Survival Analysis of Breastfeeding

STAT 639V Survival Analysis Final Project

Carson Stacy

Introduction

Breastfeeding has been a topic of academic and public interest since the invention of formula and the feeding bottle in the 19th century [cite]. Since the invention of baby formula, breastfeeding rates have reduced dramatically. The breastfeeding rate was 90% in the 20th century, but has decreased to approximately 37% in the 21st century (Gaynor, 2003; Victorial *et al.*, 2016). This trend has many scientists concerned [McFadden *et al.*, 2016; Pomeranz *et al.*, 2021; Pérez-Escamilla, 2020].

The importance of breastfeeding in low and middle-income nations is widely acknowledged. In low-income nations, unclean water in formula is a death sentence for an infant. Perhaps this partially explains why the prevalence of breastfeeding is higher in low and middle-income nations than in high-income nations. A large body of scientific research spanning public health studies to cell biology experiments show the importance of promoting infant breastfeeding everywhere []. From kick-starting the infant's gut microbiome via human milk oligosaccharides, to the transfer of important immune molecules (e.g. IgA) to transfer of stem cells (Hassiotou *et al.*, 2014) and micro-RNAs from mother to infant suggested to regulate infant gene expression (Munch *et al.*, 2013). Beyond positive impacts on the child's current and future health, benefits have been shown for the breast feeding mother as well [cite]. From the World Health Organization to the American Academy of Pediatrics, most doctors and organizations avidly support exclusive breastfeeding during the first six month of an infant's life.

It is apparent that breastfeeding is important for health – or is it? Despite the ubiquity of recommendations regarding breastfeeding, there is less high quality data on the topic than might be expected. Ethical considerations make the gold standard double-blind experimental design a non-starter, so observational studies and their confounding baggage are the norm in breastfeeding literature. A hallmark study in the 1990's in Belarus called the PROBIT trial involved 17,000 mothers which were experimentally "treated" with promotion of breastfeeding while the control group was not [cite]. The results of this trial were mixed. In the context of immediate health benefits of the child, breast feeding showed a significant reduction in: number of gastrointestinal infections, likelihood of eczema and other rashes. However, no significant differences were seen in any other considered outcomes (e.g., respiratory infections, ear infections, wheezing, mortality). Regarding long-term outcomes, the PROBIT trial found no effect on any long-term outcomes measured. Sibling studies, which compare outcomes of siblings pairs where one was breastfed while the other bottle fed, find no impact on any measured outcomes [wu 2018, "is breast truly best" 2014].

It has been argued that the differences seen in many observational studies comparing breast and bottle fed infants are the result of maternal selection. In other words, mothers are not deciding randomly whether to feed their infants with breast or bottle. In the US, mothers who breastfeed tend to be more highly educated and wealthier than mothers who bottle feed. A recent study suggests

> *"... most physical health benefits associated with breastfeeding are likely attributable to demographic characteristics such as race and socioeconomic status, and other difficult to measure unobservable characteristics."* - (Raissan and Su, 2018)

The controversy is not against breastfeeding, especially in low-income nations, rather it is promoting communication evidence-based of the magnitude of benefits of breastfeeding.

It is in the context of thinking about a mother's breastfeeding decisions through a socioeconomic lens that this project examines time to cessation of breastfeeding data of new mothers from the National Longitudinal Survey of Youth (NSLY, 1995). A finding that demographic factors have no effect on time to cessation of breastfeeding would be unexpected based on Raissan and Su's claims. A finding of significant differences does not confirm their assertions, but rather provides valuable information about relevant demographic variables related to breast feeding cessation and context for considering some observational research finding drastic benefits of breast feeding.

Data Collection and Experimental Design

This project utilizes data on breastfeeding decisions of young mothers compiled from the National Longitudinal Survey of Youth (NLSY, 1995) personal interviews. The data set was compiled as part of the 1997 text *Survival Analysis Techniques for Censored and truncated data* by Klein and Moeschberger, available in the `KMsurv` package as `bfeed`.

```
library(tidyverse) # package for data analysis
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

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

```
library(ggsci) # for graph colors
library(KMsurv) # package with data set
library(survival) # package for survival analysis
library(survminer) # for ggsurvplot viz
```

```
## Loading required package: ggpubr

##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

```
data(bfeed) # load data set into R
```

Variables

The data is comprised of 927 new mothers, with 10 variables recorded for each individual. Descriptions for each variable recorded can be seen in Table 1 below.

```r
var_names <- tibble(
  c(1:ncol(bfeed)),
  c(colnames(bfeed)),
  c("duration of breastfeeding (weeks)",
    "indicator of child weaning",
    "race of mother",
    "mother in poverty",
    "mother smoked at birth of child",
    "mother used alcohol at birth of child",
    "age of mother at birth of child",
    "year of birth",
    "education level of mother (years of school)",
    "prenatal care after 3rd month"
    )
)
knitr::kable(var_names,
             col.names = c("No.", "Variable ID", "Variable definition"),
             caption = "Table 1: List of variable IDs and their definitions"
             )
```

Table 1: Table 1: List of variable IDs and their definitions

No.	Variable ID	Variable definition
1	duration	duration of breastfeeding (weeks)
2	delta	indicator of child weaning
3	race	race of mother
4	poverty	mother in poverty
5	smoke	mother smoked at birth of child
6	alcohol	mother used alcohol at birth of child
7	agemth	age of mother at birth of child
8	ybirth	year of birth
9	yschool	education level of mother (years of school)
10	pc3mth	prenatal care after 3rd month

Censoring and Missing Values

Methods and Data Analysis

(e.g., assumptions, statistical models, parameter estimations, goodness-of-fit test, cross-group comparisons, predictions and validations, etc)

