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

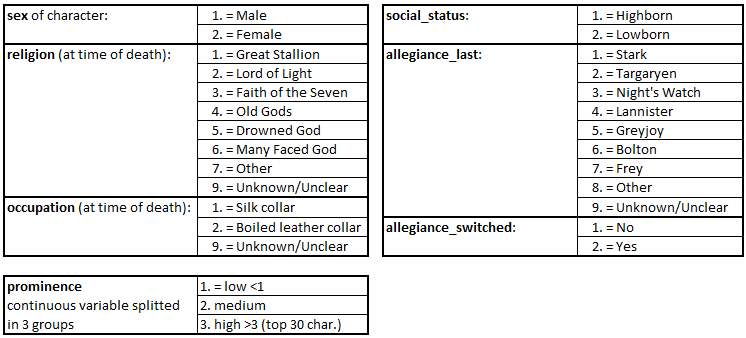
Dataset posted on 13.06.2019, 10:25 by Reidar Lystad Benjamin Brown

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)

Import datas from csv file and format output:

raw\_data = read.csv("./GoT\_dataset/character\_data\_S01-S08.csv")

dat = 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")  
 ))

# Data Exploration

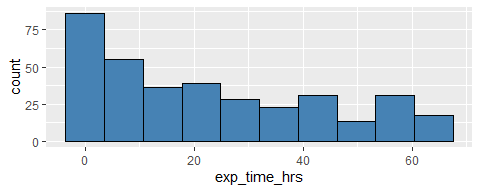
### Outcome: Survival duration

Let’s have a look at basic statistics about the survival duration.

summary(dat$exp\_time\_hrs)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 4.03 18.49 22.68 40.16 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))

