# Survival Analysis of Breastfeeding Cessation

STAT 639V Survival Analysis Final Project

Carson Stacy

2021-12-16

# Contents

# Introduction

Breastfeeding has been a topic of academic and public interest since the invention of formula and the feeding bottle in the 19th century (Stevens, Patrick, and Pickler 2009). 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; Victora et al. 2016). This trend has many scientists concerned (McFadden et al. 2016; Pérez-Escamilla 2020; Pomeranz et al. 2021).

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 (Pérez-Escamilla et al. 2012). 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 (Hoddinott, Tappin, and Wright 2008; Walters, Phan, and Mathisen 2019). 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 (Pannaraj et al. 2017; Hassiotou and Hartmann 2014) and micro-RNAs from mother to infant suggested to regulate infant gene expression (Munch et al. 2013; Esch et al. 2020). Beyond positive impacts on the child's current and future health, benefits have been shown for the breast feeding mother as well (León-Cava et al. 2002). From the World Health Organization to the American Academy of Pediatrics (Eidelman et al. 2012), 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 exists 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 (Kramer et al. 2001). 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 (Colen and Ramey 2014; Raissian and Su 2018).

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 the claims of Raissian and Su (2018). 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. Additionally, this analysis shows the utility of sfurvival analysis methodology in time-to-event scenarios such as breast feeding cessation.

## The Data

This project utilizes data on breastfeeding decisions of young mothers compiled from the National Longitudinal Survey of Youth (NLSY, 1995) personal interviews conducted by the United States Bureau of Labor Statistics branch of the US Department of Labor. All NLSY files are public access, and can be downloaded from http://www.bls.gov/nls/nlsy79.html. The data set was compiled as part of the text *Survival Analysis Techniques for Censored and truncated data* by Klein and Moeschberger (2003), available in the `KMsurv` package as `bfeed`.

### Variables

The data is comprised of data from 927 new mothers, with 10 variables recorded for each individual. Descriptions for each variable recorded can be seen in Table 1 below. There are six categorical variables, of which only `race` of mother has more than two categories. There are 4 numerical variables, all discrete integers with a sufficient number of values to loosely approximate continuity.

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 |

The *event* in this data is self-reported cessation of breastfeeding of new mothers interviewed. In the context of time-to-event analysis, the indicator variable for whether or not breastfeeding had been ceased at the time of the interview was `delta`, and the time from birth of the child to cessation of breastfeeding is coded as the variable `duration`. If the mother has not yet stopped breastfeeding the child at the time of the interview, then the `duration` variable instead represents time from birth of the child to time of the interview. In the context of survival analysis, these data are considered to be right-censored. In other words, for these patients the study stopped before the event of stopping breastfeeding had not yet occurred. It is unclear the exact definition of *completing breast feeding* utilized in the survey methodology. Whether this means the end of utilizing breast milk as the sole food source for the child vs completely removing breast milk from the infant's diet. To summarize, variables three through ten in Table 1 are candidate predictors for variable one while accounting for the censoring of some participants indicated via variable two.

### Experimental Design

Data on breastfeeding used in this study has been extracted from a large set of surveys sampling several thousand individuals, many of whom have been surveyed over decades. The NLSY79 Child and Young Adult surveys include a wide variety of information on children born to female respondents of the NLSY79 surveys. Parents reported in interviews on many aspects of the raising of their child, among that corpus of information are the data shown here.

Participants were chosen randomly from the United States population for the study, so responses from all 50 states and outlying territories are included in the sample. Detailed information about the design of the survey is available at https://www.bls.gov/nls/nlsy79.htm#intro-to-sample. Relevant surveys were conducted from 1979 through 1986 and questions related to breastfeeding were asked to mothers who had given birth in the past 12 months. Information about duration of breastfeeding was provided by mothers via memory recall.

Responses to other variables (e.g. smoking at the time of birth) were also provided by the mother. A sample of the data itself can be seen in Table 2. The `SurvObj` variable combines the `duration` and `delta` variables to give duration with participants who were still breastfeeding at last interview denoted with the `+` symbol.

Table 2: Time to cessation of breastfeeding data set

| duration | delta | race | poverty | smoke | alcohol | agemth | ybirth | yschool | pc3mth | SurvObj |
|---|---|---|---|---|---|---|---|---|---|---|
| 16 | yes | white | no | no | yes | 24 | 1982 | 14 | no | 16 |
| 1 | yes | white | no | yes | no | 26 | 1985 | 12 | no | 1 |
| 4 | no | white | no | no | no | 25 | 1985 | 12 | no | 4+ |
| 3 | yes | white | no | yes | yes | 21 | 1985 | 9 | no | 3 |
| 36 | yes | white | no | yes | no | 22 | 1982 | 12 | no | 36 |

**Censoring and Missing Values**

In this data set, a total of 35 mothers were still breastfeeding their infant at the time of their final data collection interview. Most of these censoring events occurred in the final year(s) of the study, when time from birth to final interview was significantly less than the nearly 10 years from birth of child to final interview of the earliest participants in the study. The compiled dataset does not contain any information about possible patient drop-out or missed interviews.

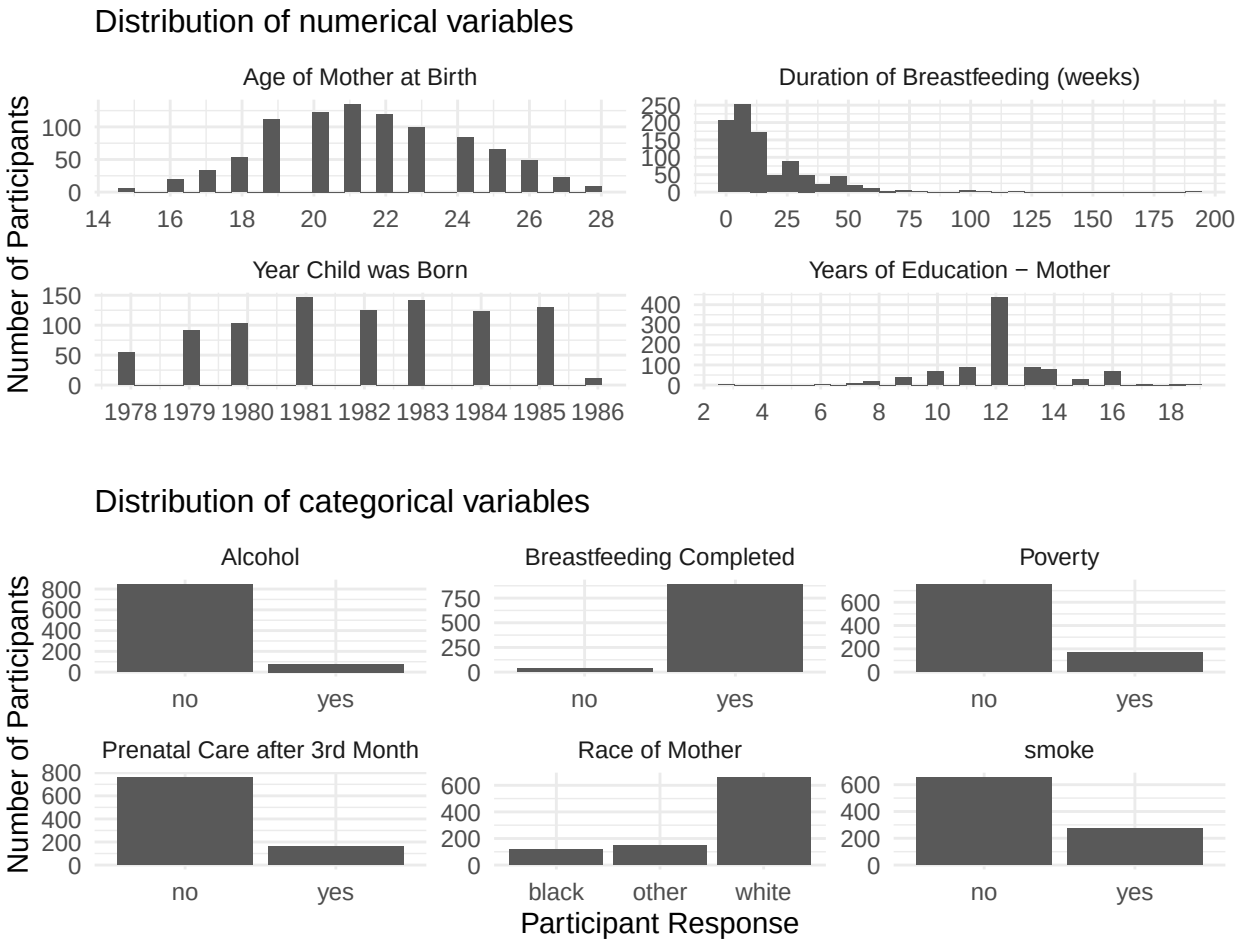



Figure 1: Distributions of categorical and numerical variables comprising the data set.

