R code: Survival Costs and Benefits of Reproduction

Peeter Hõrak and Richard Meitern

2023-11-22

Introduction

The following code is to reproduce the plots in the article and supplementary tables The hypotheses revolve around the idea that having children bears some cost in lifespan. The data is from the Estonian Population Registry and the analysis is done in R. The code is written in Rmarkdown and can be run as is.

```
source("./R/utils.R", encoding = "UTF-8")
source("./R/summary.R", encoding = "UTF-8")
source("./R/pretifySummary.R", encoding = "UTF-8")
source("./R/plot_surv_curves.R", encoding = "UTF-8")
source("./R/meanplot.R", encoding = "UTF-8")

# init directories etc
ddir <- "./data/"
rdir <- "./results/"
inputFname <- paste0(ddir, "sup_mat_ee_pop_reg_data.rds")
```

Data Import

The dataset has following columns imported from the Estonian Population Registry:

- YOB - year of birth (1905-1945)
- YOD22 - year of death (1905-2022), except for those that were alive in 2022, then 2022 is used in that column
- dead - TRUE if the person is dead, FALSE if alive at 2022
- sex - male or female
- children - total number of children born to the person

```
df <-  readRDS(inputFname)
```

Preparing data for analysis

First add variables to the data to simplify analysis.

```
#Defining cohorts (based on mean children counts).
cohort_breaks <- c(1904,1927,1945)
break_labels <- c("1905-27","1928-45")

df$cohort <- cut(df$YOB, breaks=cohort_breaks,
```

```
                 labels=break_labels)
table(df$cohort, sort = F)
```

	count
1905-27	293114
1928-45	286157
NA	0

```
#make maximum children count 9
#set all the children over 8 as 9
df$children_f <- ifelse(df$children > 8, 9, df$children)
#make a factor
cf_lvls <- setNames(0:9, c(0:8, "9+"))
df$children_f <- factor(df$children_f,
                        levels = cf_lvls,
                         labels = names(cf_lvls))
```

```
#extract alive/dead status
df$dead <- !is.na(df$YOD)
#make a variable for lifespan
df$lifespan <- df$YOD - df$YOB
#2023 is the year of the data extraction from population registry
df$YOD22 <- ifelse(df$dead, df$YOD, 2022)
df$lifespan2022 <- df$YOD22 - df$YOB
```

```
#calculate relative LRS
df$rel_children <- ave(df$children, df$sex, df$cohort)
df$rel_children <- df$children / df$rel_children
```

```
#define lists to store results
survFits <- list()
survFitDatas <- list()
plots <- list()
cohort_tbls_results <- list()
```

Select Cohort

```
#define available cohorts
cohort0 <- "1905-45"
cohort1 <- "1905-27"
cohort2 <- "1928-45"

#you can change the cohort here to display the results for
#selected  cohort
cohort <- cohort0

if(cohort %in% break_labels){
  data2use <- df[df$cohort == cohort, ]
} else {
```

```
  warning("All cohorts selected.", call. = F)
  data2use <- df
}
```