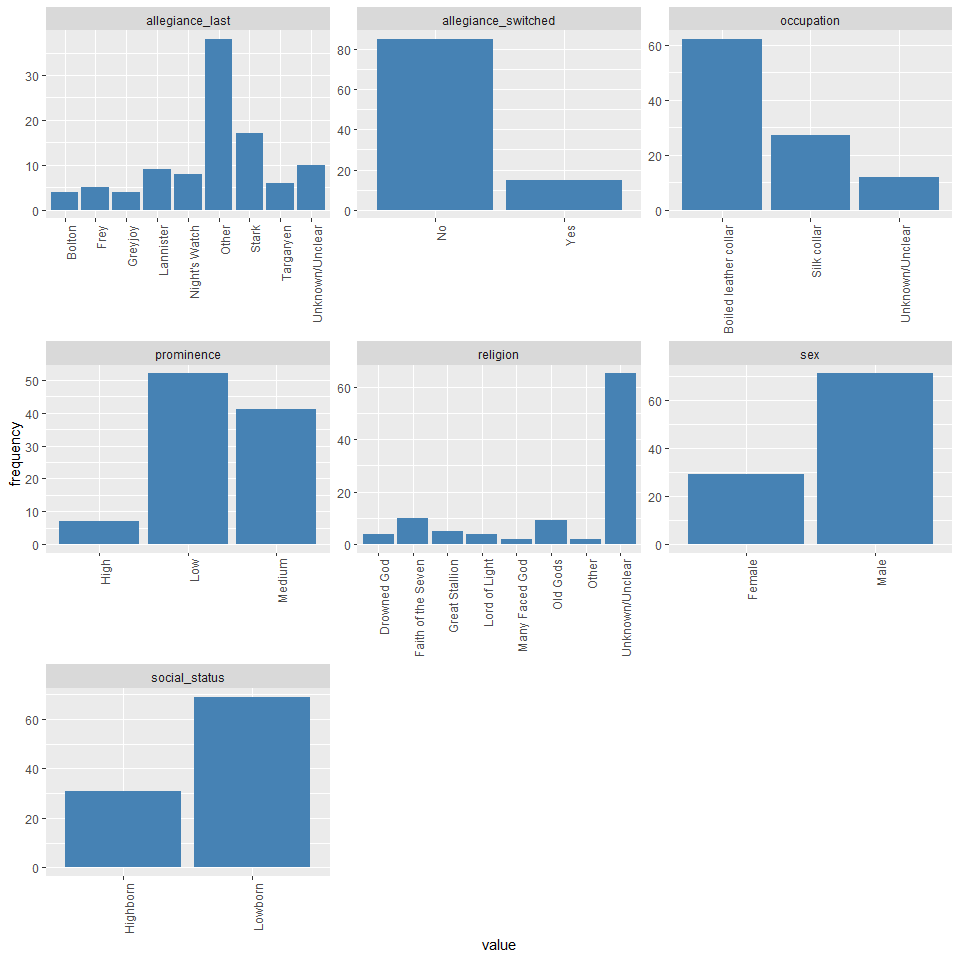
##   
## 0 1   
## 0.4094708 0.5905292

–> roughly 40% of censored datas, 60% of the characters in the study are dead before the end of the serie

### 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))



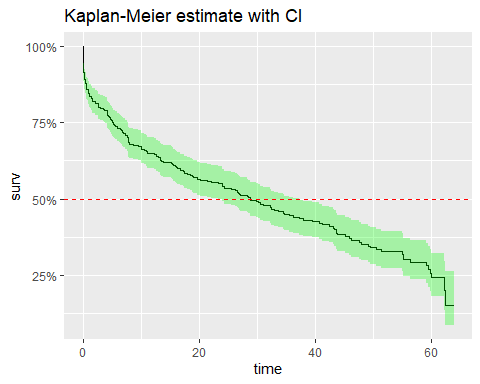
* 65% of the population have not known or unclear religion –> Careful to check if meaningful
* most are Boiled leather collar
* 70% are lowborn

# 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") +  
 geom\_hline(yintercept=.5, linetype="dashed", color = "red")+ ggtitle("Kaplan-Meier estimate with CI")



Median Survival Time: 28.8hrs - As a character, you would have 50% of change to stay alive up to 28.8hrs

fit.KM

## Call: survfit(formula = Surv(exp\_time\_hrs, dth\_flag) ~ 1, data = dat)  
##   
## n events median 0.95LCL 0.95UCL   
## 359.0 212.0 28.8 23.4 37.4

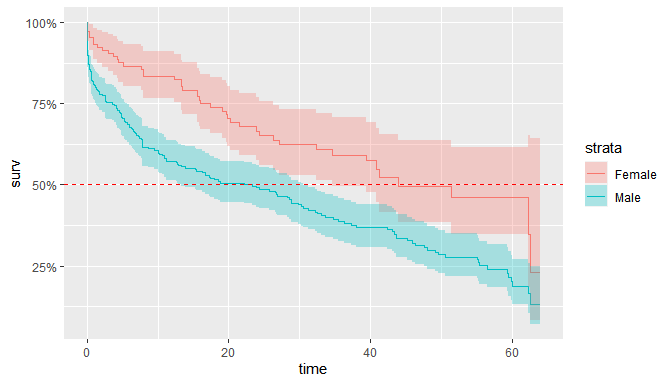
# Survival vs Explanatory variables

Used functions

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\_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 <- 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 1.87 ( 1.34;2.61 )  
##   
## Likelihood Ratio Test: 9.46e-05

Likelihood ratio test (LRT) pvalue is very small, proving that there is a significant difference between male and female survival time.

!!!! Hazard ration is 1.87, meaning that male have 1.87 more chances to be killes than women

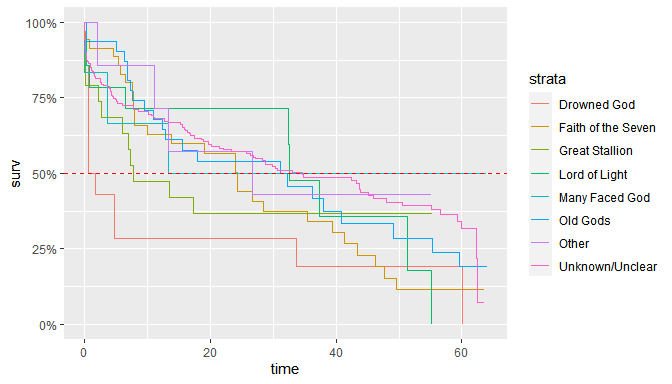
Here are the median survival time for each category:

print\_medians(dat,"sex")

