Game of Thrones - Survival Analysis

## Description

* Author: Anthony Jourdan
* Date: 5 april 2020

# Objectives

Target of this analysis is to study …

# Dataset description

Dataset downloaded from [here](https://figshare.com/articles/Game_of_Thrones_mortality_and_survival_dataset/8259680/1)

Game of Thrones mortality and survival dataset

Dataset posted on 13.06.2019, 10:25 by Reidar Lystad Benjamin Brown

This dataset includes data from Game of Thrones Seasons 1–8. The dataset comprises two separate datasets and an accompanying data dictionary. The character dataset contains 359 observations (i.e. characters) and 35 variables, including information about sociodemographics, exposures, and mortality. The episode dataset contains 73 observations (i.e. episodes) and 8 variables, including information about episode running time.

In this study we will use only the character dataset.

### Character dataset

* Number of observations: 359.
* Outcome:
  + **exp\_time\_hrs** Survival time of character (calculated as the time between first apparition and death)
* Censoring indicator:
  + **dth\_flag** = 0 if character is not dead by the end of the serie , = 1 otherwise
* Explanatory variables:

# Data Preparation

load needed libraries

library(tidyverse)  
library(survival)  
library(ggfortify)  
library(ggplot2)  
library(broom)  
library(survminer)  
library(survivalROC)

import datas from csv file:

setwd("C:/MY\_DATAS/MyGit/GoT-Survival\_Analysis")  
raw\_data = read.csv("./GoT\_dataset/character\_data\_S01-S08.csv")  
  
dat\_full = raw\_data %>%   
 select(name,  
 exp\_time\_hrs,  
 dth\_flag,  
 sex,  
 religion,  
 occupation,  
 social\_status,  
 allegiance\_last,  
 allegiance\_switched,  
 prominence  
 ) %>%   
 mutate(sex = c("Male",  
 "Female")[match(sex, c(1,2))],  
 religion = c("Great Stallion",  
 "Lord of Light",  
 "Faith of the Seven",  
 "Old Gods",  
 "Drowned God",  
 "Many Faced God",  
 "Other",  
 "Unknown/Unclear")[match(religion,c(1,2,3,4,5,6,7,9))],  
 occupation = c("Silk collar",  
 "Boiled leather collar",  
 "Unknown/Unclear")[match(occupation,c(1,2,9))],  
 social\_status = c("Highborn",  
 "Lowborn")[match(social\_status,c(1,2))],  
 allegiance\_last = c("Stark",  
 "Targaryen",  
 "Night's Watch",  
 "Lannister",  
 "Greyjoy",  
 "Bolton",  
 "Frey",  
 "Other",  
 "Unknown/Unclear")[match(allegiance\_last,c(1,2,3,4,5,6,7,8,9))],  
 allegiance\_switched = c("No",  
 "Yes")[match(allegiance\_switched,c(1,2))],  
 prominence = ifelse(prominence>3,  
 "High",  
 ifelse(prominence<1,  
 "Low",  
 "Medium")  
 )  
 )

Keep 15% of data for evaluating the final model

train\_size = 85 / 100 \* nrow(dat\_full)  
idx.dat = sample.int(nrow(dat\_full), size = train\_size, replace = FALSE)  
dat = dat\_full[idx.dat,]  
dat\_test = dat\_full[-idx.dat,]

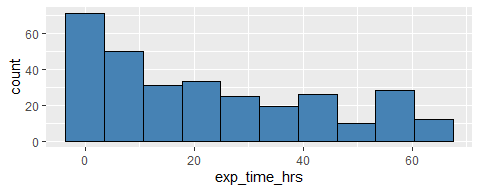
# Data Exploration

### Outcome: Survival duration

summary(dat$exp\_time\_hrs)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 4.22 17.96 22.42 38.03 63.99

ggplot(dat,aes(exp\_time\_hrs)) + geom\_histogram(bins = 10, color="black",fill="steelblue")



### Censoring indicator

Proportion of people dead before the end of the serie.