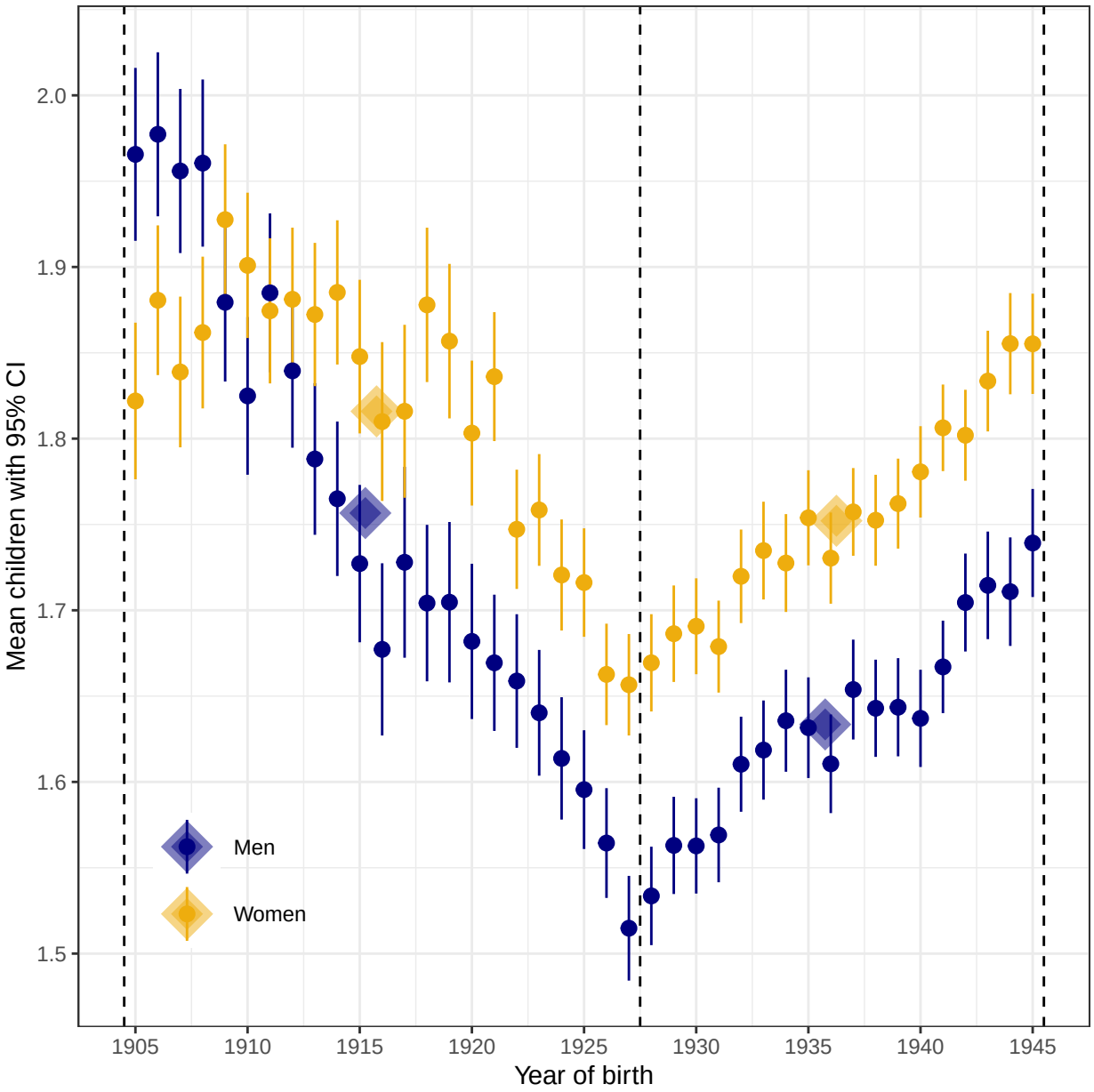
Generate Plots

The following code demonstrates how to generate the plots in the article and supplementary materials. With some minor modifications, the same code can be used to generate either plots in the supplementary material or in the main article.

```
library(ggplot2)
```

Mean Children Count by Year of Birth

```
#make a figure depicting the mean children count by parent YOB
fig1 <- ggplot(data2use, aes(x=YOB, y=children, color=sex)) +
      stat_summary(fun.data=mean_ci) +
      labs(x = "Year of birth", y="Mean children with 95% CI") +
      scale_x_continuous(n.breaks = 9) +
       geom_vline(xintercept = cohort_breaks+0.5, color="black", linetype=2) +
      theme_bw() +
      scale_color_manual(values = c("navy","darkgoldenrod2"),
                         labels = c("Men", "Women")) +
     theme(legend.position = c(0.15,0.15),
           legend.title = element_blank(),
           legend.background = element_blank())

#make table for the cohort means
tbl <- meanplot_data(data2use, y="children", by=c("cohort","sex"))
cohort_means <- (cohort_breaks[-3] + cohort_breaks[-1])/2
names(cohort_means) <-break_labels[1:2]
tbl$YOB <- cohort_means[tbl$cohort]
tbl$children <- tbl$Mean

#add cohort means to the plot
fig1 <- fig1 + geom_point(aes(y=Mean, x=YOB, color=sex),
                  data=tbl, size=10, shape=18, alpha=0.5,
                  position=position_dodge(width = 1)) +
  geom_point(aes(y=Mean, x=YOB, color=sex),
                  data=tbl, size=6, shape=18, alpha=0.5,
                  position=position_dodge(width = 1))
fig1
```




```
#export the table
margin <- theme(plot.margin = unit(c(0.2,0.2,6,0.2), "in"))
ggsave(paste0("./results/mean_children_by_YOB_",cohort,".pdf"),
       fig1 + margin, width = 8.27, height = 11.69)
```

Fit Kaplan Mayer by Sex

```
#define formula for survival analysis
lifespanFormula <- formula(paste0("survival::Surv(",
                                  "time=lifespan2022",
                                  ",",
                                  "event=dead",
```

```
                                    ") ~",
                                    "children_f"))
```

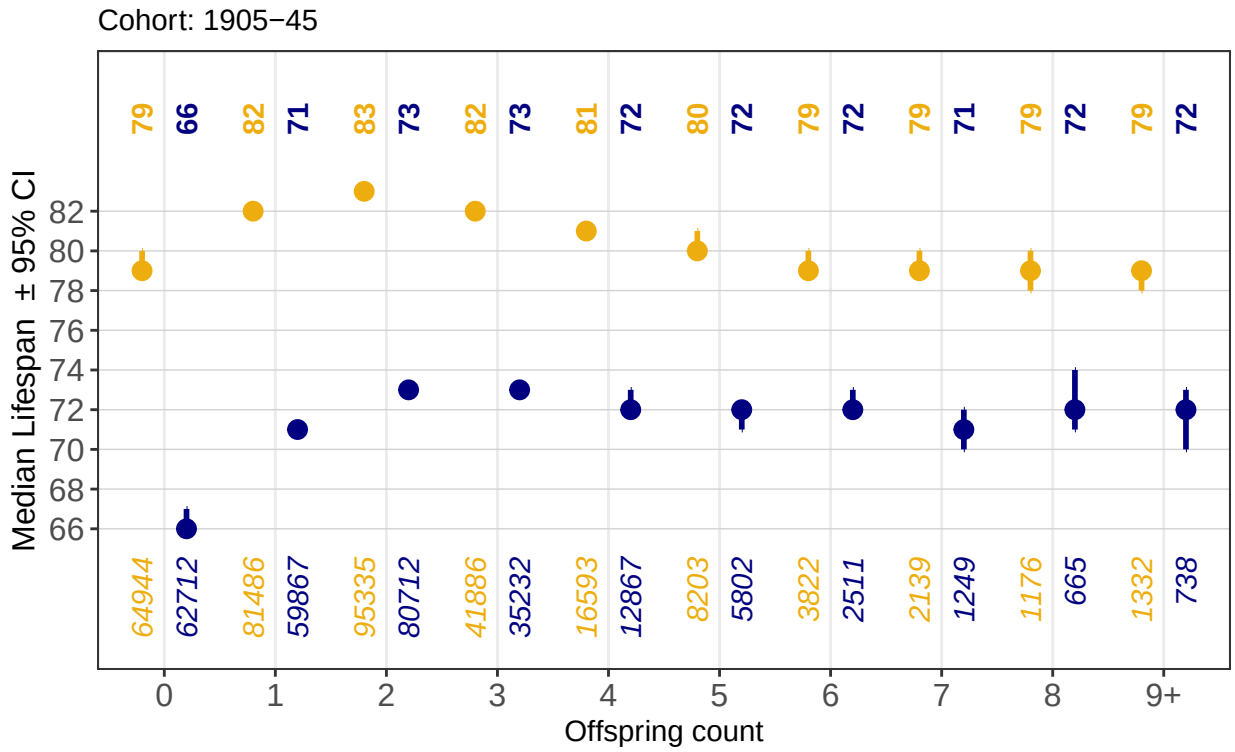
```
# select sex and cohort
selectedSex <- "F"
data2useSex <- data2use[data2use$sex == selectedSex, ]

#fit survival curves
fitSurv <- survminer::surv_fit(lifespanFormula, data=data2useSex)
lab<- paste0(selectedSex, "_", cohort)
survFits[[lab]] <- fitSurv
survFitDatas[[lab]] <- data2useSex
```

```
# select sex and cohort
selectedSex <- "M"
data2useSex <- data2use[data2use$sex == selectedSex, ]
#fit survival curves
fitSurv <- survminer::surv_fit(lifespanFormula, data=data2useSex)
lab<- paste0(selectedSex, "_", cohort)
survFits[[lab]] <- fitSurv
survFitDatas[[lab]] <- data2useSex
```

