Donner Party Survival Analysis

Joshua Barney, Emeric Szaboky, Kieren Patel, Bridget Haus

November 25, 2016

R Markdown

Introduction

The dataset we will be using for our survival data analysis comes from the University of Florida archives of Miscellaneous Datasets. This dataset consists of 89 members of the Donner Party from 1846-1847. We designed a new variable called duration which is the time until event, calculated by subtracting the join date from the death date (with the mean duration inserted for missing censored values). We also designed 3 more variables named famsize (size of nuclear family), groupsize (size of group including teamsters), and isteamster (indicator function whether subject is a teamster or not). We are censoring all observations who survived the donner party tragedy. Since many of our findings are censored we will have to adjust our findings accordingly.

Our goal for this survival data analysis is to see whether the covariates age, sex, famsize, groupsize, camp, trapped, and isteamster had an impact on the survival of an observed donner party member. We will proceed with Kaplan-Meier plots, log-rank tests, log-log plots, Cox (proportional hazard) regression models with confidence intervals, AIC to find the best fit model, and random survival forests to support our findings and conclusions. The Kaplan-Meier plots will be used to visualize the survival rates of donner party members with differing covariates. The log-rank tests will be used to test if there is a difference between two or more of our survival curves. The Cox regression models will be used to test the proportional hazards model assumptions to see which coefficients for each covariate has the most statistical significance in predicting the evolving survival proabability for different groups. The random survival forests will be used to find an alternative, non-parametric survival probability prediction model through supervised learning techniques. This model will allow us to ignore parametric constraints imposed by the Cox regression models. Additionally, we will allow for more abstract relationships between covariates.

Donner Party Background

Milford Elliott: Cannibalism

We found additional data that Milford Elliott, a man, age 28, who died Februrary 9th, 1847, was most likely the first person in the donner party to be canabalized. There was also a significant number of people, trapped in the mountains (correlating to a 1 for the 'trapped' covariate), who died during February and March of 1847 (many more died in February and March as opposed to the suspected winter months of December and January). This evidence could suggest that after the canabalism of Elliott, food rations were low, and many of the surviving people were sacrificed as food. Further evidence for this is shown in the Kaplan-Meier plot controlled for the 'trapped' covariate, since many of the members of the Donner Party trapped in the mountains survived longer than those not trapped. This was concerning at first, however, with analysis, we found it could suggest many of the trapped people who died early were sacrificed in order to help the rest of the trapped people survive.

"The Forlorn Hope"

"The Forlorn Hope" was a party within the Donner Party of 17 men, women, and children who set out on foot in the snow (12 ft. deep) to cross the mountain pass when food rations became low. They became trapped

in the mountains without food. Eventually, debate arose about sacrificing people in the party to feed the others, and it was decided that the party would wait until people fell or died on their own before anyone was eaten. Antonio, a teamster, and Franklin Graves were the first two members of the party to die. Patrick Dolan and Lemuel Murphy died next. They were preserved as food. The party agreed to not allow people to eat their own relatives; this information could provide evidence for why the children and women (wives) had better survival rates than the men. Families were cherished and protected before single individuals. This is supported in the Kaplan-Meier plots, which illustrate women and children having significantly stronger survival rates.

The Miwok Native American mountain guides Salvador and Luis, along with William Eddy refused to eat human flesh. The party members discussed killing Salvador and Luis for food, however William Eddy warned them, and the two men were able to escape. The party found Salvador and Luis near death several days later and killed them for food.

William Eddy was led out of the mountains by a Miwok Native American whom the party had met on their journey. He was able to haphazardly organize a rescue party to save the other remaining six members of "The Forlorn Hope". These survivors are documented in the data with the rescue 'party' covariate value of 0 (0 attributed to survival without rescue) because they were able to make it almost all the way out of the mountains without dying on their own.

William Eddy was later an integral member of the first rescue party ('party' covariate value of 1).

James Reed

James Reed, 45, was banished from the Donner Party and made it out of the mountains alive on his own. He was an integral part of the second rescue party ('party' covariate value of 2). It was an anomaly that all of the members of the Reed family survived.

Setup: Read In And Organize Data

```
# Read in original data
donner <- read.csv("~/Desktop/PSTAT 175/group project/donner.txt", header=FALSE, sep=";")

# Assign covariate names
names <- donner$V1
age <- donner$V2
sex <- donner$V3
survive <- donner$V4
deathdate <- donner$V5
party <- donner$V6
joindate <- donner$V7
trapped <- donner$V8
camp <- donner$V9

# Family Size
famsize <- c(rep(7,7),rep(9,9),rep(1,3),rep(6,6),2,2,rep(1,3),
             rep(9,9),1,rep(4,4),rep(13,13),rep(4,4),1,1,2,2,1,1,
             rep(3,3),rep(12,12),rep(1,7))

# Traveling Group Size (Including Teamsters)
groupsize <- c(rep(26,19),rep(11,11),rep(10,10),rep(4,4),rep(13,13),rep(6,4),25,6,
               2,2,25,6,rep(3,3),rep(13,13),rep(25,4),2,2)
```

```
# Indicator vector for whether or not the subject was a teamster
isteamster <- c(rep(0,16),rep(1,3),rep(0,6),rep(1,5),rep(0,9),1,rep(0,21),1,1,
                0,0,1,1,rep(0,15),rep(1,7))

# Trim out whitespace and convert to date format
joindate <- trimws(joindate)
joindate <- as.Date(joindate, "%m/%d/%Y")

deathdate <- trimws(deathdate)
deathdate <- as.Date(deathdate, "%m/%d/%Y")

# New vector of the amount of days the subject was in the party for
duration <- as.numeric(deathdate - joindate)

# Ammend duration vector by filling in missing observations with column mean (176)
duration <- as.numeric(duration)
duration[is.na(duration)] <-
  round(mean(duration, na.rm = T))

# Reformat the "survive" variables
survive <- (1-survive)

# Reformat the Camp Variable
camp <- as.character(camp)
camp[camp != "AC" & camp != "LC"] <- "None"
camp <- as.factor(camp)

# Create final dataset
data <- data.frame(names, age, sex,
                           survive, deathdate,
                           party, joindate,
                           trapped, camp, duration,
                           famsize, groupsize, isteamster)
```

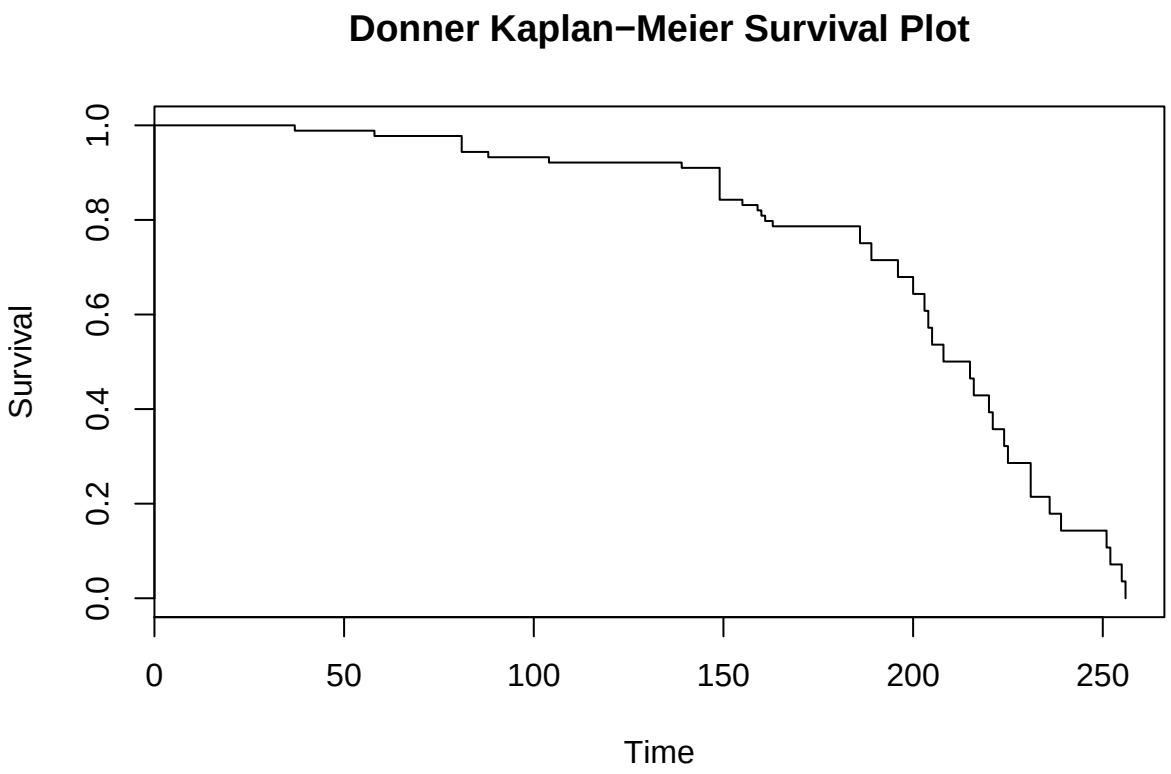
Kaplan-Meier Plots

General Kaplan-Meier

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.2.5
```

```
donner.surv <- Surv(duration, survive)
donner.fit <- survfit(donner.surv~1)
plot(donner.fit, conf.int=FALSE, xlab="Time", ylab="Survival", main="Donner Kaplan-Meier Survival Plot")
```

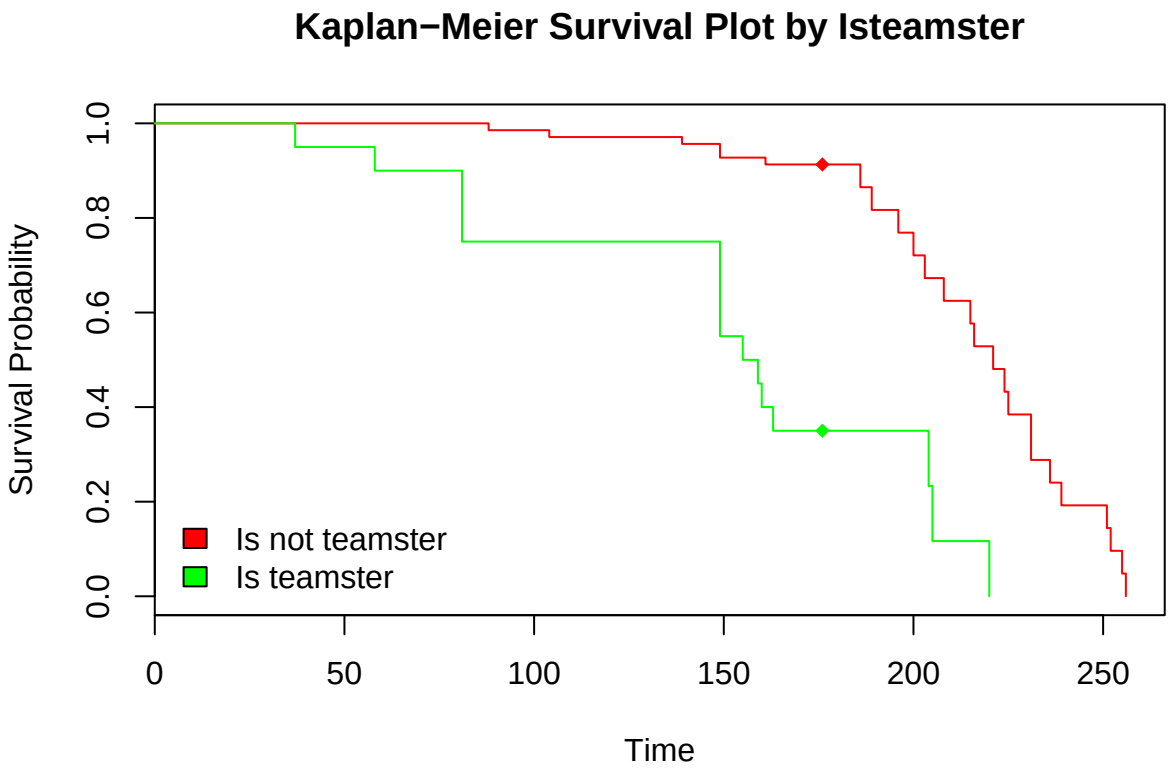


This plot describes the survival probability of all subjects across time.

Kaplan-Meier With Respect To Isteamster

