Capstone project 2: Choose Your Own

Alicia Alfonso

2022-11-20

I. Introduction

Daphniids are small crustaceans that live in freshwater ecosystems. Like many aquatic invertebrates, their body is composed of calcium (Ca) and their survival likely depends on the availability of this element in the water. The dataset used for this project are simplified results from a personal experiment (Alfonso, A. 2018 unpublished). The aim of the experiment was to assess the inter-clonal variation between two populations/clones (Northern and Southern) at different Ca concentrations (Ca 0.5 mg/l, 4 mg/l and 106 mg/l) and temperatures (15 and 25ºC). With this information, we will create a model to predict the survival of Daphniids evaluating the effect of the different variables measured.

II. Methods and analysis

Part 1: Data exploration

```
data <- read.table("CapstoneProjectData.txt", header=TRUE) # Load dataset
library(survival) # For modelling
library(survminer) # For survival curves
library(dplyr) # For analysis
library(ggplot2) # For visualization

str(data) # 123 obs. of  6 variables
```

```
## 'data.frame':    123 obs. of  6 variables:
##  $ Clone: chr  "Southern" "Southern" "Southern" "Southern" ...
##  $ Temp : int  15 15 15 15 15 15 15 15 15 15 ...
##  $ Ca   : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ ID   : int  1 10 2 3 4 5 6 7 8 9 ...
##  $ Time : int  24 32 35 39 39 46 46 8 39 35 ...
##  $ Dead : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
head(data, 10)
```

