Stat 417 Project Report
Lily Cook
Rebecca Ioffe
Wyatt De Mers
Blake Reavis
3/19/2024

Survival Analysis: Game of Thrones Character Survival Data

Introduction

The HBO show Game of Thrones is notorious for its complicated plots, its many characters, and its many character deaths. The dataset we will be examining in this report investigates time until character death in this show and how demographic variables may possibly be used to predict their time until death. The demographic variables include social status and sex of the character. This dataset also includes two variables relating to the characters' allegiance: the characters' final allegiance before death and a binary variable indicating if the character switched allegiances at some point in the series. Another variable of interest is the characters' prominence in the show, which was quantified as follows:

*prominence = ([# episodes character appeared in] / [# survived episodes]) * [# survived seasons].*

Lastly, this dataset provides us with the cumulative net running time when the character died in both hours and seconds, and a censoring variable. This dataset looks only at "important" characters, with important being defined by the original research article as:

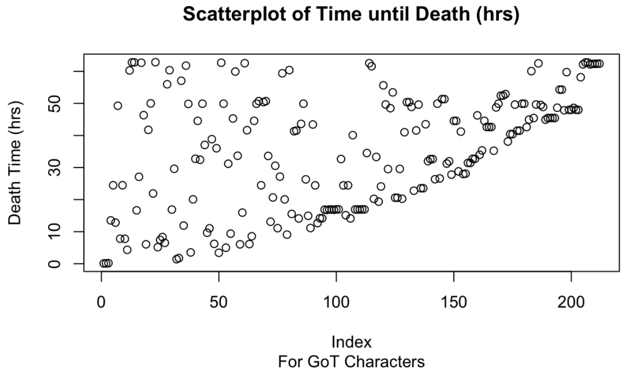
> *An important character was defined as any individual who fulfilled each of the following criteria: human; listed in either the opening or closing credits; appeared on screen during current events (i.e. excluding flashbacks); and was not already deceased when first appearing on screen. Additional non-credited characters were included if they interacted with another character in a way that was either crucial to the storyline or character development. Having a speaking role was not an essential requirement because some characters were unable to speak for medical reasons (e.g. acquired brain injury and non-elective glossectomy).*

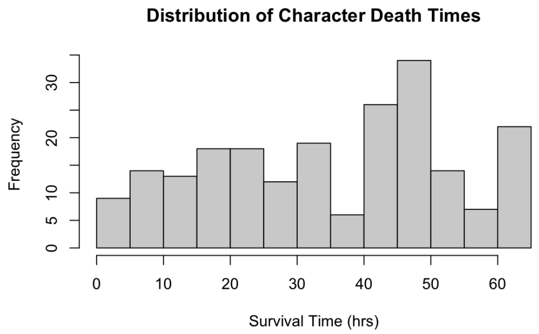
The time to event variable we will be focusing on is cumulative running time of the show until death in hours, with the beginning of time being the beginning of the series. This could be right censored if the character survives through the entire series, meaning that all censored observations will be equal to the total running time (in hours) of the series. All eight seasons are included in this dataset.

Main Report

Parametric Survival Analysis:

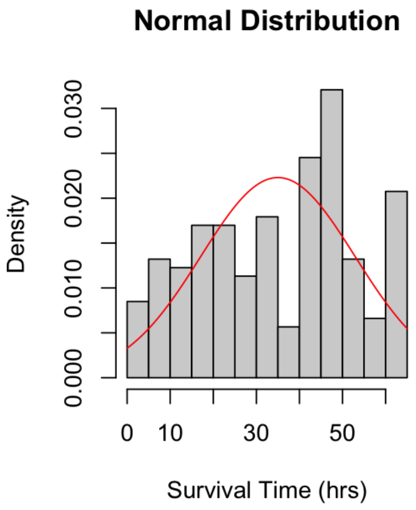
Initial Data Exploration Plots:

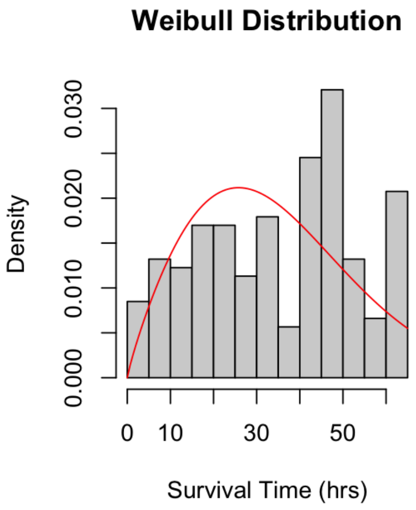




To fit a probability distribution to our time-to-event data for the time until death in hours of Game of Thrones Character, we first decided to examine the noticeable features of our response variable. With the initial data exploration of the response variable: time until death (hrs) we noticed that values for the response tended to be centered around 50 hours and a scatterplot of these response values tended to show a positive and increasing trend in time until death. The values for our response variable appear to be skewed to the left and take only non-negative values. Looking at the values that the response variable took, we saw that the variable "time until character death (hrs) was quantitative and continuous. We had a large number of recorded values for the response variable as this data set considered every Game of Thrones character and whether or not this character died during the course of the first eight seasons. Therefore, we had 212 characters with complete event times, who had died, and 147 characters with right censored death times - who had survived. Therefore, we decided to find a parametric model for our time to event variable T, or the time until death for a Game of Thrones Character, by plotting pdf functions for known distributions over our time to event data. The initial pdf's that we considered were the Normal pdf and the Weibull pdf. To find the parameters for our distribution that would best match the data, we used a function in R called "fitdistr" to best match the values of our time-until-death response variable.