# Methods and Data Analysis

The techniques of survival analysis allow for useful descriptions of time-to-event data. The primary function of survival analysis is the probability of survival beyond time `t`, called the survival function.

$$S(t) = Pr(T > t) = 1 - F(t)$$

where T is the random variable survival time, in this case T represents the duration of breast feeding in weeks. A characteristic of the survival function $S(t)$ is that it is the complement of the cumulative density function (CDF) $F(t)$ which itself is the integral of the probability density function $f(t)$ from 0 to 1.

Another essential function for analyzing time-to-event data is the hazard function, which is the instantaneous rate of event occurrence at time t given survival to time t,

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

where $h(t)$ is the hazard function.

Survival analysis involves estimating these functions based on the data. There exist non-parametric, parametric, and semi-parametric models for estimating the survival function. In this project, the non-parametric method is the Kaplan-Meier (KM) estimator. The log-rank test is used in testing for statistical differences between KM curves. The parametric method is fitting to a distribution (e.g., Weibull or exponential) and using the relationships above to estimate the survival or hazard curves. Finally, perhaps the most common technique in survival analysis is the Cox proportional hazards model, a semi-parametric approach. Each of these approaches are utilized below. Prior to analyzing the data, a visual summary of the data set is available in Figure 1 above.

## Kaplan Meier Survival Estimates

The Kaplan-Meier curve shows an estimate of the time to an event, which here represents the time in weeks until a mother stops breastfeeding her child. The Kaplan-Meier estimator for the survival function is:

$$\hat{S} = \begin{cases} 1 & t < t_1 \\ \prod_{t \geq t_i} [1 - \frac{d_i}{Y_i}] & t \geq t_1 \end{cases}$$

where $1 \leq d_i \leq Y_i$ with $t_i$ representing the distinct time at which breastfeeding ceased, $Y_i$ representing the number of individuals still breastfeeding at time $t_i$, and $d_i$ is the number of individuals who stopped breastfeeding at time $t_i$.

The Kaplan-Meier curve drops only when an individual stops breastfeeding, not when they are censored. The survival function, as well as its estimators, are bound in value from between 0 and 1. Confidence interval estimates for the KM curve can be estimated via variance. A Kaplan-Meier curve of the entire survey sample can be seen in Figure 3 below.

The red dashed line on these curves corresponds to the 6 months of breastfeeding milestone, which is the WHO recommendation for breastfeeding children (Grummer-Strawn et al. 2017). Based on the KM curve in Figure 2, only 21% of mothers interviewed reported breastfeeding their children at least 6 months CI = (18% , 24%). The KM estimate for median duration of breastfeeding for all mothers was 12 weeks, which is 14 weeks less than the WHO recommended 6 months minimum.

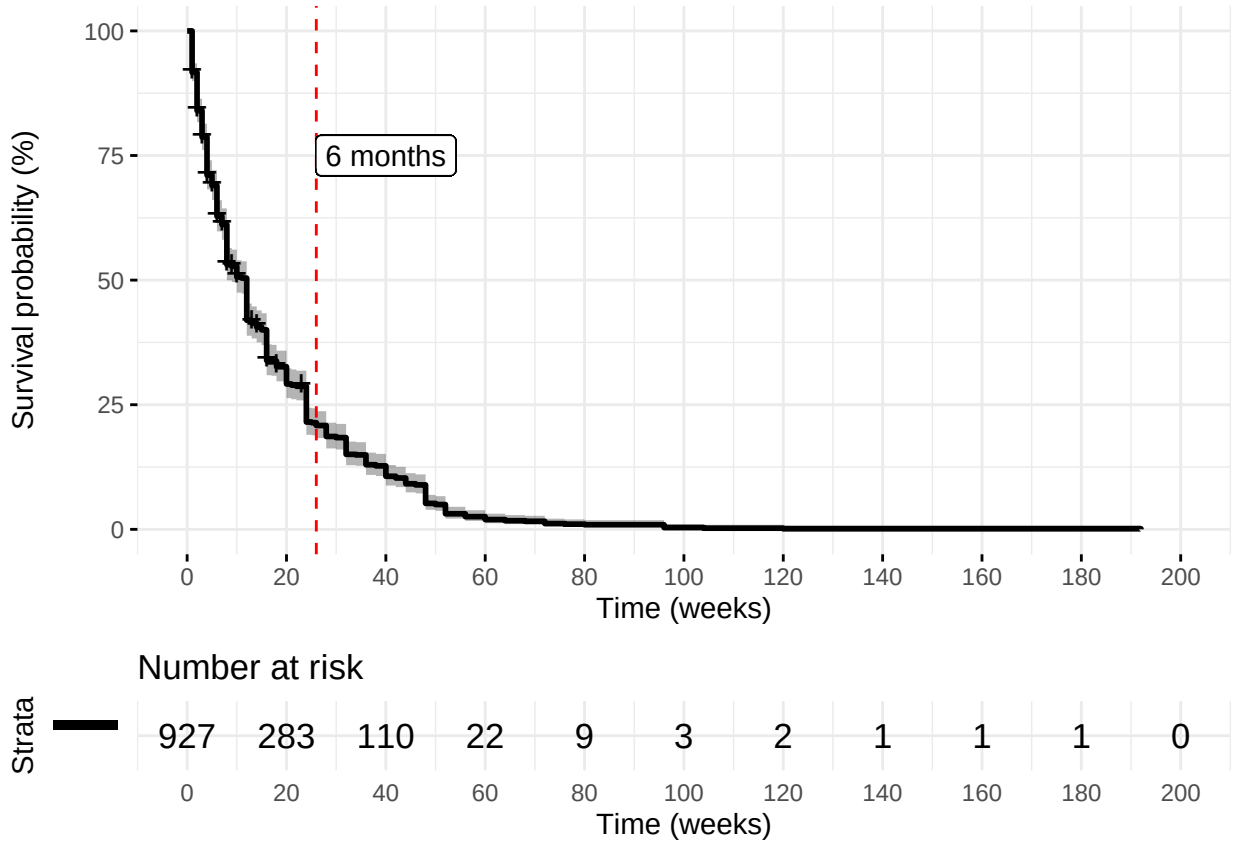



Figure 2: KM curve for all participants for duration of breastfeeding. Gray shaded region represents a pointwise 95% confidence interval for the survival curve. The symbol + corresponds to a censoring time. The number of participants still breastfeeding at a given number of weeks is shown below the survival curve plot.

**KM curve - Cross Group Comparisons**

In the survey, mothers were asked whether they were smoking at the time that they gave birth to their child. Kaplan-Meier curves for mothers who reported smoking compared to those who did not can be seen in Figure 3. It is possible to compare the point estimates for the proportion of mothers in the smoking vs nonsmoking group who were still breastfeeding at 6 months. For mothers in the smoking group, 16% of mothers interviewed reported breastfeeding their children at least 6 months CI = (12% , 21%). In the group of non-smoking mothers, 23% reported breastfeeding their children at least 6 months CI = (20% , 26%). The question then arises whether the difference between these two curves is significant. The overlapping point estimate confidence intervals at 6 months suggests the difference may be due to chance; however, survival analysis provides a more robust way to test for a difference between these two groups.