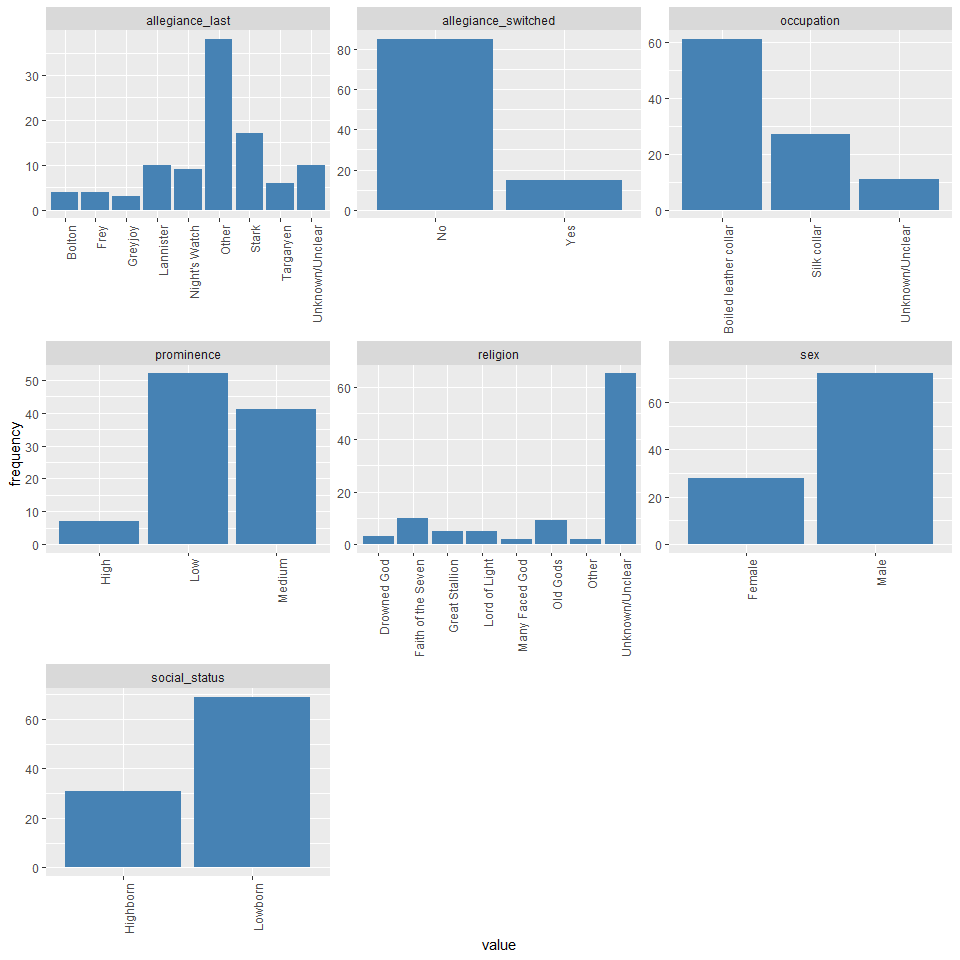
prop.table(table(dat$dth\_flag))

##   
## 0 1   
## 0.4098361 0.5901639

### Explanatory variables

Show explanatory variables composition:

d\_plot = dat %>%  
 select(-name,-exp\_time\_hrs,-dth\_flag) %>%  
 gather() %>%  
 group\_by(key) %>%  
 count(value) %>%   
 mutate(frequency=round(`n`/sum(`n`)\*100,0)) %>%  
 arrange(desc(key),desc(frequency))  
  
d\_plot %>% ggplot(aes(x=value, y=frequency)) +  
 facet\_wrap(~ key, scales = "free") +  
 geom\_bar(stat="identity", fill="steelblue") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