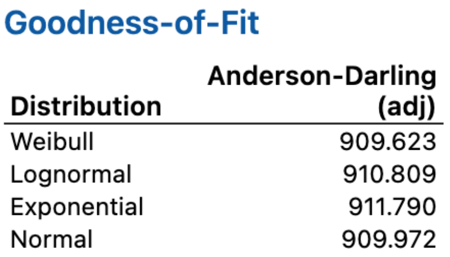
Fitting Test Distribution Pdfs:





Parameters from R output:

Distribution:	Parameters:
Normal	Shape = 1.869, Scale = 38.8437
Weibull	Mean = 35.004, sd = 17.8907



Goodness-of-Fit

Distribution	Anderson-Darling (adj)
Weibull	909.623
Lognormal	910.809
Exponential	911.790
Normal	909.972

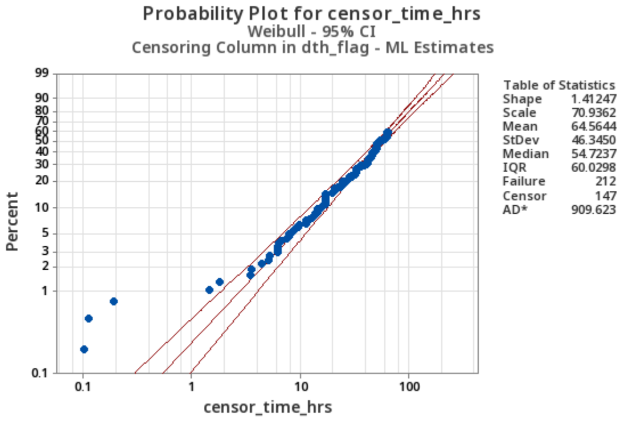
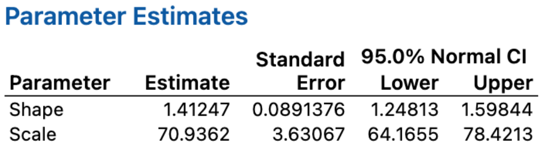


Table of Statistics	
Shape	1.41247
Scale	70.9362
Mean	64.5644
StDev	46.3450
Median	54.7237
IQR	60.0298
Failure	212
Censor	147
AD*	909.623



Parameter Estimates

Parameter	Estimate	Standard Error	95.0% Normal CI Lower	Upper
Shape	1.41247	0.0891376	1.24813	1.59844
Scale	70.9362	3.63067	64.1655	78.4213

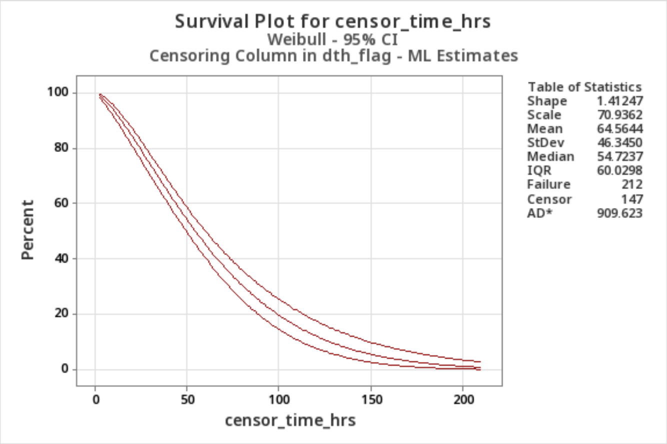
We analyzed the “Goodness-of-Fit” test statistics and the Weibull and Normal Probability Plots to make our decision about the parametric distribution. The Anderson-Darling statistics with the lowest values were the Weibull and the Normal distributions. Therefore, we decided to plot the pdfs overlaying our data, shown above, to compare the fits of the distributions. The Normal distribution pdf with the above specified parameters seems to have its peak closer to the center of the data when compared to the Weibull distribution. Furthermore, since there is still a lot of density in the response values around times closer to the center of the histogram, both the Normal and the Weibull distributions capture this. Since our data is heavily skewed left and the Weibull distribution is generally skewed to the right - this distribution does not appear to fit the data as well as the Normal distribution. However, the Goodness of Fit statistics for the Normal and the Weibull were only a few decimal places in difference, 909.96 and 909.62 respectively, so we decided to choose the Weibull distribution as the probability distribution that best fits our time-to-event data - with parameters. This distribution has the best Goodness of Fit and the Probability plot shows that most of the points are contained within the confidence limits - with the exception of a few points with extremely low character survival times. This Probability plot was more accurate than the Normal probability plot in terms of fitting the data.

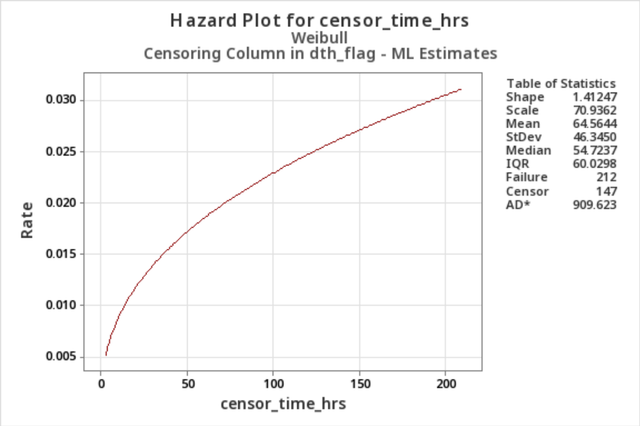
The survival function and the hazard function for all Game of Thrones characters are given below. Using the Weibull distribution with parameters $\beta = 1.412$ and $\lambda = 70.936$ from the Minitab output, the survival and hazard functions for all characters in the Game of Thrones data set are:

$$S(t)\text{: } \exp[(-\ (\frac{t}{70.936})^{1.412}]$$

$$h(t): \exp[\frac{1.412t^{0.412}}{70.936}]$$

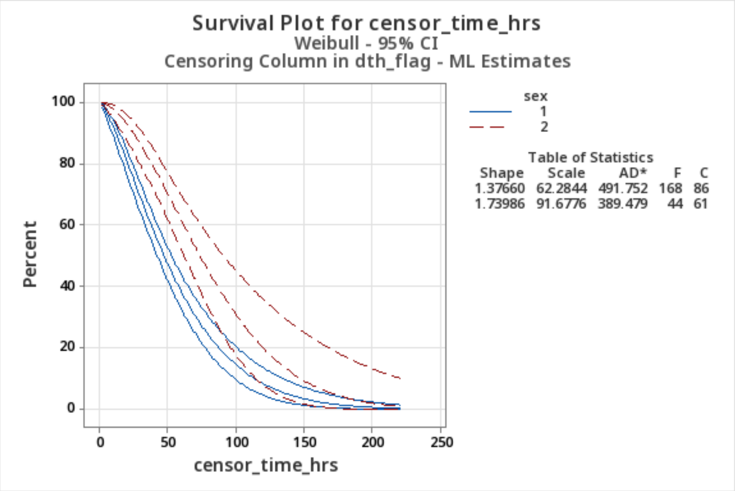
Survival and Hazard plots for all Character Death Times:

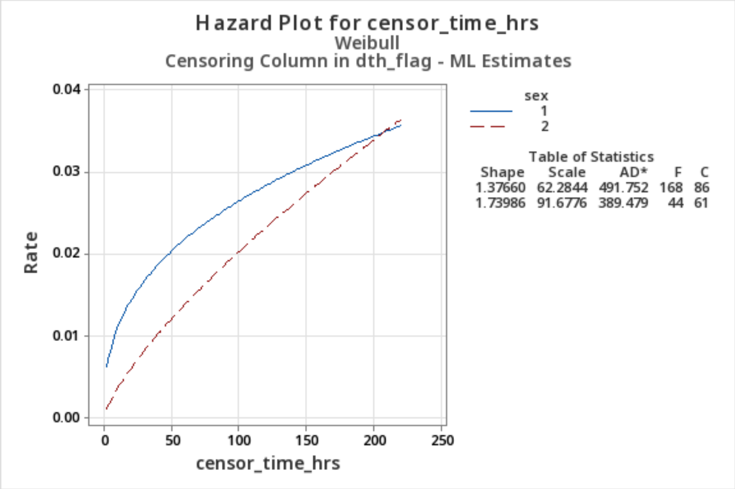




The survival curve for the time until death for all characters shows a gradual decline - with a good portion of character deaths occurring after about 55 to 100 hours of screen time. By about 100 to 150 hours of screen time - the probability that a randomly selected Game of Thrones character survives beyond this point is very low. The survival function appears to level off at around 150 - 200 hours of screen time. This means that up to 150 hours of screen time - the majority of characters who have died during the first eight seasons have already been killed. Looking at the overall hazard plot, we can see that there is a steady increase in the risk of death for Game of Thrones characters as their screen time increases. There is no point in the plot where the slope of the hazard function changes drastically, however, there appears to be a slight increase in slope between about 30 and 50 hours of screen time. This means that the risk of a character dying increases as their screen time increases. It is important to note that the hazard function does not level off - the slope appears to be non-zero towards the end of the recorded screen time value of 250 - 300 hours.

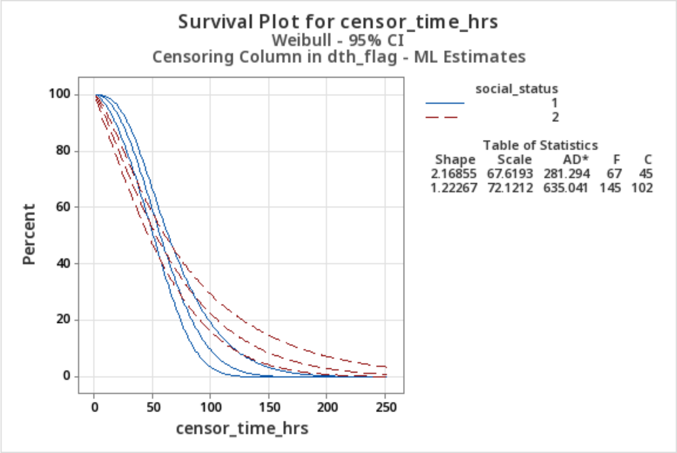
Survival and Hazard plots for Character Death Times by Sex:

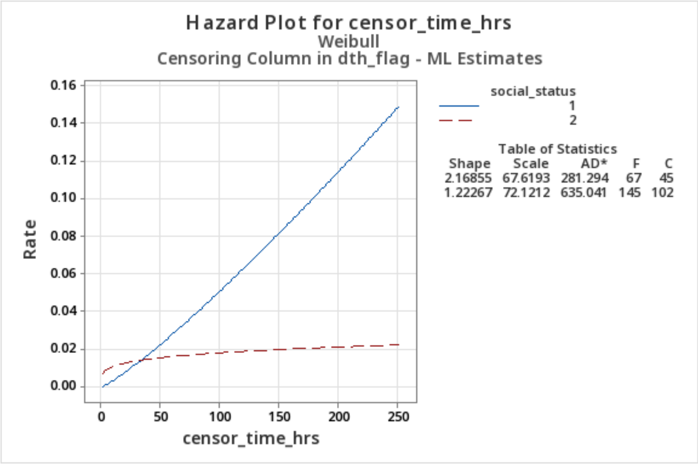




Now, we can compare the survival experiences of Characters in Game of Thrones by their sex. Game of Thrones characters who were Male tended to die before characters who were Female. The median time to death for Male characters was much lower than the median time to death for Female characters - for all screen times. The proportion of Game of Thrones characters who have not died after any particular number of screen time in hours is higher for characters who are female. Therefore, there appears to be a difference in the survival experiences between male and female Game of Thrones characters for all screen times. Now, comparing the hazard plots, we can see that the risk of character death increases for males the most around 30 to 70 hours of screen time while the risk of death increases for females rather steadily. Given that the characters, both male and female, have lived to the same amount of screen time in hours, the risk of death in the next instant is higher for males than for females.

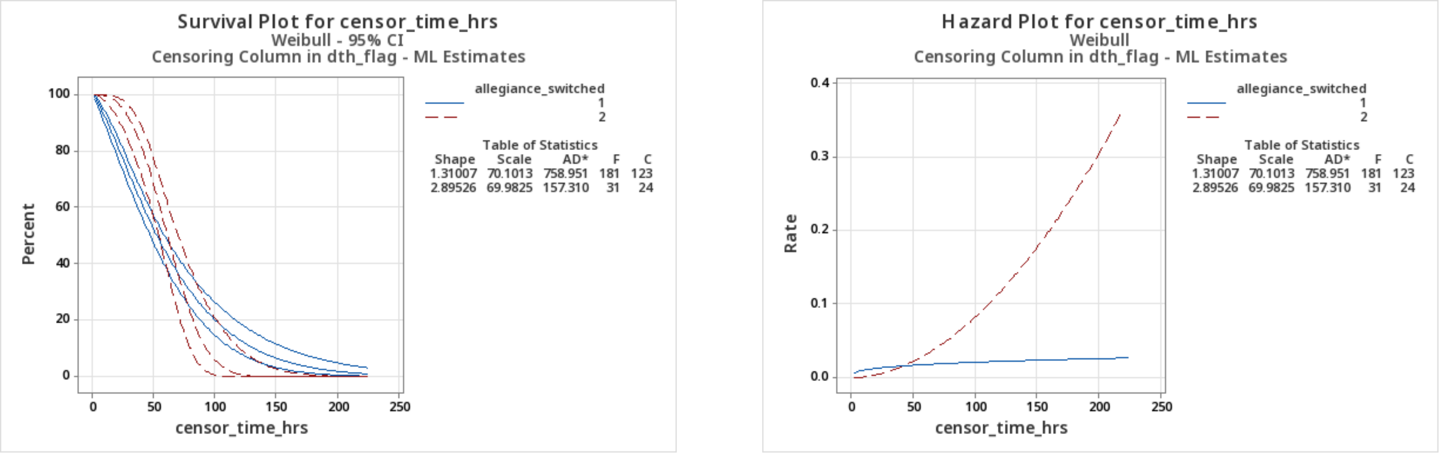
Survival and Hazard plots for Character Death Times by Social Status:





Now, we can compare the survival experiences of Characters in Game of Thrones by their social status. The survival curve for highborn characters appears to be higher than that for lowborn characters from screen times of 0 to 70 hours, then the survival curve is lower from times 70 - 250 hours. This means that from screen times 0 - 70 hours, the proportion of highborn characters who have not died is smaller than the proportion of lowborn characters still alive. The opposite is true for screen times 70 - 250 hours. Comparing the hazard plots between characters with high social status and low social status, we can see that the risk of death for characters with high social status is almost always greater than the risk of death for characters with low social status. The expectation to this is between - to zero to 30 hours of screen time - where the hazard curves and the relationships between them are reversed. In general, the hazard of death continuously increases for both highborn and lowborn characters - however- the rate of increase, or the slope, for highborn characters is much greater than that of lowborn characters. The slope of the hazard plot for lowborn characters is close to 0 - this line is almost horizontal. However, the slope for the hazard of death for highborn characters is roughly constant throughout all screen times, however, large. In general, given that both highborn and lowborn characters survive to the same number of hours in screen time, the risk of death in the next instant is higher for highborn characters.

Survival and Hazard plots for Character Death Times by Allegiance Switched:



Now, we can compare the survival experiences of Characters in Game of Thrones by whether or not they switched allegiances throughout the show. Comparing the survival curves, from screen times 0 - 60 hours, the proportion of characters who have switched allegiances and who have not died is larger than the proportion of characters who have switched allegiances and are still alive . The opposite is true for screen times 60 - 250 hours. However, the survival curve for characters who have switched allegiances levels out around 110 hours, while the curve for characters who have not switched allegiances steadily decreases through to 250 hours of screen time. Comparing the hazard plots between characters who have and have not switched allegiances, we can see that the risk of death for characters who have not is almost always greater than the risk of death for characters who have. The expectation to this is between - to zero to 30 hours of screen time - where the hazard curves and the relationships between them are reversed. In general, the hazard of death continuously increases for both characters who have and have not switched allegiances- however- the rate of increase, or the slope, for characters switching allegiances is much greater than that of characters who do not. The slope of the hazard plot for characters who have not switched allegiances is close to 0 - this line is almost horizontal. However, the slope for the hazard of death for characters who have is roughly constant throughout all screen times, however, large. In general, given that both characters who do and do not switch allegiances survive to the same number of hours in screen time, the risk of death in the next instant is higher for highborn characters.

The tables below report the mean and median survival times for all individuals and for the various groups we have examined: sex, social status, and allegiance switching.

Table of GoT Character Survival Time by Group Summaries:

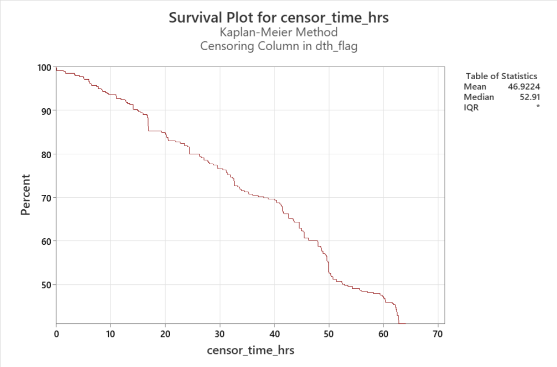
Group	Value	Mean (hrs.)	Median (hrs)
Sex	Male	56.922	50.42
Sex	Female	81.677	70.16

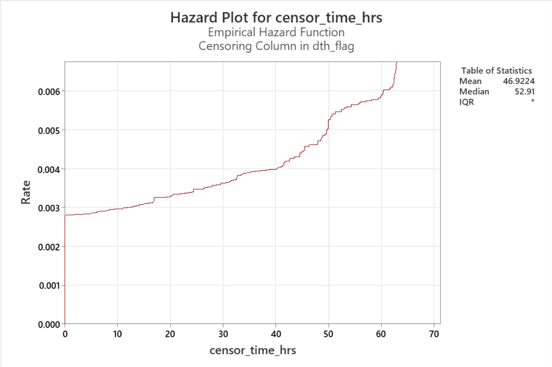
Social Status	HighBorn	59.994	57.46
Social Status	LowBorn	67.5236	54.75
Allegiance_Switch	No	64.64	54.31
Allegiance_Switch	Yes	62.399	61.84

Table of Survival Time Summary for all GoT Characters:

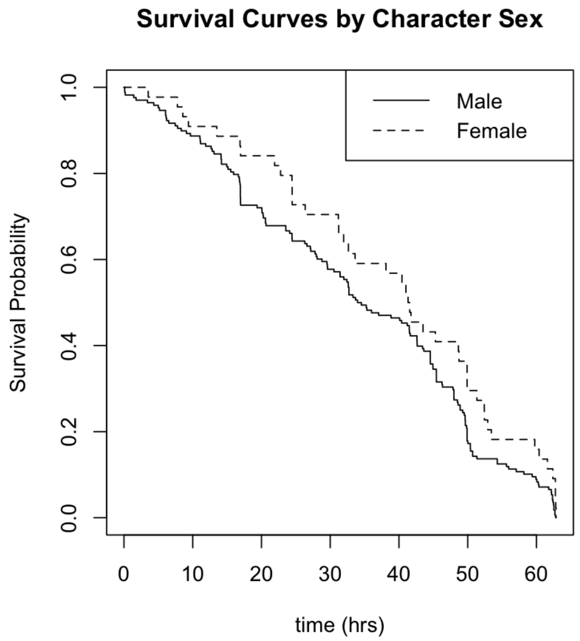
	Mean(hrs)	Median(hrs)
Overall Survival Time	64.56	54.72