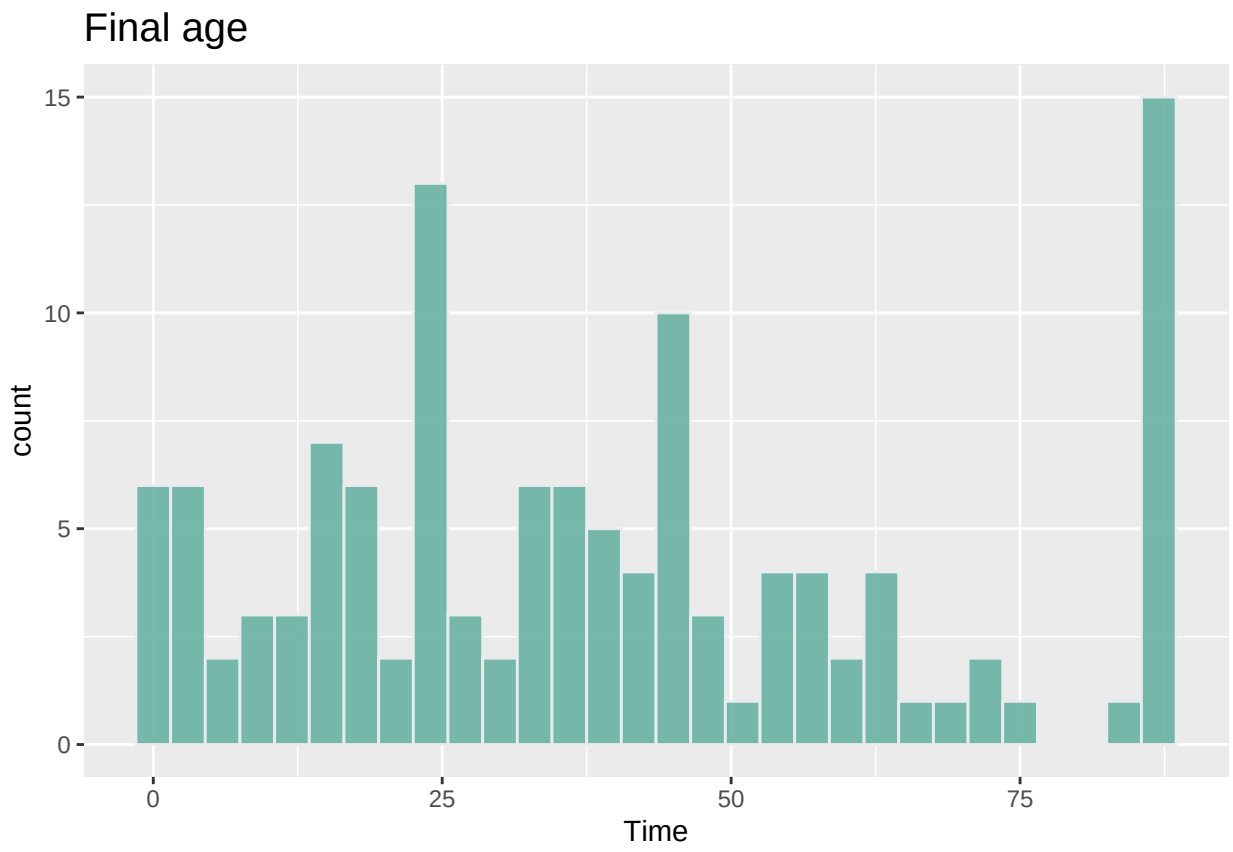
```
##       Clone Temp  Ca ID Time Dead
## 1  Southern   15 0.5  1   24 TRUE
## 2  Southern   15 0.5 10   32 TRUE
## 3  Southern   15 0.5  2   35 TRUE
## 4  Southern   15 0.5  3   39 TRUE
## 5  Southern   15 0.5  4   39 TRUE
## 6  Southern   15 0.5  5   46 TRUE
## 7  Southern   15 0.5  6   46 TRUE
## 8  Southern   15 0.5  7    8 TRUE
## 9  Southern   15 0.5  8   39 TRUE
```

```
## 10 Southern   15 0.5  9   35 TRUE
```

```
summary(data)
```

```
##     Clone                Temp             Ca              ID       
##  Length:123         Min.   :15.00   Min.   :  0.50   Min.   : 1.000  
##  Class :character   1st Qu.:15.00   1st Qu.:  0.50   1st Qu.: 3.000  
##  Mode  :character   Median :15.00   Median :  4.00   Median : 6.000  
##                     Mean   :19.72   Mean   : 33.32   Mean   : 6.195  
##                     3rd Qu.:25.00   3rd Qu.:106.00   3rd Qu.: 8.000  
##                     Max.   :25.00   Max.   :106.00   Max.   :20.000  
##       Time           Dead         
##  Min.   : 1.00   Mode :logical  
##  1st Qu.:18.00   FALSE:13       
##  Median :35.00   TRUE :110      
##  Mean   :38.94                  
##  3rd Qu.:54.50                  
##  Max.   :88.00
```

```
p <- data %>%
  ggplot(aes(x=Time)) +
    geom_histogram( binwidth=3, fill="#69b3a2", color="#e9ecef", alpha=0.9) +
    ggtitle("Final age") +
    theme(plot.title = element_text(size=15))
p
```




```
sum(data$Time == 88) # 13 alive at end
```

```
## [1] 13
```

```
data %>%
  group_by(Clone, Temp, Ca) %>%
  summarise_at(vars(Time), list(name = mean))
```

```
## # A tibble: 12 x 4
## # Groups:   Clone, Temp [4]
##    Clone     Temp    Ca  name
##    <chr>    <int> <dbl> <dbl>
##  1 Northern    15   0.5  13.1
##  2 Northern    15   4    53.4
##  3 Northern    15 106    50.4
##  4 Northern    25   0.5  14.2
##  5 Northern    25   4    51.1
##  6 Northern    25 106    40.8
##  7 Southern    15   0.5  34.3
##  8 Southern    15   4    84.8
##  9 Southern    15 106    63.4
## 10 Southern    25   0.5  13.7
## 11 Southern    25   4    43
## 12 Southern    25 106    31.3
```

The table is composed by 123 observations of 6 different variables: Clone (Northern, Southern), Temperature (15ºC, 25ºC), Ca concentration (0.5, 4, 106 mg/l), unique ID of the animal, Time (age at death or maximum survival) and Death (TRUE, FALSE). The name of the clones refer to the geographical origin of the animals, the temperature to the degrees they were grown at (Celsius) and the Ca concentration the different amounts of calcium present in the water. This dataset includes 123 animals followed through a maximum of 88 days. Time and Death then, provide information on the day of their passing or the end of the experiment. By the end of the experiment only 13 animals were alive and the mean survival was 39 days.