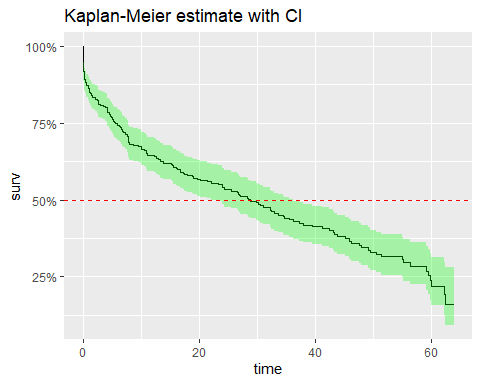


# Global survival overview

**Kaplan-Meyer estimator**

* First look at outcome:

fit.KM = survfit(Surv(exp\_time\_hrs, dth\_flag) ~ 1, data = dat)  
autoplot(fit.KM,conf.int.fill = "#00FF00", censor=FALSE) +  
 geom\_hline(yintercept=.5, linetype="dashed", color = "red") +   
 ggtitle("Kaplan-Meier estimate with CI")



fit.KM

## Call: survfit(formula = Surv(exp\_time\_hrs, dth\_flag) ~ 1, data = dat)  
##   
## n events median 0.95LCL 0.95UCL   
## 305.0 180.0 28.8 22.1 36.3

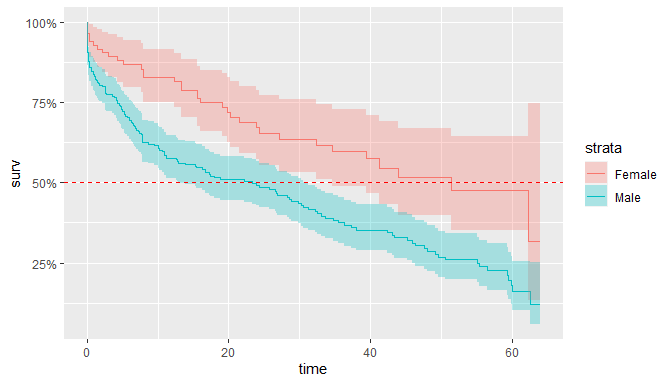
# Survival vs Explanatory variables

Used functions

# draw the KM survival curve with stratification with a given exlanatory variable  
plot\_KM <- function(df,col,CI=TRUE){  
 fit = survfit(Surv(df$exp\_time\_hrs, df$dth\_flag) ~ df[,col])  
 autoplot(fit,conf.int=CI,censor=FALSE) +   
 geom\_hline(yintercept=.5, linetype="dashed", color = "red")   
}  
  
# Print the medians for stratas (+formating)  
print\_medians <- function(df,col){  
 fit = survfit(Surv(df$exp\_time\_hrs, df$dth\_flag) ~ df[,col])  
 infos\_fit = surv\_median(fit) %>%  
 mutate(strata=substr(strata,11,100))  
 cat("Medians:\n")  
 cat(sprintf("%\*s %\*s %\*s\n",25,"Group",15,"Median",20,"Conf.Interval"))  
 fit.conf=paste("( ",infos\_fit$lower,";",infos\_fit$upper," )",sep="")  
 cat(sprintf("%\*s %\*s %\*s\n",25,infos\_fit$strata,15,infos\_fit$median,20,fit.conf))  
}  
  
# Print cox regression HR+CI and LRT for stratas (+formating)  
print\_cox <- function(df,col){  
 fit\_cox = coxph(Surv(df$exp\_time\_hrs, df$dth\_flag) ~ df[,col])  
 x = tidy(fit\_cox)  
 cox.ref = fit\_cox$xlevels[[1]][1]  
 cox.term = substr(x$term,10,100)  
 cox.hr = round(exp(x$estimate),2)  
 cox.hr.conflow = round(exp(x$conf.low),2)  
 cox.hr.confhigh = round(exp(x$conf.high),2)  
 cat("Cox Regression:\n")  
 cat(sprintf("%\*s %\*s %\*s\n",25,"Group",15,"Hazard Ratio",20,"Conf.Interval"))  
 cat(sprintf("%\*s %\*s %\*s\n",25,cox.ref,15,"(Reference)",20,"-"))  
 cox.conf=paste("( ",cox.hr.conflow,";",cox.hr.confhigh," )",sep="")  
 cat(sprintf("%\*s %\*s %\*s\n",25,cox.term,15,cox.hr,20,cox.conf))  
 y = glance(fit\_cox)  
 cox.lrt = ifelse(y$p.value.log<0.01,  
 formatC(y$p.value.log, format = "e", digits = 2),  
 formatC(y$p.value.log, digits = 2))  
 cat(paste("\nLikelihood Ratio Test:",cox.lrt))  
}

