R for US Forest Service Projects Survival Analysis

Katie Murenbeeld

5/31/2021

Introduction

The following code was used to perform the various survival analyses for this research. Please see the Python folder in this repository for a Jupyter Notebook with instructions for downloading the USFS activities data. See also the R folder in this repositroty for the R script (R__USFS__Survival__Data__Processing.R) used for further data processing of the USFS activities datasets. This R script also has the code for combining the activities datasets (FS__ACT) with the UNM-PALS data. The UMN-PALS data (Fleischman et al., 2020) is in the Data folder of this repository and can be downloaded from here: https://conservancy.umn.edu/handle/11299/211669. References can be found in the References folder of this repository.

Load libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(forcats)
library(survminer)
```

```
## Loading required package: ggpubr
```

```
library(survival)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
```

```
##
##     smiths
```

```r
library(ggplot2)
library(ggpubr)

# Set your working directory to wherever you placed your processed FS-PALS dataaset.

setwd("/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/Oct_presentation/")

# Set up to see more rows of output if desired

options(max.print = 10000)
```

Load the data

Data from post-processing. See **Download_Convert_Subset_Forest_Service_Data-for-Survival-Analysis-v02.ipynb** in the Python folder of this repository for instructions on downloading and subsetting the USFS activities data. See **R_USFS_Survival_Data_Processing.R** in the R folder of this repository for instrcutions on further data processing and combining with the PALS data.

```r
## Load the data for survival analysis

df_fin <- read.csv("df_c20201020_v02.csv")
df_fin <- df_fin %>%
  select(-X)

## Replace "Inf" in the project delay column with NA
df_fin$proj.delay[is.infinite(df_fin$proj.delay)] <- NA
## Bring in binomial and categorical data as a factor
df_fin$REGION <- as.factor(df_fin$REGION)
df_fin$LITIGATED <- as.factor(df_fin$LITIGATED)
df_fin$APPEALED <- as.factor(df_fin$APPEALED)
df_fin$size <- as.factor(df_fin$size)

## Relevel the size data in order from small to extra-large
## In this dataframe Regions are in numerical order (1->6)
df_fin <- df_fin %>%
  mutate(size = fct_relevel(size,
            "small", "medium", "large", "x-large"))

## Create different data frames with Regions releveled.
## This is required to test the impact of using different
## Regions as a baseline or reference in the Cox proportional
## analysis.
df_fin2 <- df_fin %>%
  mutate(REGION = fct_relevel(REGION,
            "2", "1", "3", "4", "5", "6"))

df_fin3 <- df_fin %>%
  mutate(REGION = fct_relevel(REGION,
            "3", "1", "2", "4", "5", "6"))

df_fin4 <- df_fin %>%
  mutate(REGION = fct_relevel(REGION,
```

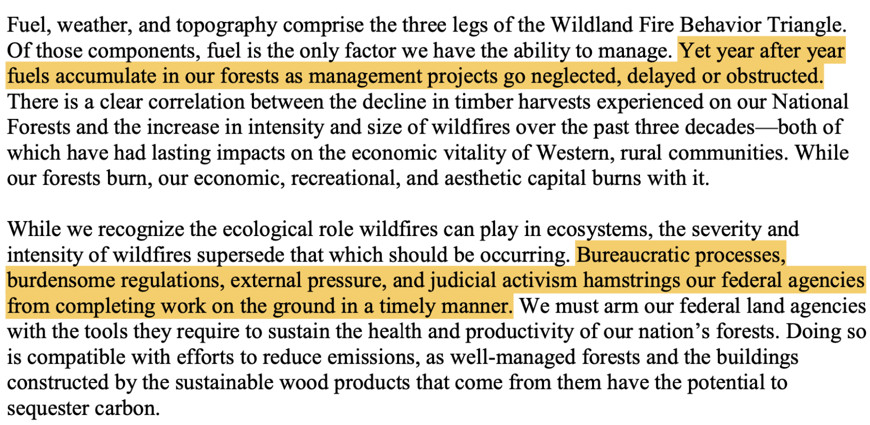
```
           "4", "1", "2", "3", "5", "6"))

df_fin5 <- df_fin %>%
  mutate(REGION = fct_relevel(REGION,
           "5", "1", "2", "3", "4", "6"))

df_fin6 <- df_fin %>%
  mutate(REGION = fct_relevel(REGION,
           "6", "1", "2", "3", "4", "5"))
```

Check for data correlations

```
df_fin_cor <- transform (df_fin, NEPA_TYPE = factor(NEPA_TYPE,
                                                    levels = c("CE", "EA", "EIS"),
                                                    labels = c(1, 2, 3)))
df_fin_cor <- transform(df_fin_cor, size = factor(size,
                                                  levels = c("small", "medium", "large", "x-large"),
                                                  label = c(1, 2, 3, 4)))

covariates <- data.frame(as.numeric(df_fin_cor$proj.delay), as.numeric(df_fin_cor$NEPA_TYPE), as.numeri
as.numeric(df_fin_cor$size))

col <- cor(covariates)

##write.csv(col, "~/Desktop/Survival_Analysis/WRITING/TABLES/correlation_20210406.csv")
```

Why care?

When it comes to USFS project planning and inititaion there is some anecdotal evidence for NEPA causing delays, but there are very few quantitative studies of delays.



Fuel, weather, and topography comprise the three legs of the Wildland Fire Behavior Triangle. Of those components, fuel is the only factor we have the ability to manage. Yet year after year fuels accumulate in our forests as management projects go neglected, delayed or obstructed. There is a clear correlation between the decline in timber harvests experienced on our National Forests and the increase in intensity and size of wildfires over the past three decades—both of which have had lasting impacts on the economic vitality of Western, rural communities. While our forests burn, our economic, recreational, and aesthetic capital burns with it.

While we recognize the ecological role wildfires can play in ecosystems, the severity and intensity of wildfires supersede that which should be occurring. Bureaucratic processes, burdensome regulations, external pressure, and judicial activism hamstrings our federal agencies from completing work on the ground in a timely manner. We must arm our federal land agencies with the tools they require to sustain the health and productivity of our nation's forests. Doing so is compatible with efforts to reduce emissions, as well-managed forests and the buildings constructed by the sustainable wood products that come from them have the potential to sequester carbon.

Figure 1: Recent letter to Congress from Western Congressional Causcus

Survival Analysis

"Time to Event"

- Time until an individual dies (or is cured!).
- Time until a kitten or puppy is adopted.
- Any time an outcome is a **duration**
- Requires a start date, end date, duration (time between), if the event occurred, and the time of observation.
- Survival package in R (Therneau, 2021)

Different approaches

- **Kaplan-Meier** –> non-parametric, predictive (Kaplan & Meier, 1958)
- **Cox Proportional Hazards** –> semi-parametric, use to estimate effect of covariate, not predictive (Cox, 1972)

Survival Analysis - Kaplan Meier

All data - Kaplan Meier (KM) Estimate

Using all the data pooled.

```
# Use the survfit function from the survival package to
# calculate the KM estimate

fit_delay_all <- survfit(Surv(proj.delay, INITIATED) ~ 1, data = df_fin)
fit_delay_all
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ 1, data = df_fin)
##
##        n  events  median 0.95LCL 0.95UCL
##     3557    3289     213     194     228
```

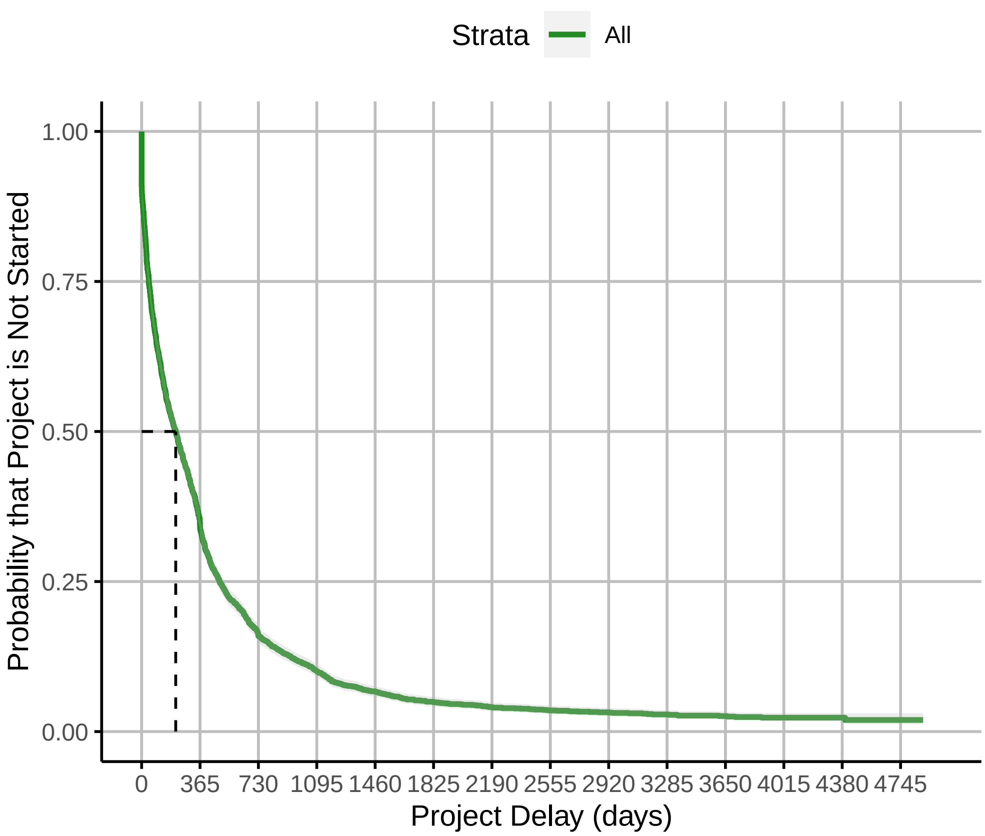
```
#summary(fit_delay_all)
```

Survival Analysis - All - Kaplan Meier Curves

All data

```
## ggsurvvplot will create survival curves from the KM model created above.