```
donner.split.isteamster <- survfit(donner.surv ~ isteamster)

plot(donner.split.isteamster, conf.int=FALSE, main="Kaplan-Meier Survival Plot by Isteamster", xlab="Ti
legend("bottomleft",c("Is not teamster","Is teamster"),fill=c("red","green"),bty="n")
```



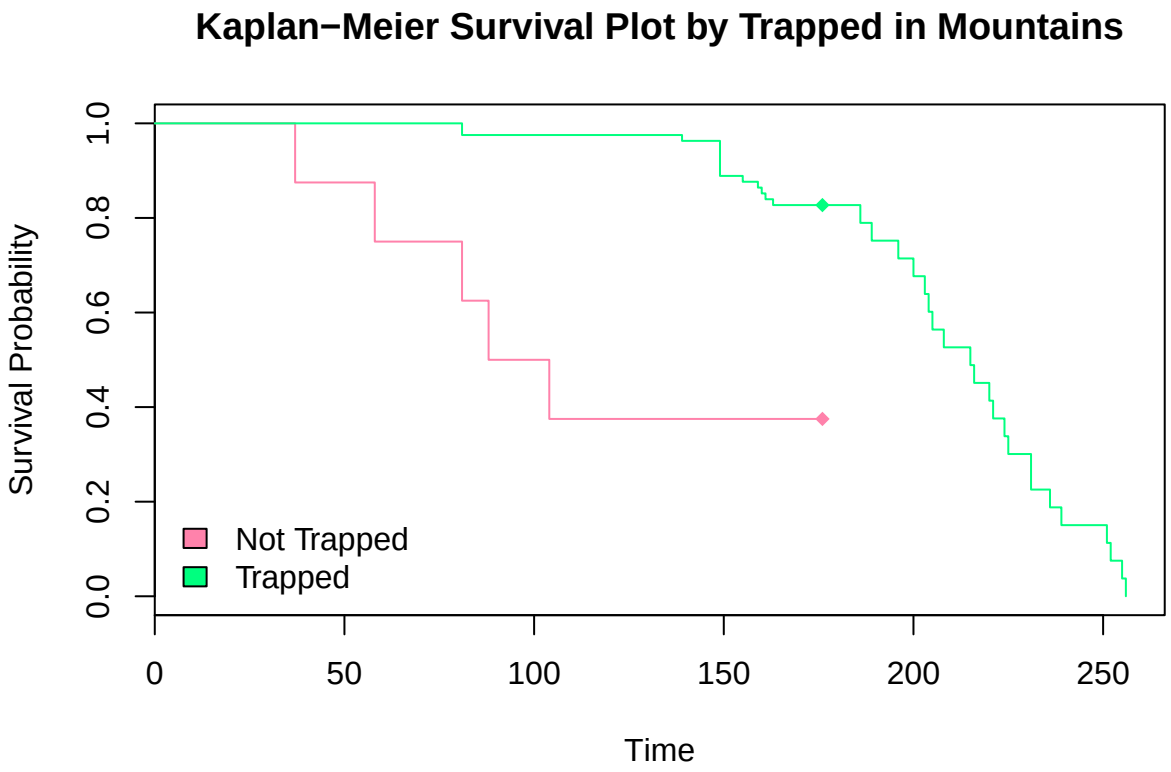
This plot describes the survival probability of teamster and non-teamster subjects across time.

The plot above illustrates teamsters having a worse survival rate than those who were not teamsters. This could potentially be explained by the fact that teamsters had to exert more of their energy than non-teamsters to take care of other creatures. Even more plausible could be the fact that teamsters were often single men without families, and thus more likely to be cannibalized early on. This is supported by the Kaplan-Meier plot for sex showing men with lower survival rates than women. It is also supported by background information on the Donner Party occurences and in the Cox proportional hazards strongest model by the interaction between the covariates sex and trapped.

Kaplan-Meier With Respect To Trapped In Mountains

```
donner.split.trap <- survfit(donner.surv ~ trapped)

plot(donner.split.trap,conf.int=FALSE,main="Kaplan-Meier Survival Plot by Trapped in Mountains",xlab="T
legend("bottomleft",c("Not Trapped","Trapped"),fill=c("palevioletred1","springgreen1"),bty="n")
```



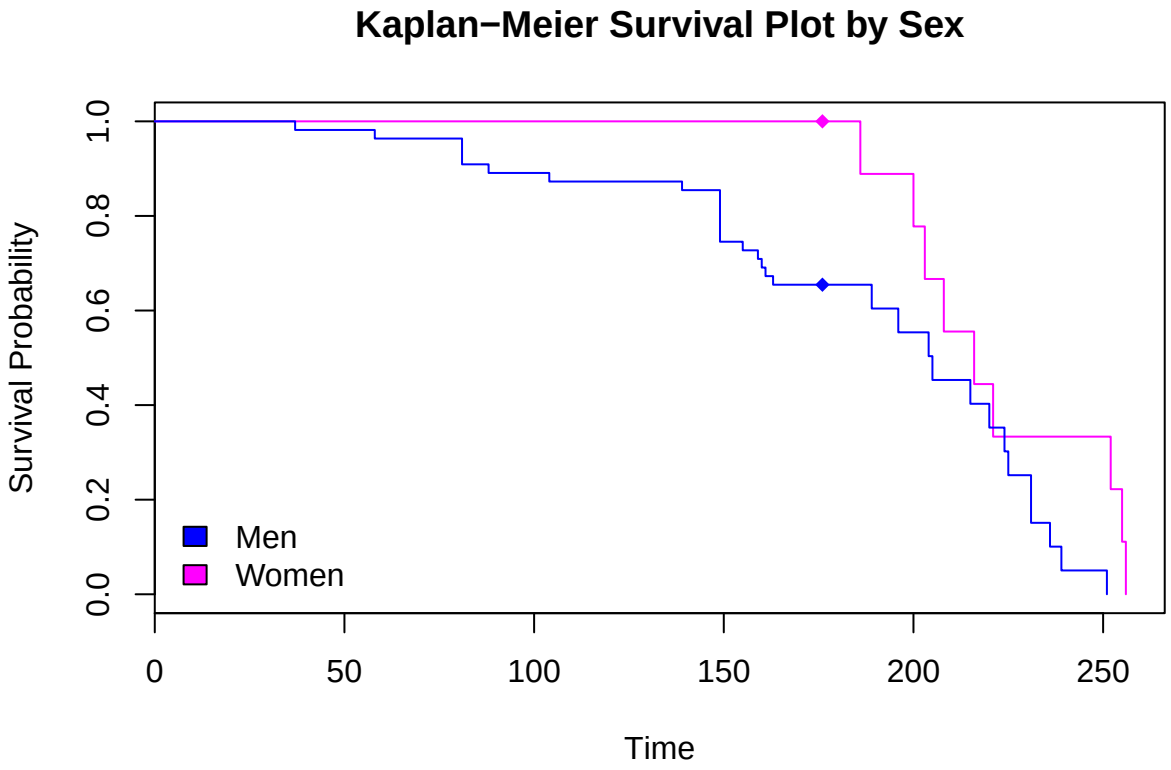
This plot describes the survival probability of trapped and not-trapped subjects across time.

The plot above illustrates those trapped in the mountains having better survival rates than those not trapped. This discrepancy is in large part due to a significantly smaller number of people having not been trapped in the mountains. The data does not provide an strongly accurate comparison between trapped and not trapped since the sizes of observations for each are so vastly different. Another reason for this relationship could be that both groups ran out of food rations around similar times and those trapped in the mountains resorted sooner to cannibalism, which kept many who were trapped surviving longer.

Kaplan-Meier With Respect To Sex

```
donner.split.sex <- survfit(donner.surv ~ sex)

plot(donner.split.sex,conf.int=FALSE,main="Kaplan-Meier Survival Plot by Sex",xlab="Time",ylab="Survival
legend("bottomleft",c("Men","Women"),fill=c("blue","magenta"),bty="n")
```



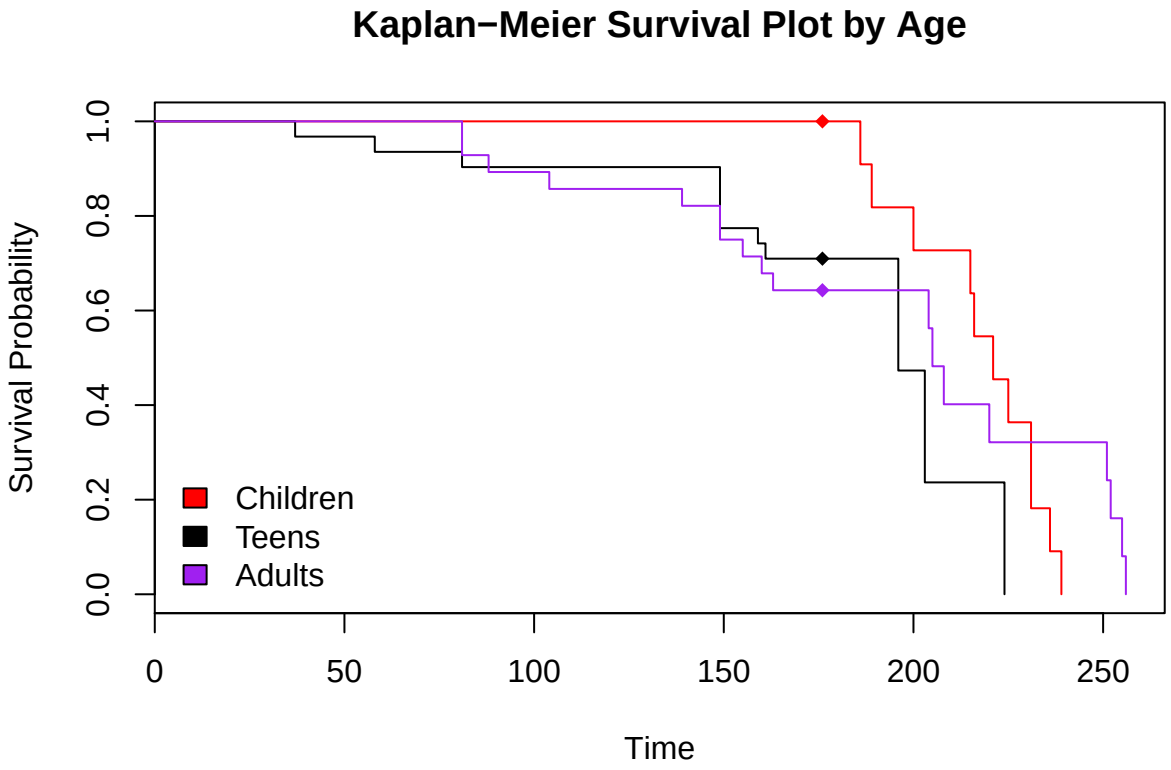
This plot describes the survival probability of male and female subjects across time.

The plot above illustrates females having better survival rates than males. This is largely due to respect for familial relations among members of the Donner Party and the effort that was given to not separate nuclear families or cannibalize women, children, and men with families. Single men (bachelors) had a much lower survival rate and a high hazard risk of being cannibalized.

Kaplan-Meier With Respect To Age