### - How is gender influencing survival time ?

plot\_KM(dat,"sex")



print\_cox(dat,"sex")

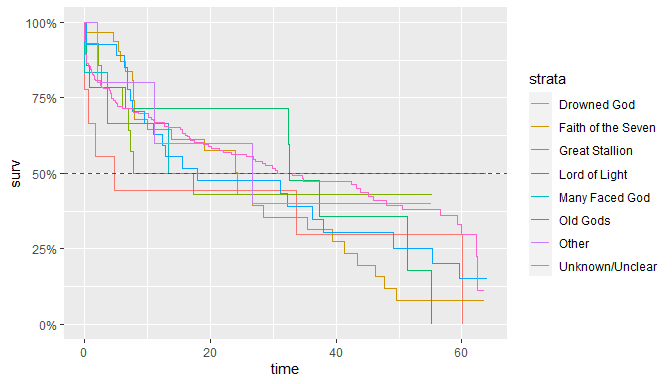
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## Female (Reference) -  
## Male 1.97 ( 1.36;2.86 )  
##   
## Likelihood Ratio Test: 1.45e-04

print\_medians(dat,"sex")

## Medians:  
## Group Median Conf.Interval  
## Female 51.42 ( 34.57;NA )  
## Male 23.38 ( 12.92;30.6 )

### - How is religion survival time ?

plot\_KM(dat,"religion",FALSE)



print\_cox(dat,"religion")

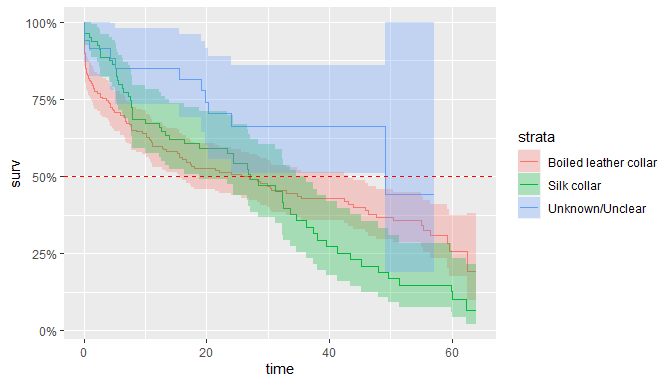
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## Drowned God (Reference) -  
## Faith of the Seven 0.89 ( 0.39;2.06 )  
## Great Stallion 0.71 ( 0.26;1.97 )  
## Lord of Light 0.7 ( 0.26;1.88 )  
## Many Faced God 0.46 ( 0.12;1.77 )  
## Old Gods 0.74 ( 0.32;1.75 )  
## Other 0.61 ( 0.16;2.38 )  
## Unknown/Unclear 0.61 ( 0.28;1.31 )  
##   
## Likelihood Ratio Test: 0.69

print\_medians(dat,"religion")

## Medians:  
## Group Median Conf.Interval  
## Drowned God 4.79 ( 0.56;NA )  
## Faith of the Seven 24.34 ( 10.05;41.33 )  
## Great Stallion 12.535 ( 6.9;NA )  
## Lord of Light 32.56 ( 32.36;NA )  
## Many Faced God 13.36 ( 3.59;NA )  
## Old Gods 17.96 ( 10.96;55.34 )  
## Other 26.63 ( 11.17;NA )  
## Unknown/Unclear 33.06 ( 23.38;47.99 )

### - How is occupation influencing ?

plot\_KM(dat,"occupation")



print\_cox(dat,"occupation")

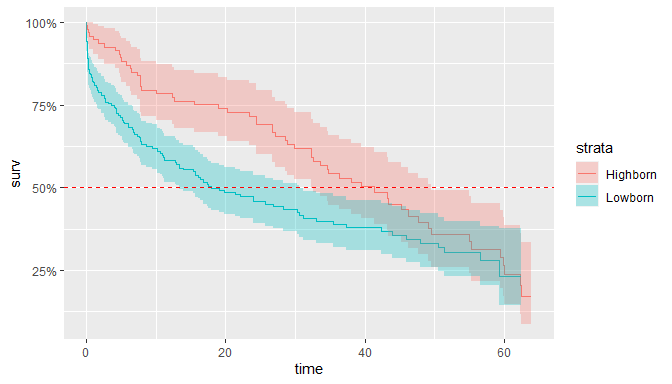
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## Boiled leather collar (Reference) -  
## Silk collar 1.14 ( 0.83;1.57 )  
## Unknown/Unclear 0.49 ( 0.26;0.92 )  
##   
## Likelihood Ratio Test: 0.019

print\_medians(dat,"occupation")