km_fit_all <- ggsurvplot(fit_delay_all,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
           break.time.by = 365,
           risk.table.y.text=FALSE,
```

```r
          censor = FALSE,
          ylab = "Probability that Project is Not Started",
          xlab = "Project Delay (days)",
          palette = c("forestgreen"),
          surv.plot.height = 1,
          ggtheme = theme(aspect.ratio = 0.75,
                          axis.line = element_line(colour = "black"),
                          panel.grid.major = element_line(colour = "grey"),
                          panel.border = element_blank(),
                          panel.background = element_blank()),
          tables.theme =  theme(aspect.ratio = 0.06)
          )
print(km_fit_all)
```



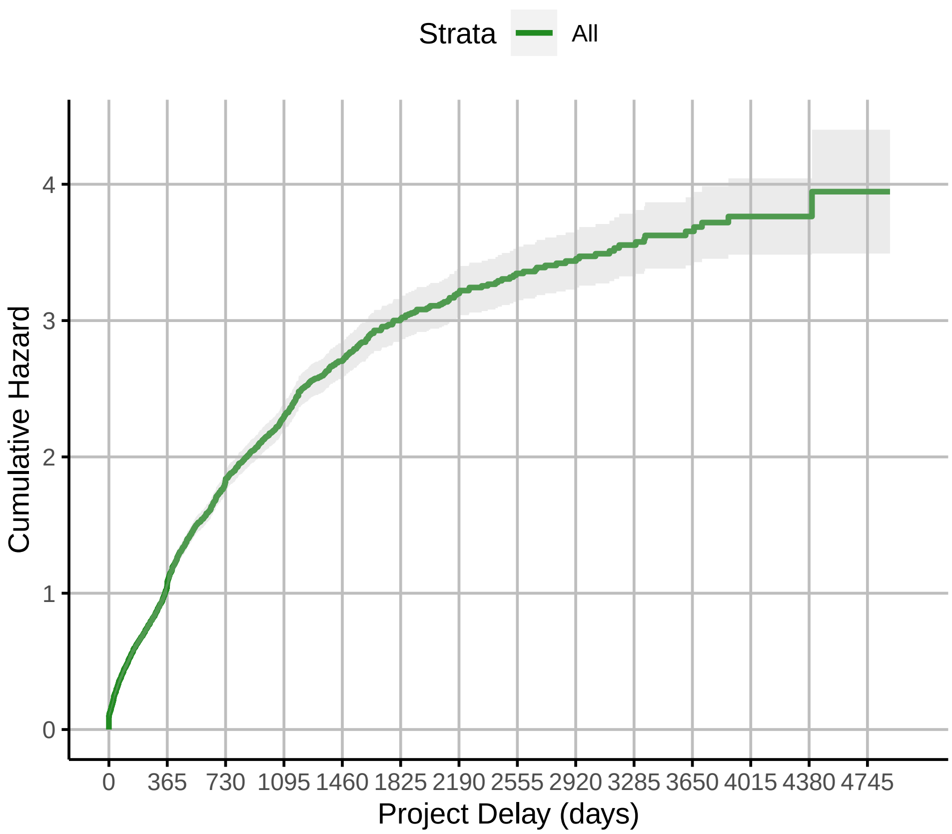

```r
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_fit_all.pdf", print(km_fit
```

Survival Analysis - All - Kaplan Meier - Cumulative Hazard

All data

```r
# ggsurvplot will create the cumulative hazard plots based on the KM model created above.

km_haz_all <- ggsurvplot(fit_delay_all,
          conf.int = TRUE,
          risk.table = FALSE,
          risk.table.col = "strata",
          break.time.by = 365,
          risk.table.y.text=FALSE,
```

```
           censor = FALSE,
           ylab = "Cumulative Hazard",
           xlab = "Project Delay (days)",
           palette = c("forestgreen"),
           fun = "cumhaz",
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                          axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_haz_all)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_all.pdf", print(km_haz_
```

Survival Analysis - Appealed? - KM Estimate

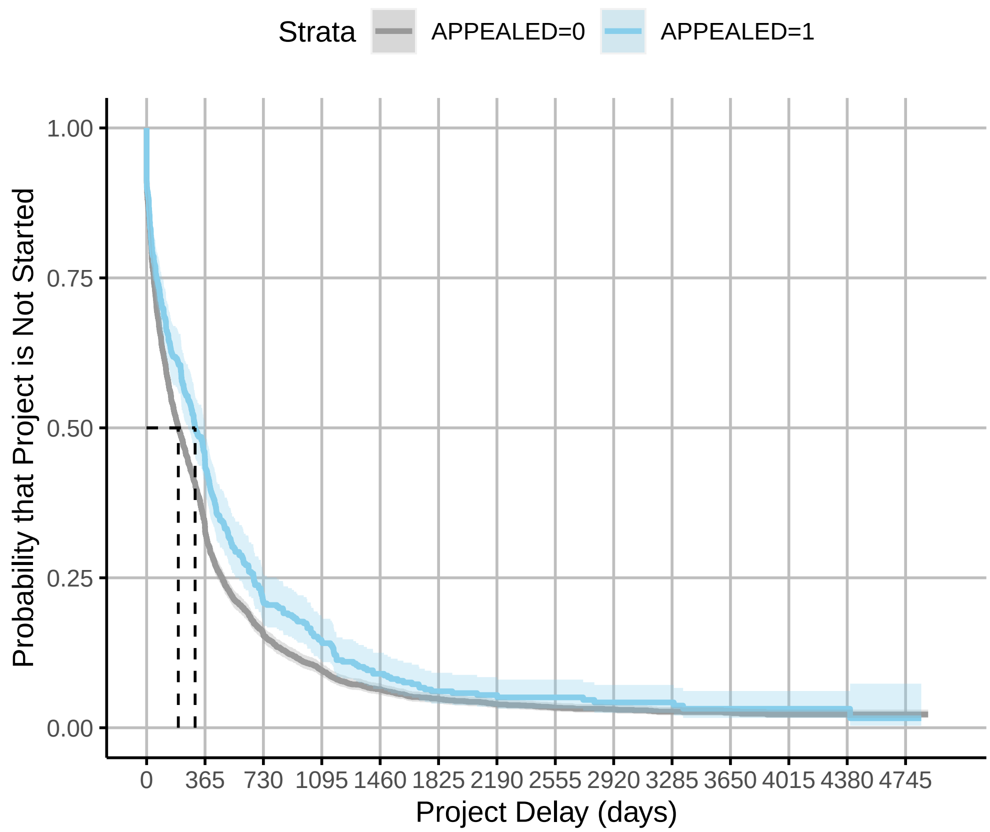
Data grouped by appealed and non-appealed projects.

```
# Use the survfit function from the survival package to
# calculate the KM estimate for data grouped by appealed or non-appealed.
# This code chunk and the next two act as a template for the K-M estimation.

fit_app <- survfit(Surv(proj.delay, INITIATED) ~ APPEALED, data = df_fin)
fit_app_table <- summary(fit_app)
fit_app
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ APPEALED, data = df_fin)
##
##                n events median 0.95LCL 0.95UCL
## APPEALED=0 3184   2937    198     182     220
## APPEALED=1  373    352    303     258     365
```

Survival Analysis - Appealed? - KM Curve

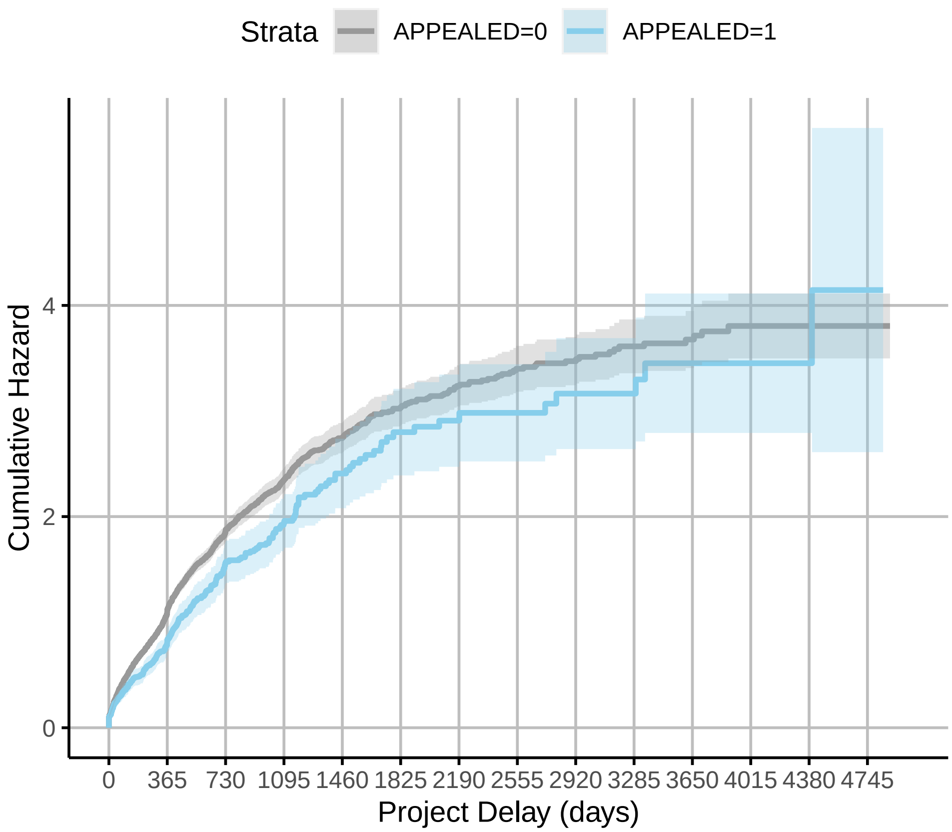
```r
km_curv_app <- ggsurvplot(fit_app,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Probability that Project is Not Started",
           xlab = "Project Delay (days)",
           palette = c("#999999", "skyblue"),
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_curv_app)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_curv_app.pdf", print(km_cu
```

Survival Analysis - Appealed? - KM Cumulative Hazard