## Medians:  
## Group Median Conf.Interval  
## Female 43.92 ( 34.57;NA )  
## Male 23.38 ( 13.32;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 0.58 ( 0.29;1.15 )  
## Great Stallion 0.62 ( 0.28;1.38 )  
## Lord of Light 0.47 ( 0.2;1.13 )  
## Many Faced God 0.3 ( 0.08;1.07 )  
## Old Gods 0.44 ( 0.21;0.88 )  
## Other 0.4 ( 0.13;1.25 )  
## Unknown/Unclear 0.4 ( 0.22;0.73 )  
##   
## Likelihood Ratio Test: 0.14

Cox regression LRT pvalue is quite large and > 5% pointing that there is no significant difference between religions

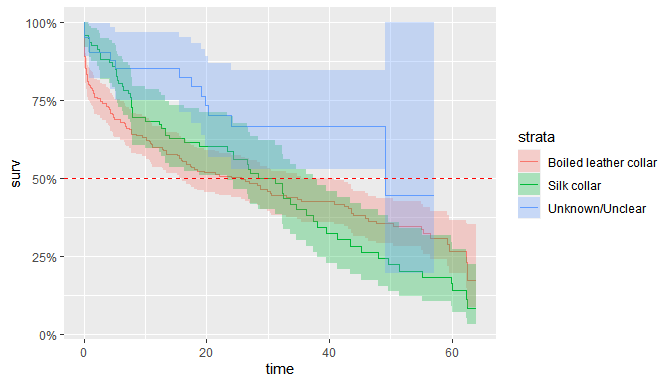
One thing that can be noted from the graph is that the “Drowned God” religion has a median survival time very low… If you were of this religion, you would have only 50% chance to survive after 1.11hrs ! (pretty scary)

print\_medians(dat,"religion")

## Medians:  
## Group Median Conf.Interval  
## Drowned God 1.11 ( 0.54;NA )  
## Faith of the Seven 24.34 ( 10.05;41.33 )  
## Great Stallion 7.77 ( 5.95;NA )  
## Lord of Light 32.56 ( 32.36;NA )  
## Many Faced God 13.36 ( 3.59;NA )  
## Old Gods 31.18 ( 12.31;55.34 )  
## Other 26.63 ( 11.17;NA )  
## Unknown/Unclear 34.57 ( 26.34;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 1.03 ( 0.76;1.39 )  
## Unknown/Unclear 0.48 ( 0.27;0.85 )  
##   
## Likelihood Ratio Test: 0.014

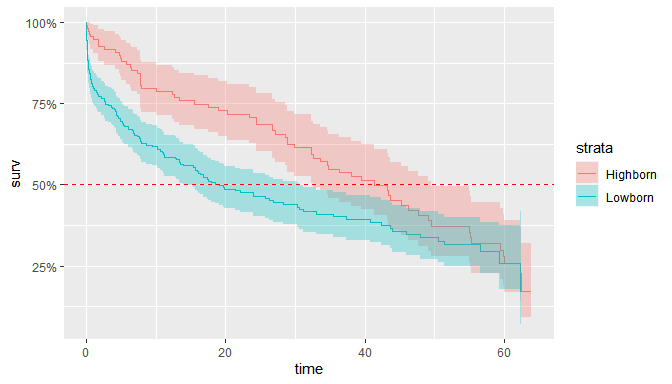
LRT pvalue is < 5%, we can say that at least one group is significantly different from other. It’s certainly due to the group ‘Unknown/Unclear’ which has an hazard ratio close to 0.5, the 2 others are very close (HR ~ 1). this can be also seen on the medians were CI are overlapping.

print\_medians(dat,"occupation")

## Medians:  
## Group Median Conf.Interval  
## Boiled leather collar 25.68 ( 15.57;40.81 )  
## Silk collar 28.51 ( 23.38;37.37 )  
## 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 ( 1.11;2.01 )  
##   
## Likelihood Ratio Test: 6.77e-03

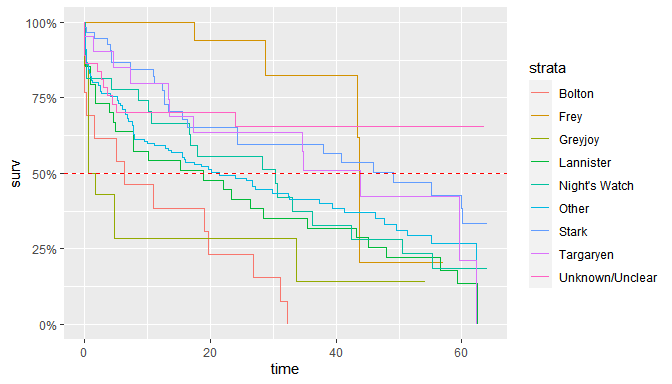
Again LRT pvalue is <5%, meaning that to be highborn or lowborn is significantly differnt in terms of survival time in GoT.