## Medians:  
## Group Median Conf.Interval  
## Boiled leather collar 25.68 ( 15.57;43.17 )  
## Silk collar 27.12 ( 18.87;34.57 )  
## Unknown/Unclear 49.15 ( 49.15;NA )

## –> Is social\_status influencing ?

plot\_KM(dat,"social\_status")



print\_cox(dat,"social\_status")

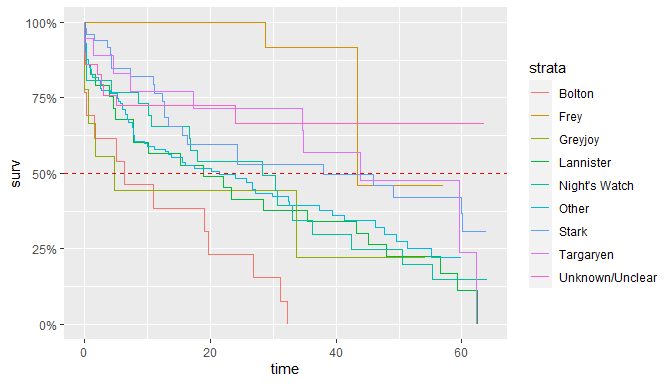
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## Highborn (Reference) -  
## Lowborn 1.49 ( 1.08;2.05 )  
##   
## Likelihood Ratio Test: 0.012

print\_medians(dat,"social\_status")

## Medians:  
## Group Median Conf.Interval  
## Highborn 41.33 ( 32.36;49.59 )  
## Lowborn 17.96 ( 13.36;30.6 )

## –> Is the last allegiance made influencing ?

plot\_KM(dat,"allegiance\_last",FALSE)



print\_cox(dat,"allegiance\_last")

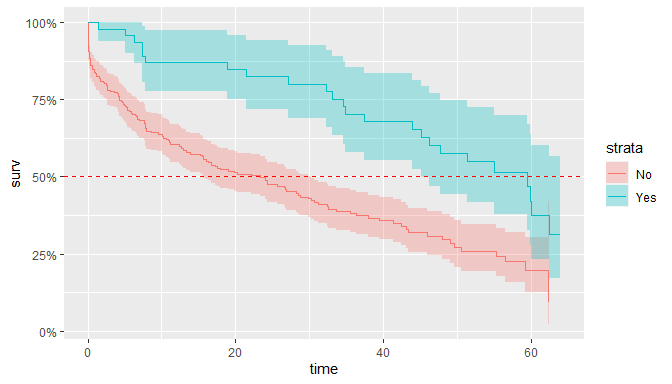
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## Bolton (Reference) -  
## Frey 0.06 ( 0.01;0.27 )  
## Greyjoy 0.58 ( 0.22;1.54 )  
## Lannister 0.47 ( 0.24;0.93 )  
## Night's Watch 0.41 ( 0.2;0.84 )  
## Other 0.41 ( 0.23;0.74 )  
## Stark 0.24 ( 0.12;0.48 )  
## Targaryen 0.25 ( 0.11;0.58 )  
## Unknown/Unclear 0.18 ( 0.08;0.42 )  
##   
## Likelihood Ratio Test: 1.02e-05

print\_medians(dat,"allegiance\_last")

## Medians:  
## Group Median Conf.Interval  
## Bolton 6.26 ( 0.28;NA )  
## Frey 43.37 ( 43.37;NA )  
## Greyjoy 4.79 ( 0.56;NA )  
## Lannister 18.87 ( 7.75;45.18 )  
## Night's Watch 28.28 ( 10.61;50.52 )  
## Other 21.45 ( 11.17;37.37 )  
## Stark 38.03 ( 15.5;NA )  
## Targaryen 43.92 ( 34.57;NA )  
## Unknown/Unclear NA ( NA;NA )