Survival time was lowest at Ca 0.5, peaking at Ca 4, and declining again at the highest Ca concentration. They also survived longer at 15C in comparison to 25C. This indicates strong effects of Ca and temperature that need to be taken into account for our model.

Part 2: Modelling survival

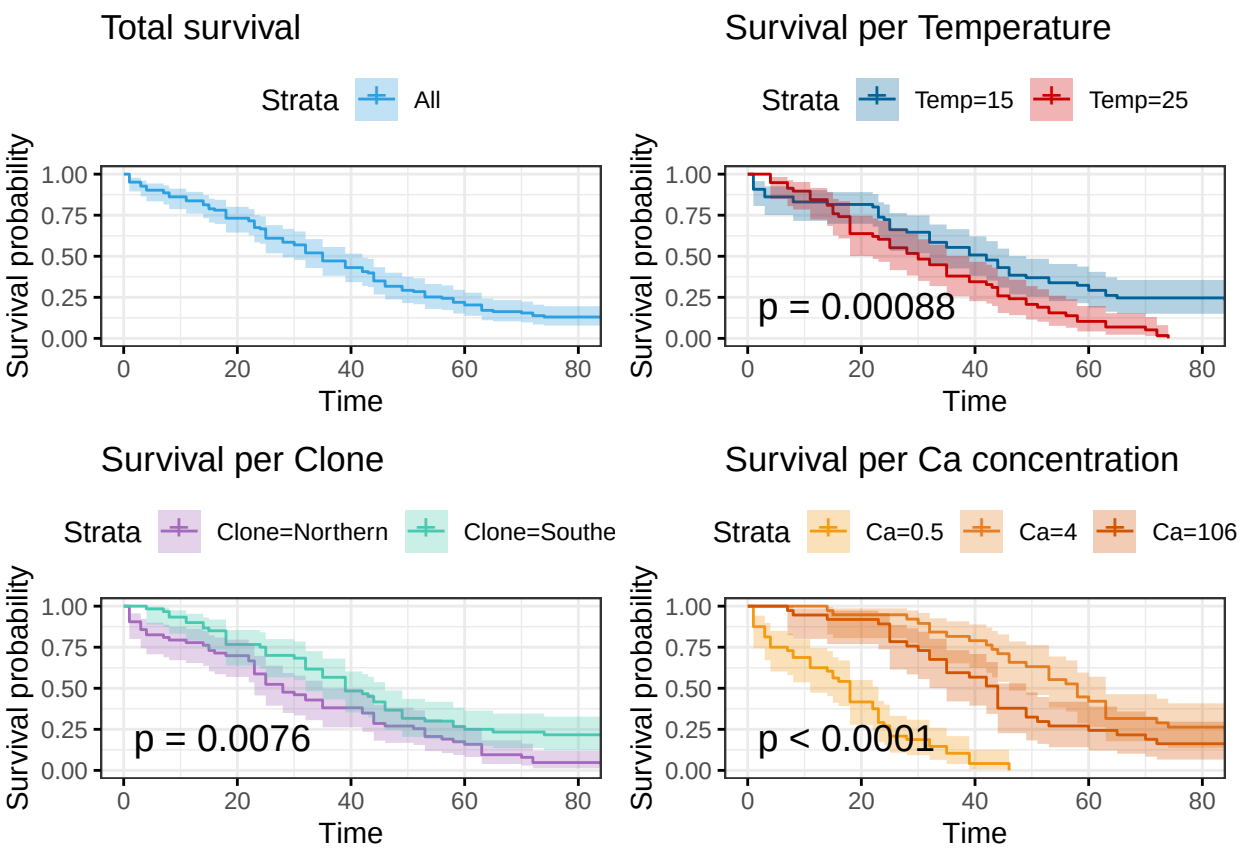
For all variables, I started models with all factors independently and all pairwise interactions (clone, temperature and Ca) to asses which ones contributed the most. This analysis was done with the packages survival (Therneau, 2015) and survminer (Kassambara & Kosinski, 2018) for data fitting and representation. A stepwise model selection discarded Ca × Temperature interaction, giving a more fitted non-parametric Cox proportional hazard model with this form:

log(hazard rate) ~ Clone + Temp + Ca + Clone × Temp + Clone × Ca

The hazard rate is the probability of an animal dying within a given time interval, given that it was alive at the start of it. As such, it corresponds to the instantaneous mortality rate in a continuous-time population model. The full model was contrasted against three other simplifications of it, from which we chose the one with the lowest Akaike Information Criterion (AIC). The final model is calculated with 95% confidence intervals of treatment effects.

```
s <- Surv(data$Time, data$Dead) # Creating our hazard rate
# Survival Plots
m1 <- survfit(s ~ 1, data = data, conf.type = "log-log")
m1p <- ggsurvplot(m1, # Our base curve for all treatments
                  data = data,
                  palette = "#2E9FDF",
                  title = "Total survival",
```

```
                  size = .5,
                  ggtheme = theme_bw())

m2 <- survfit(s ~ Clone, data = data, conf.type = "log-log")
m2p <- ggsurvplot(m2, # Our curve per clonal population
                  data = data,
                  conf.int=TRUE,
                  title = "Survival per Clone",
                  size = .5,  # change line size
                  ggtheme = theme_bw(), # Change ggplot2 theme
                  palette = c("#A569BD", "#48C9B0"), # custom color palette
                  pval = TRUE)

m3 <- survfit(s ~ Temp, data = data, conf.type = "log-log")
m3p <- ggsurvplot(m3, # Our curve by temperature
                  data = data,
                  conf.int=TRUE,
                  title = "Survival per Temperature",
                  size = .5,  # change line size
                  ggtheme = theme_bw(), # Change ggplot2 theme
                  palette = c("#006399", "#CC0000"), # custom color palette
                  pval = TRUE)

m4 <- survfit(s ~ Ca, data = data, conf.type = "log-log")
m4p <- ggsurvplot(m4, # Our curve per Ca concentration
                  data = data,
                  conf.int=TRUE,
                  title = "Survival per Ca concentration",
                  size = .5,  # change line size
                  ggtheme = theme_bw(), # Change ggplot2 theme
                  palette = c("#F39C12","#E67E22","#D35400"),
                  pval = TRUE)

survplots <- list(m1p, m2p, m3p, m4p)
arrange_ggsurvplots(survplots, # Arrange multiple survplots
                    ncol = 2, nrow = 2)
```




```
# Median survival time is 39 days
# Overlapping between clones
# Overlapping between temperatures
# No-overlapping for 0.5 mg Ca/l
```