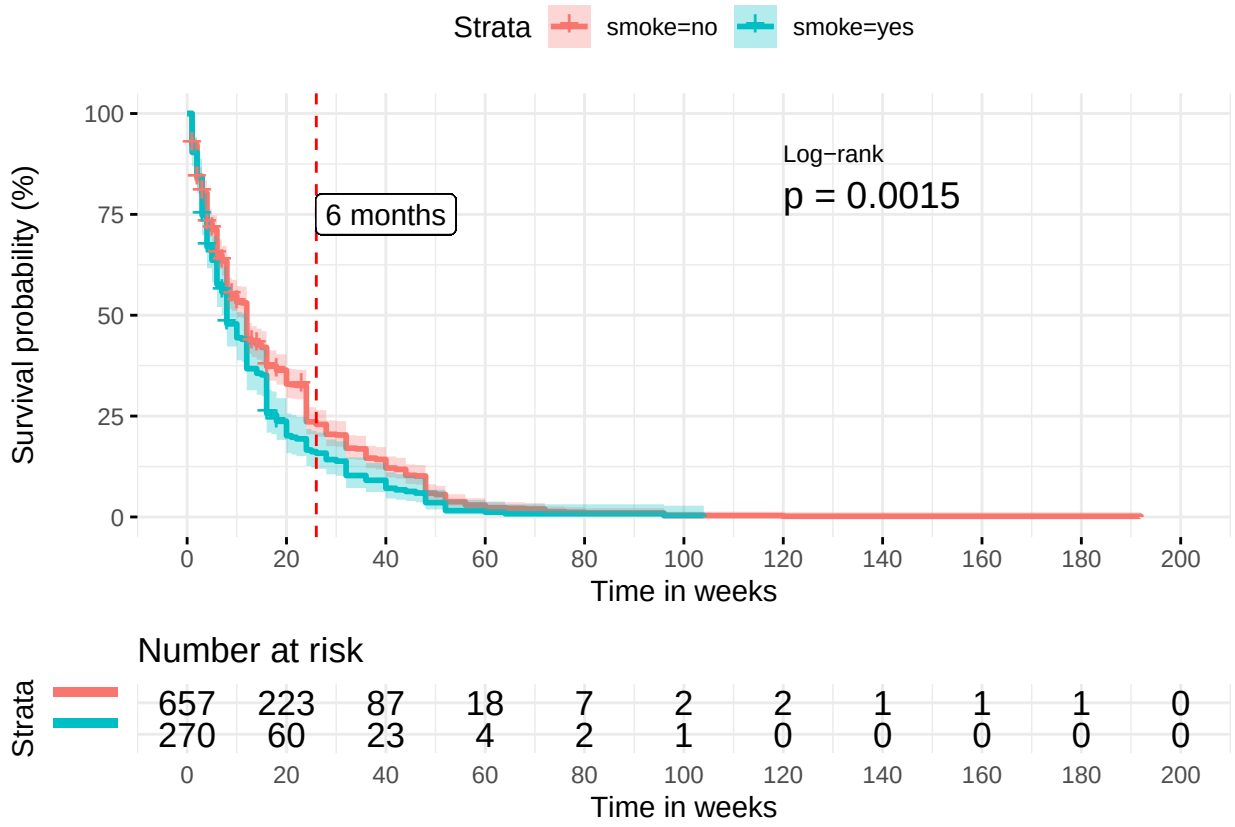



Figure 3: KM curve for duration of breastfeeding according to whether mother smoked when child was born. Shaded regions represents a pointwise 95% confidence interval for the survival curves. The symbol + corresponds to a censoring time. The number of participants still breastfeeding by group at a given number of weeks is shown below the survival curve plot.

**Log-rank test**

The log-rank test ($H_0$: no difference) can be used to compare if the difference between these KM curves is significant. We can see in Figure 3 above that there appears to be a difference in survival curves between mothers who smoked at birth of their child and mothers who did not. Given differences at early time points are considered more important from a public health perspective, the peto-peto modification would be an appropriate tool to use for testing for a difference at earlier time points in this data.

Some categorical variables are composed of more than two variables. The chi-squared generalization of the log-rank test, implemented in the R package `survival` with the function `survdiff()` can be utilized, testing

for whether at least one of the curves is significantly different. In the supplemental materials section, KM curves for other variables along with their corresponding peto-peto test p-values for difference are available as figures S? - S?.

Categorical variables that appear to have a significant effect on survival based on the log-rank test of KM curves are the race of the mother and their smoking status (Table 3).

Table 3: p-values of log-rank test for difference in breastfeeding duration according to single variable

| Variable ID | p-value | p-value.signif |
|---|---|---|
| race | 0.0177198 | * |
| poverty | 0.3984539 | ns |
| smoke | 0.0014886 | ** |
| alcohol | 0.1562283 | ns |
| pc3mth | 0.6872395 | ns |

The KM curve is not the ideal approach for resolving the effect of numerical predictor variables on the response variable breastfeeding duration. Other approaches described below are better suited to elucidating these types of variables.

#### Kaplan-Meier Curve Assumptions

There are three major assumptions of the Kaplan-Meier estimator: first, that the censored participants have the same breastfeeding duration as the participants who are not censored; second, that the time of recruitment into the study does not effect the survival outcomes; and lastly, that events happened when they are said to have happened(Goel, Khanna, and Kishore 2010). Given the nature of the data, the first assumption appears to be reasonable. The second assumption will be shown by subsequent analysis below to be perhaps a poor assumption. The final assumption is likely not entirely true, but ideally the errors in recall of participants will be randomly distributed throughout the survey sample, reducing the effect of incorrect information.

Given that earlier ages of stopping breastfeeding are considered biologically more important, the peto-peto modification would better resolve differences early in the estimated survival curves. Regardless of the version of the log-rank test used, the assumptions for these tests are based on the KM assumptions, so the use of these tests is appropriate under weak assumptions.

## Parametric Survival Estimates

Parametric modelling is a powerful tool for analyzing a wide variety. Linear regression is perhaps the most widely known parametric regression model in the scientific research community. Parametric models also exist for analyzing time-to-event data such as time to cessation of breastfeeding discussed here. Moving beyond univariate analyses, parametric survival models allow for description of numerical and multivariate models. There are several models that exist for modeling time-to-event data. Three parametric models will be fit to the data: the exponential, Weibull, and lognormal models. The exponential and Weibull models are interesting in that they can utilize either the accelerated failure time (AFT) assumption or the proportional hazards (PH) assumption for describing survival data. To summarize briefly, proportional hazards assumes that the hazard ratio for any two individuals is constant over time while the AFT model assumes that the effects of covariates are fixed and multiplicative by the acceleration factor on the time scale of $t$. The commonality across these parametric models is that they assume the outcome follows some known distribution.

**Exponential**

The exponential model is a less complicated model because its function is time-independent:

$$h(t) = \lambda$$

and

$$S(t) = e^{-\lambda t}$$

where $h(t)$ is the hazard function and $S(t)$ is the survival function. These are the AFT parameterizations of the exponential model. Note that the the hazard is a constant $\lambda$.

**Assumptions of the exponential survival model** The key assumption of the exponential survival model are that the the hazard rate is constant, derived from the memoryless property of the exponential distribution. Based on what is known about cessation of breastfeeding, this would not seem likely to be a valid assumption; however, figure 4 shows the exponential model provides a reasonable fit to the data. Trends between the different racial groups correspond to regression outputs below showing similar survival for these three groups.

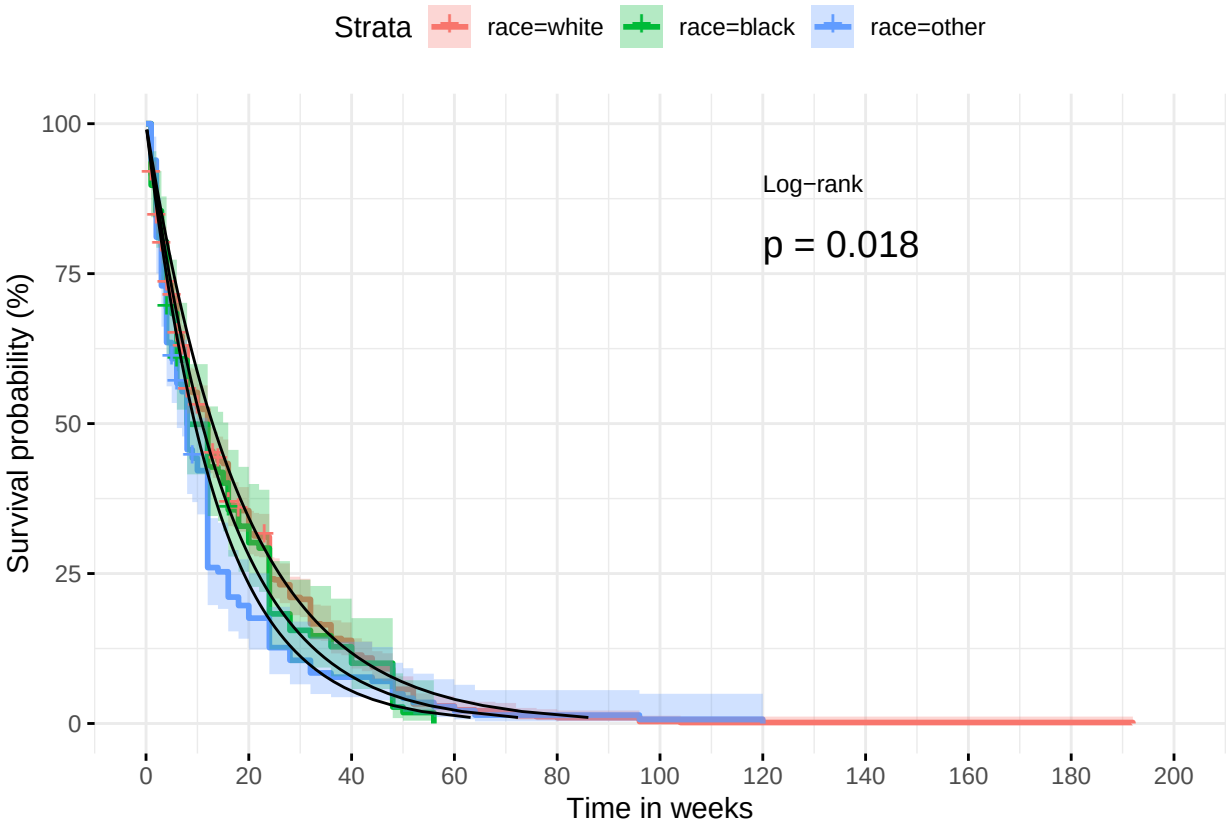



Figure 4: Comparing KM curve for duration of breastfeeding according to mothers' race with exponential regression model.