Non-parametric Survival Analysis:

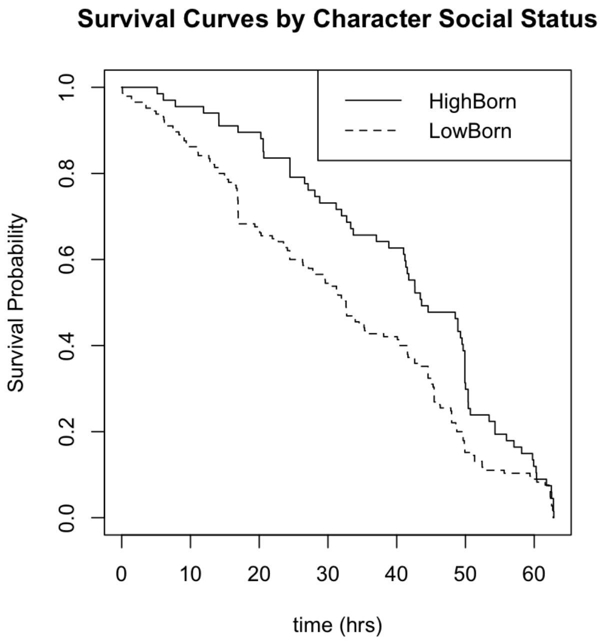




The survival experience for an average Game of Thrones character is decently steady throughout the amount of runtime, there is no specific value of lifetime where the survival rate changes, except around hour 50 where the curve levels, indicating there is a less significant change in survival rate for several hours. Similarly, the average Game of Thrones character's hazard of death is steady, except for around hour 50, where the hazard of death remains 0.005 for a few hours.

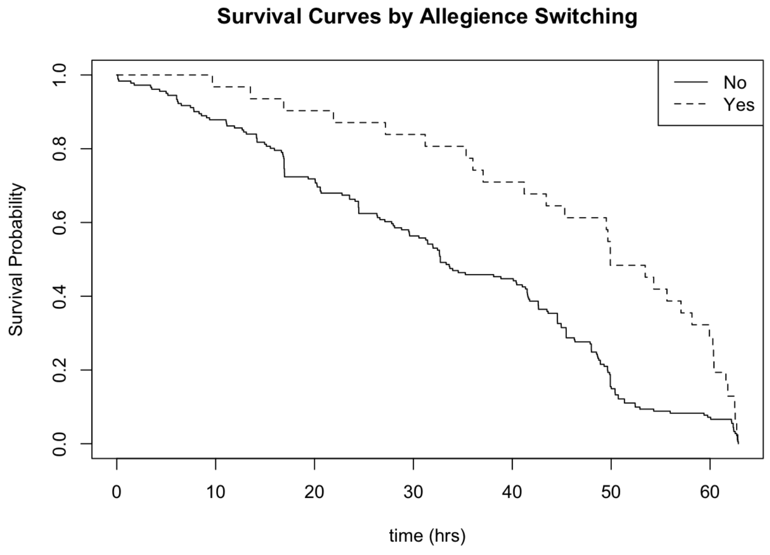


Survival Curves by Character Sex



Survival Curves by Character Social Status

When we use a survival plot to detail the survival rates between male and female characters, we can see that at any hour of runtime, female characters have a greater probability of surviving beyond any hour of runtime than male characters. In the case of using a character's social status to predict their survival rate, we can see that if a character is High Born, they are most likely to survive past any hour of runtime, for any hour of runtime. The closest these status' survival rates get is at 50 hours of runtime, indicating that many High Born characters died then.



Survival Curves by Allegience Switching

Characters who switch their allegiance at any point in the series tend to survive past any hour of runtime at a higher rate than characters who stay true to their alliance at any hour of running time.

Sex:

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	18.6586	1	0.000
Wilcoxon	18.8184	1	0.000

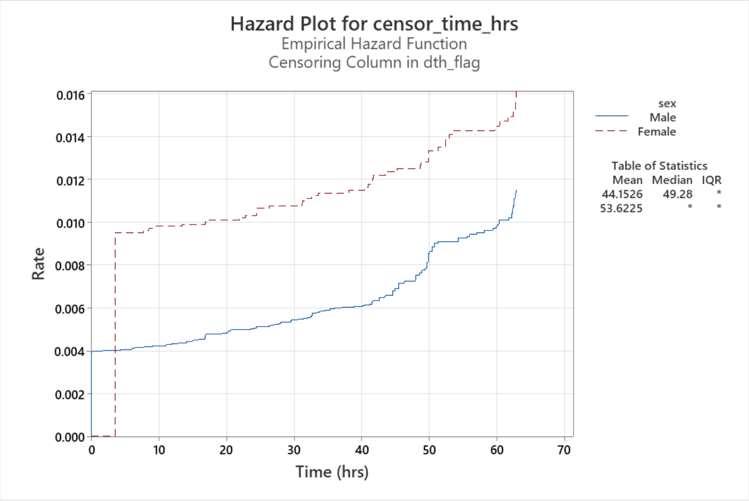


Table of Statistics

Mean	Median	IQR
44.1526	49.28	*
53.6225	*	*

The hazard of death for female characters is always higher than males after around 3 hours of runtime, it implies it is the first time a female is killed on the show, with the ratio of the amount of dead established female characters to alive females to be more extreme than males. The Log-Rank and Wilcoxon tests are both significant p-values of <.0001, but the Wilcoxon test does have the slightly higher Chi-Square value of 18.8184.