```
km_haz_app <- ggsurvplot(fit_app,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           censor = FALSE,
           ylab = "Cumulative Hazard",
           xlab = "Project Delay (days)",
           palette = c("#999999", "skyblue"),
           fun = "cumhaz",
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_haz_app)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_app.pdf", print(km_haz
```

Survival Analysis - Litigated? - KM Estimate

Data grouped by litigated and non-litigated projects.

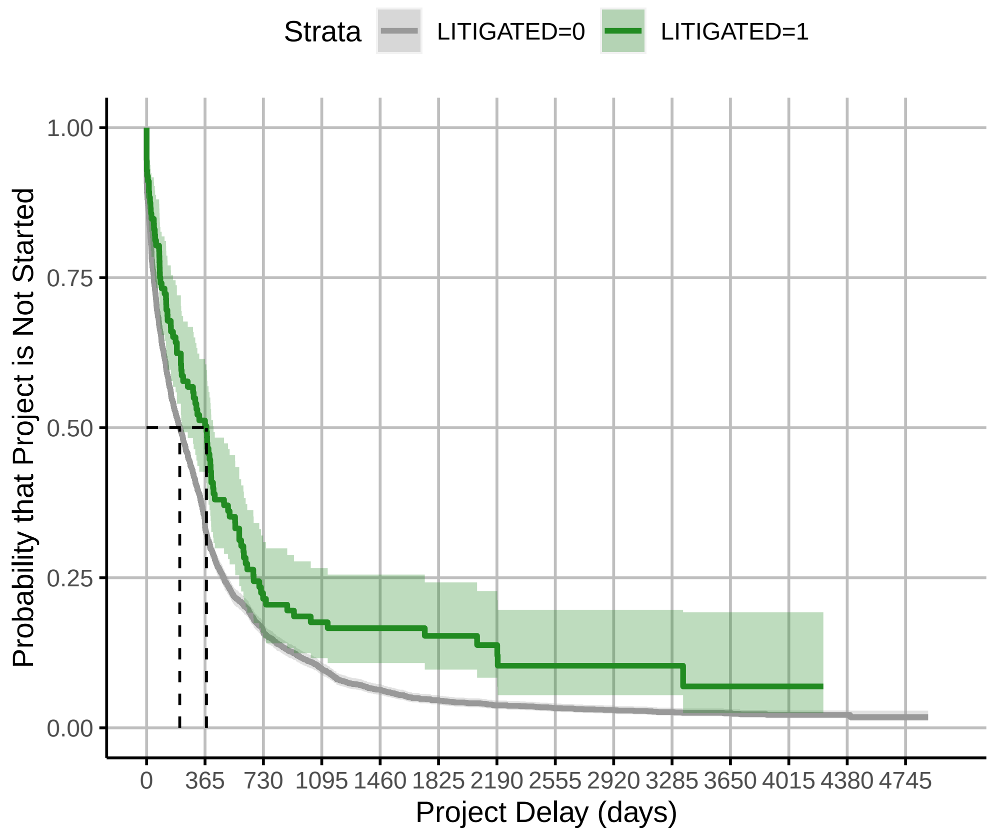
```
fit_lit <- survfit(Surv(proj.delay, INITIATED) ~ LITIGATED, data = df_fin)
fit_lit_table <- summary(fit_lit)
fit_lit
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ LITIGATED, data = df_fin)
##
##                n events median 0.95LCL 0.95UCL
## LITIGATED=0 3445   3194    207     190     226
## LITIGATED=1  112     95    374     228     418
```

Survival Analysis - Litigated? - KM Curve

```
km_curv_lit <- ggsurvplot(fit_lit,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Probability that Project is Not Started",
```

```
            xlab = "Project Delay (days)",
            palette = c("#999999", "forestgreen"),
            surv.plot.height = 1,
            ggtheme = theme(aspect.ratio = 0.75,
                            axis.line = element_line(colour = "black"),
                            panel.grid.major = element_line(colour = "grey"),
                            panel.border = element_blank(),
                            panel.background = element_blank()),
            tables.theme =  theme(aspect.ratio = 0.06)
            )
print(km_curv_lit)
```

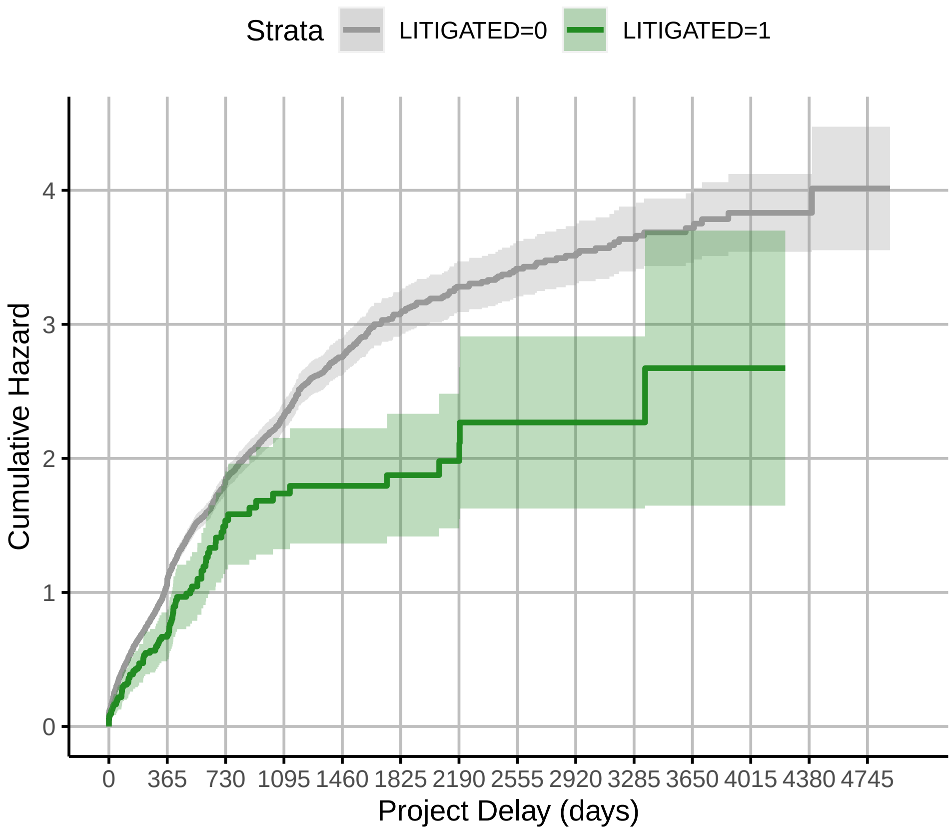



```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_curv_lit.pdf", print(km_cu
```

Survival Analysis - Litigated? - KM Cumulative Hazard

```
km_haz_lit <- ggsurvplot(fit_lit,
            conf.int = TRUE,
            risk.table = FALSE,
            risk.table.col = "strata",
            break.time.by = 365,
            risk.table.y.text=FALSE,
            censor = FALSE,
            ylab = "Cumulative Hazard",
            xlab = "Project Delay (days)",
            palette = c("#999999", "forestgreen"),
```

```
            fun = "cumhaz",
            surv.plot.height = 1,
            ggtheme = theme(aspect.ratio = 0.75,
                            axis.line = element_line(colour = "black"),
                            panel.grid.major = element_line(colour = "grey"),
                            panel.border = element_blank(),
                            panel.background = element_blank()),
            tables.theme =  theme(aspect.ratio = 0.06)
            )
print(km_haz_lit)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_lit.pdf", print(km_haz_
```

Survival Analysis - NEPA Type - KM Estimate

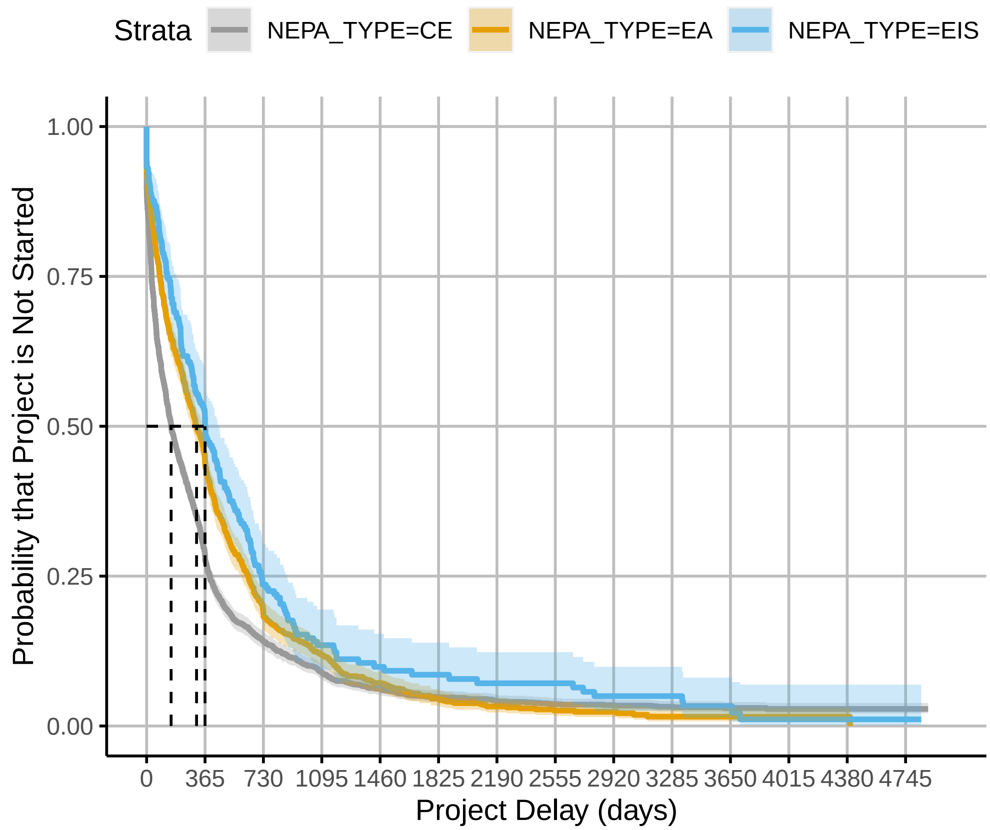
Data grouped by NEPA type for projects. *EIS* = Environmental Impact Statement, *EA* = Environmental Assessment, *CE* = Categorical Exclusion