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

## Medians:  
## Group Median Conf.Interval  
## Highborn 41.33 ( 32.56;49.59 )  
## Lowborn 19.08 ( 13.85;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 0.14 ( 0.05;0.37 )  
## Greyjoy 0.89 ( 0.4;1.99 )  
## Lannister 0.5 ( 0.26;0.96 )  
## Night's Watch 0.39 ( 0.19;0.79 )  
## Other 0.4 ( 0.22;0.72 )  
## Stark 0.22 ( 0.11;0.43 )  
## Targaryen 0.29 ( 0.13;0.64 )  
## Unknown/Unclear 0.2 ( 0.09;0.43 )  
##   
## Likelihood Ratio Test: 2.69e-06

LRT pvalue is < 5%, we can say that at least one group is significantly different from other.

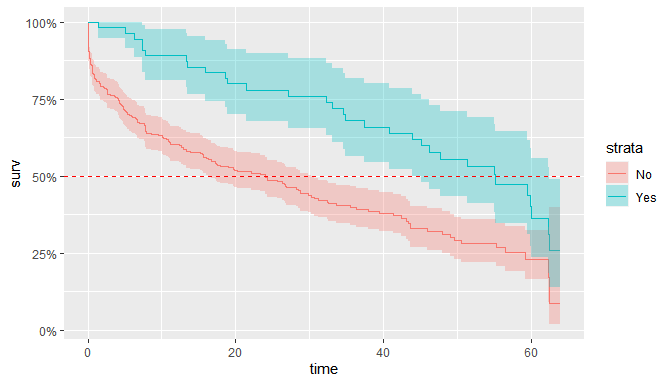
If you allegiance goes to ‘Bolton’, then you have 0% of chance to be present during all the show. But if you follow the ‘Greyjoy’, the you’re median survival time is only of 1.11hrs…

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

## Medians:  
## Group Median Conf.Interval  
## Bolton 6.26 ( 0.28;NA )  
## Frey 43.67 ( 43.37;NA )  
## Greyjoy 1.11 ( 0.54;NA )  
## Lannister 18.87 ( 4.91;43.17 )  
## Night's Watch 30.32 ( 16.73;50.52 )  
## Other 21.45 ( 13.36;37.37 )  
## Stark 49.15 ( 24.34;NA )  
## Targaryen 43.92 ( 17.3;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 0.48 ( 0.32;0.71 )  
##   
## Likelihood Ratio Test: 7.05e-05

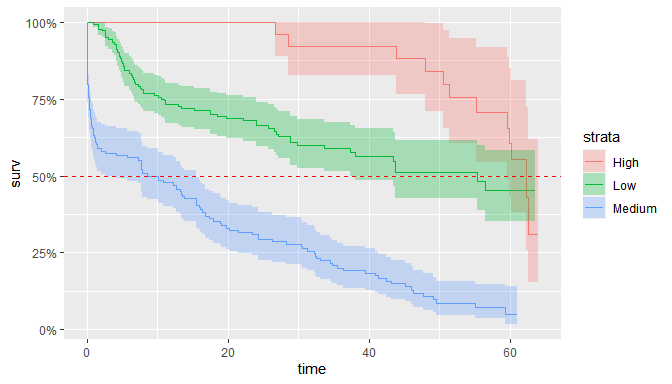
pvalue < 5%, the change in allegiance has a real impact on the characters survival times. it seems, that in GoT, if you want to maximize your chances to survive, you have to not be too strict with your allegiance.

