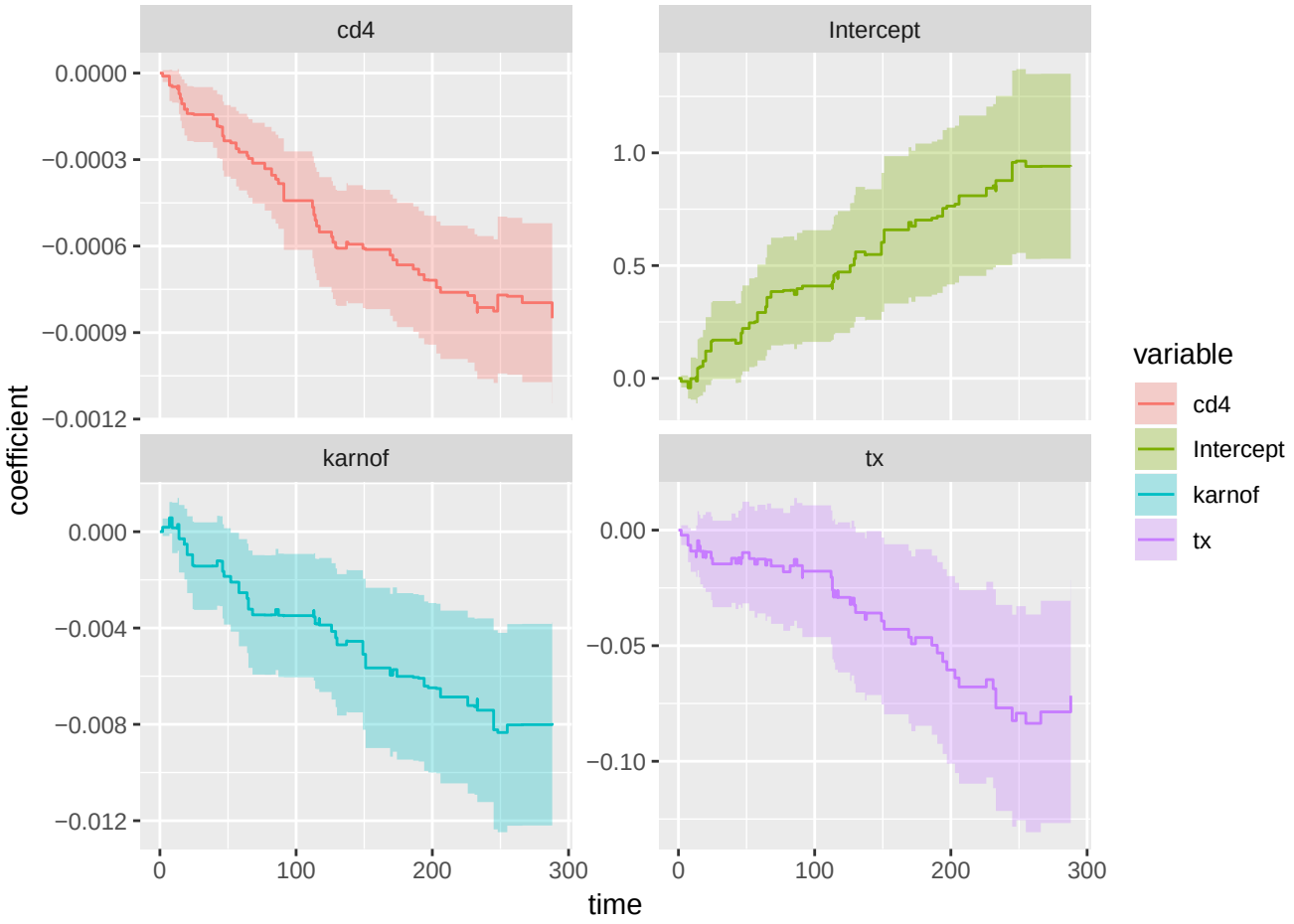
```
## [1] 0.1295928
```

To better understand how much the model fit changes with each different explanatory variable, we can use the graphical representation of the Aalen additive regression model. The Aalen model allows for time-varying covariate effects, while the Cox model allows only a common time-dependence through the baseline. In the Aalen model, we have the weighted comparisons of the crude estimate of the hazard rate of each group as compared to a baseline group, which here is defined as the estimate. As we can see, the selected explanatory variables in our model all have an inverse coefficient correlation with the baseline interceipt. The slope of an estimated cumulative regression function is positive when covariate increases and this fact correspond to an increasing hazard rate. On the other hand, if the slope is negative while the covariate increases, then this fact points to a decreasing hazard rate.

```
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")

aa_fit <-aareg(Surv(time, censor) ~ cd4 + karnof+ priorzdv +hemophil +raceth +sex +tx +ivdrug + age,
               data = aids)

autoplot(aa_fit,xlab="Coefficient", ylab="Time") + labs(x = "time", y = "coefficient")
```




```
aa_fit2 <-aareg(Surv(time, censor) ~ cd4 + karnof+ tx , data = aids)

autoplot(aa_fit2) + labs(x = "time", y = "coefficient")
```



The Aalen model assumes that the cumulative hazard $H(t)$ for a subject can be expressed as $a(t) + \mathrm{X}\ B(t)$, where $a(t)$ is a time-dependent intercept term, X is the vector of covariates for the subject possibly time-dependent, and $B(t)$ is a time-dependent matrix of coefficients.

The plots show how the effects of the covariates change over time.

Patricia's "Something New": Power analysis using simulated survival data

Power analysis is an important aspect of experimental design. Power tells us how often we can correctly reject the null hypothesis. It is useful in helping us determine how large our sample size must be in order to correctly detect if there is an effect, with a certain level of confidence. Furthermore, it can help us determine the probability with which we will correctly detect an effect given that we have a known sample size. Overall power analysis is important when conducting experiments because without a high level of power, experiments and studies would not receive funding from research centers, such as the the NIH (National Institute of Health).