```
table <- within(data, quartile <- as.integer(cut(age, quantile(age, probs=0:3/3), include.lowest=TRUE)))
q6 <- as.factor(table$quartile)

donner.split.age <- survfit(donner.surv ~ q6)
plot(donner.split.age,conf.int=FALSE,main="Kaplan-Meier Survival Plot by Age",xlab="Time",ylab="Survival
legend("bottomleft",c("Children","Teens","Adults"),fill=c("red","black","purple"),bty="n")
```



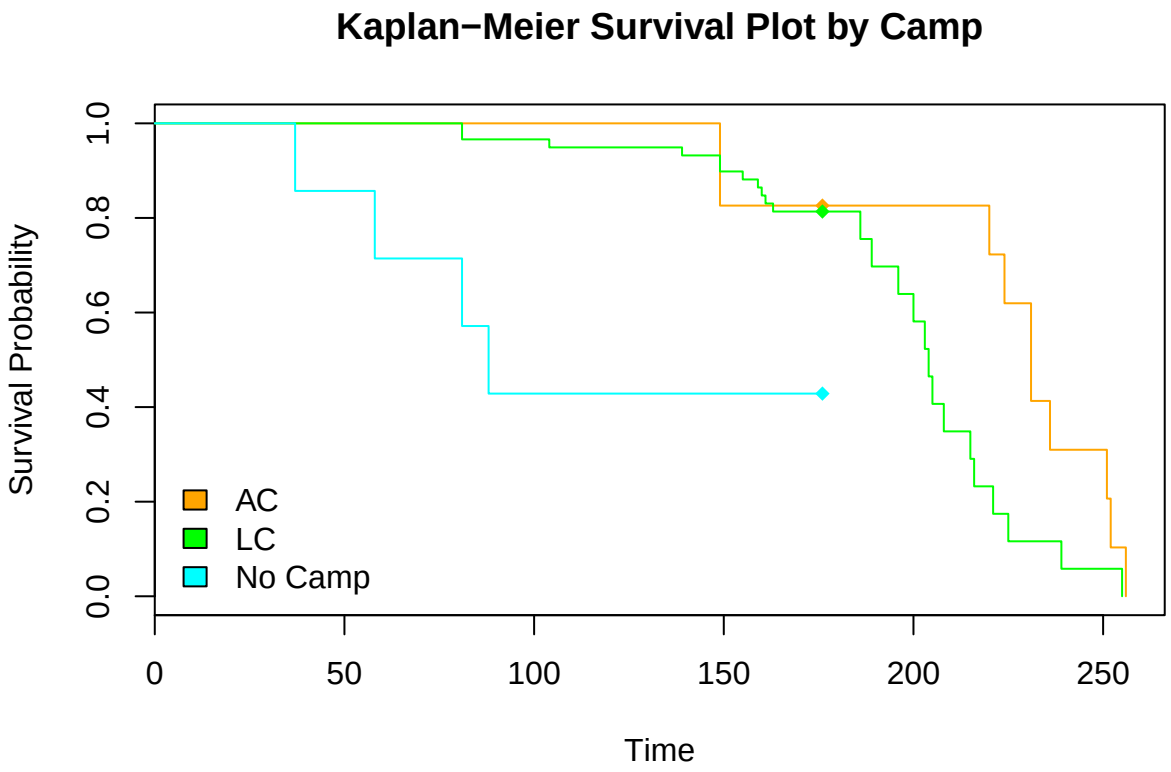
This plot describes the survival probability children, teen, and adult subjects across time.

The plot above illustrates children having the best survival rates for the majority of the recorded observations in time. There were a few adults, however, that survived longer than any children or teens. Teens and adults have relatively similar survival curves with the average survival rate for teens appearing lower than that of adults. The strong rate of survival of children is most likely due to the fact that they were cared for by adults and rescued earlier on. Teens could have had a higher hazard rate than adults because they were weaker or because they were single and more likely to be cannibalized.

Kaplan-Meier With Respect To Camp

```
donner.split.camp <- survfit(donner.surv ~ camp)

plot(donner.split.camp,conf.int=FALSE,main="Kaplan-Meier Survival Plot by Camp",xlab="Time",ylab="Survi
legend("bottomleft",c("AC","LC","No Camp"),fill=c("orange","green","cyan"),bty="n")
```



This plot describes the survival probability of Alden Creek, Lake Camp, and no camp subjects across time.

The plot above illustrates those from camp AC having the best survival rate, followed by LC, and then by No Camp. Subjects who weren't trapped generally belonged to no camp.

Cox Proportional Hazards Models With Confidence Intervals

Difference Between Sexes

```
difference_sexes <- coxph(donner.surv ~ sex, data = data)
summary(difference_sexes)
```

```
## Call:
## coxph(formula = donner.surv ~ sex, data = data)
##
##   n= 89, number of events= 41
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## sexM 1.3848    3.9940   0.4497 3.079  0.00208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sexM     3.994     0.2504     1.654     9.644
##
```

```
## Concordance= 0.688  (se = 0.052 )
## Rsquare= 0.13   (max possible= 0.948 )
## Likelihood ratio test= 12.44  on 1 df,   p=0.0004208
## Wald test            = 9.48  on 1 df,   p=0.002077
## Score (logrank) test = 10.98  on 1 df,   p=0.0009186
```

We reject the null hypothesis that the sex covariate is not significant. From the likelihood-ratio test, we have a p-value of $0.0004208 <$ alpha-level $= 0.05$. Therefore there is a significant difference between male and female. In addition, the p-value for the sex covariate is $0.00208 <$ alpha-level $= 0.05$, suggesting that in this model with the single covariate sex, sex is significant.

```
extractAIC(difference_sexes)
```

```
## [1]   1.0000 252.9325
```

1 covariate; AIC = 252.9325

Difference Between Sexes; Controlled For All Other Covariates

```
## 1 ##
difference_controlled <- coxph(donner.surv ~ sex+age+trapped+famsize+groupsize+isteamster, data = data)
summary(difference_controlled)
```

```
## Call:
## coxph(formula = donner.surv ~ sex + age + trapped + famsize +
##     groupsize + isteamster, data = data)
##
##   n= 89, number of events= 41
##
##                 coef exp(coef)  se(coef)      z Pr(>|z|)
## sexM        1.032238  2.807340  0.499135  2.068  0.03863 *
## age         0.003103  1.003108  0.012386  0.251  0.80217
## trapped    -0.943383  0.389308  0.566112 -1.666  0.09563 .
## famsize     0.003992  1.004000  0.063395  0.063  0.94979
## groupsize  -0.067927  0.934329  0.023820 -2.852  0.00435 **
## isteamster  1.809605  6.108036  0.593768  3.048  0.00231 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## sexM          2.8073     0.3562    1.0554     7.467
## age           1.0031     0.9969    0.9791     1.028
## trapped       0.3893     2.5687    0.1284     1.181
## famsize       1.0040     0.9960    0.8867     1.137
## groupsize     0.9343     1.0703    0.8917     0.979
## isteamster    6.1080     0.1637    1.9076    19.558
##
## Concordance= 0.85  (se = 0.06 )
## Rsquare= 0.383   (max possible= 0.948 )
## Likelihood ratio test= 42.92  on 6 df,   p=1.209e-07
## Wald test            = 42.55  on 6 df,   p=1.432e-07
## Score (logrank) test = 58.12  on 6 df,   p=1.083e-10
```

```r
# LRT1: p-value = p=1.209e-07
# significant covariates for model: sex, groupsize, isteamster

extractAIC(difference_controlled)
```

```
## [1]   6.0000 232.4491
```

```r
# 6 covariates; AIC(1) = 232.4491

## 2 ##
all_except_sex <- coxph(donner.surv ~ age+trapped+famsize+groupsize+isteamster, data = data)
summary(all_except_sex)
```

```
## Call:
## coxph(formula = donner.surv ~ age + trapped + famsize + groupsize +
##     isteamster, data = data)
##
##   n= 89, number of events= 41
##
##                 coef exp(coef)  se(coef)      z Pr(>|z|)
## age        -0.005012  0.995000  0.012295 -0.408 0.683511
## trapped    -1.221147  0.294892  0.558013 -2.188 0.028642 *
## famsize    -0.008823  0.991216  0.060990 -0.145 0.884977
## groupsize  -0.058428  0.943246  0.023318 -2.506 0.012220 *
## isteamster  2.077112  7.981389  0.586111  3.544 0.000394 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## age           0.9950     1.0050   0.97131    1.0193
## trapped       0.2949     3.3911   0.09878    0.8803
## famsize       0.9912     1.0089   0.87953    1.1171
## groupsize     0.9432     1.0602   0.90111    0.9874
## isteamster    7.9814     0.1253   2.53035   25.1754
##
## Concordance= 0.782  (se = 0.06 )
## Rsquare= 0.348   (max possible= 0.948 )
## Likelihood ratio test= 38.12  on 5 df,   p=3.567e-07
## Wald test            = 41.6  on 5 df,   p=7.092e-08
## Score (logrank) test = 55.72  on 5 df,   p=9.266e-11
```

```r
# LRT2: p-value = 3.567e-07
# significant covariates for model: trapped, groupsize, isteamster

# LRT1 - LRT2 = 42.92 - 38.12 = 4.8 ; df1 - df2 = 6 - 5 = 1

pchisq(4.8,1,lower.tail = F)
```

```
## [1] 0.02845974
```

```
# from pchisq(4.8,1, lower.tail = F) :
# ^found that df=1, LRT1 - LRT2= 4.8 correlates to an approximate p-value = 0.02845974

extractAIC(all_except_sex)
```

```
## [1]   5.0000 235.2482
```

```
# 5 covariates; AIC(2) = 235.24827
```

P-value (sex covariate) = 0.02845974 (rounds to 0.03) AIC(1) = 6 covariates; 232.4491 AIC(2) = 5 covariates; 235.24827

Because 0.02845974 < alpha-level = 0.05, we reject the null hypothesis that there is not a significant difference between males and females when controlling for all other covariates. Therefore, there is a significant difference between males and females in this case.

Searching For The Strongest Model: Difference Between Sexes; Adjusted For Significance

```
difference_controlled_adj <- coxph(donner.surv ~ sex+trapped+groupsize+isteamster, data = data)
summary(difference_controlled_adj)
```

```
## Call:
## coxph(formula = donner.surv ~ sex + trapped + groupsize + isteamster,
##     data = data)
##
##   n= 89, number of events= 41
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## sexM        1.00080   2.72046  0.48330  2.071  0.03838 *
## trapped    -0.97823   0.37598  0.54367 -1.799  0.07197 .
## groupsize  -0.06609   0.93605  0.02236 -2.955  0.00313 **
## isteamster  1.81681   6.15218  0.40704  4.464 8.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## sexM          2.7205     0.3676    1.0550     7.015
## trapped       0.3760     2.6597    0.1295     1.091
## groupsize     0.9361     1.0683    0.8959     0.978
## isteamster    6.1522     0.1625    2.7705    13.662
##
## Concordance= 0.841  (se = 0.06 )
## Rsquare= 0.382   (max possible= 0.948 )
## Likelihood ratio test= 42.86  on 4 df,   p=1.107e-08
## Wald test            = 42.61  on 4 df,   p=1.249e-08
## Score (logrank) test = 58.09  on 4 df,   p=7.304e-12
```

```
# LRT: p-value = 1.107e-08
# significant covariates for model: sex, groupsize, isteamster
```

We reject the null hypothesis that this model is not significant. From the likelihood-ratio test, we have a p-value of 1.107e-08 < alpha-level = 0.05. Therefore, the model is significant and there is a significant difference between males and females. In addition, the p-value for the sex covariate is 0.03838 < alpha-level = 0.05, suggesting that in this model with four covariates, sex is still significant.

```
extractAIC(difference_controlled_adj)
```

```
## [1]   4.0000 228.5112
```

4 covariates; AIC = 228.5112

After thorough testing through step-wise model selection, we found that the above model, difference-controlled-adj, possessing the lowest AIC value of 228.5112 is the strongest model. In this model, three (sex, groupsize, isteamster) out of four covariates are significant. Although the trapped covariate is not significiant, including it in the model lowers the AIC. This suggests that there is some interaction between the trapped covariate and another covariate which lowers this value. The p-value for the sex covariate is affected largely by the addition of the trapped covariate to the model. Both the sex and trapped covariates are significant when tested alone in single-covariate models. This may suggest a relationship between the sex and trapped covariates. With contextual analysis, a relationship between these two covariates could make sense, considering the lives of many women who were trapped in the mountains were spared before the lives of men trapped in the mountains. This trend is visualized in the Kaplan-Meier plot and is explained in the background information to be due to the family life component of men taking care of their nuclear families before themselves. There was a significant amount of respect amongst the members of the Donner Party for families; this is why the survival rates of families were higher than those of single men and why single men were often the first to be cannibalized. We will assume that the trapped covariate has some positive effect on the results.

Further Reduced Model To Prove Removing Trapped Covariate Raises AIC

```
coxph_reduced_signif <- coxph(donner.surv ~ sex+groupsize+isteamster, data = data)
summary(coxph_reduced_signif)
```

```
## Call:
## coxph(formula = donner.surv ~ sex + groupsize + isteamster, data = data)
##
##   n= 89, number of events= 41
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## sexM        1.08705   2.96550  0.47806  2.274 0.022975 *
## groupsize  -0.07244   0.93012  0.02164 -3.348 0.000814 ***
## isteamster  1.88653   6.59641  0.40511  4.657 3.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## sexM          2.9655     0.3372    1.1619    7.5688
## groupsize     0.9301     1.0751    0.8915    0.9704
```