```
fit_delay <- survfit(Surv(proj.delay, INITIATED) ~ NEPA_TYPE, data = df_fin)
fit_nepa_table <- summary(fit_delay)
fit_delay
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ NEPA_TYPE, data = df_fin)
##
##                   n events median 0.95LCL 0.95UCL
## NEPA_TYPE=CE   2266   2110    153     139     171
## NEPA_TYPE=EA   1073    988    312     285     350
## NEPA_TYPE=EIS   218    191    365     303     455
```

Survival Analysis - NEPA Type - KM Curve

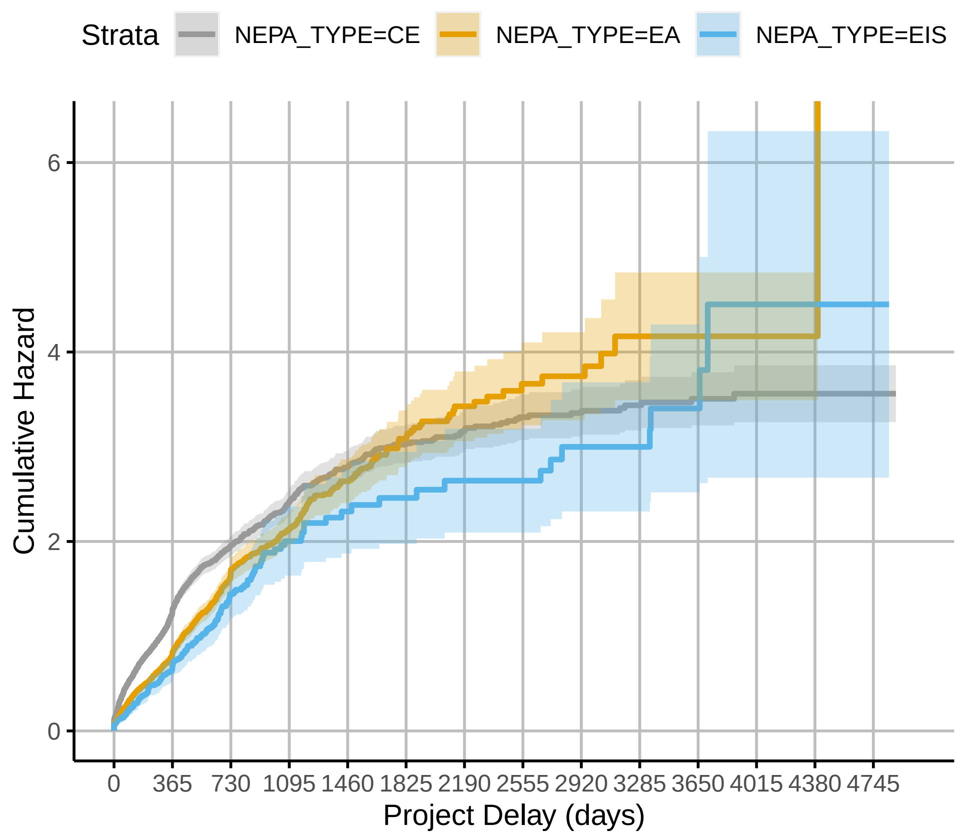
```
km_curv_nepa <- ggsurvplot(fit_delay,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Probability that Project is Not Started",
           xlab = "Project Delay (days)",
           palette = c("#999999", "#E69F00", "#56B4E9"),
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_curv_nepa)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_curv_nepa.pdf", print(km_c
```

Survival Analysis - NEAP Type - KM Cumulative Hazard

```
km_haz_nepa <- ggsurvplot(fit_delay,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Cumulative Hazard",
           xlab = "Project Delay (days)",
           palette = c("#999999", "#E69F00", "#56B4E9"),
           fun = "cumhaz",
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_haz_nepa)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_nepa.pdf", print(km_ha
```

Survival Analysis - Regions - KM Estimate

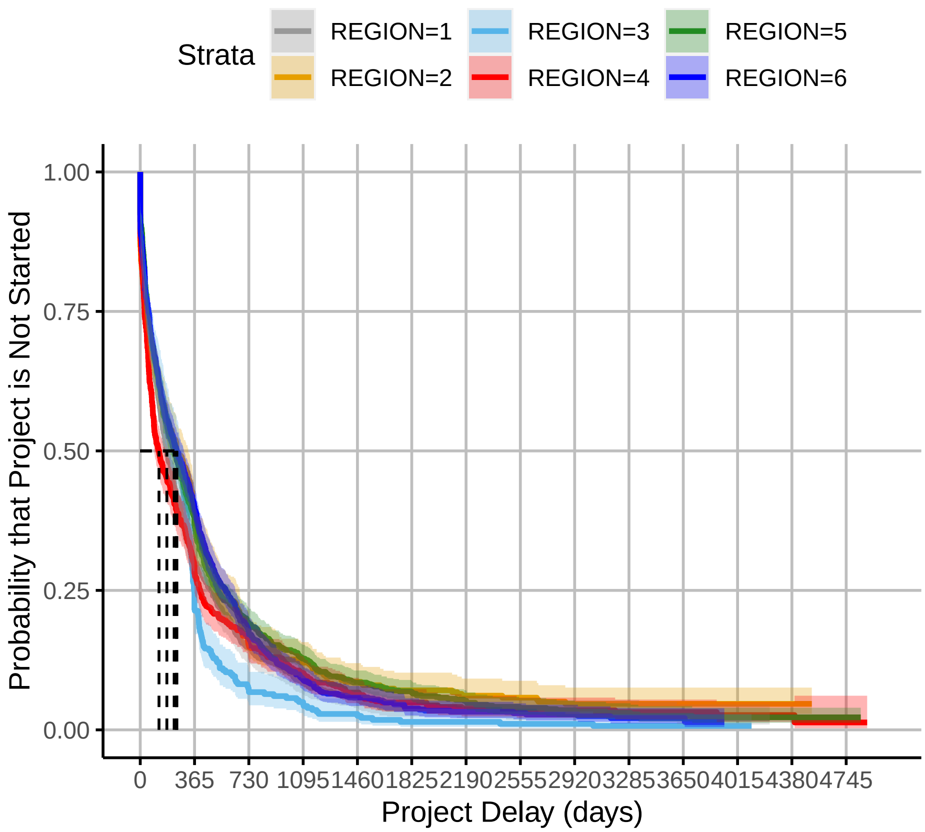
Data grouped by project location. United Stated Forest Service (USFS) Regions 1-6

```
fit_reg <- survfit(Surv(proj.delay, INITIATED) ~ REGION, data = df_fin)
fit_reg_table <- summary(fit_reg)
fit_reg
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ REGION, data = df_fin)
##
##            n events median 0.95LCL 0.95UCL
## REGION=1 538    503    179     153     218
## REGION=2 455    405    245     197     323
## REGION=3 301    293    245     183     287
## REGION=4 540    507    126      94     173
## REGION=5 928    845    228     195     261
## REGION=6 795    736    245     205     294
```

Survival Analysis - Regions - KM Curve

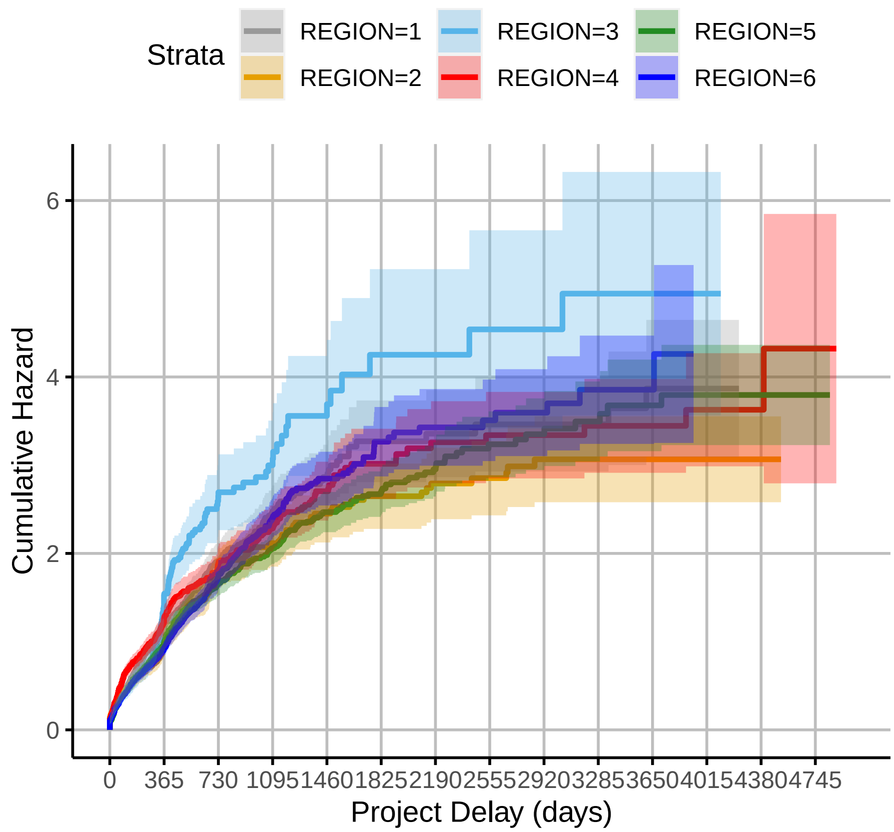
```
km_curv_reg <- ggsurvplot(fit_reg,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Probability that Project is Not Started",
           xlab = "Project Delay (days)",
           palette = c("#999999", "#E69F00", "#56B4E9", "red", "forestgreen", "blue"),
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_curv_reg)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_curv_reg.pdf", print(km_cu
```

Survival Analysis - Regions - KM Cumulative Hazard