Median Lifespan

```
m <- makeSexTbl("M", cohort, survFits, cf_lvls, cohort = F)
f <- makeSexTbl("F", cohort, survFits, cf_lvls, cohort = F)
tbl <- rbind(m,f)
cohort_tbls_results[[cohort]] <- tbl
```

```
f2c_median <- meanplot(tbl,
                       x_var="children_f",
                       printCI = F,
                       addQuantiles = F,
                       ylab = "Median Lifespan",
                       subtitle = NULL,
                       xlab="Offspring count",
                       cpositions=c(0, -0.1, -0.15),
                       errorbar_width = 0)
f2c_median + labs(subtitle=paste0("Cohort: ", cohort))
```



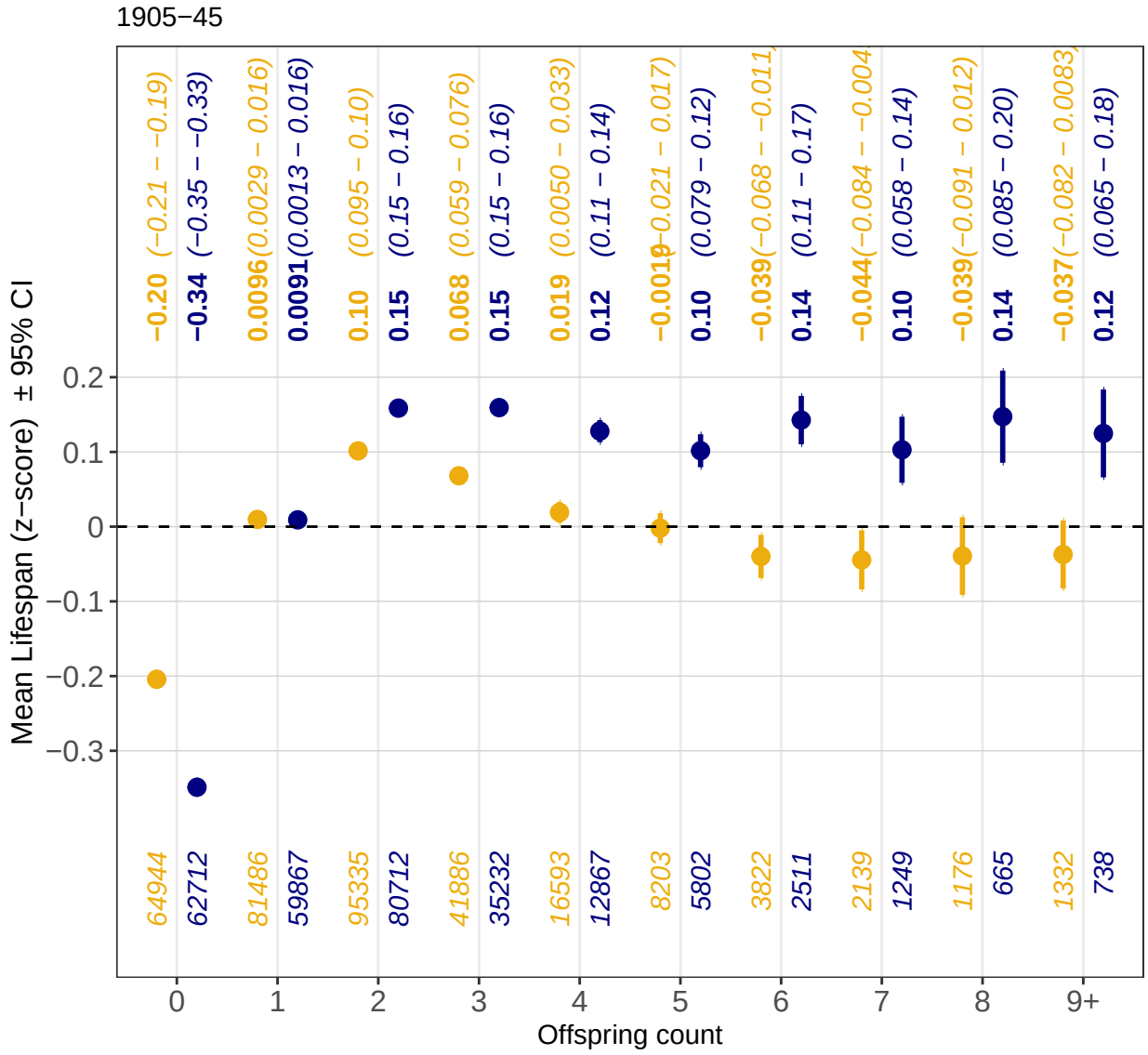

```
margin <- theme(plot.margin = unit(c(0.2,0.2,3,0.2), "in"))
fname <- paste0(rdir, "lifespan_medians_",cohort,".pdf" )
sizeFactor <- 1
ggplot2::ggsave(fname, f2c_median + margin, width=7.08*sizeFactor/2, height=6*sizeFactor)
```