In relation to survival analysis, we can simulate survival data and set it as the alternative hypothesis. Using power analysis, we can then determine how often we would be expected to correctly reject our null hypothesis. The code below simulates survival data using the beta coefficients from our Cox PH model. The Cox PH model showed us that *tx*, *karnof* and *cd4* were the most significant explanatory variables to predict time until death. Using their respective beta coefficients in the simulation, we can calculate the probability we would reject the null hypothesis given that the alternative hypothesis is the data we have at hand.

Power Analysis code and simulation

```
#Mean and standard
m <-c(mean(aids$tx), mean(aids$karnof), mean(aids$cd4))#, mean(aids$age))
s <-c(sd(aids$tx), sd(aids$karnof), sd(aids$cd4))#, sd(aids$age))

#Simulating survival data using coefficients from Cox PH model
set.seed(1234)
n.reps <- 100
simoutput <- c()
for(i in 1:n.reps){
  simdata <- sim.survdata(N=851, T=362, num.data.frames=1, censor= 0.9764,xvars=3, mu=m, sd=s, beta = c(-0
  model <- coxph(Surv(y, failed) ~ X1 + X2 + X3, data = simdata$data)
  simoutput <- rbind(simoutput, cbind(rep = rep(i, 3), model %>% tidy()))
}


#Power for the first variable: tx
simoutput%>%dplyr::filter(term=="X1")%>%dplyr::summarize(sum(p.value<0.05))
```

```
##   sum(p.value < 0.05)
## 1                  55
```

```
#Power for the second variable: karnof
simoutput%>%dplyr::filter(term=="X2")%>%dplyr::summarize(sum(p.value<0.05))
```

```
##   sum(p.value < 0.05)
## 1                   8
```

```
#Power for the third variable: cd4
simoutput%>%dplyr::filter(term=="X3")%>%dplyr::summarize(sum(p.value<0.05))
```

```
##   sum(p.value < 0.05)
## 1                   7
```

```
#simoutput%>%dplyr::filter(term=="X4")%>%dplyr::select(estimate)%>%hist(breaks=200)
```

From the output above, we can see that the power of each of the variables is not very big. The biggest power obtained is 55%, which comes from the *tx* variable. For the other two variables, power is very small. We can also see that *tx* has the largest beta coefficient and the largest power. For bigger beta coefficients we get a larger power because there is a larger effect size. In general, a power of 80% or more is desirable and given that most of our variables return an insignificant power, we would not reject our null hypothesis most of the time. To increase power, we would either have to increase our sample size, n, or increase the effect size.

References Used:

1. https://cran.r-project.org/web/packages/coxed/vignettes/simulating_survival_data.html
2. http://www.icssc.org/documents/advbiosgoa/tab%2026.00_survss.pdf

Juste's "Something New": The Schoenfeld Residuals for the Cox PH model

Cox proportional hazards (PH) model is considered a great way to identify combined effects of several covariates on the relative risk (hazard). This model assumes that the hazards of the different strata formed by the levels of the covariates are proportional at a particular point in time. This proportional hazards assumption is particularly important and can be tested via three different classes of tests. The first class is focused on the piece-wise estimation of models for subsets of data defined by stratification of time. The second one considers the interactions between covariates and some function of time. Final, third one is based on examinations of regression residuals. The

Schoenfeld Residuals are a part of the third class of proportional hazard assumption testing and I will be exploring it in order to be able to check for the validity of the PH assumption in the current and future data set analyses. This topic is particularly important in relation to survival analysis since it provides an idea of whether the model is appropriate for the data set at hand and whether some covariates should be considered as variants of time in order to supply the best model for prediction of proportional hazards. Taking a completely new model of analyzing survival data is particularly difficult since the mathematical derivations and notations are also very varied from what we have seen in class. Although, I do remember some of the ideas behind parametric functions, their applications to statistical models are much more challenging than I have expected. Therefore, it will require me a lot of time and extensive research to be able to understand and learn how to apply this model to our data and other instances of survival analysis.

Resources to learn about the topic.

I have been researching articles and scientific journals that provide insights into the Schoenfeld residuals and their use in the Cox PH model. Sources include:

1. https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12176
2. https://rstudio-pubs-static.s3.amazonaws.com/39354_34153ff19e624116bd2fbdec7d2534aa.html

Explanation of the Theory Behind Schoenfeld Residuals

Let $z_{ij}(t)$ be the j^{th} covariate of the i^{th} unit, where $i = 1, 2, ..., n$ and $j = 1, 2..., p$

This notation indicates that z_{ij} is a vector *1 x p* of covariates for unit *i*, which each can be either of fixed time or varying time, furthermore, here β is a *1 x p* vector of coefficients.

1) As we know from lecture, the Cox PH model assumes that $h(t)$ of the i^{th} individual satisfies:

- $h_i(t) = h_0(t)e^{z_i(t)\beta}$ where:
- h_0 -> baseline hazard
- $z_i(t)$ -> 1 x p vector of covariates for unit i each of which can be time fixed or time-varying.

2) However, another possibility has been presented by Therneau and Granbsh in 2000, where they proposed an idea that there could be an alternative to the current Cox model, where the coefficient of the estimate could also be varying as a function of time.

The new hazard function would look like this: $h_i(t) = h_0(t)e^{z_i(t)\beta(t)}$

Therefore, in order to examine thee two models in a case when $\beta = \beta(t)$ requires a residual analysis that could indicate whether a model should consider a covariate as a variable with time.

Due to the fact that that some observations might be censored and in particular, regarding the Cox PH model, the baseline hazard is not estimated, in order to analyse the residuals a particular score process. The risk score for unit i at time t is thought to be $r_i(t) = e^{z_i(t)\beta}$, where $Y_i(t)$ is the indicator function and $Y_i(t) = 1$ indicates a point in which i is under risk and thus observation and it is equal to 0 in other occasions.