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

## Medians:  
## Group Median Conf.Interval  
## No 23.96 ( 16.32;30.32 )  
## Yes 55.22 ( 45.18;62.52 )

### –> Is prominence influencing ?

plot\_KM(dat,"prominence")



print\_cox(dat,"prominence")

## Cox Regression:  
## Group Hazard Ratio Conf.Interval  
## High (Reference) -  
## Low 1.93 ( 1.04;3.59 )  
## Medium 6.58 ( 3.57;12.13 )  
##   
## Likelihood Ratio Test: 6.29e-21

Very significant difference, sounds logic for characters with high prominence (stars of the show), that producers decided no to kill them at the begining of the show so their survival time is higher than others. It seems more surprising to me, that people with low prominence have a higher survival time than the ones in the middle.

print\_medians(dat,"prominence")

## Medians:  
## Group Median Conf.Interval  
## High 62.31 ( 59.54;NA )  
## Low 55.34 ( 37.37;NA )  
## Medium 8.61 ( 2.67;15.57 )

# Build a model of Survival time in GoT

### - Model selection

Let’s start with a full model (using all explanatory variables) and run a step-wise model selection based on AIC.

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

## Start: AIC=2081.15  
## Surv(exp\_time\_hrs, dth\_flag) ~ sex + religion + occupation +   
## social\_status + allegiance\_last + allegiance\_switched + prominence  
##   
## Df AIC  
## - religion 7 2077.9  
## <none> 2081.2  
## - occupation 2 2082.1  
## - allegiance\_last 8 2083.2  
## - social\_status 1 2084.4  
## - sex 1 2086.7  
## - allegiance\_switched 1 2092.6  
## - prominence 2 2168.5  
##   
## Step: AIC=2077.87  
## Surv(exp\_time\_hrs, dth\_flag) ~ sex + occupation + social\_status +   
## allegiance\_last + allegiance\_switched + prominence  
##   
## Df AIC  
## <none> 2077.9  
## - occupation 2 2078.1  
## - social\_status 1 2080.7  
## - sex 1 2083.5  
## - allegiance\_last 8 2087.5  
## - allegiance\_switched 1 2088.1  
## - prominence 2 2160.6

After the step-wise selection, it appears that only religion can be removed from model.

### - Model description & explanation

MAIC

## Call:  
## coxph(formula = Surv(exp\_time\_hrs, dth\_flag) ~ sex + occupation +   
## social\_status + allegiance\_last + allegiance\_switched + prominence,   
## data = dat\_model)  
##   
## coef exp(coef) se(coef) z p  
## sexMale 0.4911 1.6342 0.1843 2.664 0.007716  
## occupationSilk collar 0.1324 1.1415 0.1960 0.676 0.499338  
## occupationUnknown/Unclear -0.5210 0.5939 0.3270 -1.594 0.111040  
## social\_statusLowborn 0.4307 1.5383 0.1976 2.179 0.029336  
## allegiance\_lastFrey -0.3664 0.6932 0.5616 -0.652 0.514091  
## allegiance\_lastGreyjoy 0.4012 1.4937 0.4343 0.924 0.355595  
## allegiance\_lastLannister -0.4836 0.6165 0.3531 -1.370 0.170766  
## allegiance\_lastNight's Watch -0.9885 0.3721 0.3740 -2.643 0.008219  
## allegiance\_lastOther -0.4750 0.6219 0.3118 -1.524 0.127623  
## allegiance\_lastStark -1.0955 0.3344 0.3581 -3.059 0.002219  
## allegiance\_lastTargaryen -0.5221 0.5933 0.4209 -1.240 0.214836  
## allegiance\_lastUnknown/Unclear -1.0134 0.3630 0.4222 -2.400 0.016378  
## allegiance\_switchedYes -0.7464 0.4741 0.2251 -3.316 0.000914  
## prominenceLow 0.0581 1.0598 0.3503 0.166 0.868295  
## prominenceMedium 1.4650 4.3275 0.3300 4.439 9.03e-06  
##   
## Likelihood ratio test=157.1 on 15 df, p=< 2.2e-16  
## n= 359, number of events= 212

# 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(d\_new)) {  
 row <- d\_new[i,]  
 p\_s = survfit(MAIC,newdata = row)  
 z = c(z,list(p\_s))  
}  
names(z)=d\_new$name  
  
ggsurvplot\_combine(z,  
 conf.int = FALSE,  
 risk.table = FALSE,  
 pval = FALSE,  
 censor = FALSE,  
 surv.median.line = "hv",  
 ggtheme = theme\_gray())