IV. Results

A Cox regression model was fitted using initially all pairwise interactions and chosen after backward selection using clone, temperature and Ca as explanatory variables. For the different Cox regression models tested, "Model 4" with lowest AIC value (769.51) was accounted for being the best fitted. Consistently with our exploration, and according to Hazard ratios, mortality probabilities are much higher for extreme Ca concentrations (0.5 and 106), 25ºC, and the Northern clone.

```
data$Clone <- as.factor(data$Clone)
data$Temp <- as.factor(data$Temp)
data$Ca <- as.factor(data$Ca)
# Non-parametric Cox proportional hazards models
# Fits a Cox proportional hazards regression model
summary(mod1 <- coxph(s ~ Clone + Temp * Ca, data=data)) # Ca:Temp interaction
```

```
## Call:
## coxph(formula = s ~ Clone + Temp * Ca, data = data)
##
##   n= 123, number of events= 110
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## CloneSouthern -0.510267  0.600335  0.206948 -2.466  0.01368 *
```

```
## Temp25         1.137991  3.120493  0.355727  3.199  0.00138 **
## Ca4           -2.984262  0.050577  0.417956 -7.140 9.32e-13 ***
## Ca106         -2.391092  0.091530  0.383478 -6.235 4.51e-10 ***
## Temp25:Ca4     0.095474  1.100180  0.536355  0.178  0.85872
## Temp25:Ca106  -0.003874  0.996134  0.515542 -0.008  0.99400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## CloneSouthern   0.60034     1.6657   0.40017    0.9006
## Temp25          3.12049     0.3205   1.55392    6.2664
## Ca4             0.05058    19.7719   0.02229    0.1147
## Ca106           0.09153    10.9254   0.04317    0.1941
## Temp25:Ca4      1.10018     0.9089   0.38452    3.1478
## Temp25:Ca106    0.99613     1.0039   0.36265    2.7362
##
## Concordance= 0.785  (se = 0.019 )
## Likelihood ratio test= 112.7  on 6 df,   p=<2e-16
## Wald test            = 94.37  on 6 df,   p=<2e-16
## Score (logrank) test = 129.2  on 6 df,   p=<2e-16
```

```
summary(mod2 <- coxph(s ~ Clone + Temp + Ca, data=data)) # No interaction
```

```
## Call:
## coxph(formula = s ~ Clone + Temp + Ca, data = data)
##
##   n= 123, number of events= 110
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## CloneSouthern -0.51231   0.59911  0.20634 -2.483    0.013 *
## Temp25         1.16378   3.20203  0.22732  5.120 3.06e-07 ***
## Ca4           -2.93271   0.05325  0.32264 -9.090  < 2e-16 ***
## Ca106         -2.40281   0.09046  0.30537 -7.868 3.59e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## CloneSouthern   0.59911     1.6691   0.39982    0.8977
## Temp25          3.20203     0.3123   2.05085    4.9994
## Ca4             0.05325    18.7785   0.02829    0.1002
## Ca106           0.09046    11.0542   0.04972    0.1646
##
## Concordance= 0.794  (se = 0.018 )
## Likelihood ratio test= 112.6  on 4 df,   p=<2e-16
## Wald test            = 94.83  on 4 df,   p=<2e-16
## Score (logrank) test = 122.4  on 4 df,   p=<2e-16
```

```
summary(mod3 <- coxph(s ~ (Clone + Temp + Ca)^2, data=data)) # Ca:Temp,
```

```
## Call:
## coxph(formula = s ~ (Clone + Temp + Ca)^2, data = data)
##
##   n= 123, number of events= 110
##
##                           coef exp(coef)  se(coef)      z Pr(>|z|)
```

```