Looking at the notation provided by Therneau and Grambsch (2000), we can provide the Schoenfeld residuals at the k^th event time t_k as:

1. $s_k = Z_{(k)} - \frac{\sum_i Y_i(t_k) r_i(t_k) Z_i(t_k)}{\sum_i Y_i(t_k) r_i(t_k)}$
2. $s_k = Z_{(k)} - \bar{z}(\hat{\beta}, t_k)$

In this case, the $Z(k)$ is the covariate vector of the particular unit that is experiencing the event at time k; $\hat{\beta}$ is the estimate of β and $\bar{z}(\hat{\beta}, t_k)$ is the weighted mean of covariate values.

Furthermore, the weighted variance can be represented by the derived equation at the k^{th} time as

$$V(\beta, t_k) = \frac{\sum_i Y_i(t_k) r_i(t_k) Z_i(t_k) - \bar{z}(\hat{\beta}, t_k)' Z_i(t_k) - \bar{z}(\hat{\beta}, t_k)}{\sum_i Y_i(t_k) r_i(t_k)}$$

From this, we can scale the Schoenfeld residuals by $V(\beta, t_k)$ of X at t_k via the equation:

$s_k^* = V^{-1}(\hat{\beta}, t_k) s_k$

The scaled Schoenfeld residuals can also be defined as follows:

$s_k^* = m \sum_{k=1}^{d} V(\hat{\beta}, t_k) s_k$

here, m is the total number of deaths in the data set.

Following the calculations, the residuals are plotted against time in order to test the proportional hazards assumption. If the assumption is correct, the residuals should be fitting around the line centered at zero (y=0). The further away this predicted line is form the horizontal of (y=0) the more likely one is to call the PH assumption to question and determine whether it is met through the model.

Interpretation of Schoenfeld Residuals: theoretical & graphical (R graphs and the p-values presented.)

– The Schoenfeld residuals are used to examine the model fit and detect outlying covariate values. Shoenfeld residuals represent the difference between the observed covariate and the expected given the risk set at that time. They should be flat, centered about zero in order for the Proportional Hazards assumption to be true. –

– Furthermore, the Schoenfeld residuals are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption in regards to time because Schienfeld individuals should be independent of them. The PH assumption is supported when there's a non-significant relationship between residuals and time. –

HIV Data Cox PH model analysis using Schoenfeld Residuals

Schoenfeld Residuals applied to our best Cox PH model for AIDS data where, we have an additive model of explanatory variables: baseline CD4 count, age,treatment group, and karnofsky performance scale score:

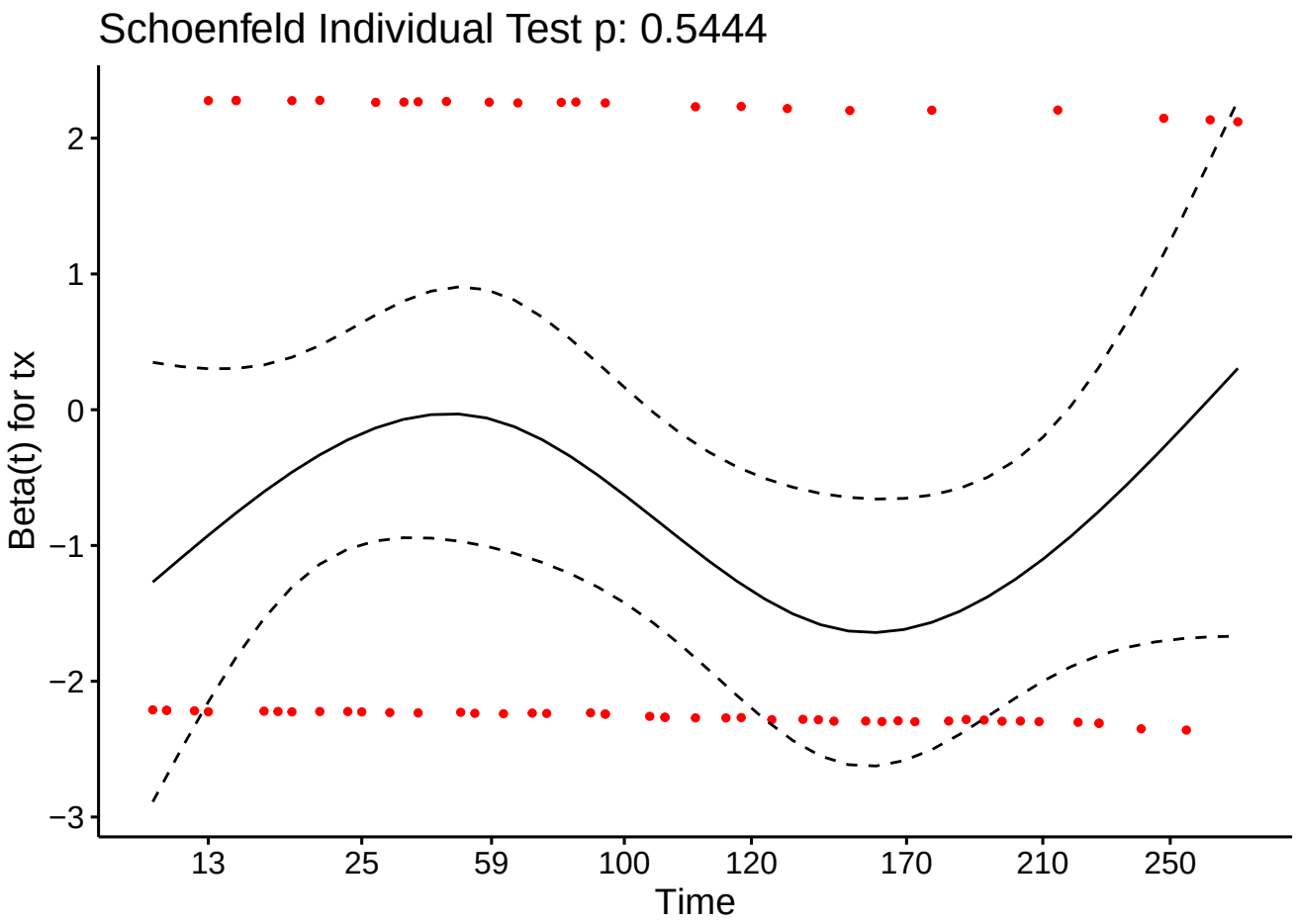
The *cox.zph()* function provides a Goodness-of-Fit (GOF) test, which tests the correlation between Schoenfeld resuduals and survival time. Here, taking with an alpha level of $\alpha = 0.05$ the p-value below this would reject the null of independence, thus indicating that the residuals are infact dependent on time and thus the PH assumption is not satisfied. If the p-value is greater than α, the null of independence is accepted and thus the PH assumption is met.