**Goodness of fit** Parametric models such as the exponential model allow for multivariate analysis. Given this, parameter selection is needed to find the best combination of parameters to predict or characterize survival. A backwards selection approach is utilized here to select parameters, in which all predictor variables

were included in the model, the Akaike Information Criterion (AIC) for the model was recorded, and then the least significant parameter was dropped from the model until no parameters remain. The results of this process are in Table 4 below. This approach selected a model with four variables: race, smoking, years of education, and year child was born.

Table 4: AIC of Backward selection of exponential survival model

| # Parameters | AIC | Equation |
|---|---|---|
| 8 | 6784.072 | SurvObj ~ ybirth + yschool + ... + agemth + pc3mth |
| 7 | 6782.440 | SurvObj ~ ybirth + yschool + ... + alcohol + agemth |
| 6 | 6781.284 | SurvObj ~ ybirth + yschool + ... + poverty + alcohol |
| 5 | 6780.807 | SurvObj ~ ybirth + yschool + smoke + race + poverty |
| 4 | 6783.687 | SurvObj ~ ybirth + yschool + smoke + race |
| 3 | 6790.684 | SurvObj ~ ybirth + yschool + smoke |
| 2 | 6795.655 | SurvObj ~ ybirth + yschool |
| 1 | 6810.122 | SurvObj ~ ybirth |
| 0 | 6820.578 | SurvObj ~ 1 |

**Parameter Estimates** The parameter estimates for the best fit exponential model are shown in Table 5. We see p-values below the 0.05 threshold for all parameters except for where race of the mother is black has a p-value of 0.0928.

Table 5: Exponential Model: SurvObj ~ race + smoke + yschool + ybirth

| | Value | Std. Error | z | p |
|---|---|---|---|---|
| (Intercept) | 159.7446 | 34.9462 | 4.5712 | 0.0000 |
| raceblack | -0.1733 | 0.1031 | -1.6806 | 0.0928 |
| raceother | -0.3096 | 0.0966 | -3.2036 | 0.0014 |
| smokeyes | -0.2669 | 0.0780 | -3.4229 | 0.0006 |
| yschool | 0.0503 | 0.0191 | 2.6370 | 0.0084 |
| ybirth | -0.0794 | 0.0177 | -4.4941 | 0.0000 |

The AFT interpretation of this model suggests that black mothers average duration of breast feeding may only be $e^{-0.173} = 0.84$ of that of white mothers (p-value = 0.093). For mothers whose race is classified as "other," the average duration of breast feeding is $e^{-0.310} = 0.73$ of white mothers. Mothers who reported smoking at time of child birth had 0.77 shorter breast feeding duration compared to non-smokers. On the other hand, every year of additional education the mother had attained corresponded to a $e^{0.05} = 1.05$. In other words, mothers showed a 5% increase in average duration of breast feeding for each additional year of schooling they completed. There also appeared to be a significant trend of child birth year effecting the duration of breastfeeding. In this model, each year later that the child was born resulted in reduction by a factor of 0.92 in average breast feeding time.

As mentioned above, the exponential model has a unique characteristic that it can be described as an AFT model or as a PH model. By dividing the negative of the coefficients of the AFT model in Table 5 by its scale parameter (1 in the case of the exponential model), one can find the proportional hazards for each variable, shown in Table S?.

### Weibull

The Weibull model is a generalization of the exponential model that is widely used in survival analysis. The hazard and survival functions for this model is

$$h(t) = \alpha\lambda t^{\alpha-1}$$

and

$$S(t) = e^{-\lambda t^{\alpha}}$$

where $h(t)$ is the hazard function and $S(t)$ is the survival function. The shape parameter $\alpha$ can be thought of as a baseline log-hazard, while $\lambda$ is the rate parameter as in the exponential distribution. These are the AFT parameterizations of the Weibull model. Note that the the hazard is monotonic for this model. When $\alpha = 1$, the Weibull model is the exponential model.

**Assumptions of the Weibull survival model** One assumption of the Weibull survival model is that the the hazard rate is monotonic. Additionally, when working as an AFT model, the differences between groups should be able to be represented by only an acceleration in aging by a constant. Part of this is that KM curves should not cross, an assumption that is reasonably held in this data set. Figure 5 shows the Weibull model provides a fit to the data comparable to that of the exponential model above. Trends between the different racial groups correspond to regression outputs below showing similar survival for these three groups.

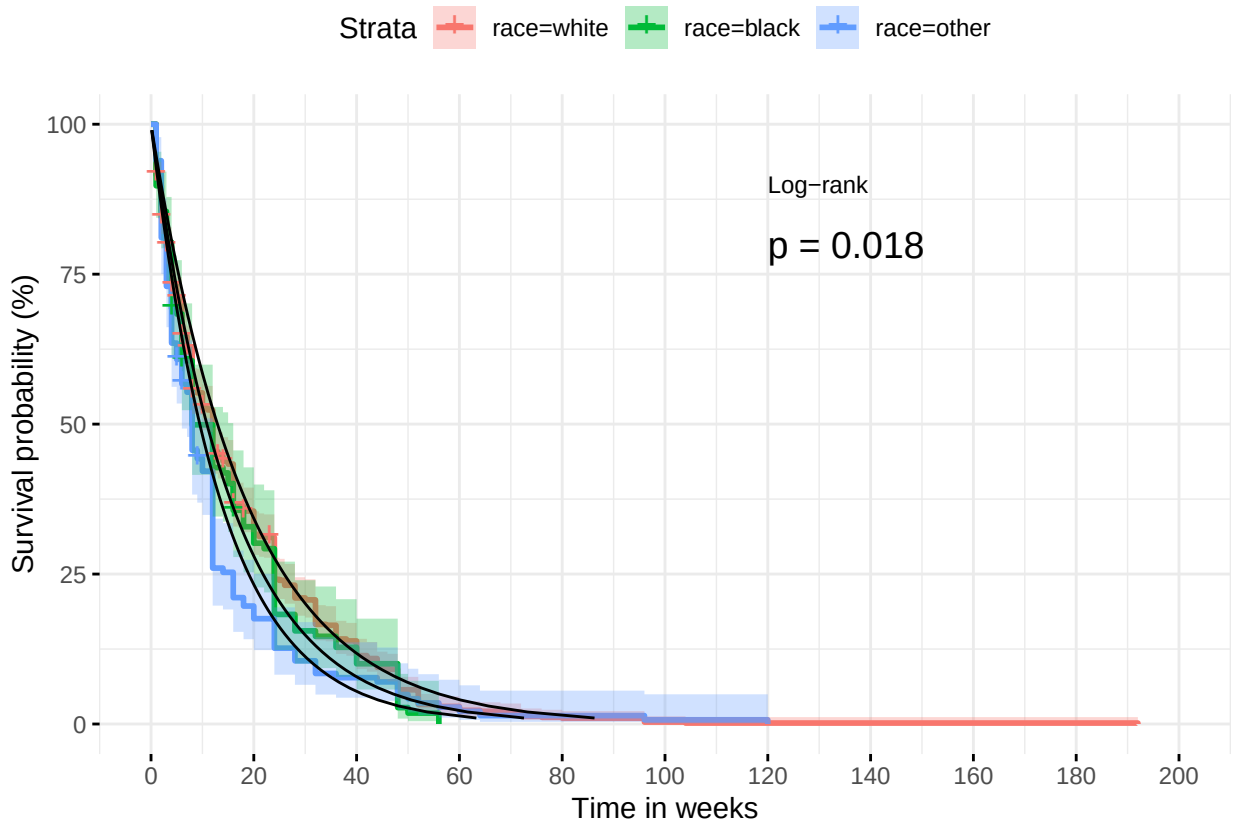



Figure 5: Comparing KM curve for duration of breastfeeding according to mothers' race with the Weibull regression model.

**Goodness of fit** The same backwards selection approach is utilized here as the exponential model to select parameters and compare AIC scores. The results of this process are in Table 6 below. The Weibull model also selected the model with the same four variables: race, smoking, years of education, and year child was born.