Average Lifespan

```
lifespanMeans <- meanplot_data(data2use,
                               y="lifespan2022",
                               by=c("children_f","sex"),
                               noDigits = 1)
rawMeansPlot <- meanplot(lifespanMeans,
                         x_var="children_f",
                         ylab = "Mean Lifespan",
                         addQuantiles = F,
                         printCI = T,
                         subtitle = cohort,
                         cpositions=c(-0.3, -0.2, -0.1),
                         xlab="Offspring count")
```

```
#scale the data within sex and YOB
lifespanMeansStd <- meanplot_data(data2use,
                                  scale_within = c("sex","YOB"),
                                  y="lifespan2022",
                                  by=c("children_f","sex"),
                                  noDigits = 2)
```

```
stdMeansPlot <- meanplot(lifespanMeansStd,
                         x_var="children_f",
                         ylab = "Mean Lifespan (z-score)",
                         addQuantiles = F,
                         printCI = T,
                         subtitle = cohort,
                         cpositions=c(-0.3, -0.2, -0.1),
                         xlab="Offspring count")
#add line at 0 for reference
stdMeansPlot <- stdMeansPlot + geom_hline(yintercept = 0, linetype=2)

stdMeansPlot
```




```
#combine the plots and save
pA <- labs(subtitle = "A")
pB <- labs(subtitle = "B")
```

```
meanPlots <- patchwork::wrap_plots(rawMeansPlot+pA, stdMeansPlot + pB +
                        theme(axis.title.y = element_blank()), ncol=2)

#add margins below the plot
meanPlots <- meanPlots + theme(plot.margin = unit(c(0.2,0.2,2.2,0.2), "in"))
```

```
fname <- paste0(rdir, "scaled_cohort_means_",cohort,".pdf" )
sizeFactor <- 3
ggplot2::ggsave(fname, meanPlots, width=7.08*sizeFactor/2, height=3*sizeFactor)
```

Summary Tables for Lifespan

```
datas <- split(data2use, data2use$children_f)
dfLife <- multiSummaryStats("lifespan2022",
                            datas,
                            by_cols=c("sex"),
                            direction = "long")
cohort_tbls_results[["lifespan2022"]] <- dfLife
knitr::kable(dfLife)
```

sex	trait	Min.	Median	Mean	Max.	SD	N
M	0lifespan2022	15	66	62.40090	107	19.56766	62712
F	0lifespan2022	15	79	74.34407	108	17.25744	64944
M	1lifespan2022	18	71	68.27757	107	15.61673	59867
F	1lifespan2022	17	80	77.23411	108	13.44797	81486
M	2lifespan2022	20	73	70.59203	105	14.10813	80712
F	2lifespan2022	19	81	78.36132	108	11.97385	95335
M	3lifespan2022	22	73	70.62965	107	13.74849	35232
F	3lifespan2022	22	80	77.97703	108	12.29707	41886
M	4lifespan2022	24	72	70.21341	103	13.60071	12867
F	4lifespan2022	22	80	77.38872	106	12.76387	16593
M	5lifespan2022	26	72	69.80386	100	13.49233	5802
F	5lifespan2022	25	79	77.03852	106	12.69435	8203
M	6lifespan2022	25	72	70.42732	104	13.02352	2511
F	6lifespan2022	29	79	76.55651	105	12.56076	3822
M	7lifespan2022	33	71	69.84147	99	12.43711	1249
F	7lifespan2022	31	79	76.40159	102	12.75868	2139
M	8lifespan2022	31	72	70.51278	99	12.71526	665
F	8lifespan2022	32	78	76.51871	106	12.50496	1176
M	9+lifespan2022	32	72	70.17480	97	12.69646	738
F	9+lifespan2022	33	78	76.51802	103	11.63503	1332

```
dfYOB <-  summary_stats(data2use, "YOB", by_cols=c("sex"))
cohort_tbls_results[["YOB"]] <- dfYOB
knitr::kable(dfYOB)
```

sex	trait	Min.	Median	Mean	Max.	SD	N
M	YOB	1905	1928	1926.497	1945	11.87099	262355
F	YOB	1905	1926	1925.661	1945	11.73052	316916

```
dfYOD  <-  summary_stats(data2use, "YOD", by_cols=c("sex"))
cohort_tbls_results[["YOD"]] <- dfYOD
knitr::kable(dfYOD)
```

sex	trait	Min.	Median	Mean	Max.	SD	N
M	YOD	1930	1994	1991.767	2023	18.59459	238072
F	YOD	1928	2000	1997.995	2023	16.86580	254481

```
data2use$dead_f <- as.factor(data2use$dead)
dfd <- summary_stats(data2use, "dead_f",
              by_cols=c("sex"))
dfd$N <- dfd$`TRUE` + dfd$`FALSE`
dfd[["dead (%)"]] <- (dfd$`TRUE` /dfd$N) *100
dfd[c("Mode", "TRUE", "FALSE","trait")] <- NULL
dfd$`dead (%)` <- round(dfd$`dead (%)`,1)
cohort_tbls_results[["dead"]] <- dfd
knitr::kable(dfd)
```