A graphical represenation is achieved using either *ggcoxzph()* or *ggcoxdiagnostics()* functions, that overall should show no pattern in the graphs in order to indicate a PH assumption. In the *ggcoxzph()* function, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a +/- 2-standard-error. Whereas, in *ggcoxdiagnostics()* function graph, the dashed blue line represents the fit to the plot, via red dashed line representing the y=0 point of reference and the grey area aroud the blue line representing thhe +/- 2-standard-error.

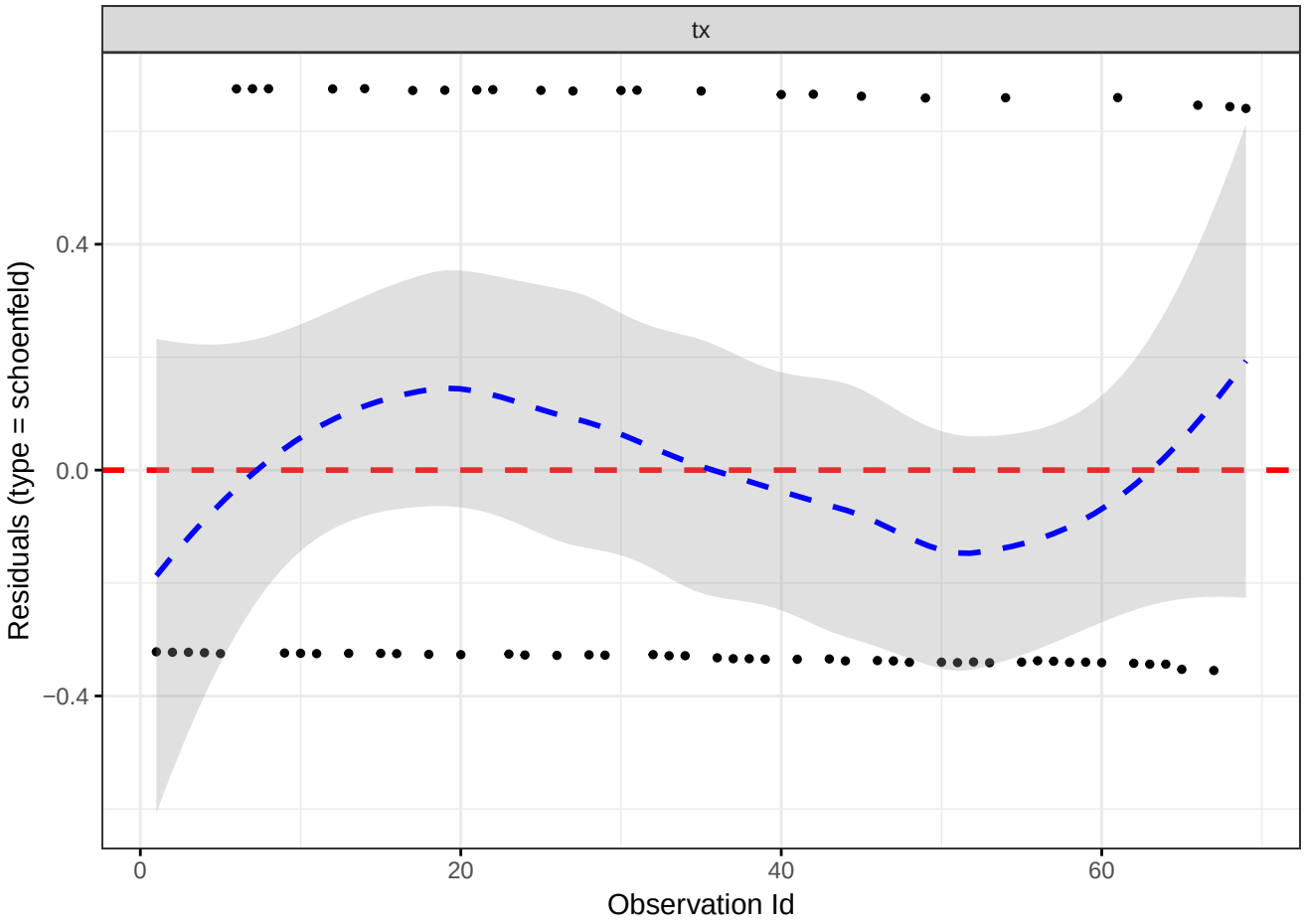
```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3, fig.width=5,
                      fig.align="center")
cph_tx <- coxph(Surv(time,censor)~ tx, data = aids)
zph_tx <- cox.zph(cph_tx)
zph_tx
```

```
##       rho chisq     p
## tx -0.073 0.367 0.544
```

```
ggcoxzph(zph_tx, point.size = 1, point.shape = 19)
```