Table 6: AIC of Backward selection of Weibull survival model

| # Parameters | AIC | Equation |
|---|---|---|
| 8 | 6786.063 | SurvObj ~ ybirth + yschool + ... + agemth + pc3mth |
| 7 | 6784.433 | SurvObj ~ ybirth + yschool + ... + alcohol + agemth |
| 6 | 6783.278 | SurvObj ~ ybirth + yschool + ... + poverty + alcohol |
| 5 | 6782.805 | SurvObj ~ ybirth + yschool + smoke + race + poverty |
| 4 | 6785.678 | SurvObj ~ ybirth + yschool + smoke + race |
| 3 | 6792.578 | SurvObj ~ ybirth + yschool + smoke |
| 2 | 6797.407 | SurvObj ~ ybirth + yschool |
| 1 | 6811.615 | SurvObj ~ ybirth |
| 0 | 6821.128 | SurvObj ~ 1 |

**Parameter Estimates** The parameter estimates for the best fit Weibull model are shown in Table 7. We see p-values below the 0.05 threshold for all parameters except for where race of the mother is black has a p-value of 0.0936.

Table 7: Weibull Model: SurvObj ~ race + smoke + yschool + ybirth

| | Value | Std. Error | z | p |
|---|---|---|---|---|
| (Intercept) | 159.6332 | 35.0449 | 4.5551 | 0.0000 |
| raceblack | -0.1732 | 0.1033 | -1.6765 | 0.0936 |
| raceother | -0.3097 | 0.0969 | -3.1972 | 0.0014 |
| smokeyes | -0.2669 | 0.0781 | -3.4154 | 0.0006 |
| yschool | 0.0504 | 0.0192 | 2.6323 | 0.0085 |
| ybirth | -0.0794 | 0.0177 | -4.4785 | 0.0000 |
| Log(scale) | 0.0023 | 0.0255 | 0.0906 | 0.9278 |

The parameter estimates for the Weibull model are extremely similar to the parameters of the exponential distribution, which one would expect given the log(scale) parameter for the fit is very close to zero corresponding to $\alpha$ value very close to one. The Weibull model can also be described as an AFT model or as a PH model. Given these values are very similar to those of the exponential model, these values have not been included in this analysis.

The finding that the generalization of the exponential distribution, the Weibull, closely matches the results of the exponential itself supports that the constant hazard assumption of the exponential is close to the optimal monotonic hazard function to fit the data.

### Lognormal

The lognormal model is useful for modeling data with a hump-shaped hazard curve. This shape of hazard curve is not the most common; however, the lognormal model is a very useful model for such scenarios. The hazard and survival functions for $X \sim lognormal(\mu, \sigma)$ the lognormal model are

$$S(t) = Pr(T > t) = Pr(ln(X) > ln(X)) = 1 - \Phi(\frac{ln(x) - \mu}{\sigma})$$

and

$$h(t) = \frac{(\frac{1}{x\sigma})\phi(\frac{ln(x)}{\sigma})}{\Phi(\frac{-ln(x)}{\sigma})}$$

with $x > 0, \sigma > 0$. Here $h(t)$ the hazard function and $S(t)$ the survival function. Here $\phi$ represents the probability density function of the normal distribution while $\Phi$ is the cumulative distribution function of the normal distribution. The distribution of $ln(X) \sim N(\mu, \sigma)$ with mean $\mu$ and standard deviation $\sigma$.

**Assumptions of the lognormal survival model**

The lognormal survival model assumes that the time-to-event variable is lognormally distributed. The duration variable of the breastfeeding data set is fairly well represented by a normal distribution following log-transformation (Supplementary Figure S1). As an AFT model, differences between groups should be able to be modelled as accerlated time. Figure 5 shows the lognormal model provides a fit to the data qualitatively similar to the other parametric models above. Trends between the different racial groups correspond to regression outputs below showing similar survival for these three groups.

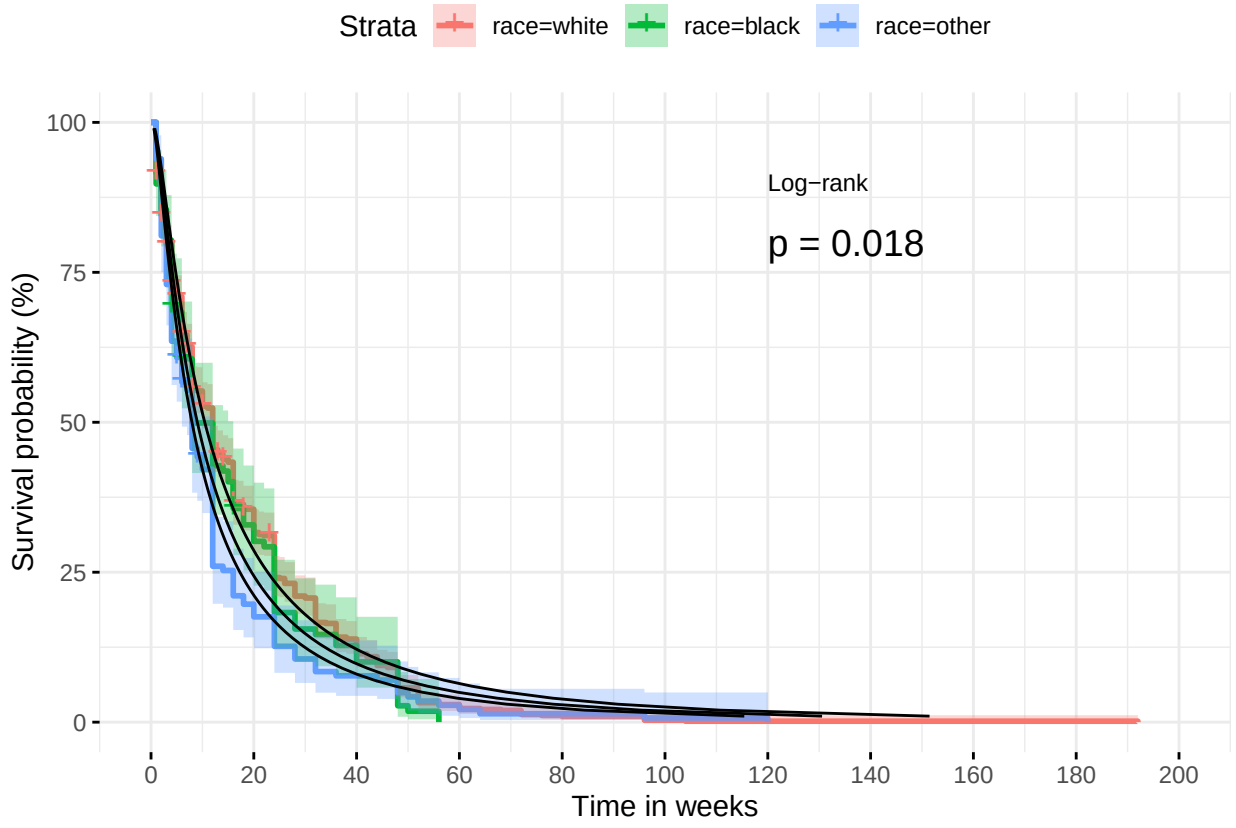



Figure 6: Comparing KM curve for duration of breastfeeding according to mothers' race with the Weibull regression model.

**Goodness of fit**

The same backwards selection approach is utilized here as the above models to select parameters and compare AIC scores. The results of this process are in Table 8 below. The lognormal model with the lowest AIC score had the same four variables: race, smoking, years of education, and year child was born.

Table 8: AIC of Backward selection of lognormal survival model

| # Parameters | AIC | Equation |
|---|---|---|
| 8 | 6782.385 | SurvObj ~ ybirth + yschool + ... + agemth + pc3mth |
| 7 | 6780.469 | SurvObj ~ ybirth + yschool + ... + alcohol + agemth |
| 6 | 6778.938 | SurvObj ~ ybirth + yschool + ... + poverty + alcohol |
| 5 | 6778.301 | SurvObj ~ ybirth + yschool + smoke + race + poverty |
| 4 | 6779.371 | SurvObj ~ ybirth + yschool + smoke + race |
| 3 | 6782.155 | SurvObj ~ ybirth + yschool + smoke |
| 2 | 6783.920 | SurvObj ~ ybirth + yschool |
| 1 | 6806.721 | SurvObj ~ ybirth |
| 0 | 6809.547 | SurvObj ~ 1 |

**Parameter Estimates**

The parameter estimates for the best fit lognormal model are shown in Table 9. We see p-values below the 0.05 threshold for all parameters except for where race of the mother is black has a p-value of 0.2062.