Allegiance Switch:

Test Statistics:

Method	Chi-Square	DF	P-Value
Log-Rank	1.88610	1	0.170
Wilcoxon	4.48723	1	0.034

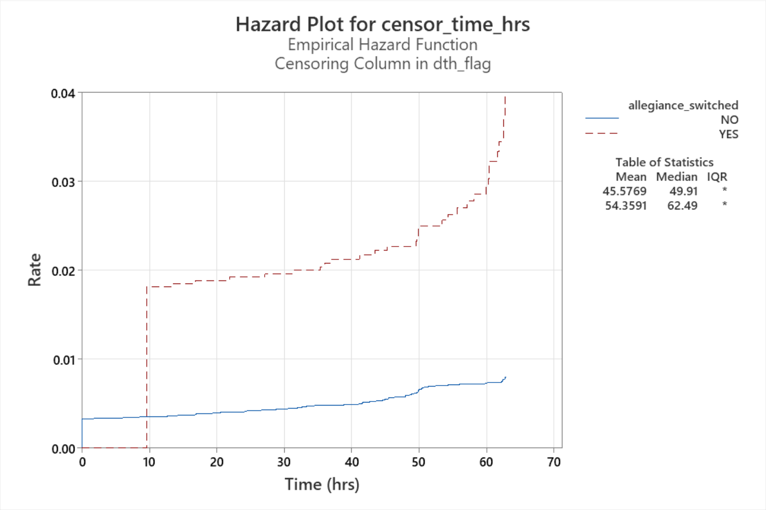


Table of Statistics

Mean	Median	IQR
45.5769	49.91	*
54.3591	62.49	*

The hazard of death for characters who switched allegiances throughout the series is greater for any hour runtime compared to characters who never switched their allegiance. The Log-Rank test is not statistically significant (p-values > 0.005) , but Wilcoxon is at the 5% level of significance (p-value = 0.034).

Social Status:

Test Statistics

Method	Chi-Square	DF	P-Value
Log-Rank	0.43389	1	0.510
Wilcoxon	2.02810	1	0.154

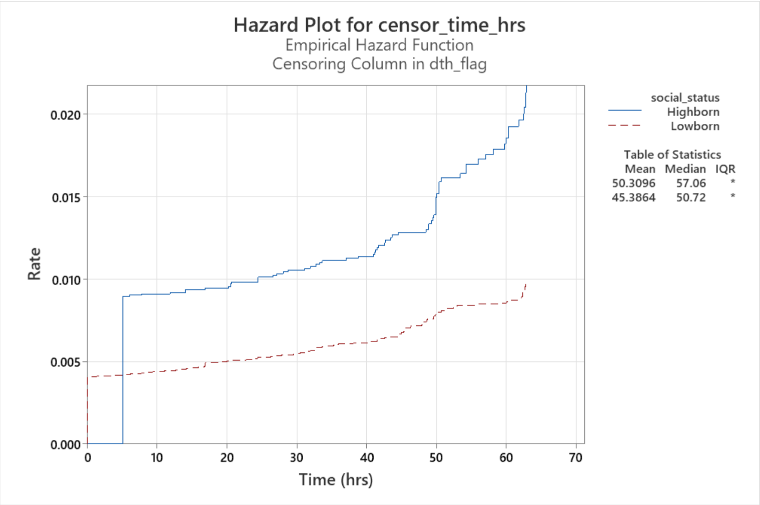


Table of Statistics

Mean	Median	IQR
50.3096	57.06	*
45.3864	50.72	*

As for the social status of characters, Highborn Characters have a higher hazard of death for any hour of runtime than Lowborn characters. However, the Log-Rank test statistic is not significant, with a p-value of 0.051 and chi-square value of 0.43389.

Group	Value	Mean (hrs.)	Median (hrs)
Sex	Male	44.15	49.28
Sex	Female	53.62	*
Social Status	HighBorn	50.31	57.06
Social Status	LowBorn	45.39	50.72
Allegiance_Switch	No	45.58	49.91
Allegiance_Switch	Yes	54.36	62.49
Allegiance_Last	Stark	50.94	*
Allegiance_Last	Targaryen	50.06	61.60
Allegiance_Last	Night's Watch	39.09	38.85
Allegiance_Last	Lannister	48.67	62.16
Allegiance_Last	Greyjoy	30.89	16.95
Allegiance_Last	Bolton	39.08	41.57
Allegiance_Last	Frey	58.13	*
Allegiance_Last	Other	45.38	48.90
Allegiance_Last	Unknown	51.72	*

The mean and median for each group/value were higher in the parametric than the non-parametric analysis. The mean and median for the overall survival time for characters is higher in the parametric than the non-parametric analysis. However, the median of the parametric and the non-parametric are far closer than the mean. The plots in the parametric analysis are easier to interpret and offer an overall indication of how the data is acting; the non-parametric plots give a more in depth and instantaneous look at how the data is acting, especially at specific time stamps.

Regression Analysis:

Our final Cox regression model uses sex of character, allegiance switch and prominence as predictor variables for hazard of character death:

$$\textit{Hazard of Death} = \beta_0 + \beta_1 \textit{Sex} + \beta_2 \textit{Allegiance Switched} + \beta_3 \textit{Prominence}$$

The following are the coefficient estimates for the model, along with standard error and Wald test results. All predictor variables are significantly associated with the hazard of death at the 0.05 significance level.

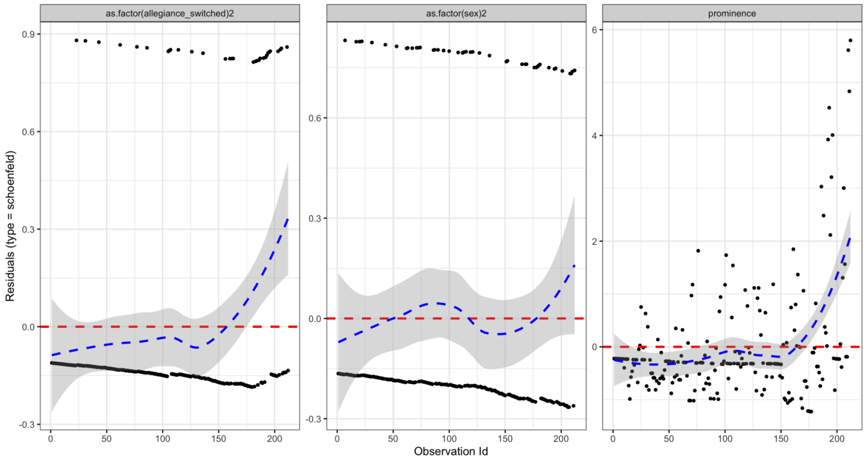
Variable	Coefficient	Standard Error	Z Statistic	P-value
Sex (Female)	-0.73023	0.17006	-4.294	<0.001
Allegiance Switched (Yes)	-0.60851	0.23333	-2.608	0.00911
Prominence	0.12693	0.05179	2.451	0.01426

Sex and Allegiance are both binary variables and prominence is a quantitative variable. We decided to include these variables as it includes some information on demographics on the character as well as decisions made. Below is the result of tests for overall model significance. All tests show that the overall model is significantly better than the null model.