```
## isteamster    6.5964     0.1516    2.9818   14.5927
##
## Concordance= 0.822  (se = 0.06 )
## Rsquare= 0.362   (max possible= 0.948 )
## Likelihood ratio test= 40.06  on 3 df,   p=1.035e-08
## Wald test            = 37.77  on 3 df,   p=3.158e-08
## Score (logrank) test = 48.11  on 3 df,   p=2.022e-10
```

```
# LRT: p-value = 1.035e-08
# significant covariates for model: all (sex, groupsize, isteamster)

extractAIC(coxph_reduced_signif)
```

```
## [1]   3.0000 229.3102
```

3 covariates; AIC = 229.3102

The p-values from stepwise AIC tell whether the two nested models differ significantly. According to AIC, the difference-controlled-adj is the strongest model with the following variables: sex, trapped, groupsize, isteamster. We cross-validated this model by removing one insignificant covariate at a time until all covariates in the model had significant p-values. In the above coxph-reduced-signif model, the insignificant trapped covariate is removed. Although all covariates (sex, groupsize, isteamster) have significant p-values, the AIC is larger than the AIC for the difference-controlled-adj model. This leads us to accept difference-controlled-adj as the strongest model.

Confidence Interval for Hazard Probability Ratio

```
# We concluded our best model was the following:

difference_controlled_adj <- coxph(donner.surv ~ sex+trapped+groupsize+isteamster, data = data)
summary(difference_controlled_adj)
```

```
## Call:
## coxph(formula = donner.surv ~ sex + trapped + groupsize + isteamster,
##     data = data)
##
##   n= 89, number of events= 41
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## sexM        1.00080   2.72046  0.48330  2.071  0.03838 *
## trapped    -0.97823   0.37598  0.54367 -1.799  0.07197 .
## groupsize  -0.06609   0.93605  0.02236 -2.955  0.00313 **
## isteamster  1.81681   6.15218  0.40704  4.464 8.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## sexM          2.7205     0.3676    1.0550     7.015
## trapped       0.3760     2.6597    0.1295     1.091
## groupsize     0.9361     1.0683    0.8959     0.978
## isteamster    6.1522     0.1625    2.7705    13.662
```

```
##
## Concordance= 0.841  (se = 0.06 )
## Rsquare= 0.382   (max possible= 0.948 )
## Likelihood ratio test= 42.86  on 4 df,   p=1.107e-08
## Wald test            = 42.61  on 4 df,   p=1.249e-08
## Score (logrank) test = 58.09  on 4 df,   p=7.304e-12
```

```
# A 95% confidence interval for the Male Sex coefficient is

1.00080 - 1.96*0.48330
```

```
## [1] 0.053532
```

```
1.00080 + 1.96*0.48330
```

```
## [1] 1.948068
```

```
# Thus the 95% confidence interval for hazard ratio of Male Sex relative to Female Sex is

exp(0.053532)
```

```
## [1] 1.054991
```

```
exp(1.948068)
```

```
## [1] 7.015121
```

Our confidence interval for the Male Sex relative to Female Sex hazard ratio is (1.001782,7.023418) therefore we would conclude that there is a significant difference between the two sexes with Male Sex having a greater hazard rate and therefore a shorter lifetime.

Log-Rank Tests & Log-Log Plots

Log-Rank Tests

For the following log-rank tests, we will assume the following null and alternative hypotheses:

$$H_0 : coefficient - for - tested - variable - equals - zero$$

vs.

$$H_a : coefficient - for - tested - variable - does - not - equal - zero.$$

Sex

```r
(donner.sex.lr <- survdiff(donner.surv ~ sex, data=data))
```

```
## Call:
## survdiff(formula = donner.surv ~ sex, data = data)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     18.8      5.12      10.9
## sex=M 55       32     22.2      4.35      10.9
##
##  Chisq= 10.9  on 1 degrees of freedom, p= 0.000962
```

```r
# Sex is significant: Chisq= 10.9 on 1 df, p-value = 0.000962 < alpha-level = 0.05
```

Trapped (In Mountains)

```r
(donner.trapped.lr <- survdiff(donner.surv ~ trapped, data=data))
```

```
## Call:
## survdiff(formula = donner.surv ~ trapped, data = data)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5    0.948     17.31      18.7
## trapped=1 81       36   40.052      0.41      18.7
##
##  Chisq= 18.7  on 1 degrees of freedom, p= 1.53e-05
```

```r
# Trapped is significant: Chisq= 18.7 on 1 df, p= 1.53e-05  < alpha-level = 0.05
```

Age

```r
age<- as.numeric(age)
age <- findInterval(age, c(0, 10, 26))
age[age == 1] <- "Under 10"
age[age == 2] <- "10 - 26"
age[age == 3] <- "Over 26"
age <- as.factor(age)

(donner.age.lr <- survdiff(donner.surv ~ age))
```

```
## Call:
## survdiff(formula = donner.surv ~ age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     7.86   2.18575    3.2239
## age=Over 26   28       18    17.71   0.00467    0.0103
## age=Under 10 29       11    15.43   1.27253    2.2957
##
##  Chisq= 3.9  on 2 degrees of freedom, p= 0.143
```

```
# Age is NOT significant: Chisq= 3.9  on 2 df, p-value = 0.143 > alpha-level = 0.05
```

Famsize

```
sizegroups <- findInterval(famsize, c(0, 5, 10))
sizegroups[sizegroups == 1] <- "Small"
sizegroups[sizegroups == 2] <- "Medium"
sizegroups[sizegroups == 3] <- "Large"

(donner.famsize.lr <- survdiff(donner.surv ~ sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ sizegroups)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large    25       10    11.46     0.186     0.269
## sizegroups=Medium   31        8    19.62     6.886    14.865
## sizegroups=Small    33       23     9.92    17.263    25.685
##
##  Chisq= 28  on 2 degrees of freedom, p= 8.46e-07
```

```
# Famsize is significant: Chisq= 28  on 2 df, p-value = 8.46e-07 < alpha-level = 0.05
```

Groupsize

```
groupsize <- findInterval(groupsize, c(0, 11, 20))
groupsize[groupsize == 1] <- "Small"
groupsize[groupsize == 2] <- "Medium"
groupsize[groupsize == 3] <- "Large"

(donner.groupsize.lr <- survdiff(donner.surv ~ groupsize))
```

```
## Call:
## survdiff(formula = donner.surv ~ groupsize)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## groupsize=Large     25       15    18.80   0.76846     1.684
## groupsize=Medium    37       14    14.31   0.00676     0.011
## groupsize=Small     27       12     7.89   2.14361     2.943
##
##  Chisq= 3.4  on 2 degrees of freedom, p= 0.183
```

```
# Groupsize is NOT significant: Chisq= 3.4  on 2 df, p-value = 0.183 > alpha-level = 0.05
```

Isteamster

```
(donner.isteamster.lr <- survdiff(donner.surv ~ isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ isteamster)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25     36.3      3.52      33.3
## isteamster=1 20       16      4.7     27.20      33.3
##
##  Chisq= 33.3  on 1 degrees of freedom, p= 7.9e-09
```

```
# Isteamster is significant: Chisq= 33.3  on 1 df, p-value = 7.9e-09 < alpha-level = 0.05
```

Joindate

```
joindate <- as.character(joindate)
joindate[joindate != "1846-07-19"] <- "After July 1846"
joindate[joindate == "1846-07-19"] <- "July 1846"
joindate <- as.factor(joindate)

(donner.joindate.lr <- survdiff(donner.surv ~ joindate))
```

```
## Call:
## survdiff(formula = donner.surv ~ joindate)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## joindate=After July 1846 19        8     4.69     2.328      2.88
## joindate=July 1846       70       33    36.31     0.301      2.88
##
##  Chisq= 2.9  on 1 degrees of freedom, p= 0.0895
```

```
# Joindate is NOT significant: Chisq= 2.9  on 1 df, p-value = 0.0895 > alpha-level = 0.05
```

Camp

```
(donner.camp.lr <- survdiff(donner.surv~camp))
```

```
## Call:
## survdiff(formula = donner.surv ~ camp)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC    23       12   18.204      2.11      4.57
## camp=LC    59       25   21.929      0.43      1.08
## camp=None   7        4    0.867     11.31     12.15
##
##  Chisq= 15.5  on 2 degrees of freedom, p= 0.000424
```

```
# Camp is significant:  Chisq= 15.5  on 2 df, p-value = 0.000424 < alpha-level = 0.05
```

After performing Log-Rank tests on each of our covariates, we conclude that our significant variables at the 0.05 level are: sex, trapped, famsize, isteamster and camp. Continous covariates were broken into factors of 2 to 3 levels so as to give more reasonable results.

Stratified Tests

Stratified On Sex

```
# Stratified on sex, all variables are significant
(survdiff(donner.surv~strata(sex) + age))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sex) + age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     6.96   3.65612    5.1798
## age=Over 26   28       18    17.60   0.00929    0.0211
## age=Under 10 29       11    16.45   1.80431    3.7444
##
##  Chisq= 6.4  on 2 degrees of freedom, p= 0.0407
```

```
(survdiff(donner.surv~strata(sex) + sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sex) + sizegroups)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large   25       10     10.1   0.00123   0.00179
## sizegroups=Medium 31        8     19.5   6.78537  15.34368
## sizegroups=Small   33       23     11.4  11.85133  19.22337
##
##  Chisq= 22.3  on 2 degrees of freedom, p= 1.4e-05
```

```
(survdiff(donner.surv~strata(sex) + isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sex) + isteamster)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25    34.48      2.61      19.7
## isteamster=1 20       16     6.52     13.77      19.7
##
##  Chisq= 19.7  on 1 degrees of freedom, p= 9.17e-06
```

```
(survdiff(donner.surv~strata(sex) + trapped))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sex) + trapped)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5     1.63     6.993      7.98
## trapped=1 81       36    39.37     0.289      7.98
##
##  Chisq= 8  on 1 degrees of freedom, p= 0.00472
```

```
(survdiff(donner.surv~strata(sex) + camp))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sex) + camp)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC   23       12    18.06      2.04      4.65
## camp=LC   59       25    21.44      0.59      1.50
## camp=None  7        4     1.49      4.21      4.76
##
##  Chisq= 8.4  on 2 degrees of freedom, p= 0.0151
```

Stratified On Famsize

```
# Stratified on famsize, only trapped, isteamster, and sex are significant

# Significant
(survdiff(donner.surv~strata(sizegroups) + trapped))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sizegroups) + trapped)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5     1.57     7.457      8.75
## trapped=1 81       36    39.43     0.298      8.75
##
##  Chisq= 8.8  on 1 degrees of freedom, p= 0.0031
```

```
(survdiff(donner.surv~strata(sizegroups) + isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sizegroups) + isteamster)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25     30.5     0.982      6.07
## isteamster=1 20       16     10.5     2.840      6.07
##
##  Chisq= 6.1  on 1 degrees of freedom, p= 0.0138
```

```
(survdiff(donner.surv~strata(sizegroups) + sex))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sizegroups) + sex)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     17.6      4.22      9.25
## sex=M 55       32     23.4      3.18      9.25
##
##  Chisq= 9.3  on 1 degrees of freedom, p= 0.00235
```

```
# NOT significant
(survdiff(donner.surv~strata(sizegroups) + age))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sizegroups) + age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     7.96    2.0466    3.3149
## age=Over 26   28       18    18.43    0.0102    0.0245
## age=Under 10 29       11    14.60    0.8894    1.8988
##
##  Chisq= 3.9  on 2 degrees of freedom, p= 0.14
```

```
(survdiff(donner.surv~strata(sizegroups) + camp))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(sizegroups) + camp)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC    23       12    11.99  1.49e-05  4.98e-05
## camp=LC    59       25    27.48  2.24e-01  1.46e+00
## camp=None   7        4     1.53  3.96e+00  4.67e+00
##
##  Chisq= 4.8  on 2 degrees of freedom, p= 0.0913
```

Stratified On Age

```
# Stratified on age, all variables are significant
(survdiff(donner.surv~strata(age) + isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(age) + isteamster)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25    34.47       2.6      20.6
## isteamster=1 20       16     6.53      13.7      20.6
##
##  Chisq= 20.6  on 1 degrees of freedom, p= 5.75e-06
```