Table 9: Lognormal Model: SurvObj ~ race + smoke + yschool + ybirth

| | Value | Std. Error | z | p |
|---|---|---|---|---|
| (Intercept) | 146.8739 | 38.1150 | 3.8534 | 0.0001 |
| raceblack | -0.1485 | 0.1175 | -1.2641 | 0.2062 |
| raceother | -0.2717 | 0.1095 | -2.4820 | 0.0131 |
| smokeyes | -0.2252 | 0.0883 | -2.5502 | 0.0108 |
| yschool | 0.0884 | 0.0222 | 3.9877 | 0.0001 |
| ybirth | -0.0735 | 0.0193 | -3.8102 | 0.0001 |
| Log(scale) | 0.1391 | 0.0236 | 5.8830 | 0.0000 |

The AFT interpretation of this lognormal model suggests that black mothers average duration of breast feeding may only be $e^{-0.1485} = 0.86$ of that of white mothers, less of a difference than predicted by the previous models. For mothers whose race is classified as "other," the average duration of breast feeding is $e^{-0.2717} = 0.76$ of white mothers, slightly less than the difference predicted from the above models. Mothers who reported smoking at time of child birth had 0.8 shorter breast feeding duration compared to non-smokers according to this model. On the other hand, every year of additional education the mother had attained corresponded to a $e^{0.0884} = 1.09$. In other words, mothers showed around a **9%** increase in average duration of breast feeding for each additional year of schooling they completed. There also appeared to be a significant trend of child birth year effecting the duration of breastfeeding. In this model, each year later that the child was born resulted in reduction by a factor of 0.93 in average breast feeding time. Unlike the exponential and Weibull models above, the exponential model can only work as a AFT model.

**Comparison of Parametric Models**

The results of the different parametric models converged on similar results for the best group of predictors and coefficients for those predictors. While any of the models would be reasonable approximations for modeling, it is possible to compare AIC values for all of the models to see which model is optimal. In Table 10 it can be seen that the model with the lowest AIC value among those tested is the lognormal model utilizing the year of birth, years of mother education, smoking, and race to predict duration of breastfeeding.

Table 10: AIC values across parametric models

| # Parameters | Exponential | Weibull | Lognormal | Equation |
|---|---|---|---|---|
| 8 | 6784 | 6786 | 6782 | SurvObj ~ ybirth + yschool + ... + agemth + pc3mth |
| 7 | 6782 | 6784 | 6780 | SurvObj ~ ybirth + yschool + ... + alcohol + agemth |
| 6 | 6781 | 6783 | 6779 | SurvObj ~ ybirth + yschool + ... + poverty + alcohol |
| 5 | 6781 | 6783 | 6778 | SurvObj ~ ybirth + yschool + smoke + race + poverty |
| 4 | 6784 | 6786 | 6779 | SurvObj ~ ybirth + yschool + smoke + race |
| 3 | 6791 | 6793 | 6782 | SurvObj ~ ybirth + yschool + smoke |
| 2 | 6796 | 6797 | 6784 | SurvObj ~ ybirth + yschool |
| 1 | 6810 | 6812 | 6807 | SurvObj ~ ybirth |
| 0 | 6821 | 6821 | 6810 | SurvObj ~ 1 |

While this model is ideal, we can see in Table S3 that the parameter estimates across the 3 models are highly similar. In theory, these models have the benefit of allowing for predicting survival beyond the times shown in this study. In the context of breastfeeding; however, the time of greatest interest is the first 6 months of life so this aspect of parametric models is not as useful in this scenario. Furthermore, the baseline hazard functions are by definition parametric in these models, so if the data do not closely match any of these distributions, a non-parametric baseline hazard functions would better elucidate patterns in the data.

## Cox Proportional Hazards Model

In contrast to the parametric models above, the Cox PH model is semi-parametric. The baseline hazard of the model is nonparametric. While the parametric models above fit the data reasonably well, it is worth comparing their findings to that of perhaps the gold standard of time-to-event analysis: the Cox Proportional Hazards model, specifically this model will be the Cox PH model of time-independent variables. The general hazard function for p predictors is

$$h(t, X, \beta) = h_0(t) \times e^{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}$$

where $h(t)$ is the hazard function, $h_0(t)$ is the non-parametric baseline hazard function, $beta_i$ is the coefficients for the $i^{th}$ predictor $X_i$. In the context of this dataset, $X_1$ could be the smoking variable, where $x_1 = 0$ is nonsmoking and $x_1 = 1$ is smoking, the effect of this variable on the hazard function would then be determined by $beta_1$. When working with proportional hazards, the hazard ratio (HR) is a useful tool. The hazard ratio is defined as

$$HR(t, x_1, x_2) = \frac{h(t, X = 1, \beta)}{h(t, X = 0, \beta)} = \frac{h_0(t)\Gamma(X = 1, \beta)}{h_0(t)\Gamma(X = 0, \beta)} = \frac{\Gamma(X = 1, \beta)}{\Gamma(X = 0, \beta)} = e^{\beta}$$

Note that the HR depends only on $\beta$ and amount of increase in $X$, not time nor the baseline hazard. A hazard ratio close to one means the variable has little effect on survival, while an HR $> 1$ corresponds to an increased risk of event occurence. An HR value close to zero means the factor increases time until the event.

### Goodness of fit

### Parameter Estimates

### Assumptions of the Cox PH survival model

There are several important assumptions for appropriate use of the Cox proportional hazards regression model, including

```
independence of survival times between distinct individuals in the sample,
a multiplicative relationship between the predictors and the hazard (as opposed to a linear one as was
a constant hazard ratio over time.
```

**Predictions and Validations**

# Discussion

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

weaknesses:

As the backward paramter selection in Table 4 shows, birth year was the most significant predictor of all predictor variables. This weakens the assumption of the Kaplan-Meier curve that time of entry has no effect on risk. Fortunately, other models shown here account for this pattern.

The approach to data collection for this study relies on accurate recall and participant truthfulness. Given that mothers' choices around raising their child are often stigmatized and human recall of previous events is suboptimal, the data is likely imperfect.

The proportion of this data that is censored is relatively small, only about 3% of the total sample. In this case it is possible that alternative approaches such as the Mann-Whitney U test would be useful, but the ability of survival analysis to estimate risk can be exceptionally useful.

# Conclusion

# References


Colen, Cynthia G, and David M Ramey. 2014. "Is Breast Truly Best? Estimating the Effects of Breastfeeding on Long-Term Child Health and Wellbeing in the United States Using Sibling Comparisons." *Social Science & Medicine* 109: 55–65.

Eidelman, Arthur I, Richard J Schanler, Margreete Johnston, Susan Landers, Larry Noble, Kinga Szucs, and Laura Viehmann. 2012. "Breastfeeding and the Use of Human Milk." *Pediatrics* 129 (3): e827–41.

Esch, Betty CAM van, Mojtaba Porbahaie, Suzanne Abbring, Johan Garssen, Daniel P Potaczek, Huub FJ Savelkoul, and RJ Neerven. 2020. "The Impact of Milk and Its Components on Epigenetic Programming of Immune Function in Early Life and Beyond: Implications for Allergy and Asthma." *Frontiers in Immunology* 11: 2141.

Gaynor, G. 2003. "Breastfeeding Advocacy." *Maine Nurse* 5 (2): 13.

Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore. 2010. "Understanding Survival Analysis: Kaplan-Meier Estimate." *International Journal of Ayurveda Research* 1 (4): 274.

Grummer-Strawn, Laurence M, Elizabeth Zehner, Marcus Stahlhofer, Chessa Lutter, David Clark, Elisabeth Sterken, Susanna Harutyunyan, Elizabeth I Ransom, and WHO/UNICEF NetCode. 2017. "New World Health Organization Guidance Helps Protect Breastfeeding as a Human Right." *Maternal & Child Nutrition* 13 (4): e12491.

Hassiotou, Foteini, and Peter E Hartmann. 2014. "At the Dawn of a New Discovery: The Potential of Breast Milk Stem Cells." *Advances in Nutrition* 5 (6): 770–78.

Hoddinott, Pat, David Tappin, and Charlotte Wright. 2008. "Breast Feeding." *Bmj* 336 (7649): 881–87.

Klein, John P, and Melvin L Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. Vol. 1230. Springer.

Kramer, Michael S, Beverley Chalmers, Ellen D Hodnett, Zinaida Sevkovskaya, Irina Dzikovich, Stanley Shapiro, Jean-Paul Collet, et al. 2001. "Promotion of Breastfeeding Intervention Trial (PROBIT): A Randomized Trial in the Republic of Belarus." *Jama* 285 (4): 413–20.

León-Cava, Natalia, Chessa Lutter, Jay Ross, and Luann Martin. 2002. "Quantifying the Benefits of Breastfeeding: A Summary of the Evidence." *Pan American Health Organization, Washington DC* 3.

McFadden, Alison, Frances Mason, Jean Baker, France Begin, Fiona Dykes, Laurence Grummer-Strawn, Natalie Kenney-Muir, Heather Whitford, Elizabeth Zehner, and Mary J Renfrew. 2016. "Spotlight on Infant Formula: Coordinated Global Action Needed." *The Lancet* 387 (10017): 413–15.