## –> Is the fact to have switched allegiance during the serie influencing ?

plot\_KM(dat,"allegiance\_switched")



print\_cox(dat,"allegiance\_switched")

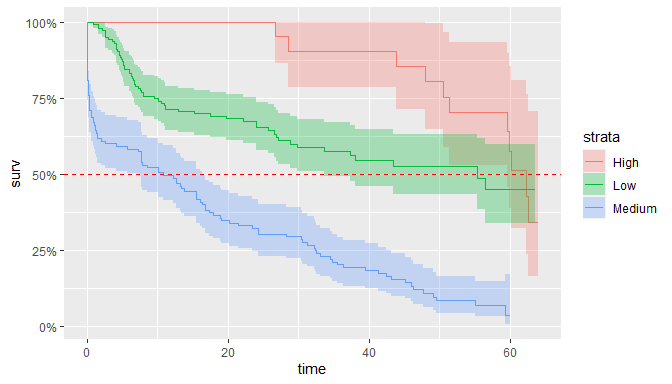
## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## No (Reference) -  
## Yes 0.41 ( 0.26;0.64 )  
##   
## Likelihood Ratio Test: 1.58e-05

print\_medians(dat,"allegiance\_switched")

## Medians:  
## Group Median Conf.Interval  
## No 23.38 ( 15.57;30.32 )  
## Yes 59.54 ( 45.18;NA )

## –> Is prominence influencing ?

plot\_KM(dat,"prominence")



print\_cox(dat,"prominence")

## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## High (Reference) -  
## Low 1.94 ( 0.98;3.83 )  
## Medium 6.28 ( 3.21;12.32 )  
##   
## Likelihood Ratio Test: 1.08e-16

print\_medians(dat,"prominence")

## Medians:  
## Group Median Conf.Interval  
## High 62.35 ( 59.54;NA )  
## Low 55.34 ( 33.64;NA )  
## Medium 10.89 ( 7.34;16.32 )

# Build a model of Survival time in GoT

dat\_model = select(dat,-name)  
Model\_Full = coxph(Surv(exp\_time\_hrs,dth\_flag)~.,data=dat\_model)  
MAIC = step(Model\_Full)

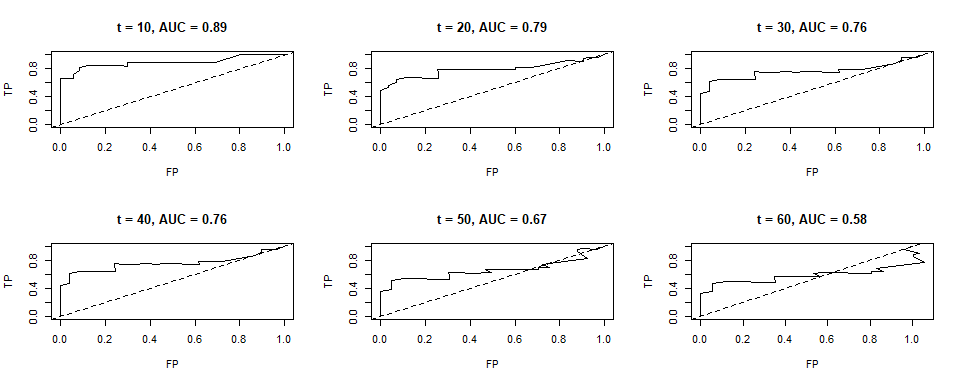
## Start: AIC=1723.26  
## Surv(exp\_time\_hrs, dth\_flag) ~ sex + religion + occupation +   
## social\_status + allegiance\_last + allegiance\_switched + prominence  
##   
## Df AIC  
## - religion 7 1716.2  
## - allegiance\_last 8 1721.9  
## - occupation 2 1722.0  
## <none> 1723.3  
## - social\_status 1 1725.1  
## - sex 1 1727.1  
## - allegiance\_switched 1 1736.0  
## - prominence 2 1780.3  
##   
## Step: AIC=1716.24  
## Surv(exp\_time\_hrs, dth\_flag) ~ sex + occupation + social\_status +   
## allegiance\_last + allegiance\_switched + prominence  
##   
## Df AIC  
## - occupation 2 1714.8  
## <none> 1716.2  
## - social\_status 1 1717.2  
## - allegiance\_last 8 1718.8  
## - sex 1 1720.5  
## - allegiance\_switched 1 1727.7  
## - prominence 2 1771.3  
##   
## Step: AIC=1714.76  
## Surv(exp\_time\_hrs, dth\_flag) ~ sex + social\_status + allegiance\_last +   
## allegiance\_switched + prominence  
##   
## Df AIC  
## <none> 1714.8  
## - social\_status 1 1715.4  
## - allegiance\_last 8 1719.6  
## - sex 1 1721.2  
## - allegiance\_switched 1 1727.6  
## - prominence 2 1769.1