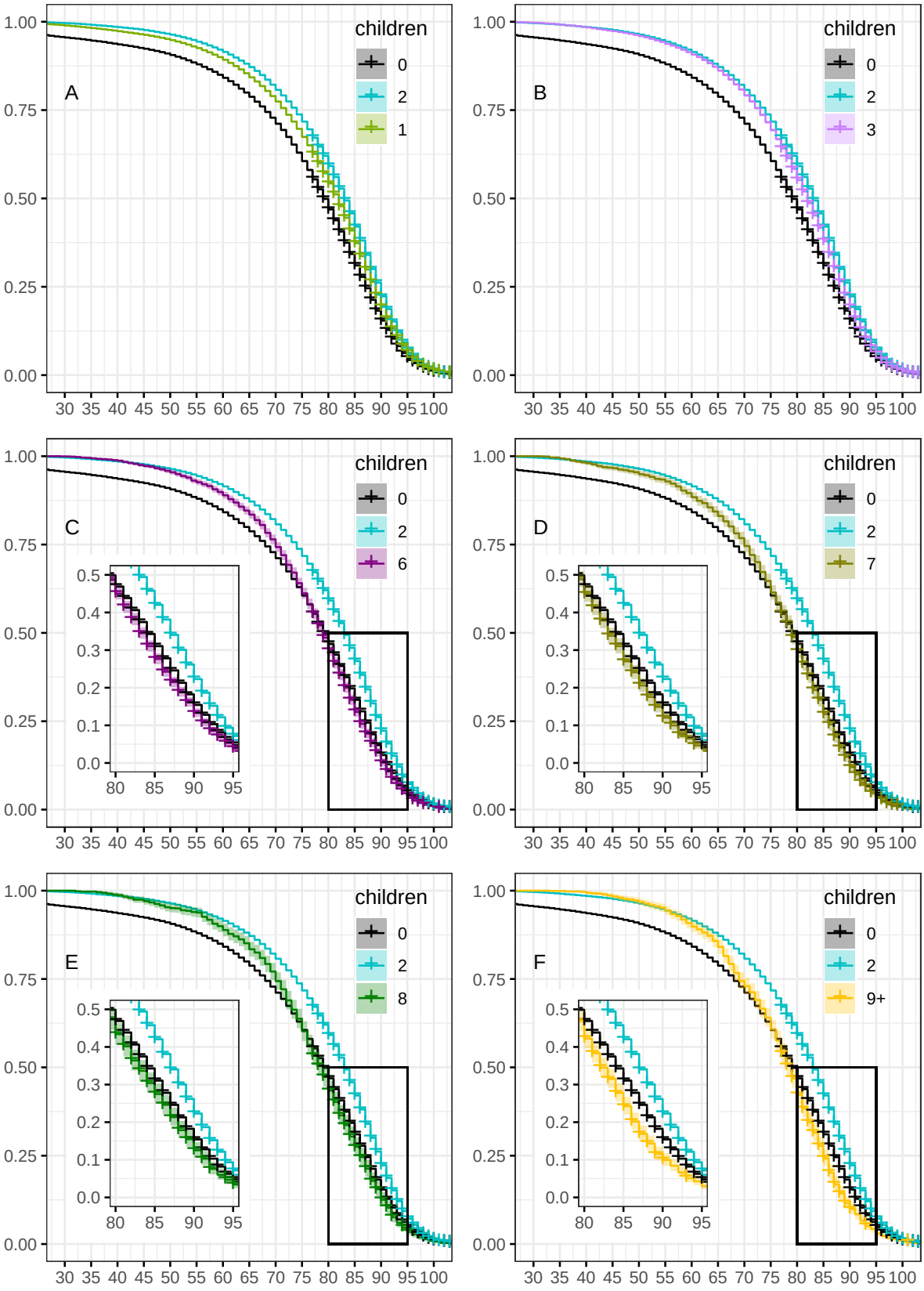
sex	N	dead (%)
M	262355	90.7
F	316916	80.3

Survival Curves by Children Count

```
zoomF <- c(xmin = 80, xmax=95, ymin= 0,ymax=0.5)
counterLabs <- c(1,2,9,9,3:6)
dName <- paste0("F_",cohort)
#generate plots
plotsF <- plot_surv_curves(fit=survFits[[dName]],
                           data=survFitDatas[[dName]],
                           zoom = zoomF,
                           counterLabs= counterLabs)
```

```
#180x254mm
plotSub <- paste0("Females, ", cohort)
patchwork::wrap_plots(plotsF[c(1,2,5:8)], ncol=2)  +
  patchwork::plot_annotation(subtitle = plotSub)
```