```
km_haz_reg <- ggsurvplot(fit_reg,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Cumulative Hazard",
           xlab = "Project Delay (days)",
           palette = c("#999999", "#E69F00", "#56B4E9", "red", "forestgreen", "blue"),
           fun = "cumhaz",
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_haz_reg)
```




```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_reg.pdf", print(km_haz
```

Survival Analysis - Size - KM Estimate

Data grouped by a project's cumulative size. Size categories = small, medium, large, and extra-large.

Sizes are categorized based on quartile ranges + x-large > 2670 acres + large 768-2670 acres + medium 174-768 acres + small < 174 acres

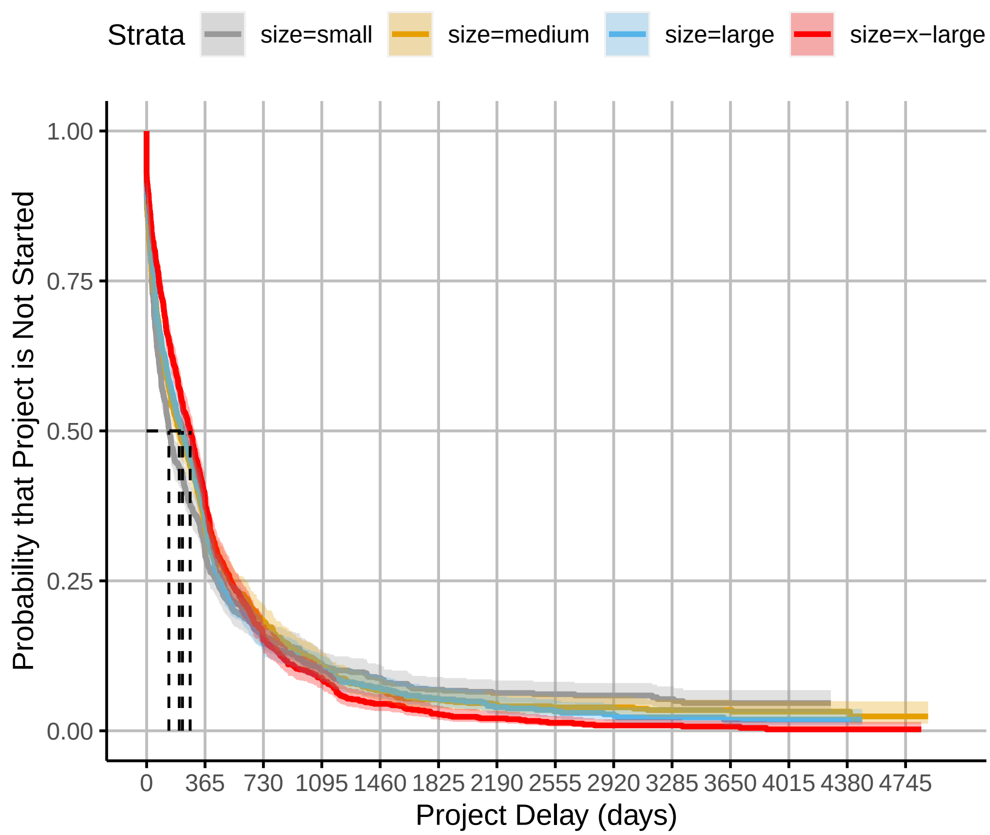
```
fit_size <- survfit(Surv(proj.delay, INITIATED) ~ size, data = df_fin)
fit_size_table <- summary(fit_size)
fit_size
```

```
## Call: survfit(formula = Surv(proj.delay, INITIATED) ~ size, data = df_fin)
##
##                  n events median 0.95LCL 0.95UCL
## size=small     888    803    139     123     167
## size=medium    890    811    204     171     245
## size=large     888    814    224     187     261
## size=x-large   891    861    272     235     300
```

Survival Analysis - Size - KM Curve

```
km_curv_size <- ggsurvplot(fit_size,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           surv.median.line = "hv",
```

```
           break.time.by = 365,
           risk.table.y.text=FALSE,
           tables.height = 0.3,
           censor = FALSE,
           ylab = "Probability that Project is Not Started",
           xlab = "Project Delay (days)",
           palette = c("#999999", "#E69F00", "#56B4E9", "red"),
           surv.plot.height = 1,
           ggtheme = theme(aspect.ratio = 0.75,
                           axis.line = element_line(colour = "black"),
                           panel.grid.major = element_line(colour = "grey"),
                           panel.border = element_blank(),
                           panel.background = element_blank()),
           tables.theme =  theme(aspect.ratio = 0.06)
           )
print(km_curv_size)
```

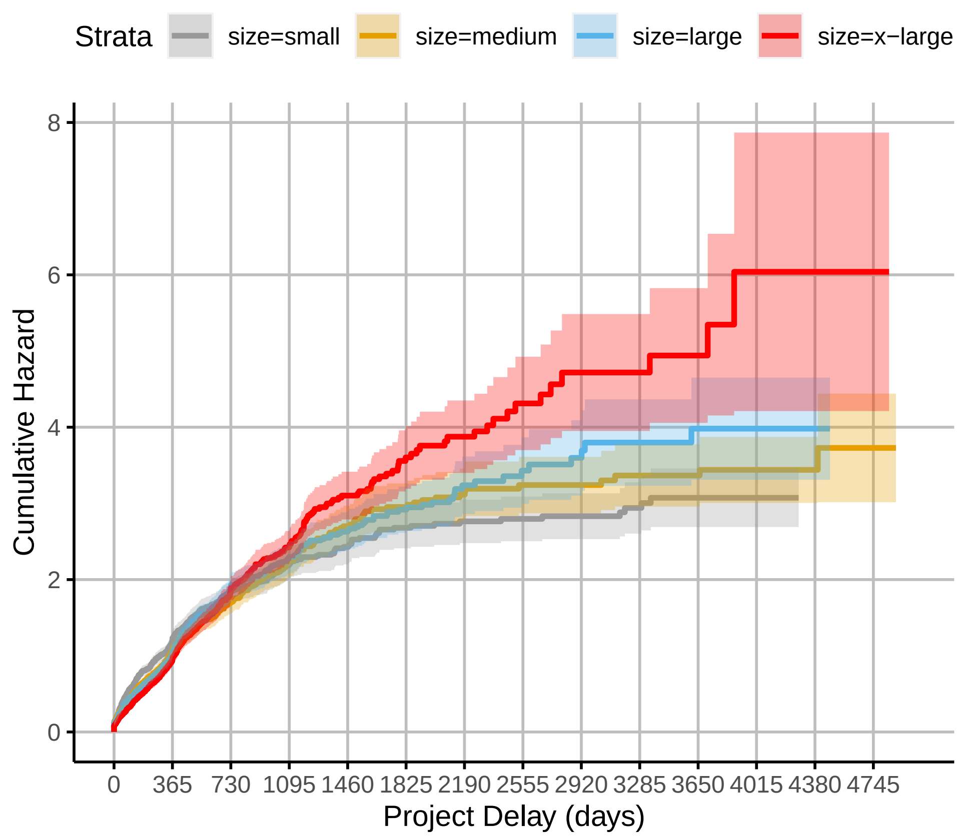



```
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_curv_size.pdf", print(km_c
```

Survival Analysis - Size - KM Cumulative Hazard

```
km_haz_size <- ggsurvplot(fit_size,
           conf.int = TRUE,
           risk.table = FALSE,
           risk.table.col = "strata",
           break.time.by = 365,
           risk.table.y.text=FALSE,
```

```r
          tables.height = 0.3,
          censor = FALSE,
          ylab = "Cumulative Hazard",
          xlab = "Project Delay (days)",
          palette = c("#999999", "#E69F00", "#56B4E9", "red"),
          fun = "cumhaz",
          surv.plot.height = 1,
          ggtheme = theme(aspect.ratio = 0.75,
                     axis.line = element_line(colour = "black"),
                          panel.grid.major = element_line(colour = "grey"),
                          panel.border = element_blank(),
                          panel.background = element_blank()),
          tables.theme =  theme(aspect.ratio = 0.06)
          )
print(km_haz_size)
```




```r
#ggsave( "/Users/kathrynmurenbeeld/Desktop/Survival_Analysis/WRITING/FIGS/km_haz_size.pdf", print(km_ha
```

Comparing survival curves using a log-rank approach

- Null hypothesis: no difference in survival between the groups
- Non-parametric test

Log-Rank - Appealed?

```r
# The survdiff function in the survival package does this for you.

surv_diff_delay_app <- survdiff(Surv(proj.delay, INITIATED) ~ APPEALED, data = df_fin)
```

```r
print(surv_diff_delay_app)
```

```
## Call:
## survdiff(formula = Surv(proj.delay, INITIATED) ~ APPEALED, data = df_fin)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## APPEALED=0 3184     2937     2870      1.58      12.6
## APPEALED=1  373      352      419     10.83      12.6
##
##  Chisq= 12.6  on 1 degrees of freedom, p= 4e-04
```

Log-Rank - Litigated?