Test	Statistic	P-value
Likelihood Ratio Test	28.9	<0.001
Wald Test	25.92	<0.001
Log Rank Test	26.85	<0.001

To determine if model assumptions are met, we investigated the Schoenfeld residuals. Below are the plots.

Schoenfeld residuals vs character residuals

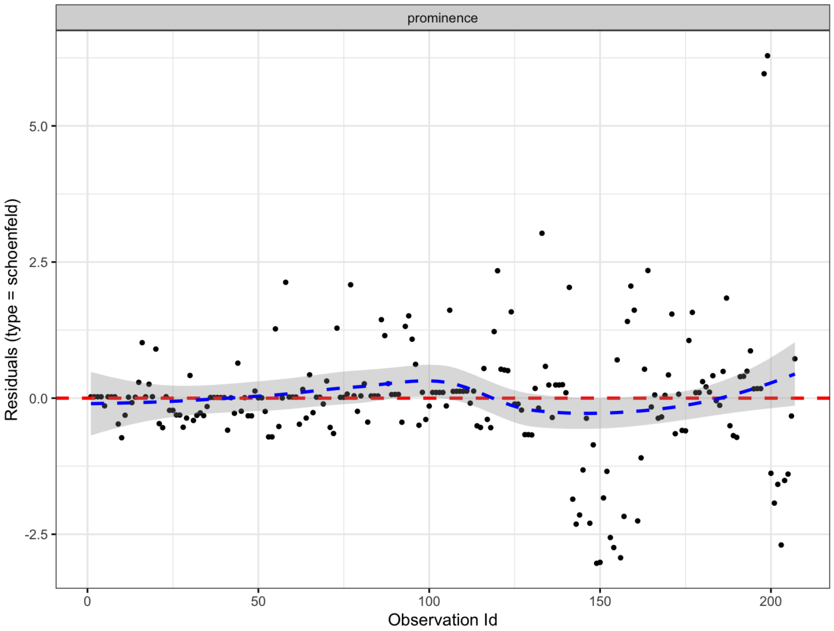


The Schoenfeld residuals reveal that the proportionality of hazards assumption is violated for all three predictor variables, as the Loess fit is not flat and has obvious deviations from the red line. This means that the relationship between these variables and hazard of death depends on time. To address this, we took out influential observations (1, 2, 3, 38 ,46, 617, 620, 634, 635, 636, for a total of 10 observations removed) and stratified the model based on sex and allegiance switched variables. After doing this, our model provided the following estimates, with prominence still being significant at the 0.05 significance level:

Variable	Coefficient	Standard Error	Z Statistic	P-value
Prominence	0.11006	0.05343	2.06	0.0394

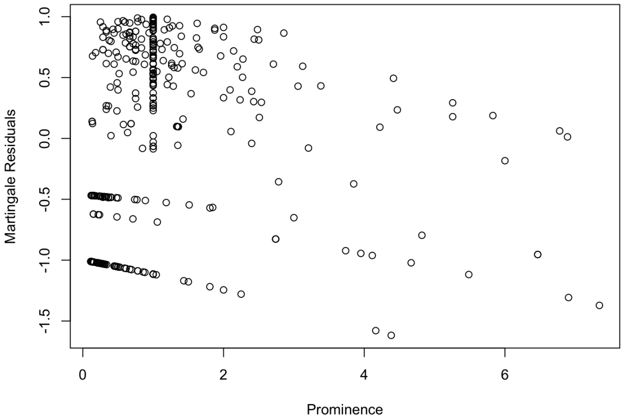
After these adjustments, the proportionality of hazards looks much better, as seen with the Schoenfeld residuals:

Schoenfeld Residuals vs Character ID



Additionally, the Martingale residuals do not reveal any major violations with linearity.

Martingale Residuals vs Prominence



There are some weird patterns present, which we will keep in mind when generalizing results, but do not reveal an issue large enough to completely scrap the model.

To investigate this model further, we will look at the hazard ratios for a one point increase in prominence, the maximum prominence value vs the minimum prominence value, as well as the mean value of prominence vs the median value. The hazard ratio for a one point increase in prominence is 1.11635 (95% CI: 1.005, 1.24), meaning that the risk of death for a Game of Thrones character increases by 11% for each additional point increase in prominence beyond any time t, after adjusting for sex and allegiance switch. The maximum value of prominence is 7.3425 and the minimum is 0.1111, with a corresponding hazard ratio of 2.2164 (95% CI: 1.037, 4.738) meaning that the a character of the highest prominence has a 121.6% higher risk of death compared to a character with the lowest prominence beyond any time t, adjusting for sex and allegiance switch. The mean value of prominence is 1.13104 while the median prominence is 0.8409, giving us a corresponding hazard ratio of 1.0325 (95% CI: 1.0014, 1.0644), meaning a character of average prominence has a 3.25% higher risk of death compared to a character of median prominence beyond anytime t, adjusting for sex and allegiance switch. All of these hazard ratios are significant at the 0.05 significance level, though the last one is barely significant.

Conclusion:

To begin the parametric analysis, we decided that the Weibull Distribution with parameters $\beta = 1.412$ and $\lambda = 70.936$ would fit our data the best. We decided this from overlaid pdfs, the Goodness of Fit statistics, and Probability plot for the Weibull Distribution. From the parametric analysis, there were several differences in the survival experiences and hazard of character death between the groups that we analyzed: sex, social status, and whether or not the character switched allegiances. Since our data was skewed, the median time (in screentime hours) until death for all characters in Game of Thrones was around 54.72 hours. The median time until death for Female characters, highborn characters, and characters who have switched allegiances is higher than the overall median. This supports the observations we made from the survival and hazard plots.