## CloneSouthern        -1.950414  0.142215  0.377640 -5.165 2.41e-07 ***
## Temp25                0.403028  1.496348  0.395744  1.018   0.3085
## Ca4                  -3.436752  0.032169  0.481403 -7.139 9.40e-13 ***
## Ca106                -3.269079  0.038041  0.486588 -6.718 1.84e-11 ***
## CloneSouthern:Temp25  2.069768  7.922984  0.423351  4.889 1.01e-06 ***
## CloneSouthern:Ca4     0.073427  1.076190  0.505842  0.145   0.8846
## CloneSouthern:Ca106   1.022727  2.780768  0.486486  2.102   0.0355 *
## Temp25:Ca4            0.014259  1.014361  0.539022  0.026   0.9789
## Temp25:Ca106         -0.003878  0.996130  0.532423 -0.007   0.9942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## CloneSouthern          0.14222     7.0316   0.06784   0.29812
## Temp25                 1.49635     0.6683   0.68893   3.25005
## Ca4                    0.03217    31.0858   0.01252   0.08264
## Ca106                  0.03804    26.2871   0.01466   0.09873
## CloneSouthern:Temp25   7.92298     0.1262   3.45567  18.16543
## CloneSouthern:Ca4      1.07619     0.9292   0.39931   2.90043
## CloneSouthern:Ca106    2.78077     0.3596   1.07169   7.21543
## Temp25:Ca4             1.01436     0.9858   0.35268   2.91749
## Temp25:Ca106           0.99613     1.0039   0.35085   2.82823
##
## Concordance= 0.805  (se = 0.018 )
## Likelihood ratio test= 143.8  on 9 df,   p=<2e-16
## Wald test            = 103.6  on 9 df,   p=<2e-16
## Score (logrank) test = 157.2  on 9 df,   p=<2e-16
```

```
# Clone:Temp, Clone:Ca interaction

anova(mod1, mod2, mod3) # (!) Model 3 is already significant
```

```
## Analysis of Deviance Table
##  Cox model: response is  s
##  Model 1: ~ Clone + Temp * Ca
##  Model 2: ~ Clone + Temp + Ca
##  Model 3: ~ (Clone + Temp + Ca)^2
##    loglik   Chisq Df P(>|Chi|)
## 1 -393.33
## 2 -393.35  0.0439  2    0.9783
## 3 -377.76 31.1907  5 8.589e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(mod1, mod2, mod3)   #
```

```
##      df      AIC
## mod1  6 798.6575
## mod2  4 794.7014
## mod3  9 773.5107
```

```
# Using the step method to refine our best model (mod3) into a better one (mod4)
summary(mod4 <- step(mod3))
```

```
## Start:  AIC=773.51
## s ~ (Clone + Temp + Ca)^2
```

```
##
##              Df    AIC
## - Temp:Ca     2 769.51
## <none>          773.51
## - Clone:Ca    2 774.63
## - Clone:Temp  1 795.89
##
## Step:  AIC=769.51
## s ~ Clone + Temp + Ca + Clone:Temp + Clone:Ca
##
##              Df    AIC
## <none>          769.51
## - Clone:Ca    2 770.72
## - Clone:Temp  1 792.26
```

```
## Call:
## coxph(formula = s ~ Clone + Temp + Ca + Clone:Temp + Clone:Ca,
##     data = data)
##
##   n= 123, number of events= 110
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## CloneSouthern        -1.95062   0.14219  0.37130 -5.253 1.49e-07 ***
## Temp25                0.40661   1.50172  0.27723  1.467   0.1425
## Ca4                  -3.43004   0.03239  0.43263 -7.928 2.22e-15 ***
## Ca106                -3.27293   0.03790  0.43865 -7.461 8.56e-14 ***
## CloneSouthern:Temp25  2.06862   7.91386  0.42141  4.909 9.17e-07 ***
## CloneSouthern:Ca4     0.07567   1.07861  0.49636  0.152   0.8788
## CloneSouthern:Ca106   1.02290   2.78124  0.48070  2.128   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## CloneSouthern          0.14219     7.0331   0.06868   0.29438
## Temp25                 1.50172     0.6659   0.87219   2.58564
## Ca4                    0.03239    30.8778   0.01387   0.07562
## Ca106                  0.03790    26.3885   0.01604   0.08953
## CloneSouthern:Temp25   7.91386     0.1264   3.46482  18.07572
## CloneSouthern:Ca4      1.07861     0.9271   0.40772   2.85344
## CloneSouthern:Ca106    2.78124     0.3596   1.08409   7.13533
##
## Concordance= 0.805  (se = 0.018 )
## Likelihood ratio test= 143.8  on 7 df,   p=<2e-16
## Wald test            = 103.7  on 7 df,   p=<2e-16
## Score (logrank) test = 151.9  on 7 df,   p=<2e-16
```

```
# Lowest AIC: Clone:Temp & Clone:Ca retained, no Temp:Ca interaction
AIC(mod1, mod2, mod3, mod4)
```