```
(survdiff(donner.surv~strata(age) + trapped))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(age) + trapped)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5     1.71     6.313      7.65
## trapped=1 81       36    39.29     0.275      7.65
##
##  Chisq= 7.7  on 1 degrees of freedom, p= 0.00567
```

```
(survdiff(donner.surv~strata(age) + sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(age) + sizegroups)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large   25       10     11.8     0.264     0.448
## sizegroups=Medium  31        8     17.0     4.732    10.095
## sizegroups=Small   33       23     12.3     9.360    16.293
##
##  Chisq= 17.7  on 2 degrees of freedom, p= 0.000143
```

```
(survdiff(donner.surv~strata(age) + sex))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(age) + sex)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     16.1      3.10      7.12
## sex=M 55       32     24.9      1.99      7.12
##
##  Chisq= 7.1  on 1 degrees of freedom, p= 0.00763
```

```
(survdiff(donner.surv~strata(age) + camp))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(age) + camp)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC   23       12    17.05     1.498     3.353
## camp=LC   59       25    22.38     0.306     0.832
## camp=None  7        4     1.56     3.803     4.512
##
##  Chisq= 7.1  on 2 degrees of freedom, p= 0.0289
```

Stratified On Camp

```r
# Stratified on camp, all variables are significant
(survdiff(donner.surv~strata(camp) + isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(camp) + isteamster)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25     35.5       3.1      28.3
## isteamster=1 20       16      5.5      20.0      28.3
##
##  Chisq= 28.3  on 1 degrees of freedom, p= 1.05e-07
```

```r
(survdiff(donner.surv~strata(camp) + trapped))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(camp) + trapped)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5     4.05    0.2221        18
## trapped=1 81       36    36.95    0.0244        18
##
##  Chisq= 18  on 1 degrees of freedom, p= 2.21e-05
```

```r
(survdiff(donner.surv~strata(camp) + sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(camp) + sizegroups)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large   25       10     13.4     0.846      2.04
## sizegroups=Medium 31        8     14.3     2.755     11.08
## sizegroups=Small  33       23     13.4     6.940     13.93
##
##  Chisq= 17  on 2 degrees of freedom, p= 0.000204
```

```r
(survdiff(donner.surv~strata(camp) + sex))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(camp) + sex)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     18.5      4.87      11.3
## sex=M 55       32     22.5      4.00      11.3
##
##  Chisq= 11.3  on 1 degrees of freedom, p= 0.000769
```

```r
(survdiff(donner.surv~strata(camp) + age))
```

```
## Call:
```

```
## survdiff(formula = donner.surv ~ strata(camp) + age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     6.77   4.03489    6.1531
## age=Over 26   28       18    18.40   0.00883    0.0233
## age=Under 10 29       11    15.82   1.47082    3.1764
##
##  Chisq= 6.8  on 2 degrees of freedom, p= 0.0332
```

Stratified on isteamster

```
# Stratified on isteamster, age is not significant

# Significant
(survdiff(donner.surv~strata(isteamster) + trapped))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(isteamster) + trapped)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## trapped=0  8        5     1.61     7.149      8.55
## trapped=1 81       36    39.39     0.292      8.55
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.00345
```

```
(survdiff(donner.surv~strata(isteamster) + sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(isteamster) + sizegroups)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large   25       10     7.61     0.749      1.13
## sizegroups=Medium 31        8    14.16     2.678      6.84
## sizegroups=Small   33       23    19.23     0.739      5.54
##
##  Chisq= 8.8  on 2 degrees of freedom, p= 0.0121
```

```
(survdiff(donner.surv~strata(isteamster) + sex))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(isteamster) + sex)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     14.1     1.855      4.35
## sex=M 55       32     26.9     0.974      4.35
##
##  Chisq= 4.4  on 1 degrees of freedom, p= 0.037
```

```
(survdiff(donner.surv~strata(isteamster) + camp))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(isteamster) + camp)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC   23       12    19.23      2.72      6.30
## camp=LC   59       25    20.19      1.14      2.67
## camp=None  7        4     1.58      3.71      4.45
##
##  Chisq= 9.4  on 2 degrees of freedom, p= 0.00905
```

```
# NOT significant
(survdiff(donner.surv~strata(isteamster) + age))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(isteamster) + age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     7.92     2.102     3.318
## age=Over 26   28       18    22.80     1.012     3.341
## age=Under 10 29       11    10.28     0.051     0.105
##
##  Chisq= 4.2  on 2 degrees of freedom, p= 0.123
```

Stratified on Trapped

```
# Stratified on trapped, all other variables are still significant, except camp

# Significant
(survdiff(donner.surv~strata(trapped) + age))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(trapped) + age)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## age=10 - 26   32       12     6.72     4.158     6.114
## age=Over 26   28       18    20.31     0.262     0.754
## age=Under 10 29       11    13.98     0.635     1.183
##
##  Chisq= 6.1  on 2 degrees of freedom, p= 0.047
```

```
(survdiff(donner.surv~strata(trapped) + sizegroups))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(trapped) + sizegroups)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sizegroups=Large   25       10     11.1    0.0999     0.144
```

```
## sizegroups=Medium 31        8     18.8    6.1780    13.361
## sizegroups=Small  33       23     11.2   12.4921    21.289
##
##  Chisq= 23.7  on 2 degrees of freedom, p= 7.14e-06
```

```
(survdiff(donner.surv~strata(trapped) + isteamster))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(trapped) + isteamster)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## isteamster=0 69       25     35.7      3.21      28.4
## isteamster=1 20       16      5.3     21.61      28.4
##
##  Chisq= 28.4  on 1 degrees of freedom, p= 9.97e-08
```

```
(survdiff(donner.surv~strata(trapped) + sex))
```

```
## Call:
## survdiff(formula = donner.surv ~ strata(trapped) + sex)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=F 34        9     17.1      3.85      8.57
## sex=M 55       32     23.9      2.76      8.57
##
##  Chisq= 8.6  on 1 degrees of freedom, p= 0.00342
```