Data Exploration

```r
# raw data processing
bfeed <- bfeed %>%
  mutate(race = factor(
             recode(race, "1" = "white", "2" = "black", "3"="other"),
             levels = c("white", "black", "other")
             ),
```

```
        delta = factor(
             recode(delta, "1" = "yes", "0" = "no"),
             levels = c("no", "yes")
             ),
        poverty = factor(
             recode(poverty, "1" = "yes", "0" = "no"),
             levels = c("no", "yes")
             ),
        smoke = factor(
             recode(smoke, "1" = "yes", "0" = "no"),
             levels = c("no", "yes")
             ),
        alcohol = factor(
             recode(alcohol, "1" = "yes", "0" = "no"),
             levels = c("no", "yes")
             ),
        pc3mth = factor(
             recode(pc3mth, "1" = "yes", "0" = "no"),
             levels = c("no", "yes")
             ),
        #convert years of education to an ordered categorical variable
        education = factor(
          recode(
               cut(yschool,
                     breaks = c(0,11.5, 12.5, max(yschool)),
                        labels = F
                   ),
              "1" = "<HS", "2" = "HS",
                           "3" = ">HS"),
          levels = c("<HS", "HS", ">HS")

          ),
        ybirth = ybirth + 1900 , # making data more exact for visualization
        SurvObj = Surv(duration, delta == "yes") #add Survival Obj variable

        )

str(bfeed)
```

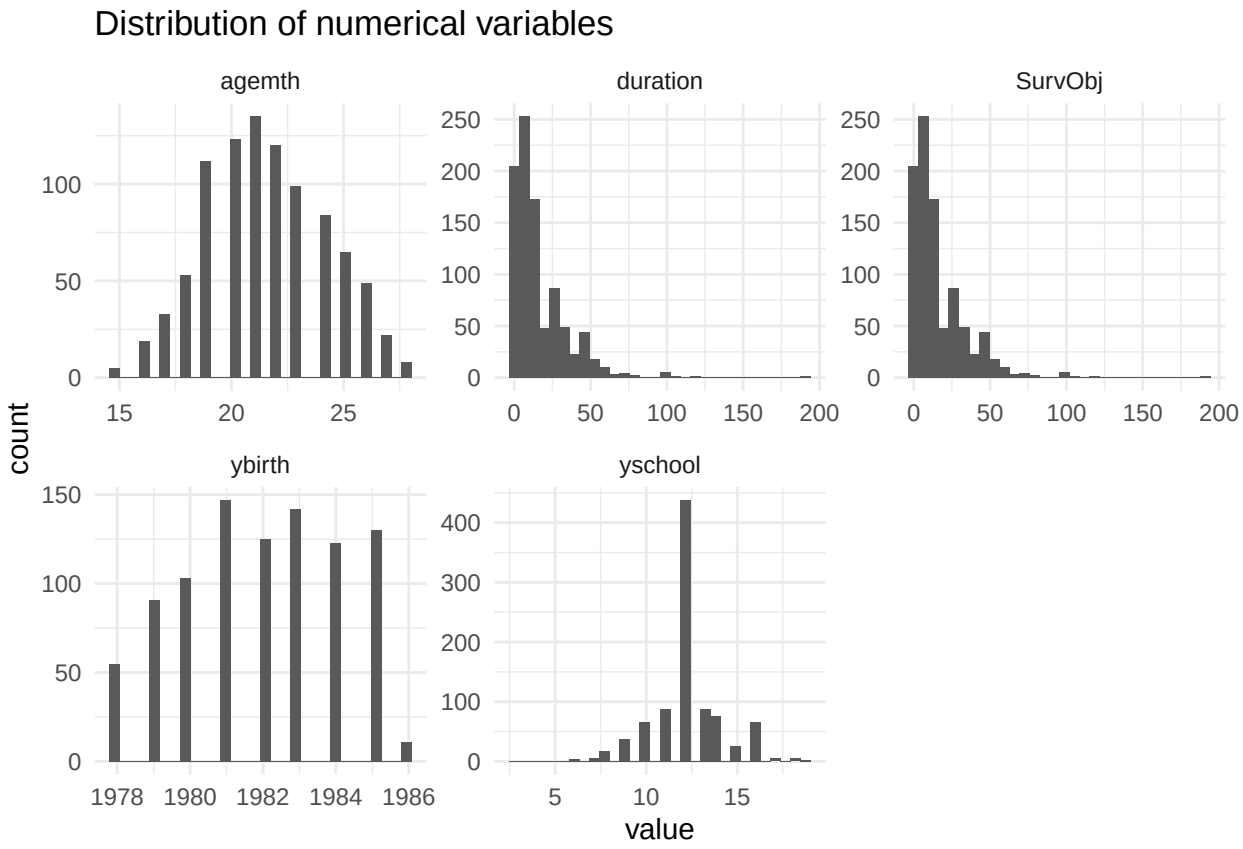
```
## 'data.frame':    927 obs. of  12 variables:
##  $ duration : int  16 1 4 3 36 36 16 8 20 44 ...
##  $ delta    : Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 1 2 2 ...
##  $ race     : Factor w/ 3 levels "white","black",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ poverty  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 2 1 2 1 ...
##  $ smoke    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 1 2 2 1 1 ...
##  $ alcohol  : Factor w/ 2 levels "no","yes": 2 1 1 2 1 1 1 1 1 1 ...
##  $ agemth   : int  24 26 25 21 22 18 20 24 24 24 ...
##  $ ybirth   : num  1982 1985 1985 1985 1982 ...
##  $ yschool  : int  14 12 12 9 12 11 9 12 12 14 ...
##  $ pc3mth   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ education: Factor w/ 3 levels "<HS","HS",">HS": 3 2 2 1 2 1 1 2 2 3 ...
##  $ SurvObj  : 'Surv' num [1:927, 1:2]  16    1    4+   3   36   36   16    8+  20   44  ...
##   ..- attr(*, "dimnames")=List of 2
```

```
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "time" "status"
##   ..- attr(*, "type")= chr "right"
```

Exploration of the data

```
bfeed %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value))+
  geom_histogram(bins = 30)+
  facet_wrap(~key, scales = "free") +
  theme_minimal() +
  scale_color_npg() +
  ggtitle("Distribution of numerical variables");
```