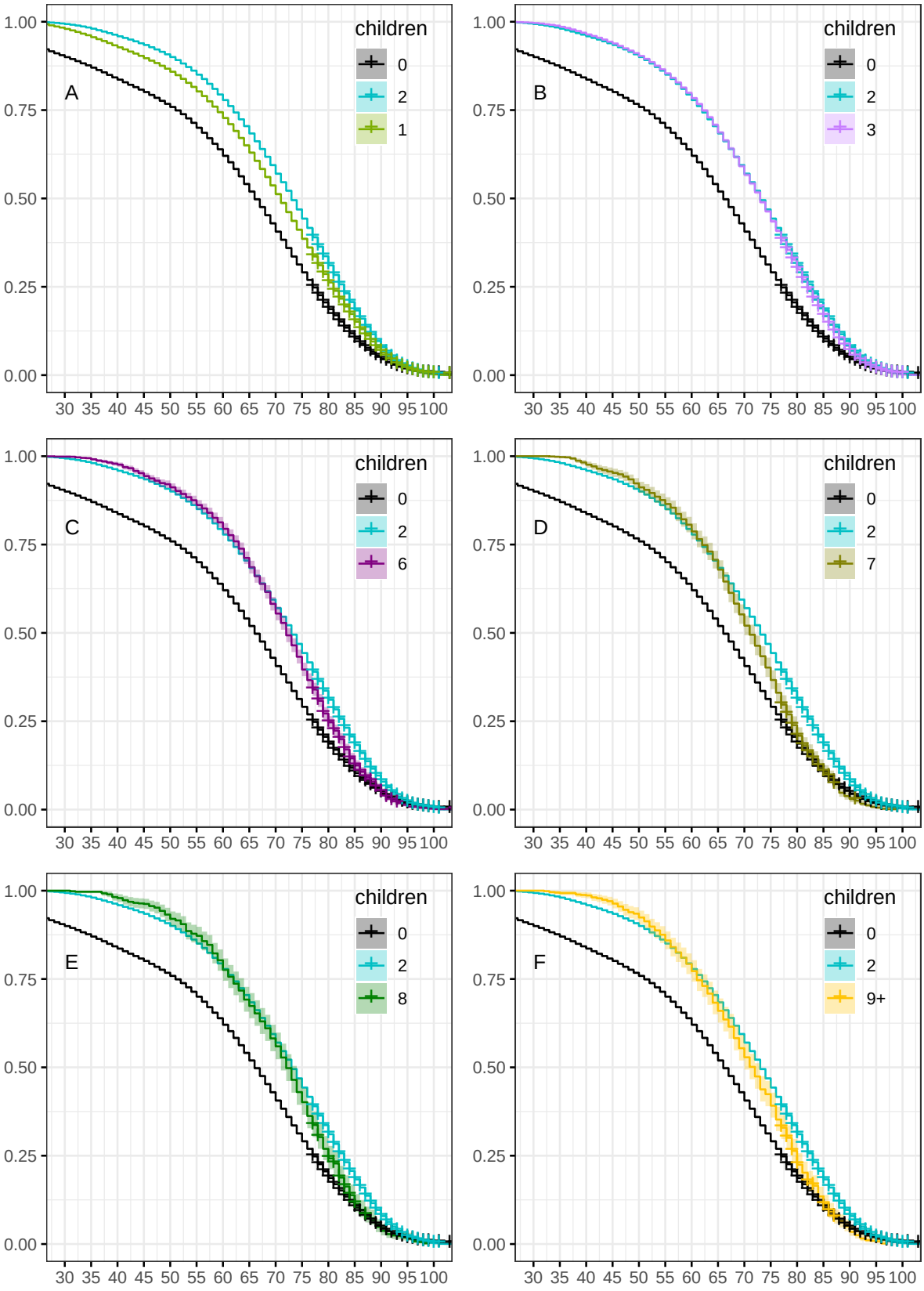
Females, 1905−45



```
zoomM <- c(xmin = 72, xmax=82, ymin= 0.15,ymax=0.45)
counterLabs <- c(1,2,9,9,3:6)
dName <- paste0("M_",cohort)
#generate plots
plotsM <- plot_surv_curves(fit=survFits[[dName]],
                           data=survFitDatas[[dName]],
                           zoom = zoomF,
                           addZoomFrom = Inf, #dont add zoom
                           counterLabs= counterLabs)
```

```
#180x254mm
plotSub <- paste0("Males, ", cohort)
patchwork::wrap_plots(plotsM[c(1,2,5:8)], ncol=2) +
  patchwork::plot_annotation(subtitle = plotSub)
```