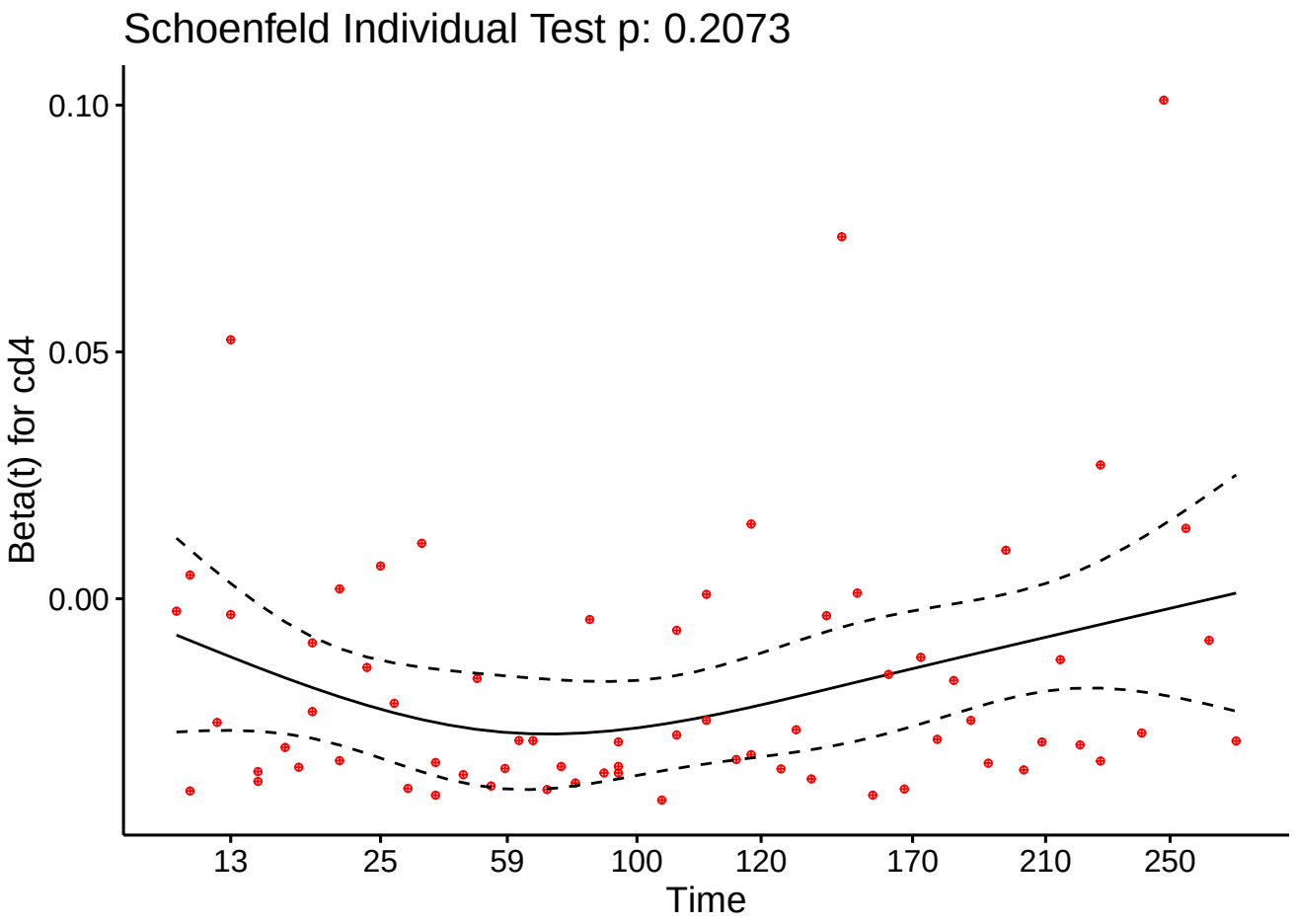

```
ggcoxdiagnostics(cph_tx, type="schoenfeld")
```



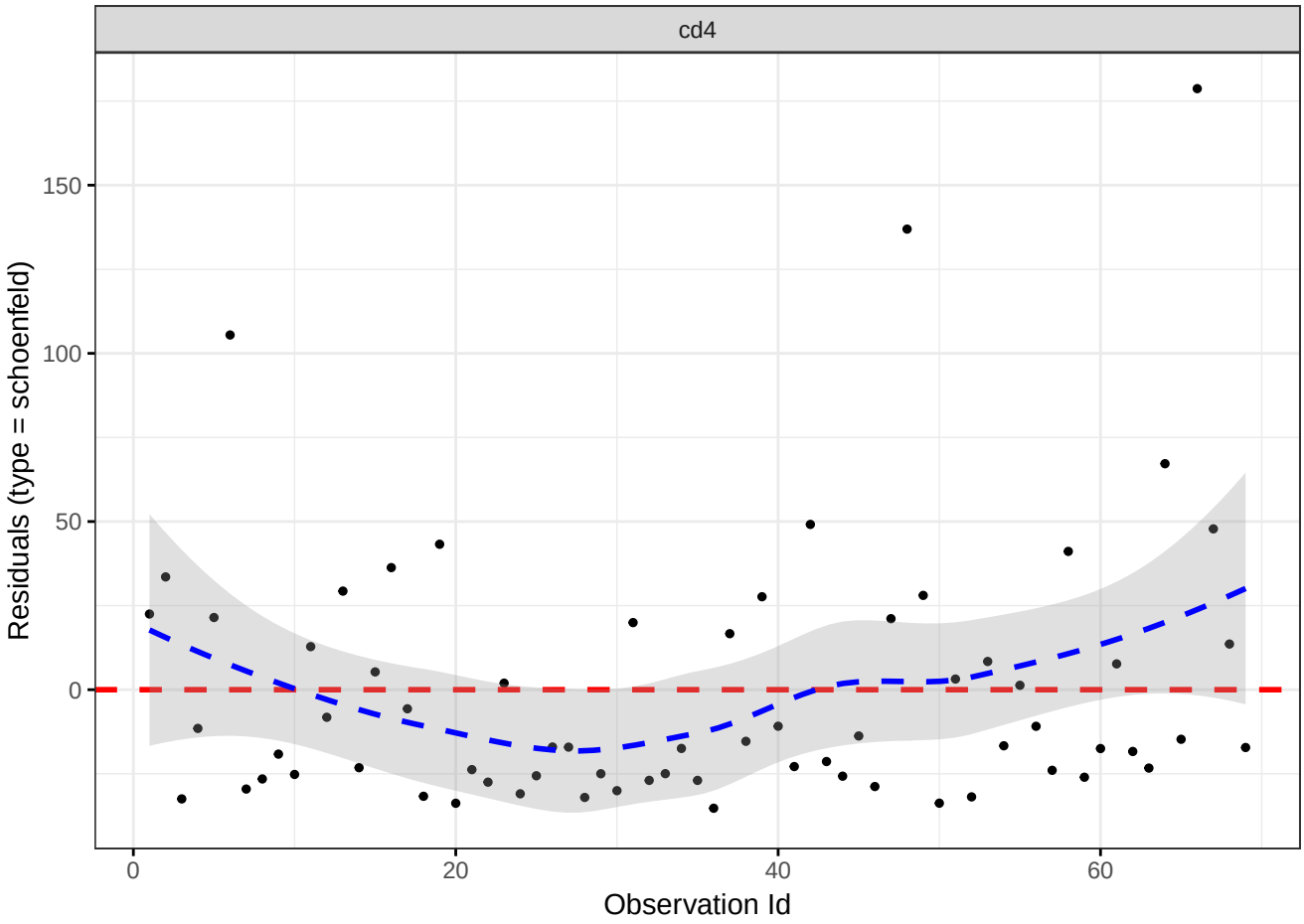

```
cph_cd4 <- coxph(Surv(time,censor)~ cd4, data = aids)
zph_cd4 <- cox.zph(cph_cd4)
zph_cd4
```

```
##      rho chisq     p
## cd4 0.15  1.59 0.207
```

```
ggcoxzph(zph_cd4, point.size = 1, point.shape = 10)
```