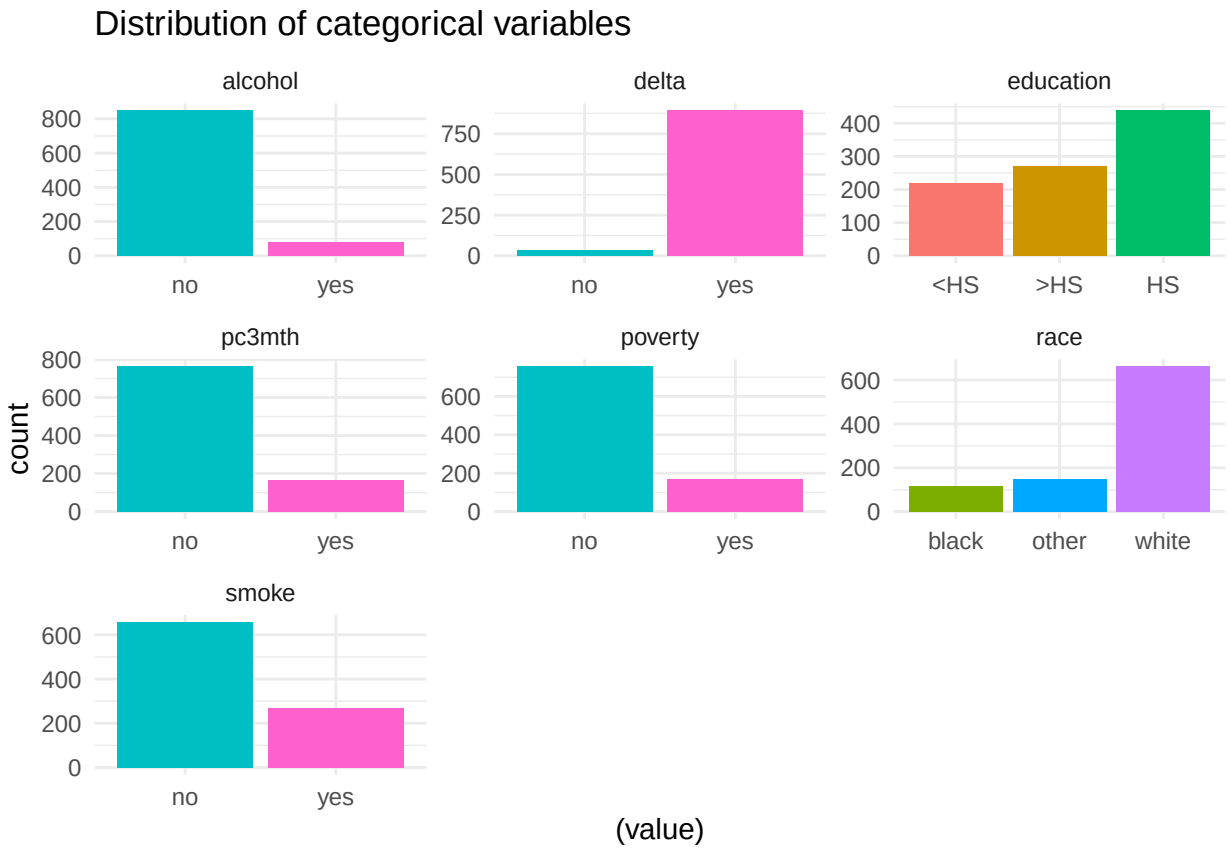
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```




```
bfeed %>%
  select_if(negate(is.numeric)) %>%
  gather() %>%
  ggplot(aes((value), fill = value))+
  geom_bar()+
```

```
  facet_wrap(~key, scales = "free")+
  theme_minimal() +
  # scale_fill_uchicago() +
  theme(legend.position = "none") +
  ggtitle("Distribution of categorical variables")
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



Assumptions

Given that (1) earlier ages of stopping breastfeeding are often considered more important and (2) ages of stopping breastfeeding likely are right skewed, the peto-peto

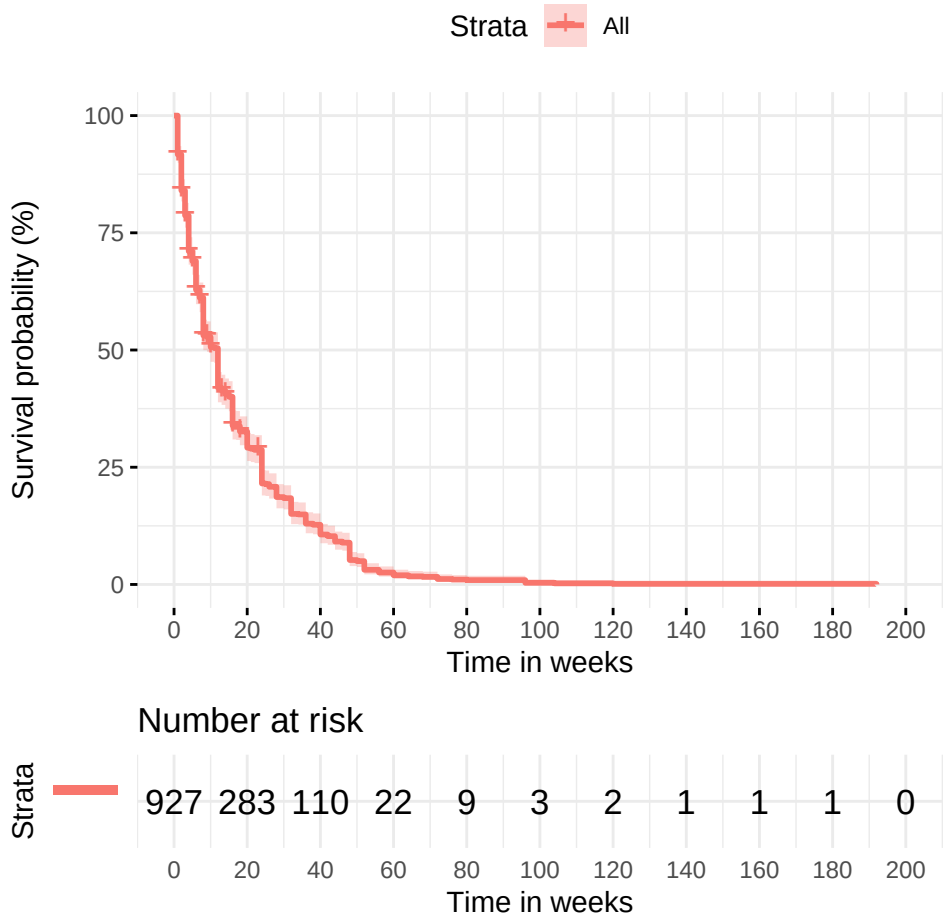
Statistical Models

Cross-Group Comparisons

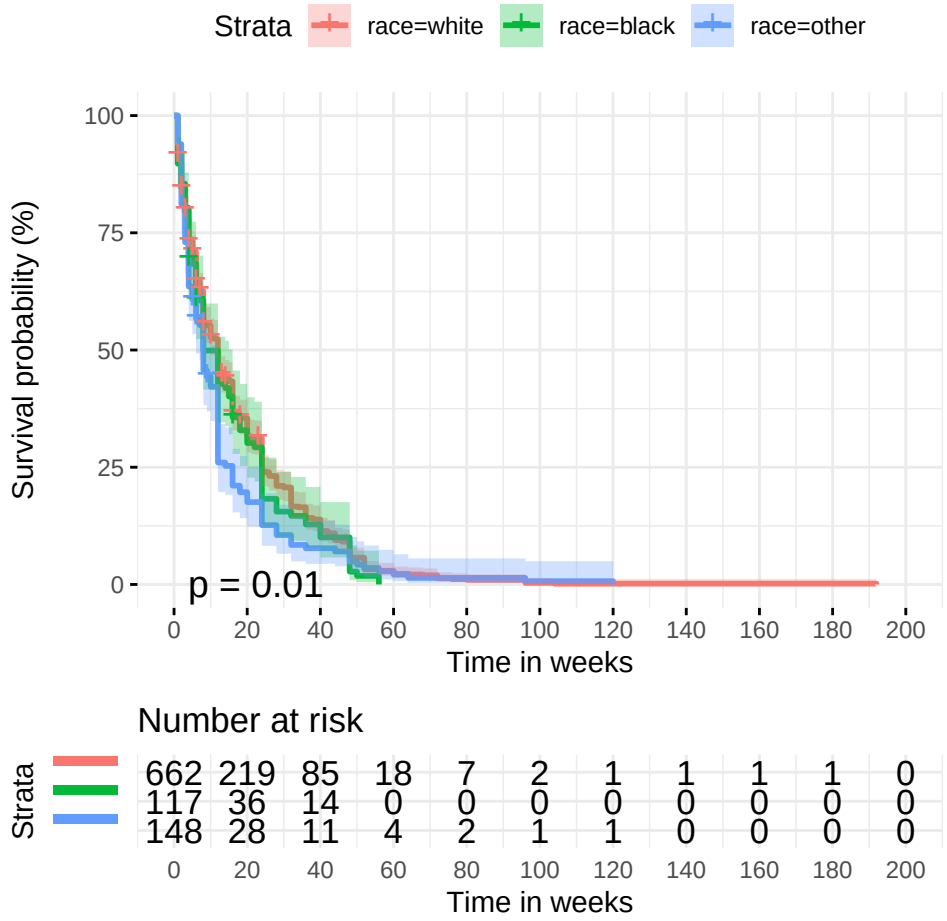
Kaplan Meier Curve of entire dataset combined:

```
# create survival object:
km.as.one <- survfit(SurvObj ~ 1, data = bfeed)
```

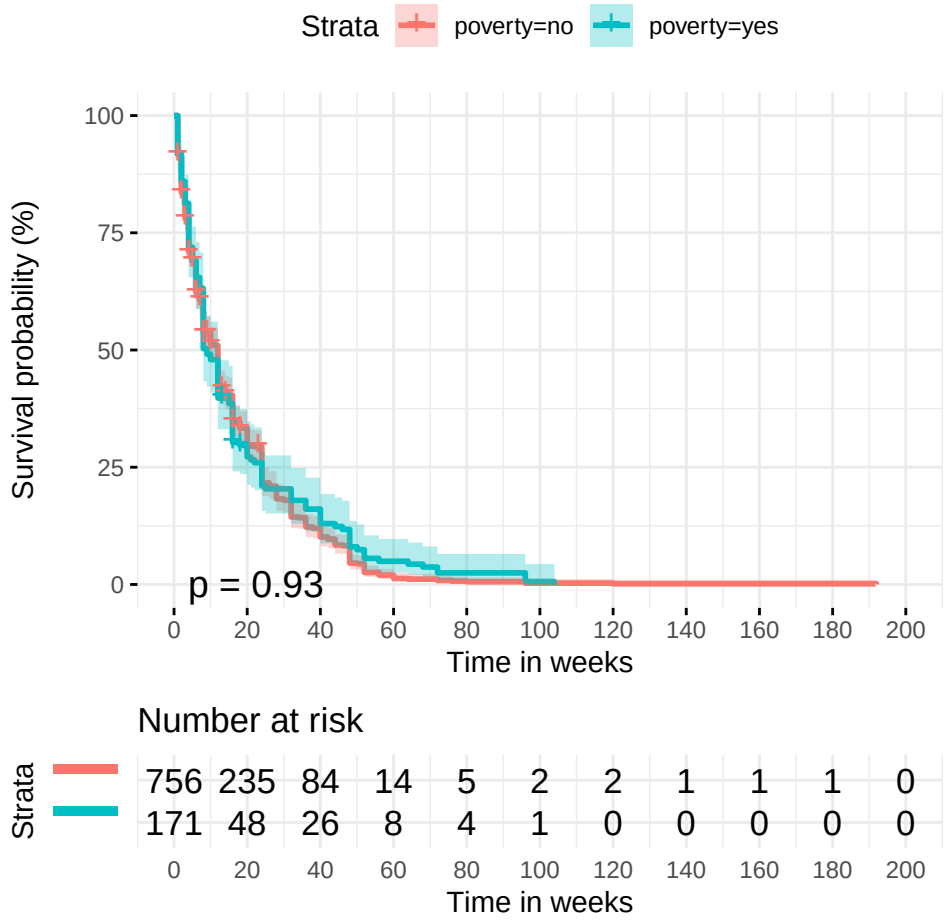
```r
# summary(km.as.one)