```r
surv_diff_delay_lit <- survdiff(Surv(proj.delay, INITIATED) ~ LITIGATED, data = df_fin)
print(surv_diff_delay_lit)
```

```
## Call:
## survdiff(formula = Surv(proj.delay, INITIATED) ~ LITIGATED, data = df_fin)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## LITIGATED=0 3445     3194     3154     0.504      12.4
## LITIGATED=1  112       95      135    11.787      12.4
##
##  Chisq= 12.4  on 1 degrees of freedom, p= 4e-04
```

Log-Rank - NEPA

```r
surv_diff_delay_nepa <- survdiff(Surv(proj.delay, INITIATED) ~ NEPA_TYPE, data = df_fin)
print(surv_diff_delay_nepa)
```

```
## Call:
## survdiff(formula = Surv(proj.delay, INITIATED) ~ NEPA_TYPE, data = df_fin)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## NEPA_TYPE=CE  2266     2110     1898      23.7      57.1
## NEPA_TYPE=EA  1073      988     1135      19.1      29.7
## NEPA_TYPE=EIS  218      191      256      16.4      18.0
##
##  Chisq= 60.3  on 2 degrees of freedom, p= 8e-14
```

Log-Rank - Region

```r
surv_diff_delay_reg <- survdiff(Surv(proj.delay, INITIATED) ~ REGION, data = df_fin)
print(surv_diff_delay_reg)
```

```
## Call:
## survdiff(formula = Surv(proj.delay, INITIATED) ~ REGION, data = df_fin)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## REGION=1 538      503      477      1.47      1.74
## REGION=2 455      405      441      2.93      3.43
## REGION=3 301      293      241     11.44     12.59
## REGION=4 540      507      457      5.41      6.37
## REGION=5 928      845      905      3.95      5.52
## REGION=6 795      736      769      1.41      1.87
##
##  Chisq= 27.1  on 5 degrees of freedom, p= 6e-05
```

Log-Rank - size

```
surv_diff_delay_size <- survdiff(Surv(proj.delay, INITIATED) ~ size, data = df_fin)
print(surv_diff_delay_size)
```

```
## Call:
## survdiff(formula = Surv(proj.delay, INITIATED) ~ size, data = df_fin)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## size=small   888      803      772   1.25516  1.666100
## size=medium  890      811      827   0.29103  0.393744
## size=large   888      814      813   0.00047  0.000633
## size=x-large 891      861      877   0.30047  0.415975
##
##  Chisq= 1.9  on 3 degrees of freedom, p= 0.6
```

Cox Proportional Hazards (Cox ph) Model

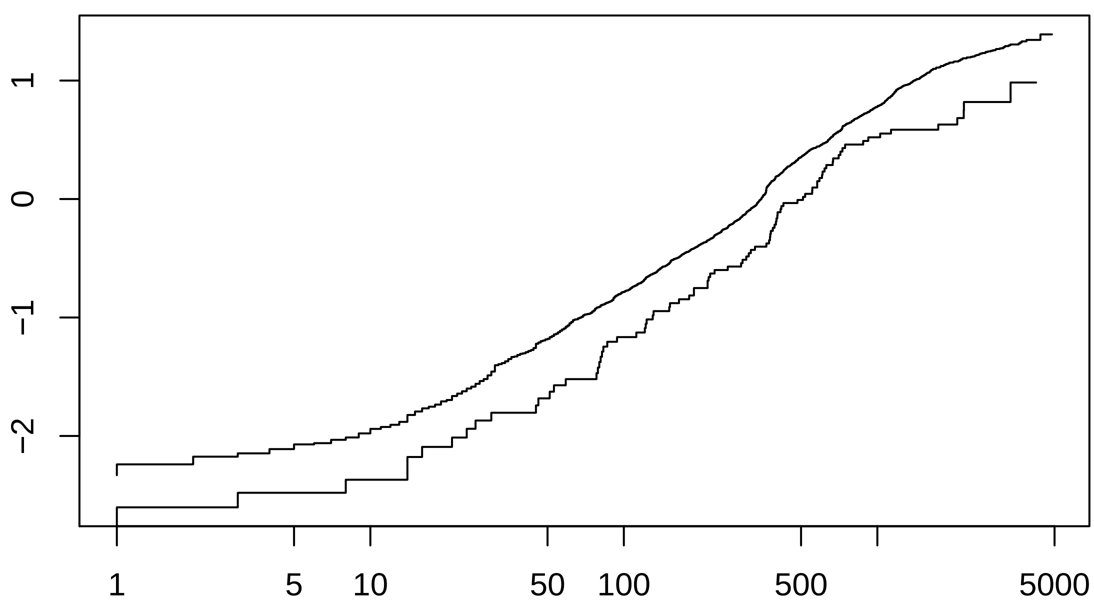
Test Cox proportional hazards assumptions

There are 5 assumptions for using a Cox PH model

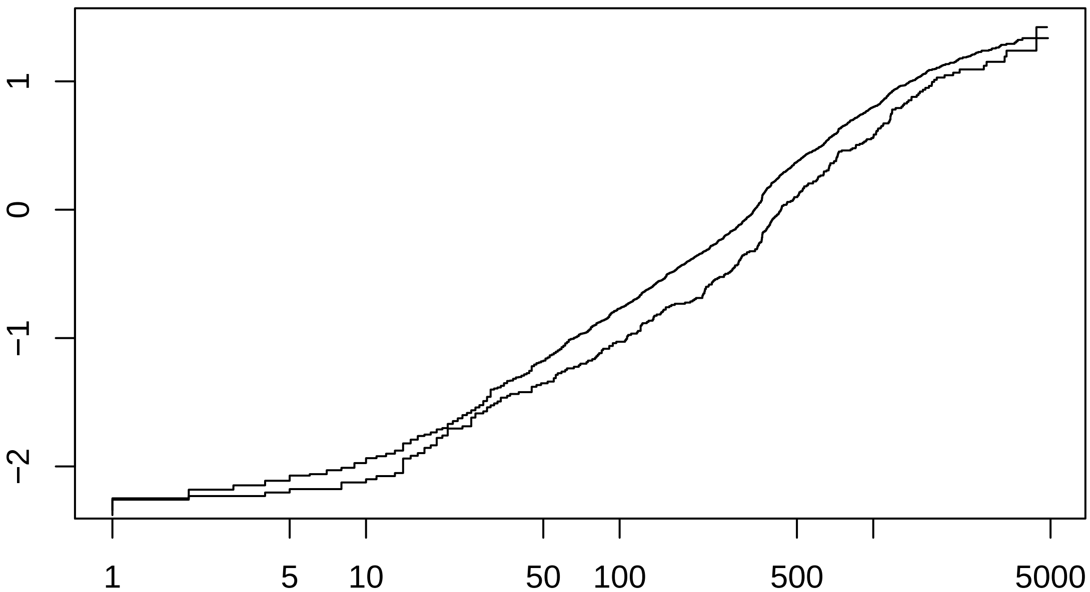
1. Non-informative censoring
2. Survival times, t, are independent
3. Hazards are proportional
4. ln(HR) is a linear function of *numerical* covariates (non in this study)
5. Covariate values don't change over time (eg. changing a treatment or dosage)
6. The baseline hazard (ie. hazard if all covariates were 0) is unspecified

Cox PH - Test proportionality assumption

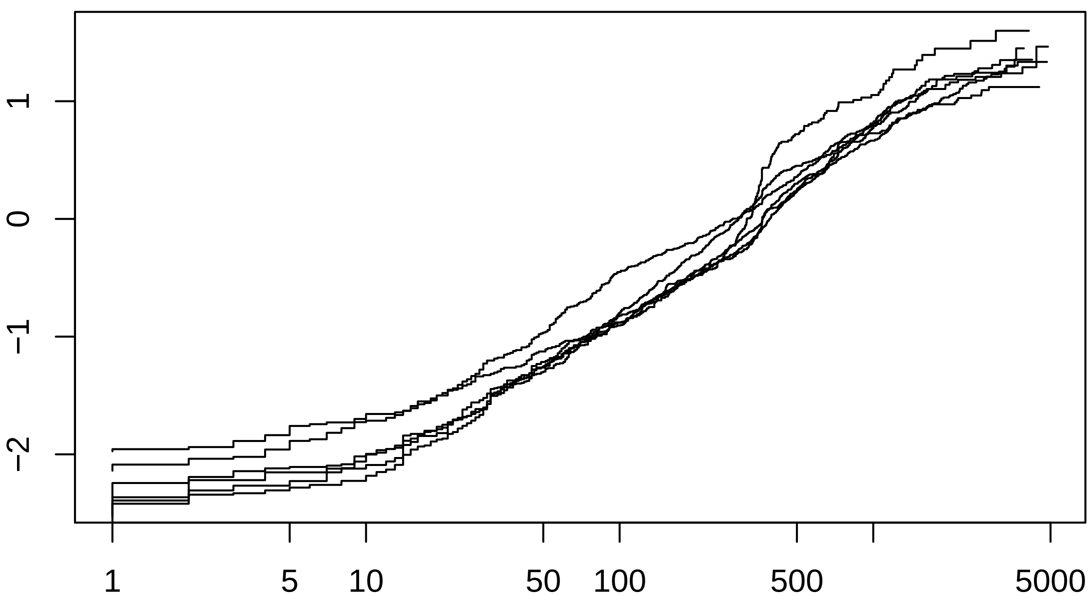
c-log-log test Litigated?



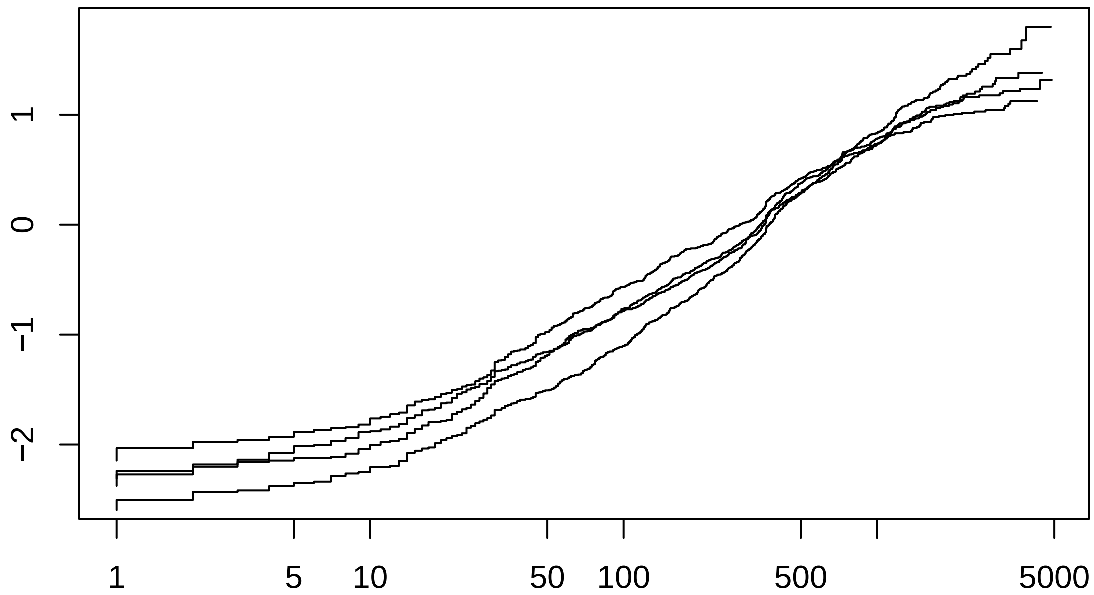
c-log-log test Appealed?



c-log-log test Region?



c-log-log test Size?



Determining the Cox Proportional Hazards Model

Here we used a likelihood ratios test (LRT) to determine the final form of the Cox proportional hazards model. If p-values are <0.05 the covariate should be included in the model.