summary(MAIC)

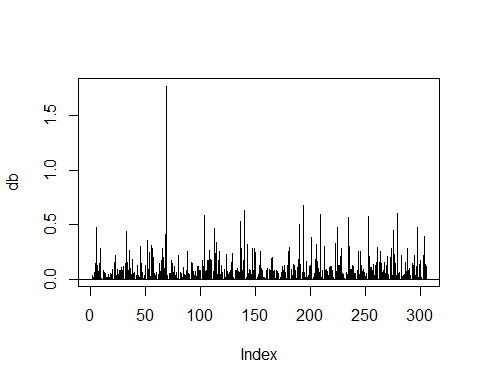
## Call:  
## coxph(formula = Surv(exp\_time\_hrs, dth\_flag) ~ sex + social\_status +   
## allegiance\_last + allegiance\_switched + prominence, data = dat\_model)  
##   
## n= 305, number of events= 180   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## sexMale 0.56158 1.75344 0.20130 2.790 0.005275 \*\*   
## social\_statusLowborn 0.29359 1.34124 0.18394 1.596 0.110453   
## allegiance\_lastFrey -1.54787 0.21270 0.80873 -1.914 0.055625 .   
## allegiance\_lastGreyjoy 0.11117 1.11758 0.51004 0.218 0.827461   
## allegiance\_lastLannister -0.62418 0.53570 0.35607 -1.753 0.079603 .   
## allegiance\_lastNight's Watch -0.98427 0.37371 0.36170 -2.721 0.006503 \*\*   
## allegiance\_lastOther -0.50191 0.60537 0.31087 -1.615 0.106414   
## allegiance\_lastStark -1.06295 0.34544 0.36307 -2.928 0.003415 \*\*   
## allegiance\_lastTargaryen -0.70029 0.49644 0.43507 -1.610 0.107482   
## allegiance\_lastUnknown/Unclear -1.27573 0.27923 0.44821 -2.846 0.004423 \*\*   
## allegiance\_switchedYes -0.89655 0.40797 0.24958 -3.592 0.000328 \*\*\*  
## prominenceLow -0.04301 0.95790 0.38290 -0.112 0.910557   
## prominenceMedium 1.20346 3.33163 0.36157 3.328 0.000873 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## sexMale 1.7534 0.5703 1.18180 2.6016  
## social\_statusLowborn 1.3412 0.7456 0.93527 1.9234  
## allegiance\_lastFrey 0.2127 4.7014 0.04359 1.0379  
## allegiance\_lastGreyjoy 1.1176 0.8948 0.41127 3.0369  
## allegiance\_lastLannister 0.5357 1.8667 0.26659 1.0765  
## allegiance\_lastNight's Watch 0.3737 2.6759 0.18393 0.7593  
## allegiance\_lastOther 0.6054 1.6519 0.32916 1.1134  
## allegiance\_lastStark 0.3454 2.8949 0.16956 0.7037  
## allegiance\_lastTargaryen 0.4964 2.0143 0.21161 1.1647  
## allegiance\_lastUnknown/Unclear 0.2792 3.5813 0.11600 0.6722  
## allegiance\_switchedYes 0.4080 2.4511 0.25014 0.6654  
## prominenceLow 0.9579 1.0440 0.45227 2.0288  
## prominenceMedium 3.3316 0.3002 1.64017 6.7675  
##   
## Concordance= 0.75 (se = 0.019 )  
## Likelihood ratio test= 124.2 on 13 df, p=<2e-16  
## Wald test = 112.2 on 13 df, p=<2e-16  
## Score (logrank) test = 129.9 on 13 df, p=<2e-16