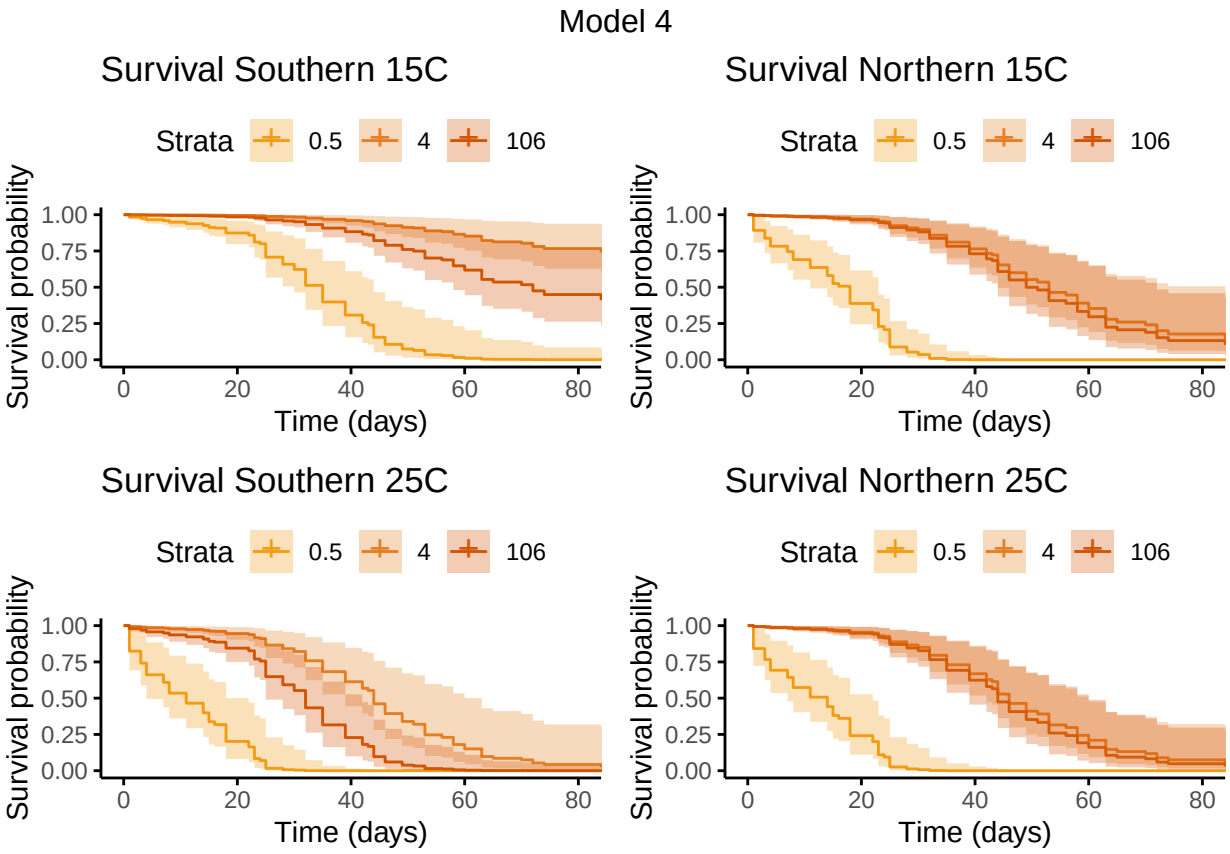
```
##      df      AIC
## mod1  6 798.6575
## mod2  4 794.7014
## mod3  9 773.5107
## mod4  7 769.5120
```

```r
# Creating the better fitted model adding interactions Clone x Temp and
# Clone x Ca
mod4 <- coxph(s ~ Clone + Temp + Ca + Clone:Temp + Clone:Ca, data=data)
summary(mod4)
```

```
## Call:
## coxph(formula = s ~ Clone + Temp + Ca + Clone:Temp + Clone:Ca,
##     data = data)
##
##   n= 123, number of events= 110
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## CloneSouthern        -1.95062   0.14219  0.37130 -5.253 1.49e-07 ***
## Temp25                0.40661   1.50172  0.27723  1.467   0.1425
## Ca4                  -3.43004   0.03239  0.43263 -7.928 2.22e-15 ***
## Ca106                -3.27293   0.03790  0.43865 -7.461 8.56e-14 ***
## CloneSouthern:Temp25  2.06862   7.91386  0.42141  4.909 9.17e-07 ***
## CloneSouthern:Ca4     0.07567   1.07861  0.49636  0.152   0.8788
## CloneSouthern:Ca106   1.02290   2.78124  0.48070  2.128   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## CloneSouthern          0.14219     7.0331   0.06868   0.29438
## Temp25                 1.50172     0.6659   0.87219   2.58564
## Ca4                    0.03239    30.8778   0.01387   0.07562
## Ca106                  0.03790    26.3885   0.01604   0.08953
## CloneSouthern:Temp25   7.91386     0.1264   3.46482  18.07572
## CloneSouthern:Ca4      1.07861     0.9271   0.40772   2.85344
## CloneSouthern:Ca106    2.78124     0.3596   1.08409   7.13533
##
## Concordance= 0.805  (se = 0.018 )
## Likelihood ratio test= 143.8  on 7 df,   p=<2e-16
## Wald test            = 103.7  on 7 df,   p=<2e-16
## Score (logrank) test = 151.9  on 7 df,   p=<2e-16
```

```r
# Model 4 visualization separated by treatment
mod4S15 <- survfit(mod4, newdata=data.frame(Clone="Southern", Temp="15",
                                            Ca=c("0.5", "4", "106")))
pmod4S15 <- ggsurvplot(mod4S15, conf.int=TRUE,
                       title = "Survival Southern 15C",
                       legend.labs = c("0.5", "4", "106"),
                       data=data.frame(Clone="Southern", Temp="15",
                                       Ca=c("0.5", "4", "106")),
                       size = .5,  # change line size
                       xlab = "Time (days)",
                       ggtheme = theme_classic(), # Change ggplot2 theme
                       palette = c("#F39C12","#E67E22","#D35400"))

mod4S25 <- survfit(mod4, newdata=data.frame(Clone="Southern",Temp="25",
                                            Ca=c("0.5", "4", "106")))
pmod4S25 <- ggsurvplot(mod4S25, conf.int=TRUE,
                       title = "Survival Southern 25C",
                       legend.labs = c("0.5", "4", "106"),
```

```
                     data=data.frame(Clone="Southern", Temp="25",