```
## Call:  coxph(formula = Surv(proj.delay, INITIATED) ~ 1, data = df_fin)
##
## Null model
##   log likelihood= -23950.14
##   n= 3557
```

Should we add NEPA type?

```
## Call:
## coxph(formula = Surv(proj.delay, INITIATED) ~ NEPA_TYPE, data = df_fin)
##
##                  coef exp(coef) se(coef)      z        p
## NEPA_TYPEEA  -0.24888   0.77967  0.03865 -6.439 1.21e-10
## NEPA_TYPEEIS -0.40285   0.66841  0.07563 -5.327 1.00e-07
##
## Likelihood ratio test=62.32  on 2 df, p=2.934e-14
## n= 3557, number of events= 3289
```

```
# Here we use the anova function to complete the LRT test.
anova(cox.cat2, cox.test, test="LRT")
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(proj.delay, INITIATED)
##  Model 1: ~ NEPA_TYPE
##  Model 2: ~ 1
##   loglik Chisq Df P(>|Chi|)
## 1 -23919
## 2 -23950 62.32  2 2.934e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes

Should we include Litigated?

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(proj.delay, INITIATED)
##  Model 1: ~ NEPA_TYPE + LITIGATED
##  Model 2: ~ NEPA_TYPE
##   loglik  Chisq Df P(>|Chi|)
## 1 -23916
## 2 -23919 6.2551  1   0.01238 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes

Should we include Appealed?