km.by.race <- survfit(SurvObj ~ race, data = bfeed)

km.by.poverty <- survfit(SurvObj ~ poverty, data = bfeed)

km.by.education <- survfit(SurvObj ~ education, data = bfeed)

km.by.smoke <- survfit(SurvObj ~ smoke, data = bfeed)

km.by.alcohol <- survfit(SurvObj ~ alcohol, data = bfeed)

km.by.agemth <- survfit(SurvObj ~ agemth, data = bfeed)

km.by.pc3mth <- survfit(SurvObj ~ pc3mth, data = bfeed)

#KM plot combining all participants
ggsurvplot(
   km.as.one,                # survfit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = TRUE,        # show risk table.
   #pval = TRUE,              # show p-value of log-rank test.
   #conf.int = TRUE,          # show confidence intervals for
                             # point estimates of survival curves.
   xlim = c(0,200),          # present narrower X axis, but not affect
                             # survival estimates.
   xlab = "Time in weeks",   # customize X axis label.
   break.time.by = 20,       # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                             # in legend of risk table
  # palette = "uchicago",      # change colors to be pretty
  log.rank.weights = "S1",   # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,      # p-val text size
  fun = "pct"                #show survival function as percentage
)
```



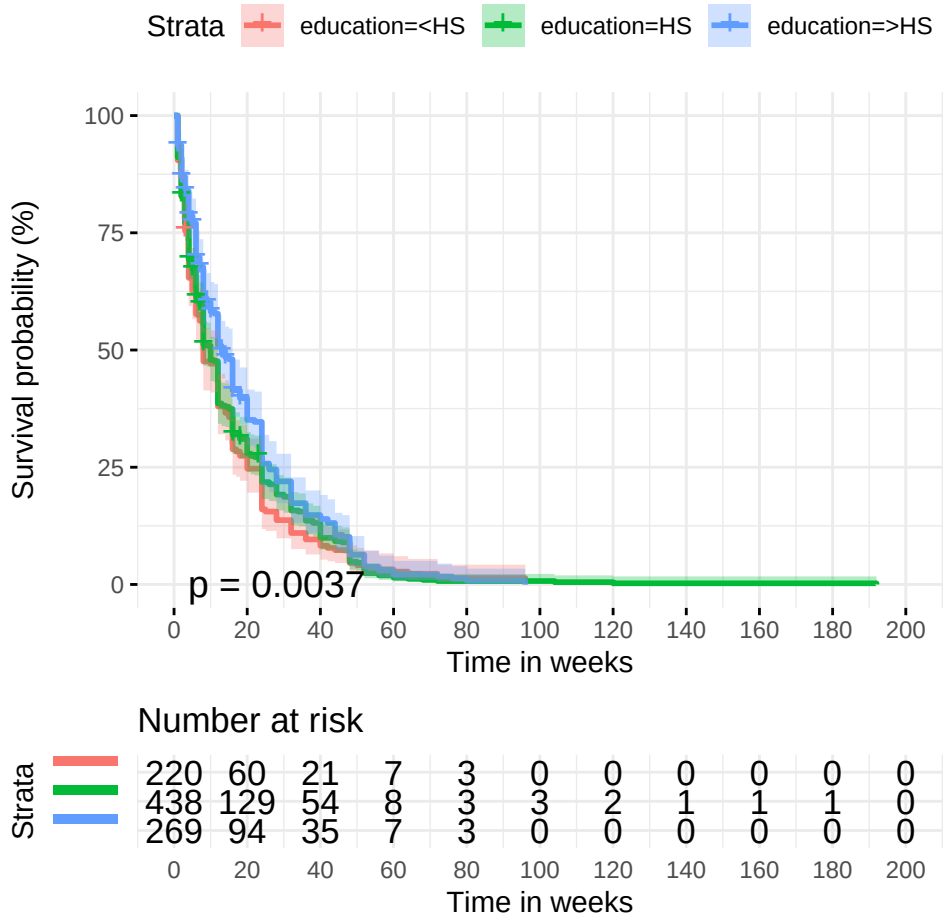