```
# NOT significant
(survdiff(donner.surv~strata(trapped) + camp))
```

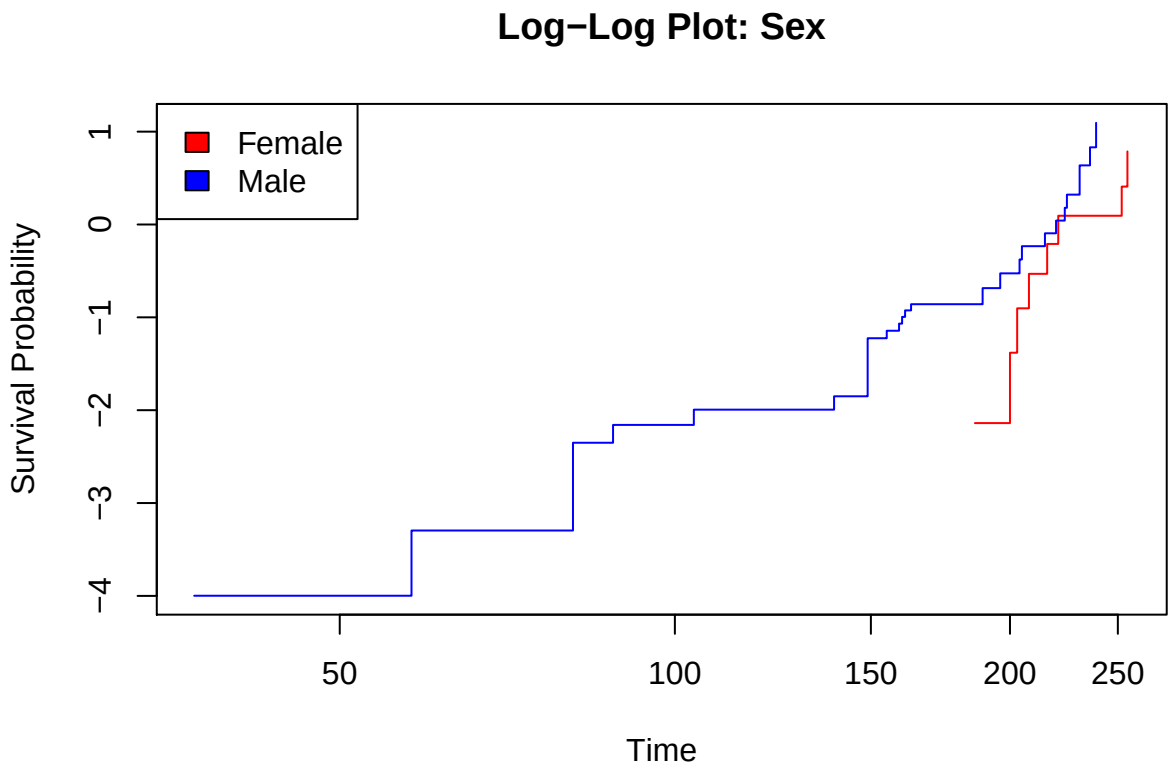
```
## Call:
## survdiff(formula = donner.surv ~ strata(trapped) + camp)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## camp=AC   23       12    17.04   1.49061    3.3776
## camp=LC   59       25    19.84   1.33925    3.2260
## camp=None  7        4     4.12   0.00324    0.0186
##
##  Chisq= 3.4  on 2 degrees of freedom, p= 0.183
```

From our stratified log-rank tests, we see that age is not significant when stratified on isteamster and famsize, and is barely significant at the 0.05 level when stratified on trapped. We aso see that camp is not significant when stratified on famsize and trapped. From this information, we can see that age may not be something we wish to include in our model, but as for camp being insignificant when stratified on trapped, we can assume that this is due to the the fact that subjects who weren't trapped generally belonged to no camp.

Log-Log Plots

Sex

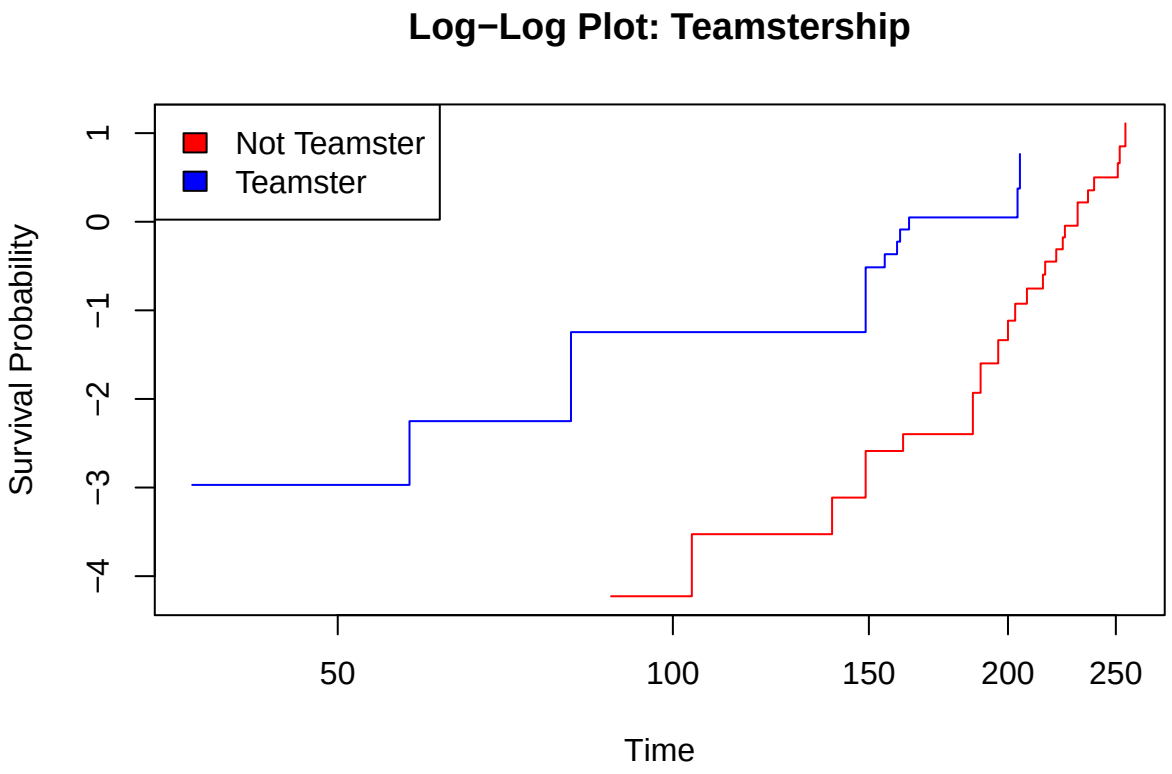
```
plot(survfit(donner.surv ~ sex), col=c("red", "blue"), xlab="Time", ylab="Survival Probability",
     main="Log-Log Plot: Sex", fun="cloglog")
legend("topleft",c("Female", "Male"), fill = c("red", "blue"))
```



The curves intersect at 225 days, suggesting a violation of the PH assumptions.

Isteamster

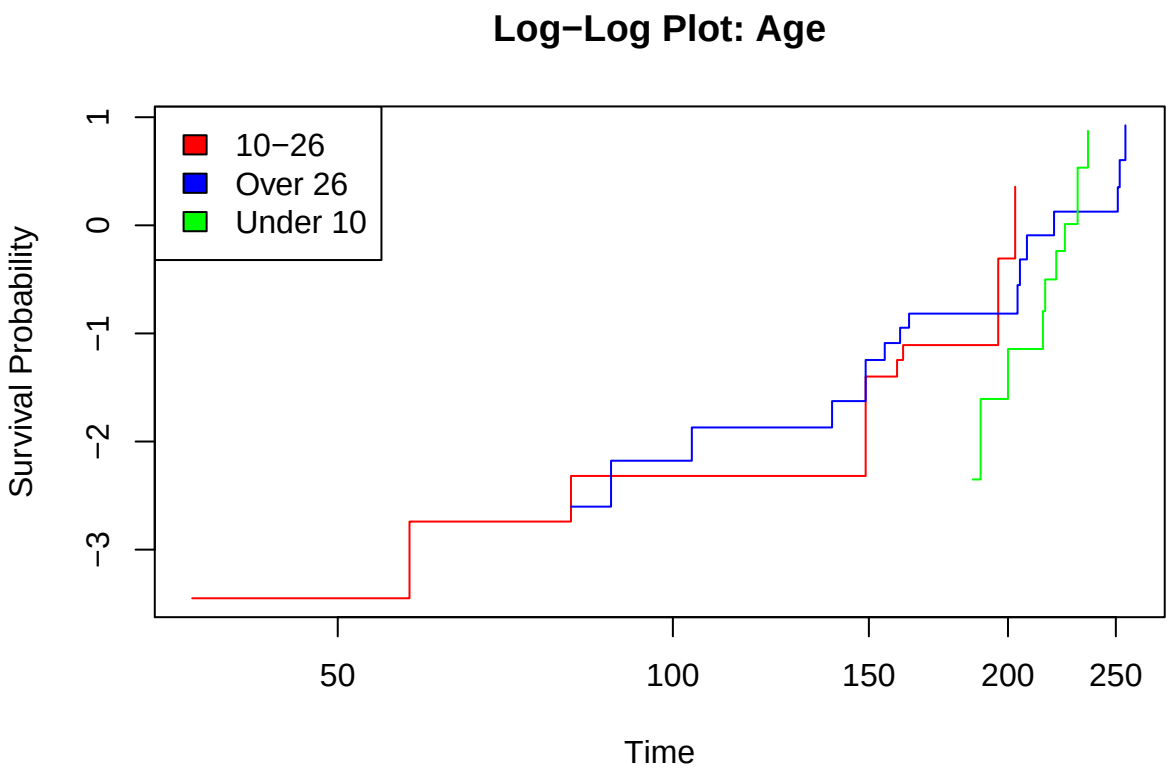
```
isteamster <- as.factor(isteamster)
plot(survfit(donner.surv ~ isteamster), col=c("red", "blue"), xlab="Time", ylab="Survival Probability",
     main="Log-Log Plot: Teamstership", fun="cloglog")
legend("topleft",c("Not Teamster", "Teamster"), fill = c("red", "blue"))
```



The curves do not violate the PH assumptions.

Age

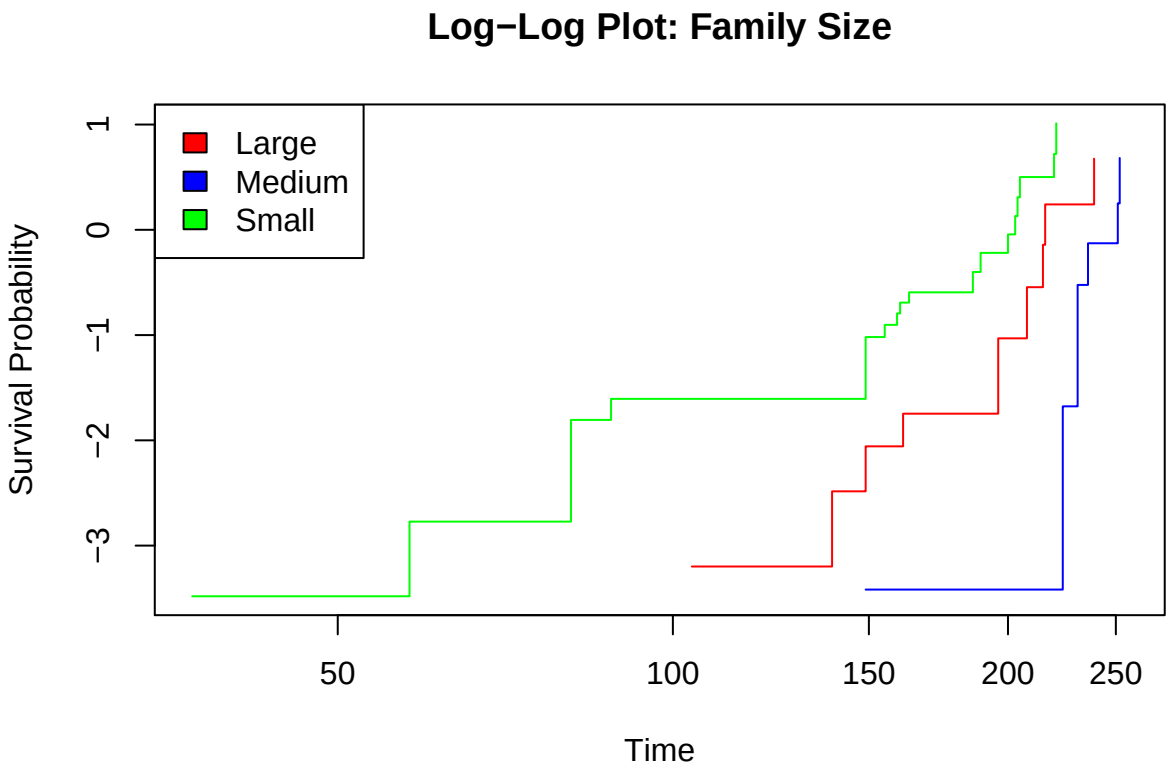
```
plot(survfit(donner.surv ~ age), col=c("red", "blue", "green"), xlab="Time", ylab="Survival Probability"
     main="Log-Log Plot: Age", fun="cloglog")
legend("topleft",c("10-26", "Over 26", "Under 10"), fill = c("red", "blue", "green"))
```



All 3 the curves intersect with at least one of the others, and thus violate the PH assumptions.

Famsize

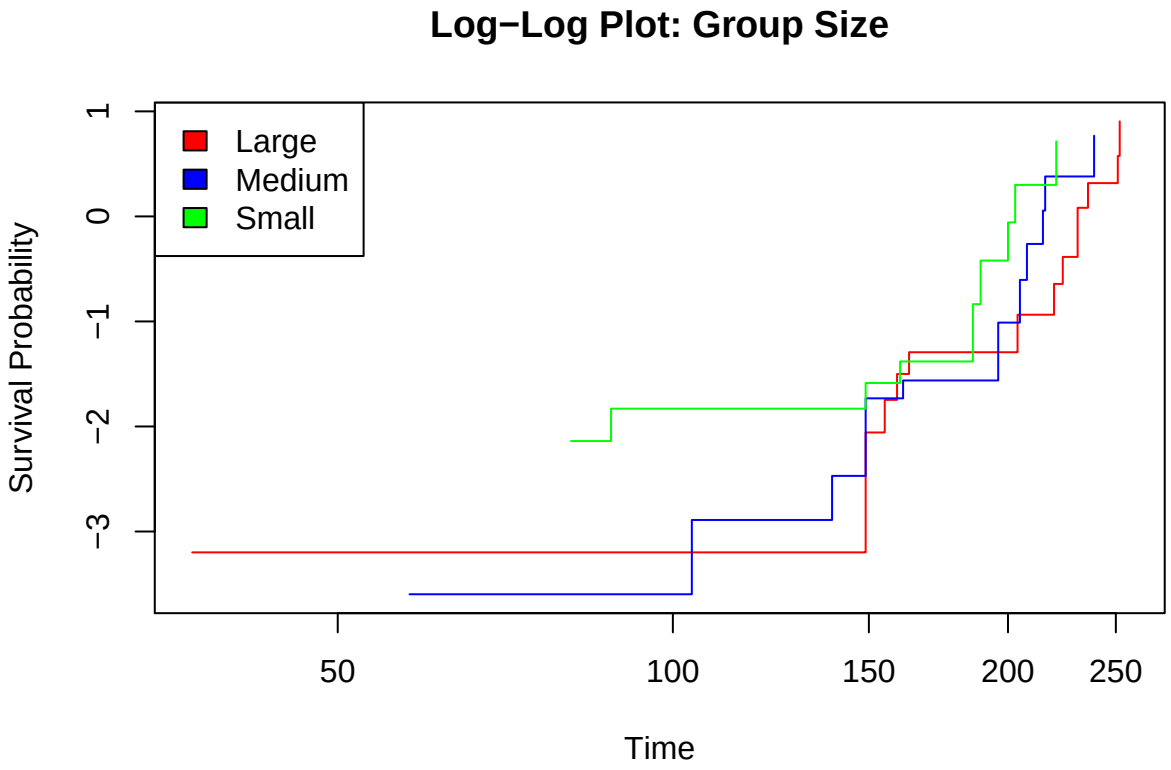
```
sizegroups <- as.factor(sizegroups)
plot(survfit(donner.surv ~ sizegroups), col=c("red", "blue", "green"), xlab="Time", ylab="Survival Proba
     main="Log-Log Plot: Family Size", fun="cloglog")
legend("topleft",c("Large", "Medium", "Small"), fill = c("red", "blue", "green"))
```



The curves do not violate the PH assumptions.

Groupsize

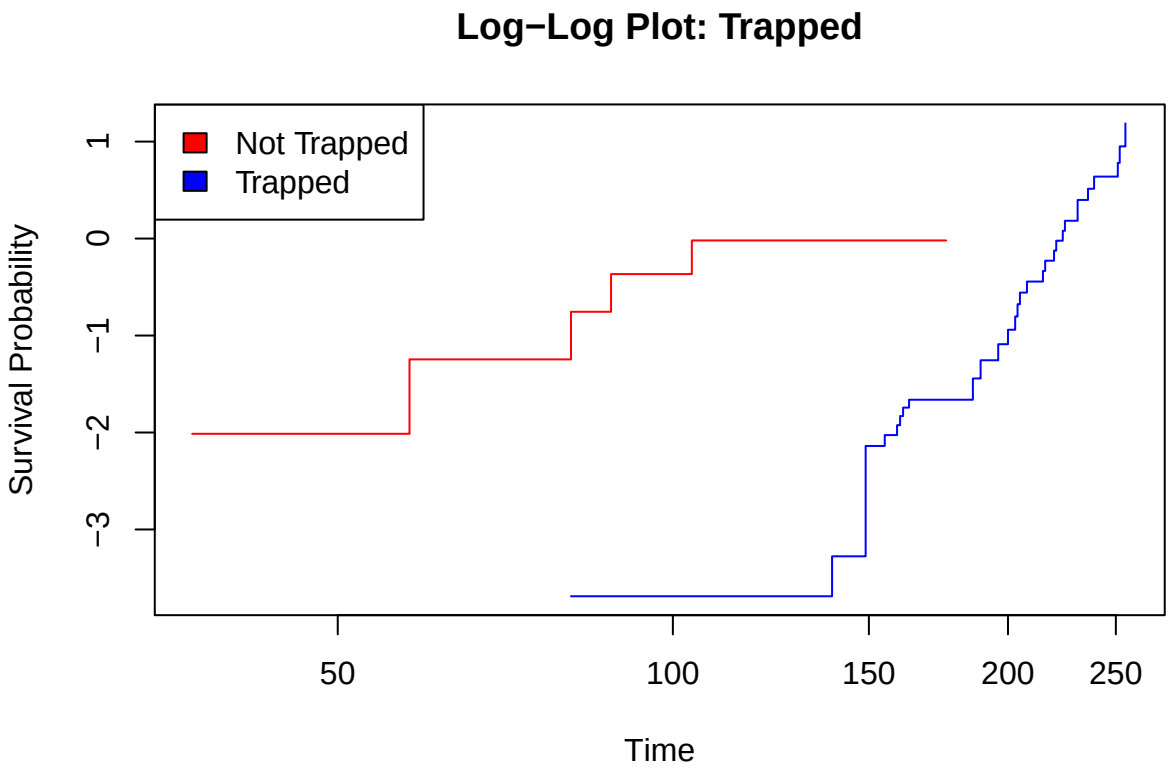
```
groupsize <- as.factor(groupsize)
plot(survfit(donner.surv ~ groupsize), col=c("red", "blue", "green"), xlab="Time", ylab="Survival Probal
     main="Log-Log Plot: Group Size", fun="cloglog")
legend("topleft",c("Large", "Medium", "Small"), fill = c("red", "blue", "green"))
```



All 3 the curves intersect with each other and therefore violate the PH assumptions.

Trapped (In Mountains)

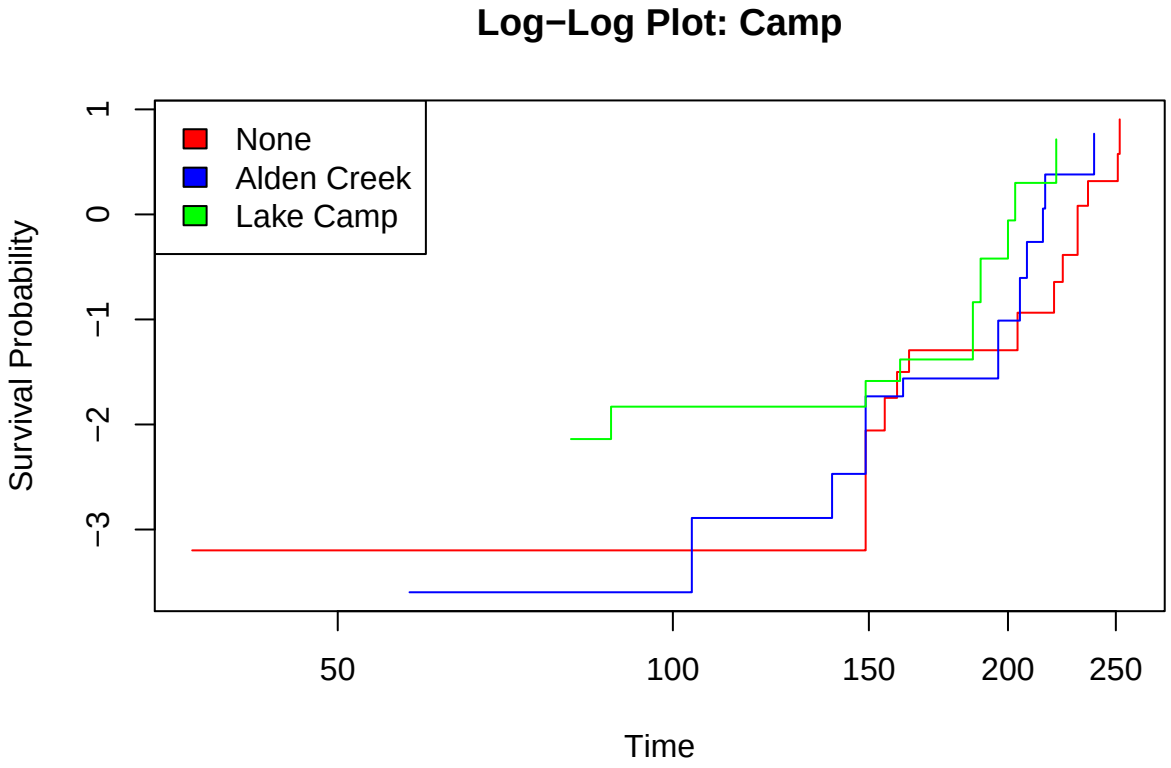
```
trapped <- as.factor(trapped)
plot(survfit(donner.surv ~ trapped), col=c("red", "blue"), xlab="Time", ylab="Survival Probability",
     main="Log-Log Plot: Trapped", fun="cloglog")
legend("topleft",c("Not Trapped", "Trapped"), fill = c("red", "blue"))
```



The curves do not violate the proportional hazards assumptions.

Camp