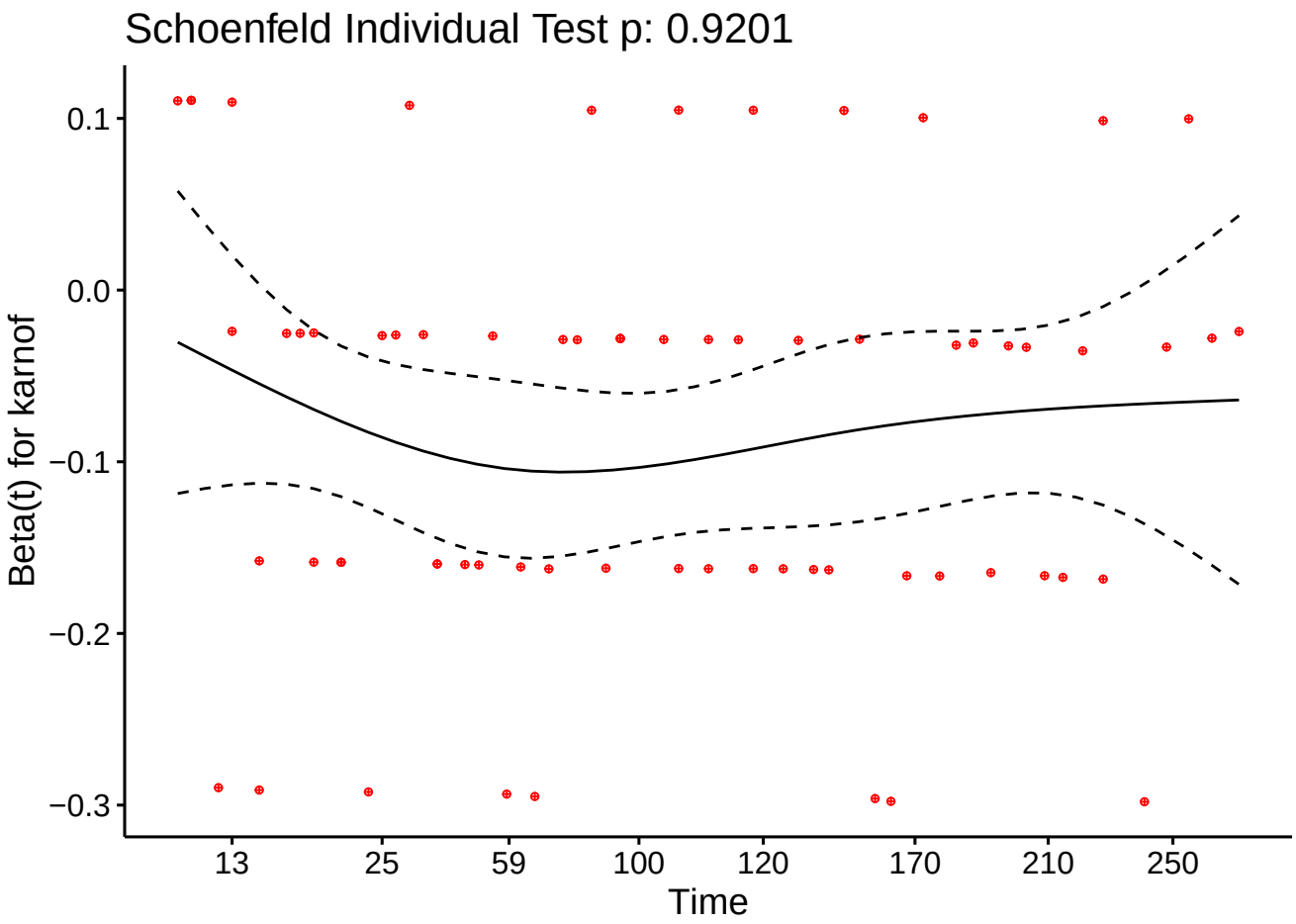

```
ggcoxdiagnostics(cph_cd4, type="schoenfeld")
```