```
#KM curve according to race
ggsurvplot(
   km.by.race,                 # survfit object with calculated statistics.
   data = bfeed,               # data used to fit survival curves.
   risk.table = TRUE,          # show risk table.
   pval = TRUE,                # show p-value of log-rank test.
   conf.int = TRUE,            # show confidence intervals for
                               # point estimates of survival curves.
   xlim = c(0,200),            # present narrower X axis, but not affect
                               # survival estimates.
   xlab = "Time in weeks",     # customize X axis label.
   break.time.by = 20,         # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                               # in legend of risk table
  # palette = "uchicago",       # change colors to be pretty
  log.rank.weights = "S1",    # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,       # p-val text size
  fun = "pct"                 #show survival function as percentage
)
```



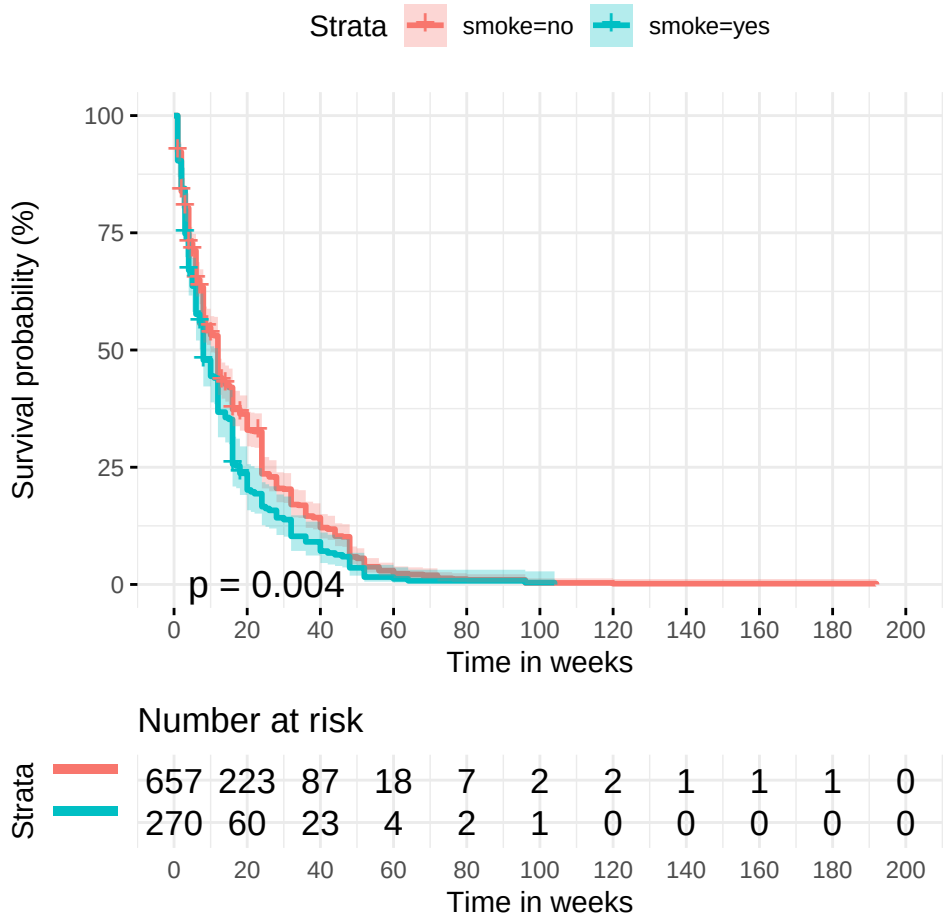

```
#KM curve according to poverty
ggsurvplot(
   km.by.poverty,                # survfit object with calculated statistics.
   data = bfeed,              # data used to fit survival curves.
   risk.table = TRUE,         # show risk table.
   pval = TRUE,               # show p-value of log-rank test.
   conf.int = TRUE,           # show confidence intervals for
                              # point estimates of survival curves.
   xlim = c(0,200),           # present narrower X axis, but not affect
                              # survival estimates.
   xlab = "Time in weeks",    # customize X axis label.
   break.time.by = 20,        # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                              # in legend of risk table
  # palette = "uchicago",       # change colors to be pretty
  log.rank.weights = "S1",   # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,      # p-val text size
  fun = "pct"                #show survival function as percentage
)
```



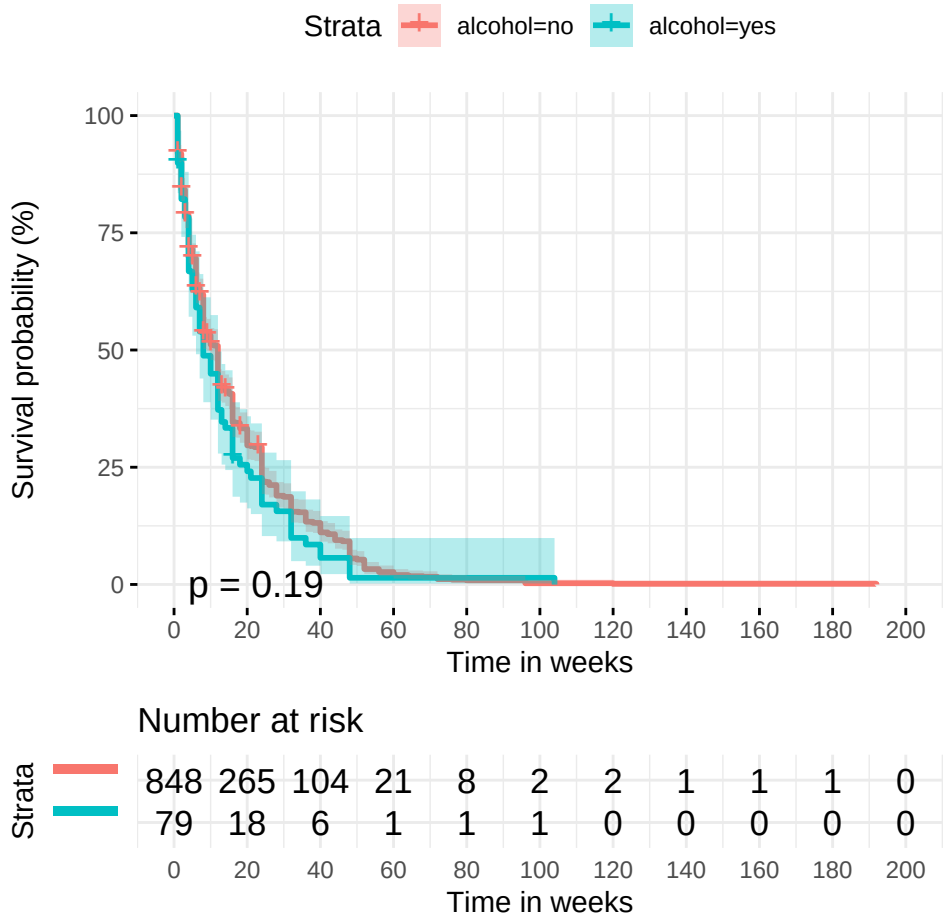