```
plot(survfit(donner.surv ~ groupsize), col=c("red", "blue", "green"), xlab="Time", ylab="Survival Probal
     main="Log-Log Plot: Camp", fun="cloglog")
legend("topleft",c("None", "Alden Creek", "Lake Camp"), fill = c("red", "blue", "green"))
```



All 3 curves violate the proportional hazards assumptions.

From our log-log plots we see that the only significant covariates that don't violate the proportional hazards assumptions are: isteamster, famsize, and trapped.

Random Survival Forest

```
data$deathdate<-NULL
data$joindate<-NULL
install.packages("Hmisc",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/xc/ny8v2yrx2dz8lgrmyhx0n5w00000gn/T//RtmpKmin6Q/downloaded_packages
```

```
library(ggplot2)
library(randomForestSRC)
```

```
## Warning: package 'randomForestSRC' was built under R version 3.2.5
```

```
##
##  randomForestSRC 2.4.1
##
##  Type rfsrc.news() to see new features, changes, and bug fixes.
##
```

```
library(ggRandomForests)
```

```
## Warning: package 'ggRandomForests' was built under R version 3.2.5
```

```
##
## Attaching package: 'ggRandomForests'
```

```
## The following object is masked from 'package:randomForestSRC':
##
##     partial.rfsrc
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.2.5
```

```
## Loading required package: lattice
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:randomForestSRC':
##
##     impute
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
library(risksetROC)
```

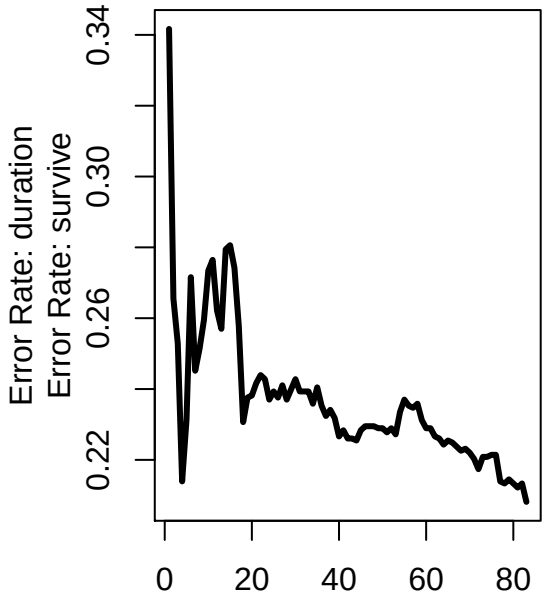
```
## Loading required package: MASS
```

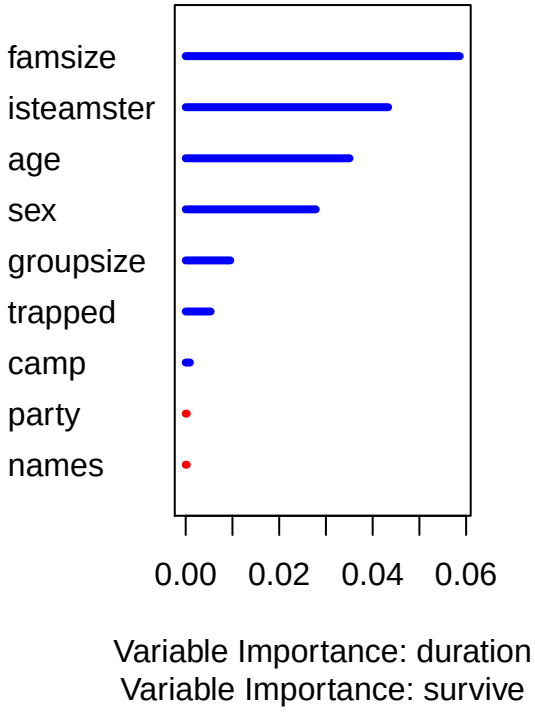
We want to use a non-parametric model to predict the survival status of a subject during the median survival time. We will find the prediction error rate of the model and find the most important variables in the model.

```
rf <- rfsrc(Surv(duration,survive)~.,data=data, na.action = "na.impute",nimpute=1,ntree=83,importance=T
rf
```

```
##                          Sample size: 89
##                     Number of deaths: 41
##                     Was data imputed: yes
##                      Number of trees: 83
##           Minimum terminal node size: 3
##        Average no. of terminal nodes: 25.56627
## No. of variables tried at each split: 3
##               Total no. of variables: 9
##                                Analysis: RSF
##                                  Family: surv
##                      Splitting rule: logrank
##                          Error rate: 20.81%
```

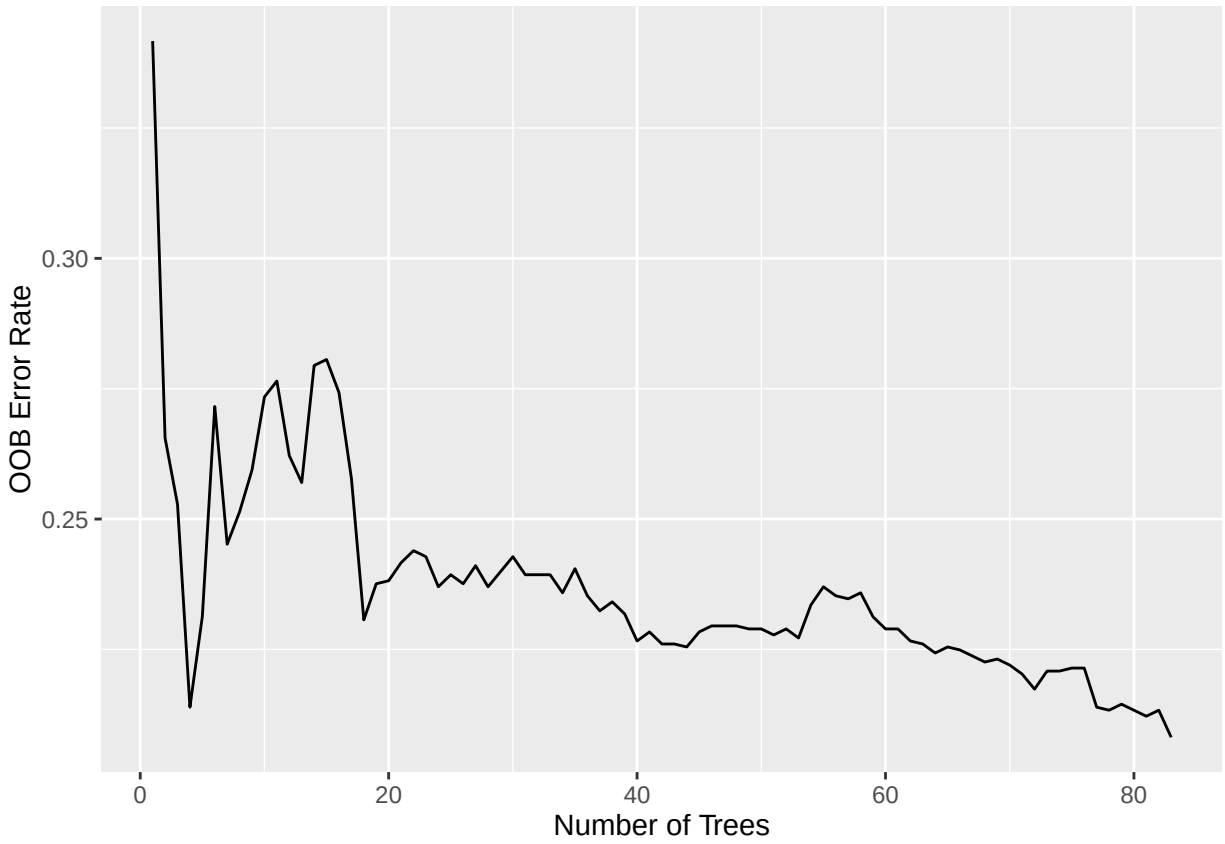
```
plot(rf)
```