                                     Ca=c("0.5", "4", "106")),
                     size = .5,  # change line size
                    xlab = "Time (days)",
                    ggtheme = theme_classic(), # Change ggplot2 theme
                     palette = c("#F39C12","#E67E22","#D35400"))

mod4N15 <- survfit(mod4, newdata=data.frame(Clone="Northern",Temp="15",
                                     Ca=c("0.5", "4", "106")))
pmod4N15 <- ggsurvplot(mod4N15, conf.int=TRUE,
                     title = "Survival Northern 15C",
                     legend.labs = c("0.5", "4", "106"),
                     data=data.frame(Clone="Northern", Temp="15",
                                     Ca=c("0.5", "4", "106")),
                     size = .5,  # change line size
                    xlab = "Time (days)",
                    ggtheme = theme_classic(), # Change ggplot2 theme
                     palette = c("#F39C12","#E67E22","#D35400"))

mod4N25 <- survfit(mod4, newdata=data.frame(Clone="Northern",Temp="25",
                                     Ca=c("0.5", "4", "106")))
pmod4N25 <- ggsurvplot(mod4N25, conf.int=TRUE,
                     title = "Survival Northern 25C",
                     legend.labs = c("0.5", "4", "106"),
                     data=data.frame(Clone="Northern", Temp="25",
                                     Ca=c("0.5", "4", "106")),
                     size = .5,  # change line size
                    xlab = "Time (days)",
                    ggtheme = theme_classic(), # Change ggplot2 theme
                     palette = c("#F39C12","#E67E22","#D35400"))

mod4plots <- list(pmod4S15, pmod4S25, pmod4N15, pmod4N25)
arrange_ggsurvplots(mod4plots, # Arrange multiple survplots
                    title = "Model 4",
                    ncol = 2, nrow = 2)
```



Survival curves for each Ca treatment (0.5, 4 and 106 mg Ca l-1), temperature (15 and 25ºC) and clone population (Northern and Southern). Shaded error bands represent 95% confidence intervals around the fitted model (solid lines).

V. Conclusion

Modelling survival of organisms can give us a deep insight into which factors contribute to their success in an ecosystem. Using R for this purpose proves extremely useful and practical as we are able to generate predictions from smaller datasets. In our case, we establish the strong negative effect that low calcium concentrations have in the survival of Daphniids. We can conclude that the mortality will highly increase if there is a depletion of this element and that northern populations will be the most affected. Our model also hints towards strong interactions between variables that leads to more interesting questions regarding phenotypic plasticity or even toxicity for even higher calcium concentrations.