Males, 1905−45

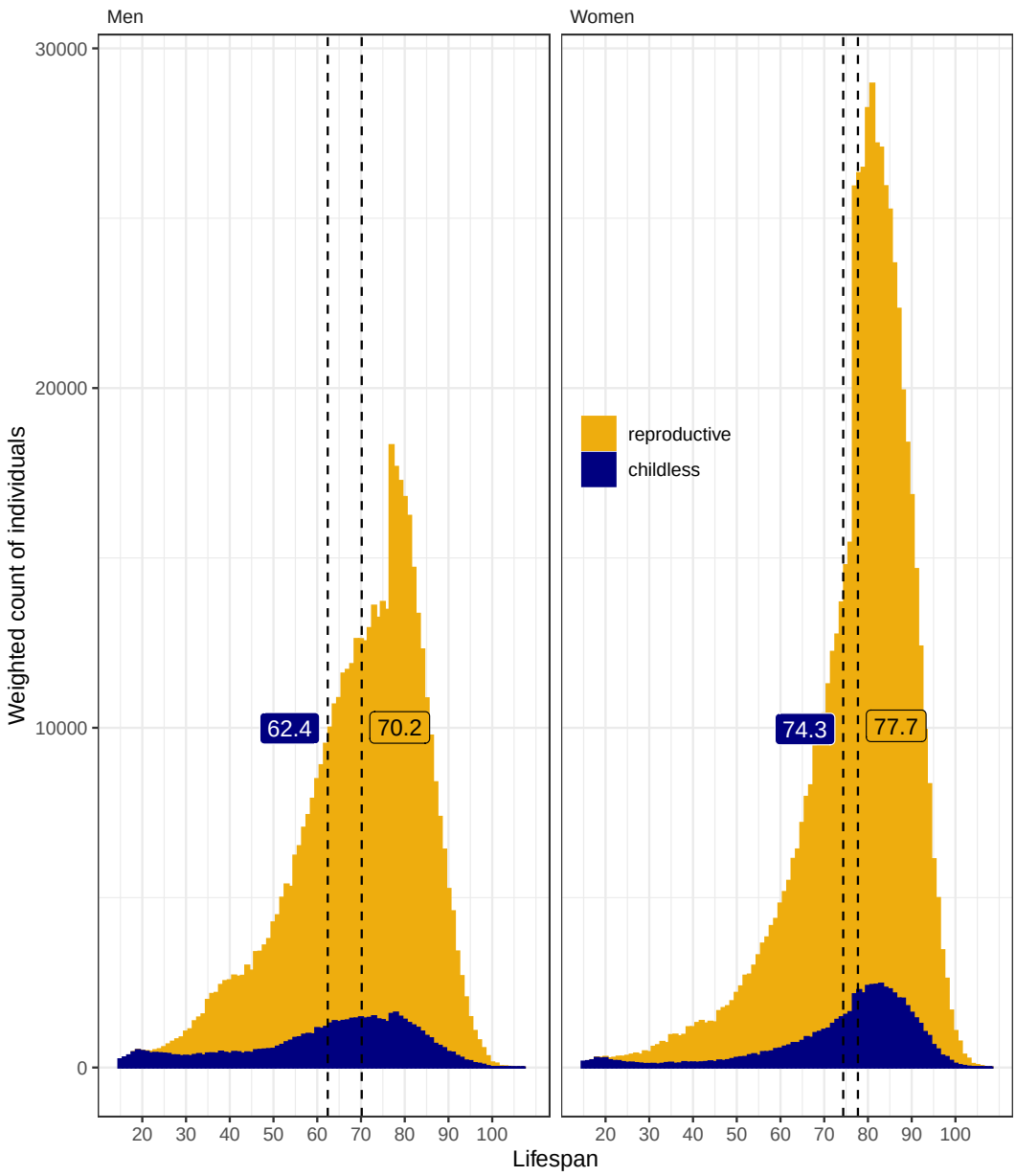


Natural selection on lifespan

```
colors <- c('darkgoldenrod2','navy')


theme  <-  ggplot2::theme_bw() +
  ggplot2::theme(legend.position = c(0.61,0.62),
                 strip.background = element_blank(),
                 legend.background = element_blank(),
                 legend.title = element_blank(),
                 strip.text = element_text(hjust = 0))
fillLabs <- c("reproductive","childless")
textColors  <- c("black","black","white","white")
```

```
#set up childless and reproductive groups
data2use$childless <- data2use$children == 0
data2use$weight <- data2use$children + data2use$childless
# Calculate the weighted sums and the sum of weights for each group
data2use$weighted_sums <- data2use$lifespan2022 * data2use$weight
sums <- aggregate(weighted_sums ~ sex + childless, data = data2use, sum)
weights <- aggregate(weight ~ sex + childless, data = data2use, sum)

# Calculate the weighted mean for each group
meansWeighted <- data.frame(sex = sums$sex,
                            childless = sums$childless,
                            lifespan2022 = sums$weighted_sums/weights$weight)
meansWeighted$children  <-1
```

```
margin <- theme(plot.margin = unit(c(0.2,0.2,3,0.2), "in"))
pWeighted <- ggplot(data2use,
                    aes(x=lifespan2022,
                        fill=childless,
                        color=childless,
                        weight=children + childless)) +
  geom_bar() +
  facet_wrap(~ifelse(sex=="F","Women","Men")) +
  geom_vline(data = meansWeighted,
             aes(xintercept=lifespan2022, color=childless),
             linetype = "dashed", color="black") +
  ggrepel::geom_label_repel(data = meansWeighted,
                            aes(label=round(lifespan2022,1),
                                x=lifespan2022, y=5000*2),
                            color=textColors, show.legend = F) +
  scale_color_manual(values = colors, labels=fillLabs) +
  scale_fill_manual(values = colors, labels=fillLabs)+
  scale_x_continuous(breaks = seq(20, 100, by=10))+
  labs(y="Weighted count of individuals", x="Lifespan") +
theme + margin

pWeighted
```


Men
Women
30000
20000
10000
0
Weighted count of individuals
reproductive
childless
62.4
70.2
74.3
77.7
20
30
40
50
60
70
80
90
100
Lifespan

```
#save the weighted plot with bottom margin extended to pdf
fname <- paste0(rdir, "weighted_LRS_by_sex_",cohort,".pdf" )
sizeFactor <- 1
ggplot2::ggsave(fname, pWeighted,
                width=7.08*sizeFactor,
                height=11*sizeFactor)
```