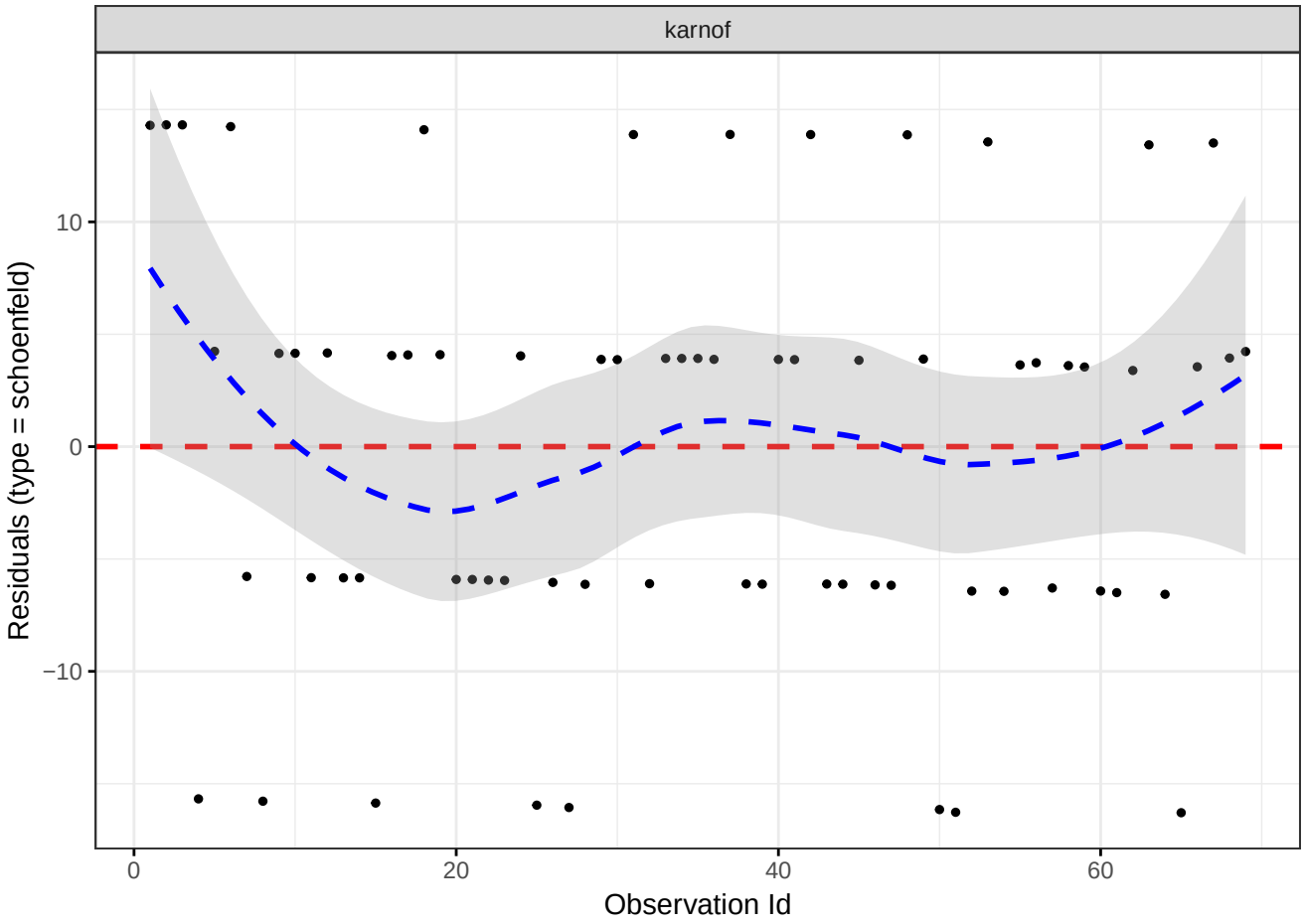

```
cph_k <- coxph(Surv(time,censor)~ karnof, data = aids)
zph_k <- cox.zph(cph_k)
zph_k
```

```
##            rho  chisq    p
## karnof -0.0116 0.0101 0.92
```

```
ggcoxzph(zph_k, point.size = 1, point.shape = 10)
```



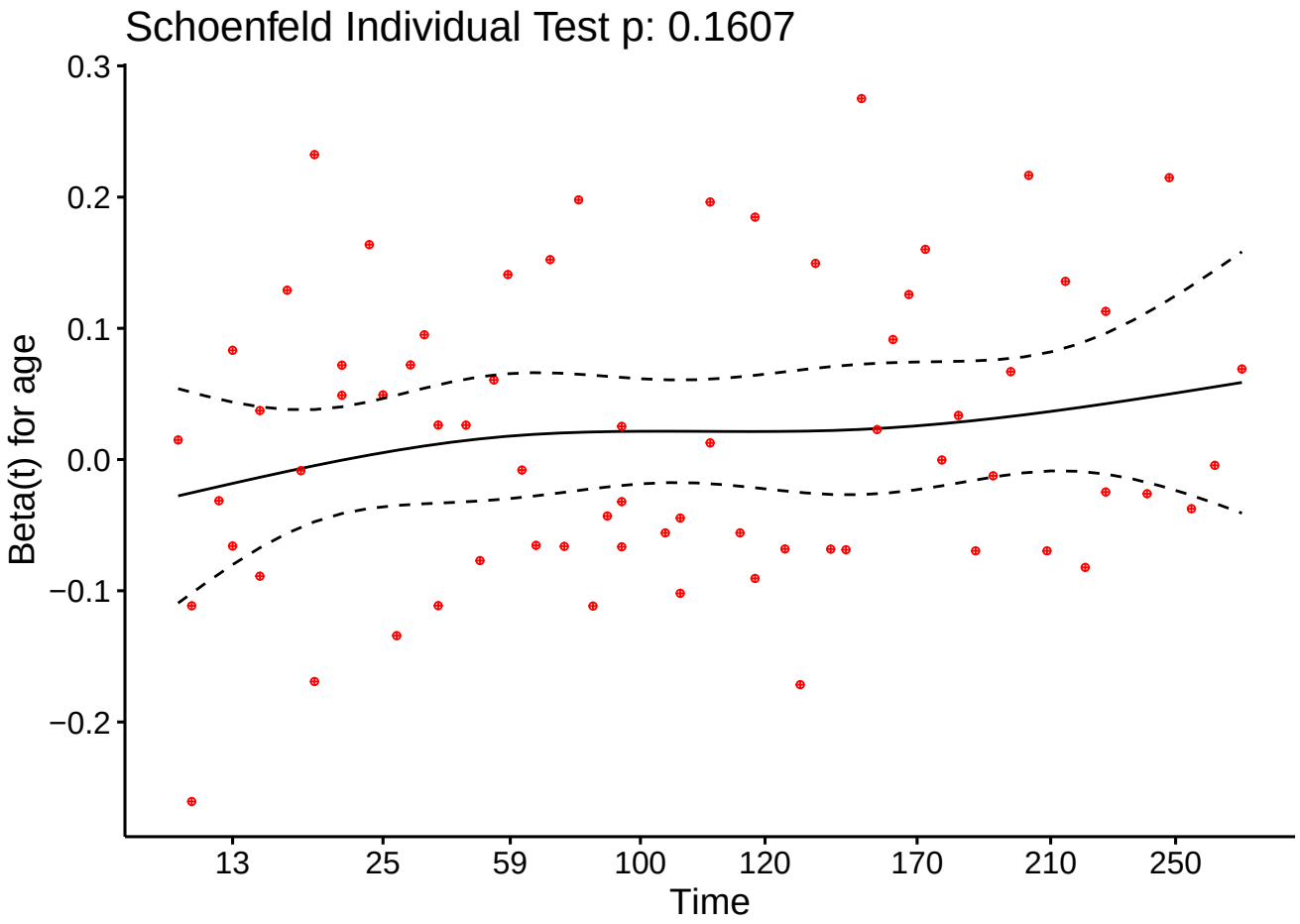

```
ggcoxdiagnostics(cph_k, type="schoenfeld")
```