The proportion of Game of Thrones characters who have not died after any particular number of screen time in hours is higher for characters who are female. Furthermore, given that both highborn and lowborn characters survive to the same number of hours in screen time, the risk of death in the next instant is higher for highborn characters. Finally, the risk of death for characters who have not switched allegiances is almost always greater than the risk of death for characters who have.

Most important findings from CR model include the finding that more prominent characters are at higher risk of death, sex and allegiance switch are significantly associated with hazard of death, though the stratification on these variables lends itself to the model better. All predictor variables are significantly associated with the hazard of death at the 0.05 significance level and the relationship between our variables of interest and hazard of death depends on time.

The non-parametric analysis yielded similar and different results. The median screentime until death for all characters was approximately 52.91 hours. Female characters showed higher survival probabilities than male characters, similarly, Highborn characters had higher survival rates than Lowborn characters, and characters who switched allegiances survived longer than those who did not. Furthermore, Log-rank tests p-values showed that sex and allegiance switching were significantly associated with the screentime until the death of Game of Thrones characters. Additionally, a character's final allegiance was significantly associated with the screentime until the death for the characters.

Appendix:

Parametric Analysis:

~~~~
got <- read.csv("character_data.csv")

got2 <- got[!is.na(got$dth_time_hrs), ]

fitdistr(got2$dth_time_hrs, "weibull")
fitdistr(got2$dth_time_hrs, "lognormal")
fitdistr(got2$dth_time_hrs, "normal")

par(mfrow = c(1,2))
plot(got2$dth_time_hrs, main = "Scatterplot of Time until Death (hrs)", sub = "For GoT Characters", ylab = "Death Time (hrs)")

hist(got2$dth_time_hrs, freq = TRUE, main = "Distribution of Character Death Times", xlab = "Survival Time (hrs)")
```


par(mfrow = c(1,2))
hist(got2$dth_time_hrs, freq = FALSE, main = "Normal Distribution", xlab = "Survival Time (hrs)")
curve(dnorm(x, mean=35.004,sd=17.89072), add = T, col = "red")

hist(got2$dth_time_hrs, freq = FALSE, main = "Weibull Distribution", xlab = "Survival Time (hrs)")
curve(dweibull(x, shape =  1.8690105, scale = 38.8437382 ), add = T, col = "red")

#fit <- survfit(cr.object,newdata=pred.vals)

par(mfrow = c(1,2))
fit2 <- survfit(Surv(dth_time_hrs,dth_flag)~sex, data=got2)
plot(fit2, lty = 1:2, xlab = "time (hrs)", ylab = "Survival Probability", main = "Survival Curves by Character Sex")
legend("topright",c("Male","Female"),lty=1:2)


fit3 <- survfit(Surv(dth_time_hrs,dth_flag)~social_status, data=got2)
plot(fit3, lty = 1:2, xlab = "time (hrs)", ylab = "Survival Probability", main = "Survival Curves by Character Social Status")
legend("topright",c("HighBorn","LowBorn"),lty=1:2)

fit4 <- survfit(Surv(dth_time_hrs,dth_flag)~allegiance_switched, data=got2)
plot(fit4, lty = 1:2, xlab = "time (hrs)", ylab = "Survival Probability", main = "Survival Curves by Allegiance Switching")
legend("topright",c("No","Yes"),lty=1:2)

plot.haz(fit2)

fit <- survfit(Surv(dth_time_hrs,dth_flag)~1, data=got2)
plot(fit)

fit2 <- survfit(Surv(dth_time_hrs,dth_flag)~sex, data=got2)
plot(fit2)
~~~~

```
lines(fit2$dth_time_hrs,fit2$dth_flag,type="s",lty=2)
legend(20,.8,c("Male","Female"),lty=1:2)
```

Cox Regression Model:

```
library(survminer)
library(tidyverse)
library(survival)
got<- read.csv('character_data_S01-S08.csv', header = T)

null<-coxph(Surv(censor_time_hrs,dth_flag)~1,data = got)
summary(null)

model1 <- coxph(Surv(censor_time_hrs,dth_flag)~ prominence + as.factor(social_status)+as.factor(sex), data=got)
summary(model1)

model2 <- coxph(Surv(censor_time_hrs,dth_flag)~ prominence +as.factor(sex), data=got)
summary(model2)

model3 <- coxph(Surv(censor_time_hrs,dth_flag)~ as.factor(social_status)+as.factor(sex), data=got)
summary(model3)

model4 <- coxph(Surv(censor_time_hrs,dth_flag)~ prominence + as.factor(occupation)+as.factor(sex), data=got)
summary(model4)

model5 <- coxph(Surv(censor_time_hrs,dth_flag)~  as.factor(occupation)+as.factor(sex)+as.factor(allegiance_switched) +
intro_season, data=got)
summary(model5)

model6 <- coxph(Surv(censor_time_hrs,dth_flag)~ as.factor(sex)+as.factor(allegiance_switched) + intro_season, data=got)
summary(model6)

finalmodel <- coxph(Surv(censor_time_hrs,dth_flag)~ as.factor(sex)+as.factor(allegiance_switched) + prominence, data=got)
summary(finalmodel)

ggcoxdiagnostics(finalmodel, type='schoenfeld')

deviance <- residuals(finalmodel,type="deviance")
which(deviance >= abs(2.5))

schoen <- residuals(finalmodel,type = "schoenfeld")
which(schoen >= 4)

revisedmodel <- coxph(Surv(censor_time_hrs,dth_flag)~ strata(as.factor(sex))+strata(as.factor(allegiance_switched)) +
prominence, subset = (-c(1,2,3,38,46,617,620,634,635,636)), data=got)
summary(revisedmodel)

ggcoxdiagnostics(revisedmodel, type='schoenfeld')
martin <- residuals(revisedmodel, type = "martingale")
prominence <- got$prominence[-c(1,2,3,38,46,617,620,634,635,636)]
plot(prominence,martin, main = 'Martingale Residuals vs Prominence', xlab = 'Prominence', ylab = 'Martingale Residuals')

minprom<- min(prominence)
maxprom<- max(prominence)
```

```
medprom<- median(prominence)
meanprom<- mean(prominence)

exp((maxprom-minprom)*0.11006)
1.005^7.2314
1.24^7.2314
exp((meanprom-medprom)*0.11006)
1.005^0.290139
1.24^0.290139
```