```
##
##               Importance   Relative Imp
## famsize           0.0586         1.0000
## isteamster        0.0433         0.7387
## age               0.0350         0.5976
## sex               0.0278         0.4745
## groupsize         0.0095         0.1625
## trapped           0.0053         0.0908
## camp              0.0009         0.0151
## party             0.0000         0.0000
## names             0.0000         0.0000
```

```
plot(gg_error(rf))
```

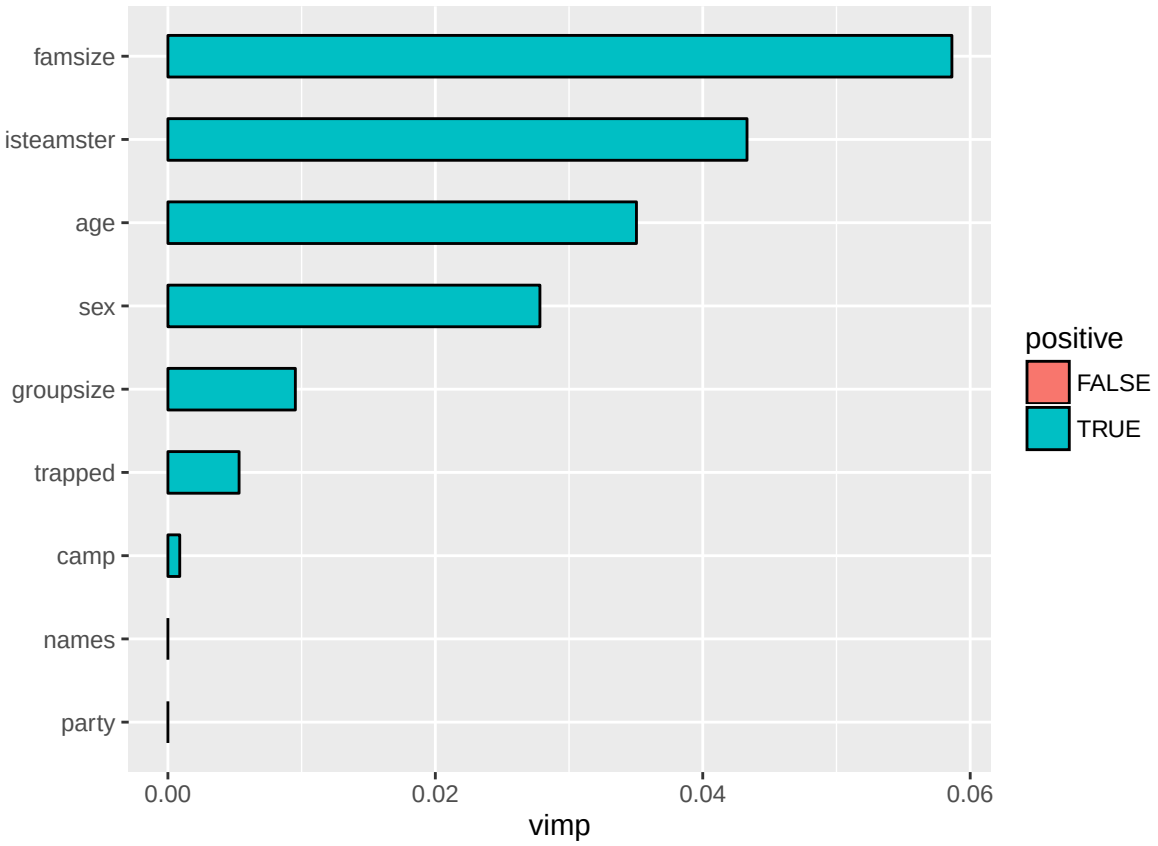


Prediction error rate with the increased number of decision trees used for our model.

```
importance.rf <- sort(rf$importance,decreasing = T)
importance.rf
```

```
##      famsize   isteamster          age          sex    groupsize
## 0.0586322364 0.0433096492 0.0350413029 0.0278226895 0.0095303373
##      trapped         camp        names        party
## 0.0053220896 0.0008855735 0.0000000000 0.0000000000
```

```
plot(gg_vimp(rf))
```



Famsize is the most important covariate in our random survival forest model.

```
rf.predict <- predict(rf,data,type="prob")
rf.predict
```

```
##   Sample size of test (predict) data: 48
##                 Number of grow trees: 83
##   Average no. of grow terminal nodes: 25.56627
##           Total no. of grow variables: 9
##                               Analysis: RSF
##                                 Family: surv
```

We want to compute the C Index (which is the area under the receiver operating characteristic) for our classification for the censored response variables (status).

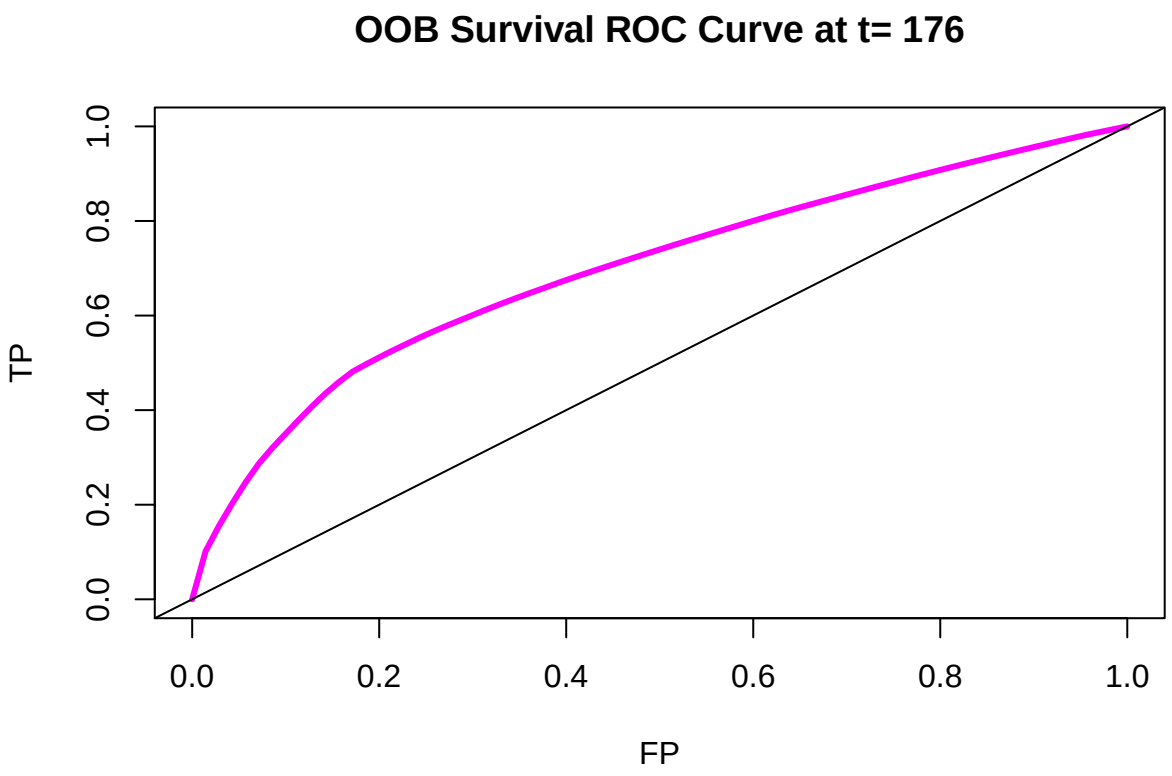
```
rcorr.cens(rf$predicted.oob,Surv(duration,survive))["C Index"]
```

```
##   C Index
## 0.2049037
```

```
error.rate.rf=rf$err.rate[rf$ntree]
error.rate.rf
```

```
## [1] 0.2081409
```

```
#ROC (Receiver operating characteristic) for true positive (TP) to false positive (FP)
donner.ROC = risksetROC(Stime = data$duration, status = data$survive,
                        marker = rf$predicted.oob,
                        predict.time = median(data$duration), method = "Cox",
                        main = paste("OOB Survival ROC Curve at t=",
                                     median(data$duration)),lwd = 3,
                        col = "magenta")
```



The area under the curve is 0.6889838 which is the expected uniform arbitrary positive, ranked before a uniform arbitrary negative. 0.6889838 is an okay AUC value so this classification model is moderately effective at separating those who died before the median survival time against those who died after the median survival time.

```
donner.ROC$AUC
```

```
## [1] 0.6951223
```

Conclusion

To effectively analyze the demise of the Donner Party, we tested the covariates age, sex, famsize, groupsize, camp, trapped, and isteamster to see if these variables had an impact on the survival of an observed member. Using a combination of Cox regression and backwards AIC selection we determined our best model was difference-controlled-adj, a survival model in relation to sex, controlling for the covariates trapped, groupsize, and isteamster. This model produced the lowest AIC value of 228.5112. In the strongest model, there was a key relationship between the covariates, sex and trapped. Although the trapped covariate was insignficant, the relationship between sex and trapped resulted in a smaller AIC, when trapped was added. This led us to

accept the model including sex, groupsize, isteamster, in addition to trapped as the strongest model. The relationship between the two covariates, sex and trapped, could be explained by familial relations influencing the survival rate of women trapped in the mountains. In the groups trapped in the mountains during the winter, considerable effort was taken to avoid separating families if unnecessary. Because of this, women were protected and single men (bachelors) and people with no familial connections were at higher hazard risk of being eaten. These trends are supported in the Kaplan-Meier plot for sex, which illustrates women having much better survival rates. According to our Log-Rank tests, we concluded that the variables sex, trapped, famsize, isteamster and camp all had an impact on the survival probability at a 0.05 significance. Next, we created log-log plots in order to check if proportional hazard assumptions were satisfied. The log-log plots for sex, age, groupsize, and camp contained curves that were not proportionally consistent over time and therefore did not satisfy the conditions of a proportional Cox model. Given that these models did not satisfy proportional hazard assumptions, we stratified each variable to recheck significance. Stratified on sex, age and camp, all variables remained significant. Because many of our covariates violated PH assumptions, we decided to implement a non-parametric approach by creating a random survival forest to test for significant variables. The most accurate model in predicting the survival status of a subject during the median survival time contained 83 trees. In the random survival forest model, famsize was the most important covariate in our prediction. In order to test the accuracy of our forest, an OOB error was calculated to measure the balance of false positives and true positives. Our AUC (area under the curve) was 0.6889838, which displays moderate effectiveness of our forest. In conclusion, many variables contributed to the survival probability of the Donner Party. Famsize may have reigned most significant because larger families could share rationings, take care of each other and provide emotional support. Females of the Donner Party may have been more likely to survive because they contain more body fat and a lower metabolic rate, making them more likely to survive disaster.

Works Cited

AIC Computation Reference: "Compute AIC in Survival Analysis (survfit/coxph)." StackOverflow, 2013. Web. 22 Nov. 2016. http://stackoverflow.com/questions/19679183/compute-aic-in-survival-analysis-survfit-coxph

Background Information Reference: "Donner Party." Wikipedia. Web. 22 Nov. 2016. https://en.wikipedia.org/wiki/Donner_Party

Text Reference: Ishwaran, Hemant, and Udaya B. Kogalur. "Random Survival Forests for R." 7.2 (2007): n. pag. Web. 22 Nov. 2016. http://www.ccs.miami.edu/~hishwaran/papers/randomSurvivalForests.pdf.

Dataset Reference: "Members and Survival of the Donner Party." Journal of Statistical Education Data Archive. University of Florida, n.d. Web. 22 Nov. 2016. http://www.stat.ufl.edu/~winner/data/donner.dat.

Richardson, Mary, Terry Wright, and Eric Daly. "Donner Party." Donner Party. WikiTree, 2008. Web. 22 Nov. 2016. https://www.wikitree.com/wiki/Space:Donner_Party.

Background Information Reference: "The Donner Party." WikiTree. Web. 22 Nov. 2016. https://www.wikitree.com/wiki/Space:Donner_Party

AIC/P-value Analysis Reference: "When I use AIC (akaike information criterion) to find the model of the best fit, do I need to consider p-values?" ResearchGate, 2013. Web. 22 Nov. 2016. https://www.researchgate.net/post/When_I_use_AIC_akaike_information_criterion_to_find_the_model_of_the_best_fit_do_I_need_to_consider_p-values

R Packages References: randomSurvivalForest package (Ishwaran and Kogalur 2007)