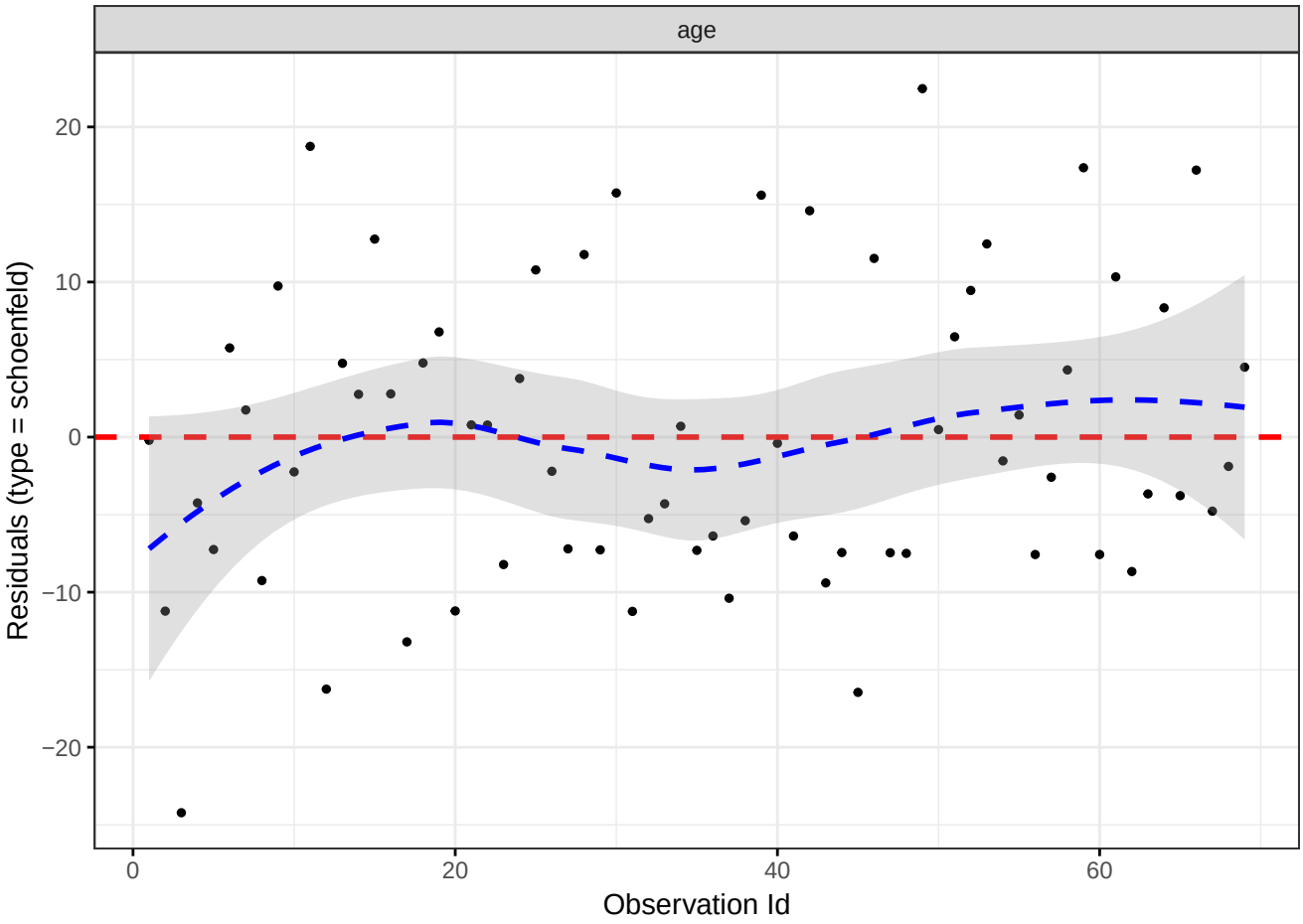

```
cph_a <- coxph(Surv(time,censor)~ age, data = aids)
zph_a <- cox.zph(cph_a)
zph_a
```

```
##       rho chisq     p
## age 0.165  1.97 0.161
```

```
ggcoxzph(zph_a, point.size = 1, point.shape = 10)
```




```
ggcoxdiagnostics(cph_a, type="schoenfeld")
```



Looking at the output results and the graphs for all the variables selected in our Cox PH model: *tx*, *cd4*,*karnof*, and *age*. The p-values for the variables are 0.544, 0.207, 0.92, and 0.161 respectively. All of these values are greater than the $\alpha = 0.05$, which indicates that the proportional hazards assumption is met for all variables involved in the model and that the Schoenfeld residuals of the explanatory variables are independent of time.

Discussion

Conclusion

References:

1. http://www.ukm.my/jsm/pdf_files/SM-PDF-46-3-2017/15%20Aditif%20Aalen.pdf

Acknowledgements