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(proj.delay, INITIATED)
##  Model 1: ~ NEPA_TYPE + LITIGATED + APPEALED
##  Model 2: ~ NEPA_TYPE + LITIGATED
##   loglik  Chisq Df P(>|Chi|)
## 1 -23916
## 2 -23916 0.4222  1    0.5159
```

No

Should we include Region?

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(proj.delay, INITIATED)
##  Model 1: ~ NEPA_TYPE + LITIGATED + REGION
##  Model 2: ~ NEPA_TYPE + LITIGATED
##   loglik  Chisq Df P(>|Chi|)
## 1 -23905
## 2 -23916 22.477  5 0.0004248 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes

Should we include size?

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(proj.delay, INITIATED)
##  Model 1: ~ NEPA_TYPE + LITIGATED + REGION + size
##  Model 2: ~ NEPA_TYPE + LITIGATED + REGION
##   loglik  Chisq Df P(>|Chi|)
## 1 -23898
## 2 -23905 12.284  3   0.00647 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

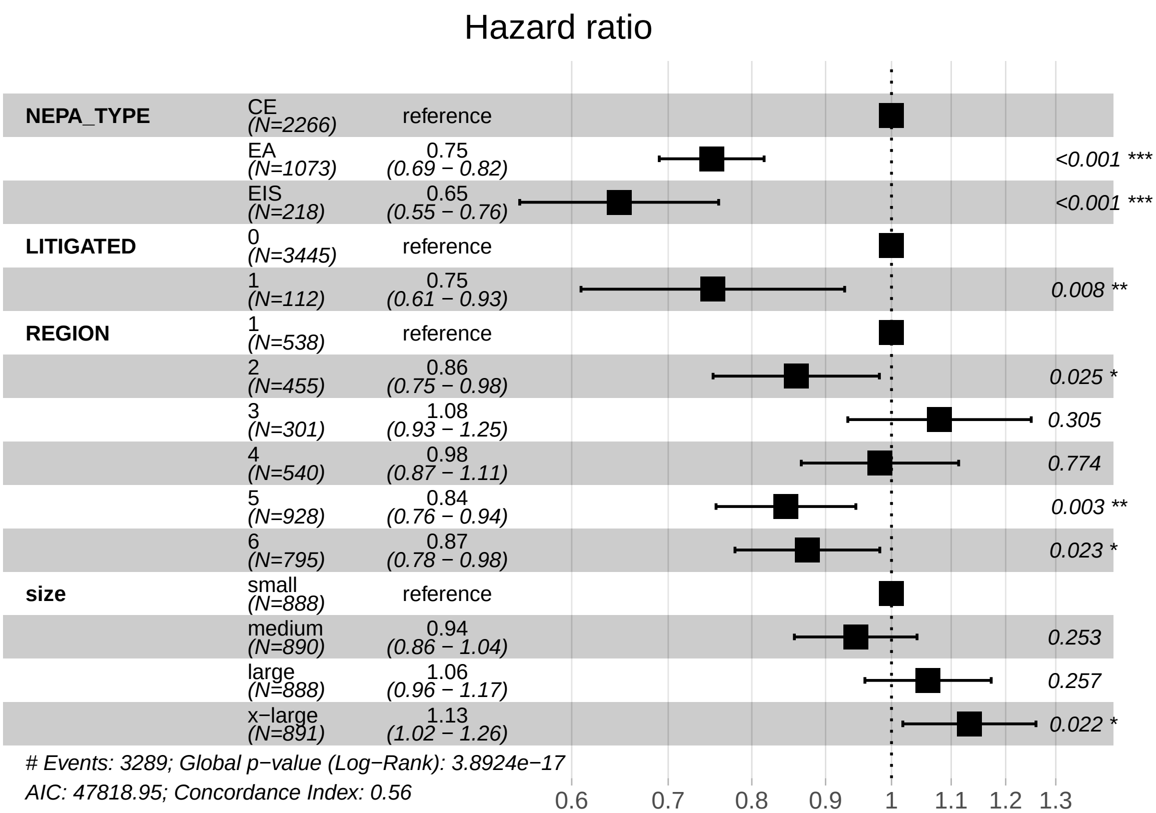
The final Cox proportional hazards model should include all covariates exceot for appealed.

Need to create Cox models using each of the dataframes made with a different region set as the reference.

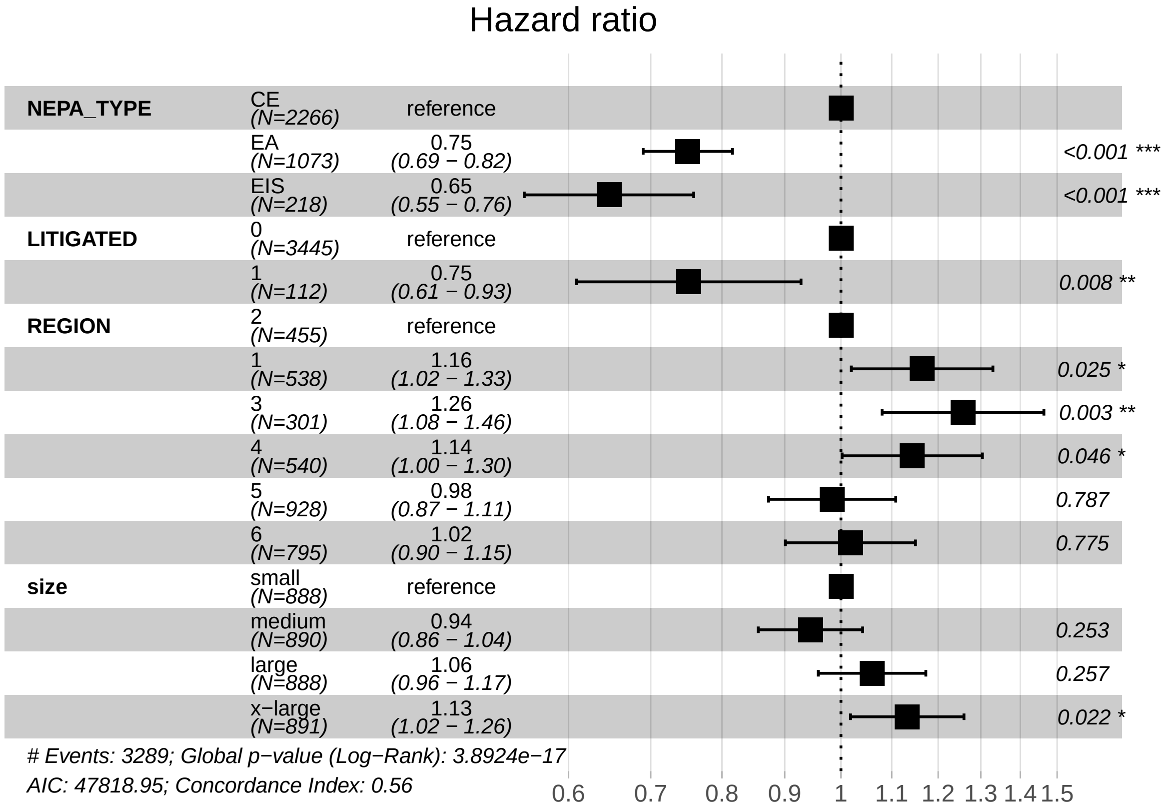
Create forest plots of Cox proportional hazards models.

Cox PH Model - Hazard Ratios - Region 1 as Ref

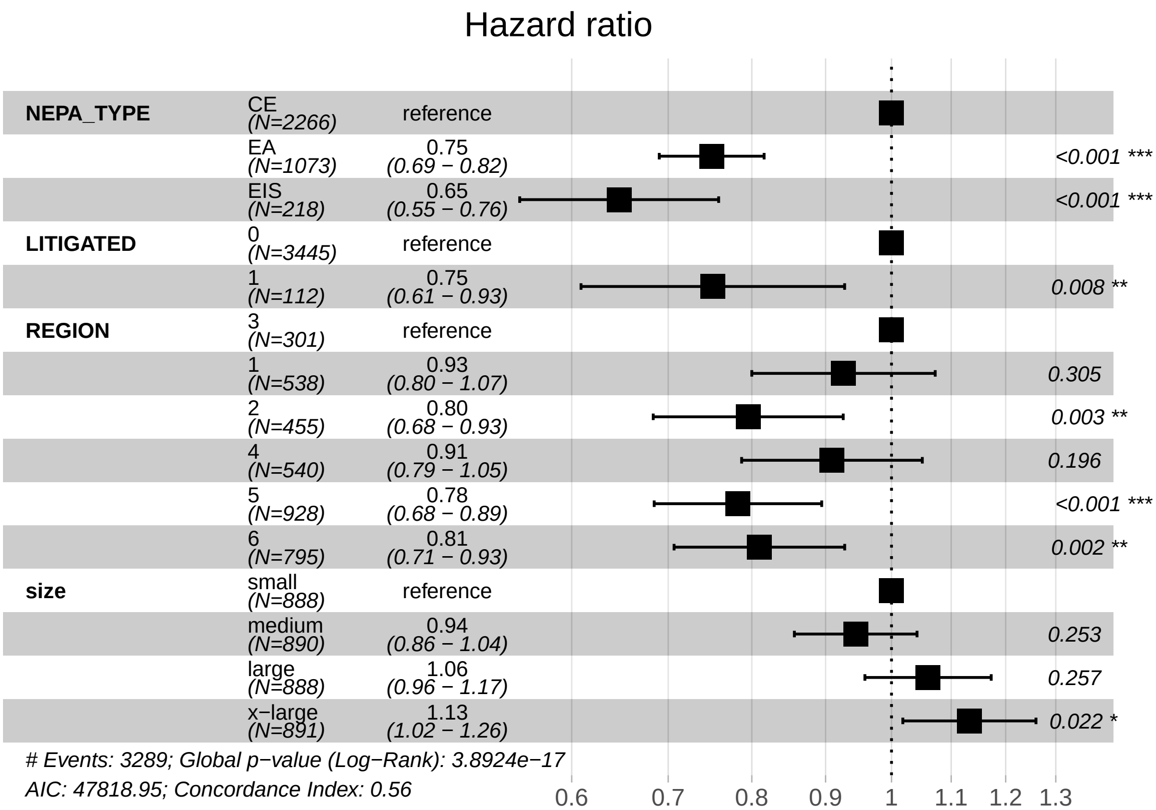
```
## Call:
## coxph(formula = Surv(proj.delay, INITIATED) ~ NEPA_TYPE + LITIGATED +
##     REGION + size, data = df_fin)
##
##                  coef exp(coef) se(coef)      z        p
## NEPA_TYPEEA  -0.28722   0.75035  0.04272 -6.722 1.79e-11
## NEPA_TYPEEIS -0.43501   0.64726  0.08107 -5.366 8.06e-08
## LITIGATED1   -0.28542   0.75170  0.10743 -2.657  0.00789
## REGION2      -0.15222   0.85880  0.06776 -2.246  0.02468
## REGION3       0.07668   1.07970  0.07474  1.026  0.30493
## REGION4      -0.01845   0.98172  0.06413 -0.288  0.77357
## REGION5      -0.16863   0.84482  0.05698 -2.960  0.00308
## REGION6      -0.13439   0.87425  0.05901 -2.277  0.02277
## sizemedium   -0.05720   0.94441  0.04999 -1.144  0.25251
## sizelarge     0.05836   1.06010  0.05146  1.134  0.25677
## sizex-large   0.12441   1.13248  0.05425  2.293  0.02183
##
## Likelihood ratio test=103.3  on 11 df, p=< 2.2e-16
## n= 3557, number of events= 3289
```



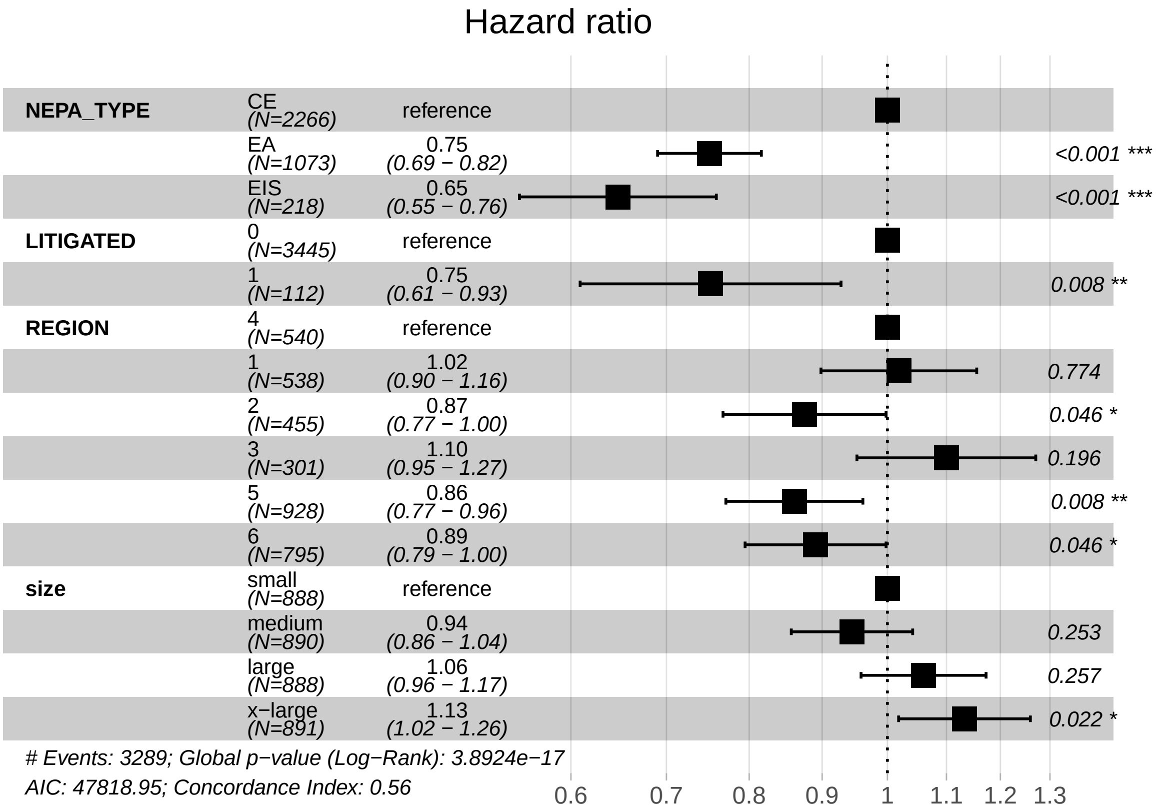
Cox PH Model - Hazard Ratios - Region 2 as Ref



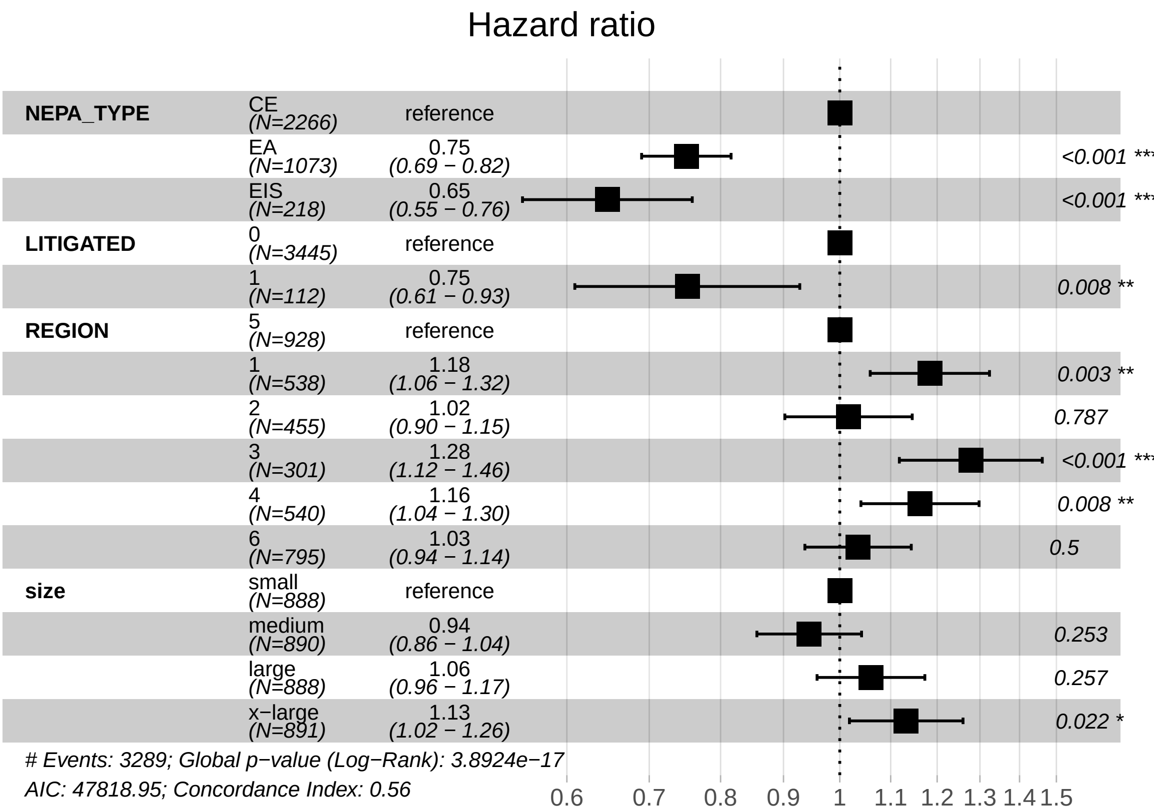
Cox PH Model - Hazard Ratios - Region 3 as Ref



Cox PH Model - Hazard Ratios - Region 4 as Ref



Cox PH Model - Hazard Ratios - Region 5 as Ref



Cox PH Model - Hazard Ratios - Region 6 as Ref