Munch, Erika M, R Alan Harris, Mahmoud Mohammad, Ashley L Benham, Sasha M Pejerrey, Lori Showalter, Min Hu, et al. 2013. "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 (2): e50564.

Pannaraj, Pia S, Fan Li, Chiara Cerini, Jeffrey M Bender, Shangxin Yang, Adrienne Rollie, Helty Adisetiyo, et al. 2017. "Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome." *JAMA Pediatrics* 171 (7): 647–54.

Pérez-Escamilla, Rafael. 2020. "Breastfeeding in the 21st Century: How We Can Make It Work." *Social Science & Medicine* 244: 112331.

Pérez-Escamilla, Rafael, Leslie Curry, Dilpreet Minhas, Lauren Taylor, and Elizabeth Bradley. 2012. "Scaling up of Breastfeeding Promotion Programs in Low-and Middle-Income Countries: The 'Breastfeeding Gear' Model." *Advances in Nutrition* 3 (6): 790–800.

Pomeranz, Jennifer L, Xiangying Chu, Oana Groza, Madeline Cohodes, and Jennifer L Harris. 2021. "Breastmilk or Infant Formula? Content Analysis of Infant Feeding Advice on Breastmilk Substitute Manufacturer Websites." *Public Health Nutrition*, 1–9.

Raissian, Kerri M, and Jessica Houston Su. 2018. "The Best of Intentions: Prenatal Breastfeeding Intentions and Infant Health." *SSM-Population Health* 5: 86–100.

Stevens, Emily E, Thelma E Patrick, and Rita Pickler. 2009. "A History of Infant Feeding." *The Journal of Perinatal Education* 18 (2): 32–39.

Victora, Cesar G, Rajiv Bahl, Alu'sio JD Barros, Giovanny VA França, Susan Horton, Julia Krasevec, Simon Murch, et al. 2016. "Breastfeeding in the 21st Century: Epidemiology, Mechanisms, and Lifelong Effect." *The Lancet* 387 (10017): 475–90.

Walters, Dylan D, Linh TH Phan, and Roger Mathisen. 2019. "The Cost of Not Breastfeeding: Global Results from a New Tool." *Health Policy and Planning* 34 (6): 407–17.