```
#KM curve according to education
ggsurvplot(
   km.by.education,                  # survfit object with calculated statistics.
   data = bfeed,                  # data used to fit survival curves.
   risk.table = TRUE,             # show risk table.
   pval = TRUE,                   # show p-value of log-rank test.
   conf.int = TRUE,               # show confidence intervals for
                                  # point estimates of survival curves.
   xlim = c(0,200),               # present narrower X axis, but not affect
                                  # survival estimates.
   xlab = "Time in weeks",        # customize X axis label.
   break.time.by = 20,            # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                                  # in legend of risk table
  # palette = "uchicago",          # change colors to be pretty
  log.rank.weights = "S1",       # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,          # p-val text size
  fun = "pct"                    #show survival function as percentage
)
```



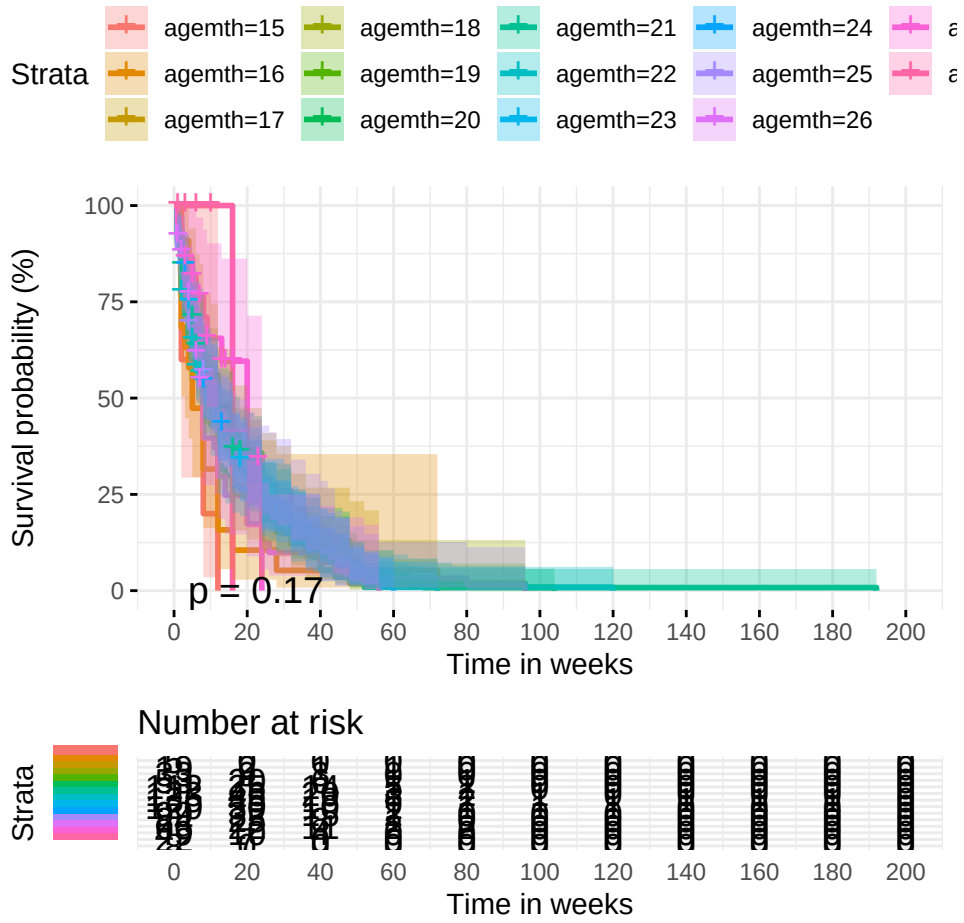

```
#KM curve according to smoking
ggsurvplot(
   km.by.smoke,                 # survfit object with calculated statistics.
   data = bfeed,               # data used to fit survival curves.
   risk.table = TRUE,          # show risk table.
   pval = TRUE,                # show p-value of log-rank test.
   conf.int = TRUE,            # show confidence intervals for
                               # point estimates of survival curves.
   xlim = c(0,200),            # present narrower X axis, but not affect
                               # survival estimates.
   xlab = "Time in weeks",     # customize X axis label.
   break.time.by = 20,         # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                              # in legend of risk table
  # palette = "uchicago",       # change colors to be pretty
  log.rank.weights = "S1",    # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,       # p-val text size
  fun = "pct"                 #show survival function as percentage
)
```



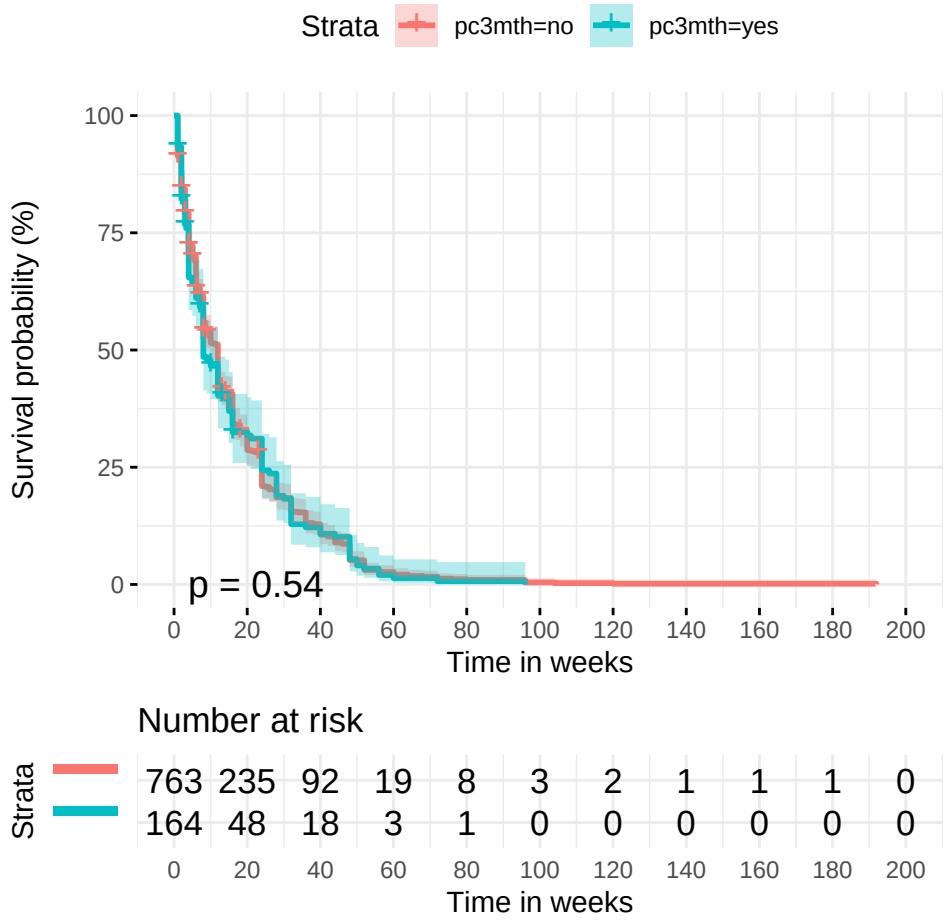