Regression models by Sex

Females

```
# select sex and cohort
selectedSex <- "F"
data2useSex <- data2use[data2use$sex == selectedSex, ]
```

```
#absolute LRS
modelLRS <- lm(children ~ lifespan2022 + I(lifespan2022*lifespan2022),
               data2useSex)
cohort_tbls_results[[paste0("LRS_",selectedSex)]] <-
lmSummary(modelLRS, "LRS", c(sex=selectedSex))
pander::pander(summary(modelLRS))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2092	0.03931	-5.323	1.022e-07
lifespan2022	0.05526	0.001179	46.88	0
I(lifespan2022 * lifespan2022)	-0.0003693	8.719e-06	-42.36	0

Table 7: Fitting linear model: children ~ lifespan2022 + I(lifespan2022 * lifespan2022)

Observations	Residual Std. Error	R^2	Adjusted R^2
316916	1.537	0.008516	0.008509

```
#relative LRS
modelLRS <- lm(rel_children ~ lifespan2022 + I(lifespan2022*lifespan2022),
               data2useSex)
cohort_tbls_results[[paste0("relLRS_",selectedSex)]] <-
 lmSummary(modelLRS, "relative LRS", c(sex=selectedSex))
pander::pander(summary(modelLRS))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1416	0.02189	-6.467	1.002e-10
lifespan2022	0.03173	0.0006566	48.33	0
I(lifespan2022 * lifespan2022)	-0.0002128	4.857e-06	-43.81	0

Table 9: Fitting linear model: rel_children ~ lifespan2022 + I(lifespan2022 * lifespan2022) ### Males

Observations	Residual Std. Error	R^2	Adjusted R^2
316916	0.8563	0.008928	0.008922

```
# select sex and cohort
selectedSex <- "M"
data2useSex <- data2use[data2use$sex == selectedSex, ]
```

```
#absolute LRS
modelLRS <- lm(children ~ lifespan2022 + I(lifespan2022*lifespan2022),
               data2useSex)
cohort_tbls_results[[paste0("LRS_",selectedSex)]] <-
lmSummary(modelLRS, "LRS", c(sex=selectedSex))
pander::pander(summary(modelLRS))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.7334	0.03111	-23.57	1.053e-122
lifespan2022	0.06716	0.001036	64.83	0
I(lifespan2022 * lifespan2022)	-0.0004384	8.357e-06	-52.46	0

Table 11: Fitting linear model: children ~ lifespan2022 + I(lifespan2022 * lifespan2022)

Observations	Residual Std. Error	R^2	Adjusted R^2
262355	1.451	0.03235	0.03235

```
#relative LRS
modelLRS <- lm(rel_children ~ lifespan2022 + I(lifespan2022*lifespan2022),
               data2useSex)
cohort_tbls_results[[paste0("relLRS_",selectedSex)]] <-
 lmSummary(modelLRS, "relative LRS", c(sex=selectedSex))
pander::pander(summary(modelLRS))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4501	0.01823	-24.69	1.762e-134
lifespan2022	0.04027	0.0006068	66.37	0
I(lifespan2022 * lifespan2022)	-0.0002638	4.895e-06	-53.9	0

Table 13: Fitting linear model: rel_children ~ lifespan2022 + I(lifespan2022 * lifespan2022)

Observations	Residual Std. Error	R^2	Adjusted R^2
262355	0.8498	0.03329	0.03328

Export tables

This Exports the generated tables to Excel

```
#export to excel
writexl::write_xlsx(cohort_tbls_results,
                    paste0(rdir, "S01_summary_tbls_",cohort,".xlsx"))
```

```
#print R and package versions
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=Estonian_Estonia.utf8  LC_CTYPE=Estonian_Estonia.utf8
## [3] LC_MONETARY=Estonian_Estonia.utf8 LC_NUMERIC=C
## [5] LC_TIME=Estonian_Estonia.utf8
##
## time zone: Europe/Tallinn
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.4.4
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    tidyr_1.3.0       rstatix_0.7.2
##  [5] lattice_0.22-5    digest_0.6.33     magrittr_2.0.3    evaluate_0.23
##  [9] grid_4.3.2        fastmap_1.1.1     Matrix_1.6-3      writexl_1.4.2
## [13] ggrepel_0.9.4     backports_1.4.1   survival_3.5-7    gridExtra_2.3
## [17] pander_0.6.5      purrr_1.0.2       fansi_1.0.5       scales_1.2.1
## [21] textshaping_0.3.7 abind_1.4-5       cli_3.6.1         KMsurv_0.1-5
## [25] rlang_1.1.2       splines_4.3.2     munsell_0.5.0     withr_2.5.2
## [29] yaml_2.3.7        tools_4.3.2       ggsignif_0.6.4    dplyr_1.1.4
## [33] colorspace_2.1-0  ggpubr_0.6.0      km.ci_0.5-6       broom_1.0.5
## [37] vctrs_0.6.4       R6_2.5.1          zoo_1.8-12        lifecycle_1.0.4
## [41] car_3.1-2         ragg_1.2.6        pkgconfig_2.0.3   survminer_0.4.9
## [45] pillar_1.9.0      gtable_0.3.4      Rcpp_1.0.11       data.table_1.14.8
## [49] glue_1.6.2        systemfonts_1.0.5 xfun_0.41         tibble_3.2.1
```

```
## [53] tidyselect_1.2.0  highr_0.10        rstudioapi_0.15.0 knitr_1.45
## [57] xtable_1.8-4      farver_2.1.1      survMisc_0.5.6    patchwork_1.1.3
## [61] htmltools_0.5.7   rmarkdown_2.25    carData_3.0-5     labeling_0.4.3
## [65] compiler_4.3.2
```

```
#END
```