Check model on test set

lp = predict(MAIC, newdata = dat\_test, type="lp")  
ROC\_func <- function(t){  
 res = survivalROC(Stime = dat\_test$exp\_time\_hrs,  
 status = dat\_test$dth\_flag,  
 marker = lp,  
 predict.time = t,  
 method = "KM")  
 with(res, plot(TP ~ FP, type = "l", main = sprintf("t = %.0f, AUC = %.2f", t, AUC)))  
 abline(a = 0, b = 1, lty = 2)  
 res  
}  
layout(matrix(1:6, byrow = TRUE, ncol = 3))  
res.survivalROC.age.sex <- lapply(1:6 \* 10, function(t) {  
 ROC\_func(t)  
})



dfbetas = residuals(MAIC, type='dfbetas')  
db = sqrt(rowSums(dfbetas^2))  
plot(db,type = 'h')  
abline(h=0)

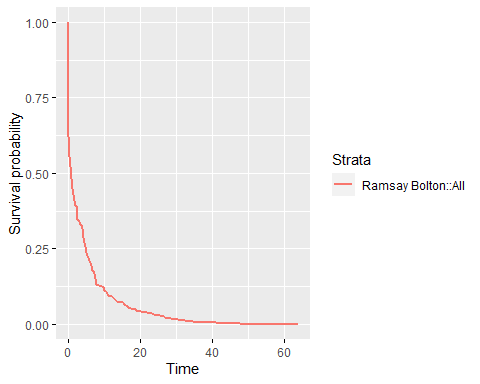


find who is the peak

idx=names(db[db>1])  
dat[idx,]

## name exp\_time\_hrs dth\_flag sex religion occupation social\_status  
## 165 Ramsay Bolton 31.18 1 Male Old Gods Silk collar Lowborn  
## allegiance\_last allegiance\_switched prominence  
## 165 Bolton No Medium

dat\_new = dat[idx,]  
  
z = list()  
for(i in 1:nrow(dat\_new)) {  
 row <- dat\_new[i,]  
 p\_s = survfit(MAIC,newdata = row)  
 z = c(z,list(p\_s))  
}  
names(z)=dat\_new$name  
  
ggsurvplot\_combine(z,  
 conf.int = FALSE,  
 risk.table = FALSE,  
 pval = FALSE,  
 censor = FALSE,  
 surv.median.line = "hv",  
 ggtheme = theme\_gray(),  
 legend="right")



# Predict from model for some characters and compare with observed datas

d\_new = dat %>%  
 filter(name %in% c("Arya Stark",  
 "Jaime Lannister",  
 "Theon Greyjoy",  
 "Jon Snow",  
 "Eddard Stark",  
 "Ramsay Bolton",  
 "Samwell Tarly",  
 "Illyrio Mopatis",  
 "Mhaegen",  
 "Todder",  
 "Merry Frey")) %>%  
 select(-exp\_time\_hrs,-dth\_flag)  
  
z = list()  
  
for(i in 1:nrow(dat\_test)) {  
 row <- dat\_test[i,]  
 p\_s = survfit(MAIC,newdata = row)  
 z = c(z,list(p\_s))  
}  
names(z)=dat\_test$name  
  
ggsurvplot\_combine(z,  
 conf.int = FALSE,  
 risk.table = FALSE,  
 pval = FALSE,  
 censor = FALSE,  
 surv.median.line = "hv",  
 ggtheme = theme\_gray())

## Warning: Vectorized input to `element\_text()` is not officially supported.  
## Results may be unexpected or may change in future versions of ggplot2.  
  
## Warning: Vectorized input to `element\_text()` is not officially supported.  
## Results may be unexpected or may change in future versions of ggplot2.  
  
## Warning: Vectorized input to `element\_text()` is not officially supported.  
## Results may be unexpected or may change in future versions of ggplot2.