```r
#KM curve according to alcohol
ggsurvplot(
   km.by.alcohol,                 # survfit object with calculated statistics.
   data = bfeed,                # data used to fit survival curves.
   risk.table = TRUE,           # show risk table.
   pval = TRUE,                 # show p-value of log-rank test.
   conf.int = TRUE,             # show confidence intervals for
                                # point estimates of survival curves.
   xlim = c(0,200),             # present narrower X axis, but not affect
                                # survival estimates.
   xlab = "Time in weeks",      # customize X axis label.
   break.time.by = 20,          # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                                # in legend of risk table
  # palette = "uchicago",        # change colors to be pretty
  log.rank.weights = "S1",     # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,        # p-val text size
  fun = "pct"                  #show survival function as percentage
)
```




```
#KM curve according to age of mother at birth of child
ggsurvplot(
   km.by.agemth,                 # survfit object with calculated statistics.
   data = bfeed,                # data used to fit survival curves.
   risk.table = TRUE,           # show risk table.
   pval = TRUE,                 # show p-value of log-rank test.
   conf.int = TRUE,             # show confidence intervals for
                                # point estimates of survival curves.
   xlim = c(0,200),             # present narrower X axis, but not affect
                                # survival estimates.
   xlab = "Time in weeks",      # customize X axis label.
   break.time.by = 20,          # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                                # in legend of risk table
  # palette = "uchicago",        # change colors to be pretty
  log.rank.weights = "S1",     # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,        # p-val text size
  fun = "pct"                  #show survival function as percentage
)
```




```
#KM curve according to prenatal care after 3rd month
print(ggsurvplot(
   km.by.pc3mth,                 # survfit object with calculated statistics.
   data = bfeed,               # data used to fit survival curves.
   risk.table = TRUE,          # show risk table.
   pval = TRUE,                # show p-value of log-rank test.
   conf.int = TRUE,            # show confidence intervals for
                               # point estimates of survival curves.
   xlim = c(0,200),            # present narrower X axis, but not affect
                               # survival estimates.
   xlab = "Time in weeks",     # customize X axis label.
   break.time.by = 20,         # break X axis in time intervals by 500.
   ggtheme = theme_minimal(),# customize plot and risk table with a theme.
  risk.table.y.text.col = T, # colour risk table text annotations.
  risk.table.y.text = FALSE,  # show bars instead of names in text annotations
                              # in legend of risk table
  # palette = "uchicago",       # change colors to be pretty
  log.rank.weights = "S1",    # Peto Peto test for log-rank test
  pval.method.coord = c(150, 80), # location of p-value text
  pval.method.size = 3,       # p-val text size
  fun = "pct"                 #show survival function as percentage
)
)
```



Goodness-of-Fit Test

```
 # log rank test survdiff
survdiff(SurvObj ~ race, bfeed)
```

```
## Call:
## survdiff(formula = SurvObj ~ race, data = bfeed)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## race=white 662      634      664     1.397     6.306
## race=black 117      113      108     0.247     0.324
## race=other 148      145      120     5.347     7.128
##
##  Chisq= 8.1  on 2 degrees of freedom, p= 0.02
```

Discussion

(e.g., strengths and shortcomings of your model, and possible improvements)

Methods and data analysis

Conclusion

References

Gaynor G. Breastfeeding advocacy. Maine Nurse. 2003;5(2):13.

Victora, C.G., Bahl, R., Barros, A.J., França, G.V., Horton, S., Krasevec, J., Murch, S., Sankar, M.J., Walker, N., Rollins, N.C. and Group, T.L.B.S., 2016. Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. The Lancet, 387(10017), pp.475-490.

McFadden, A., Mason, F., Baker, J., Begin, F., Dykes, F., Grummer-Strawn, L., Kenney-Muir, N., Whitford, H., Zehner, E. and Renfrew, M.J., 2016. Spotlight on infant formula: coordinated global action needed. *The Lancet*, *387*(10017), pp.413-415.

Pérez-Escamilla, R., 2020. Breastfeeding in the 21st century: How we can make it work. *Social Science & Medicine*, *244*, p.112331.

Pomeranz, J. L., Chu, X., Groza, O., Cohodes, M., & Harris, J. L. (2021). Breastmilk or infant formula? Content analysis of infant feeding advice on breastmilk substitute manufacturer websites. *Public Health Nutrition*, 1-9.

EM Munch, RA Harris, M Mohammad, *et al.***Transcriptome profiling of microRNA by Next-Gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk**

PLoS One, 8 (2013), p. e50564

F Hassiotou, PE Hartmann**At the dawn of a new discovery: the potential of breast milk stem cells**

Adv Nutr, 5 (2014), pp. 770-778

Is breast truly best? Estimating the effects of breastfeeding on long-term child health and wellbeing in the United States using sibling comparisons

Klein and Moeschberger (1997) *Survival Analysis Techniques for Censored and truncated data*, Springer. *National Longitudinal Survey of Youth Handbook*The Ohio State University, 1995.

<!–# It is in this context

The prevalence of breastfeeding behaviors is lower

Researchers and public health organizations have touted the benefits of breastfeeding for infants [cite][][].

Social determinants of health have profound impacts on life outcomes.

There is a preponderance of evidence that childhood experiences significantly impact life trajectories. Improving child health has been a top priority for the World Health Organization for decades [cite].

Previous research has shown that the age at which a child stops breastfeeding has significant effects on later development [CITE]. An increased understanding of how to –>