## Statistical Models

### Parametric model: Exponential vs Weibull

```
##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool + pc3mth, data = bfeed, dist = "exponential")
##                Value Std. Error     z       p
## (Intercept) 175.2188    39.6410  4.42 9.9e-06
## raceblack    -0.1944     0.1050 -1.85 0.06408
## raceother    -0.3253     0.0967 -3.37 0.00076
## povertyyes    0.2165     0.0932  2.32 0.02014
## smokeyes     -0.2679     0.0792 -3.38 0.00072
## alcoholyes   -0.1565     0.1227 -1.28 0.20217
## agemth        0.0174     0.0188  0.93 0.35395
## ybirth       -0.0875     0.0201 -4.35 1.3e-05
## yschool       0.0571     0.0231  2.48 0.01332
## pc3mthyes     0.0543     0.0900  0.60 0.54599
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3382   Loglik(intercept only)= -3409.3
##  Chisq= 54.51 on 9 degrees of freedom, p= 1.5e-08
## Number of Newton-Raphson Iterations: 4
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool, data = bfeed, dist = "exponential")
##                Value Std. Error     z       p
## (Intercept) 173.2055    39.5333  4.38 1.2e-05
## raceblack    -0.1890     0.1046 -1.81 0.07080
## raceother    -0.3237     0.0966 -3.35 0.00081
## povertyyes    0.2200     0.0930  2.37 0.01798
## smokeyes     -0.2652     0.0791 -3.35 0.00080
## alcoholyes   -0.1549     0.1227 -1.26 0.20679
## agemth        0.0172     0.0188  0.92 0.35931
## ybirth       -0.0865     0.0200 -4.31 1.6e-05
## yschool       0.0559     0.0230  2.43 0.01517
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3382.2   Loglik(intercept only)= -3409.3
##  Chisq= 54.14 on 8 degrees of freedom, p= 6.5e-09
## Number of Newton-Raphson Iterations: 4
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
```

```
##     yschool + ybirth, data = bfeed, dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept) 156.1664    34.9804  4.46 8.0e-06
## raceblack    -0.2007     0.1039 -1.93 0.05334
## raceother    -0.3233     0.0967 -3.34 0.00083
## povertyyes    0.2119     0.0926  2.29 0.02203
## smokeyes     -0.2643     0.0791 -3.34 0.00083
## alcoholyes   -0.1543     0.1226 -1.26 0.20818
## yschool       0.0658     0.0202  3.26 0.00112
## ybirth       -0.0777     0.0177 -4.39 1.1e-05
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3382.6   Loglik(intercept only)= -3409.3
##  Chisq= 53.29 on 7 degrees of freedom, p= 3.2e-09
## Number of Newton-Raphson Iterations: 4
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + yschool +
##     ybirth, data = bfeed, dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept) 157.0951    34.9186  4.50 6.8e-06
## raceblack    -0.2023     0.1039 -1.95 0.05152
## raceother    -0.3246     0.0968 -3.36 0.00079
## povertyyes    0.2003     0.0921  2.18 0.02960
## smokeyes     -0.2796     0.0780 -3.58 0.00034
## yschool       0.0633     0.0201  3.15 0.00162
## ybirth       -0.0782     0.0177 -4.43 9.5e-06
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3383.4   Loglik(intercept only)= -3409.3
##  Chisq= 51.77 on 6 degrees of freedom, p= 2.1e-09
## Number of Newton-Raphson Iterations: 4
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + smoke + yschool + ybirth,
##     data = bfeed, dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept) 159.7446    34.9462  4.57 4.9e-06
## raceblack    -0.1733     0.1031 -1.68 0.09283
## raceother    -0.3096     0.0966 -3.20 0.00136
## smokeyes     -0.2669     0.0780 -3.42 0.00062
## yschool       0.0503     0.0191  2.64 0.00837
## ybirth       -0.0794     0.0177 -4.49 7.0e-06
##
## Scale fixed at 1
##
```

```
## Exponential distribution
## Loglik(model)= -3385.8   Loglik(intercept only)= -3409.3
##  Chisq= 46.89 on 5 degrees of freedom, p= 6e-09
## Number of Newton-Raphson Iterations: 4
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ smoke + yschool + ybirth, data = bfeed,
##     dist = "exponential")
##               Value Std. Error     z       p
## (Intercept) 161.8842    34.7731  4.66 3.2e-06
## smokeyes     -0.2017     0.0754 -2.67 0.00750
## yschool       0.0632     0.0188  3.37 0.00075
## ybirth       -0.0806     0.0176 -4.59 4.5e-06
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3391.3   Loglik(intercept only)= -3409.3
##  Chisq= 35.89 on 3 degrees of freedom, p= 7.9e-08
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ yschool + ybirth, data = bfeed, dist = "exponential")
##               Value Std. Error     z       p
## (Intercept) 165.9710    34.8736  4.76 1.9e-06
## yschool       0.0741     0.0182  4.06 4.8e-05
## ybirth       -0.0828     0.0176 -4.69 2.7e-06
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3394.8   Loglik(intercept only)= -3409.3
##  Chisq= 28.92 on 2 degrees of freedom, p= 5.2e-07
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ ybirth, data = bfeed, dist = "exponential")
##               Value Std. Error     z       p
## (Intercept) 118.2389    32.7880  3.61 0.00031
## ybirth       -0.0582     0.0165 -3.52 0.00043
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -3403.1   Loglik(intercept only)= -3409.3
##  Chisq= 12.46 on 1 degrees of freedom, p= 0.00042
## Number of Newton-Raphson Iterations: 4
## n= 927
```

```
##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool + pc3mth, data = bfeed, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept) 175.29620   39.55209  4.43 9.3e-06
## raceblack    -0.19448    0.10474 -1.86 0.06336
## raceother    -0.32521    0.09644 -3.37 0.00075
## povertyyes    0.21657    0.09294  2.33 0.01980
## smokeyes     -0.26789    0.07903 -3.39 0.00070
## alcoholyes   -0.15640    0.12245 -1.28 0.20153
## agemth        0.01737    0.01872  0.93 0.35363
## ybirth       -0.08752    0.02006 -4.36 1.3e-05
## yschool       0.05705    0.02302  2.48 0.01321
## pc3mthyes     0.05433    0.08975  0.61 0.54501
## Log(scale)   -0.00244    0.02555 -0.10 0.92382
##
## Scale= 0.998
##
## Weibull distribution
## Loglik(model)= -3382   Loglik(intercept only)= -3408.6
##  Chisq= 53.07 on 9 degrees of freedom, p= 2.8e-08
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool, data = bfeed, dist = "weibull")
##                Value Std. Error     z       p
## (Intercept) 173.2760    39.4547  4.39 1.1e-05
## raceblack    -0.1890     0.1044 -1.81 0.07012
## raceother    -0.3236     0.0964 -3.36 0.00079
## povertyyes    0.2201     0.0928  2.37 0.01771
## smokeyes     -0.2652     0.0790 -3.36 0.00078
## alcoholyes   -0.1548     0.1224 -1.26 0.20620
## agemth        0.0172     0.0187  0.92 0.35901
## ybirth       -0.0865     0.0200 -4.32 1.5e-05
## yschool       0.0558     0.0230  2.43 0.01507
## Log(scale)   -0.0022     0.0255 -0.09 0.93139
##
## Scale= 0.998
##
## Weibull distribution
## Loglik(model)= -3382.2   Loglik(intercept only)= -3408.6
##  Chisq= 52.7 on 8 degrees of freedom, p= 1.2e-08
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     yschool + ybirth, data = bfeed, dist = "weibull")
##                 Value Std. Error     z       p
```

```
## (Intercept) 156.25987   34.93343  4.47 7.7e-06
## raceblack    -0.20070    0.10365 -1.94 0.05283
## raceother    -0.32319    0.09650 -3.35 0.00081
## povertyyes    0.21201    0.09238  2.30 0.02173
## smokeyes     -0.26429    0.07893 -3.35 0.00081
## alcoholyes   -0.15421    0.12237 -1.26 0.20761
## yschool       0.06571    0.02016  3.26 0.00111
## ybirth       -0.07777    0.01766 -4.40 1.1e-05
## Log(scale)   -0.00201    0.02555 -0.08 0.93720
##
## Scale= 0.998
##
## Weibull distribution
## Loglik(model)= -3382.6   Loglik(intercept only)= -3408.6
##  Chisq= 51.85 on 7 degrees of freedom, p= 6.2e-09
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + yschool +
##     ybirth, data = bfeed, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept) 157.15980   34.89373  4.50 6.7e-06
## raceblack    -0.20230    0.10374 -1.95 0.05118
## raceother    -0.32457    0.09663 -3.36 0.00078
## povertyyes    0.20031    0.09194  2.18 0.02935
## smokeyes     -0.27958    0.07792 -3.59 0.00033
## yschool       0.06328    0.02007  3.15 0.00162
## ybirth       -0.07822    0.01764 -4.43 9.3e-06
## Log(scale)   -0.00137    0.02556 -0.05 0.95721
##
## Scale= 0.999
##
## Weibull distribution
## Loglik(model)= -3383.4   Loglik(intercept only)= -3408.6
##  Chisq= 50.32 on 6 degrees of freedom, p= 4e-09
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + smoke + yschool + ybirth,
##     data = bfeed, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept) 159.63315   35.04486  4.56 5.2e-06
## raceblack    -0.17322    0.10333 -1.68 0.09365
## raceother    -0.30970    0.09687 -3.20 0.00139
## smokeyes     -0.26688    0.07814 -3.42 0.00064
## yschool       0.05042    0.01915  2.63 0.00848
## ybirth       -0.07937    0.01772 -4.48 7.5e-06
## Log(scale)    0.00231    0.02552  0.09 0.92784
##
## Scale= 1
```

```
##
## Weibull distribution
## Loglik(model)= -3385.8   Loglik(intercept only)= -3408.6
##  Chisq= 45.45 on 5 degrees of freedom, p= 1.2e-08
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ smoke + yschool + ybirth, data = bfeed,
##     dist = "weibull")
##                Value Std. Error     z       p
## (Intercept) 161.50221   35.06995  4.61 4.1e-06
## smokeyes     -0.20175    0.07607 -2.65 0.00800
## yschool       0.06355    0.01894  3.36 0.00079
## ybirth       -0.08044    0.01773 -4.54 5.7e-06
## Log(scale)    0.00832    0.02554  0.33 0.74463
##
## Scale= 1.01
##
## Weibull distribution
## Loglik(model)= -3391.3   Loglik(intercept only)= -3408.6
##  Chisq= 34.55 on 3 degrees of freedom, p= 1.5e-07
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ yschool + ybirth, data = bfeed, dist = "weibull")
##                Value Std. Error     z       p
## (Intercept) 165.2997    35.3232  4.68 2.9e-06
## yschool       0.0746     0.0185  4.03 5.5e-05
## ybirth       -0.0825     0.0179 -4.62 3.9e-06
## Log(scale)    0.0127     0.0255  0.50    0.62
##
## Scale= 1.01
##
## Weibull distribution
## Loglik(model)= -3394.7   Loglik(intercept only)= -3408.6
##  Chisq= 27.72 on 2 degrees of freedom, p= 9.6e-07
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ ybirth, data = bfeed, dist = "weibull")
##                Value Std. Error     z       p
## (Intercept) 116.8541    33.4165  3.50 0.00047
## ybirth       -0.0575     0.0169 -3.41 0.00064
## Log(scale)    0.0182     0.0256  0.71 0.47841
##
## Scale= 1.02
##
```

```
## Weibull distribution
## Loglik(model)= -3402.8   Loglik(intercept only)= -3408.6
##  Chisq= 11.51 on 1 degrees of freedom, p= 0.00069
## Number of Newton-Raphson Iterations: 5
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool + pc3mth, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 163.0387    45.0965  3.62 0.00030
## raceblack    -0.1703     0.1195 -1.42 0.15434
## raceother    -0.2858     0.1095 -2.61 0.00906
## povertyyes    0.1926     0.1044  1.85 0.06500
## smokeyes     -0.2228     0.0894 -2.49 0.01267
## alcoholyes   -0.1660     0.1383 -1.20 0.23006
## agemth        0.0151     0.0215  0.70 0.48232
## ybirth       -0.0818     0.0229 -3.58 0.00035
## yschool       0.0915     0.0260  3.52 0.00043
## pc3mthyes     0.0296     0.1024  0.29 0.77283
## Log(scale)    0.1365     0.0236  5.77 7.7e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3380.2   Loglik(intercept only)= -3402.8
##  Chisq= 45.16 on 9 degrees of freedom, p= 8.6e-07
## Number of Newton-Raphson Iterations: 3
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 161.7216    44.8658  3.60 0.00031
## raceblack    -0.1692     0.1195 -1.42 0.15674
## raceother    -0.2844     0.1094 -2.60 0.00935
## povertyyes    0.1961     0.1036  1.89 0.05842
## smokeyes     -0.2221     0.0893 -2.49 0.01292
## alcoholyes   -0.1653     0.1383 -1.20 0.23191
## agemth        0.0147     0.0215  0.69 0.49312
## ybirth       -0.0811     0.0228 -3.56 0.00036
## yschool       0.0910     0.0259  3.51 0.00045
## Log(scale)    0.1365     0.0236  5.78 7.7e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3380.2   Loglik(intercept only)= -3402.8
##  Chisq= 45.08 on 8 degrees of freedom, p= 3.6e-07
## Number of Newton-Raphson Iterations: 3
## n= 927
```

```
##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     yschool + ybirth, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 145.3860    38.0284  3.82 0.00013
## raceblack    -0.1792     0.1186 -1.51 0.13068
## raceother    -0.2847     0.1094 -2.60 0.00926
## povertyyes    0.1874     0.1028  1.82 0.06851
## smokeyes     -0.2202     0.0893 -2.47 0.01366
## alcoholyes   -0.1614     0.1382 -1.17 0.24281
## yschool       0.0995     0.0228  4.37 1.3e-05
## ybirth       -0.0728     0.0192 -3.78 0.00015
## Log(scale)    0.1367     0.0236  5.78 7.4e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3380.5   Loglik(intercept only)= -3402.8
##  Chisq= 44.61 on 7 degrees of freedom, p= 1.6e-07
## Number of Newton-Raphson Iterations: 3
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + poverty + smoke + yschool +
##     ybirth, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 145.7593    38.0517  3.83 0.00013
## raceblack    -0.1799     0.1187 -1.52 0.12948
## raceother    -0.2834     0.1095 -2.59 0.00965
## povertyyes    0.1801     0.1027  1.75 0.07954
## smokeyes     -0.2357     0.0884 -2.67 0.00765
## yschool       0.0976     0.0227  4.29 1.8e-05
## ybirth       -0.0730     0.0192 -3.79 0.00015
## Log(scale)    0.1374     0.0236  5.81 6.2e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3381.2   Loglik(intercept only)= -3402.8
##  Chisq= 43.25 on 6 degrees of freedom, p= 1e-07
## Number of Newton-Raphson Iterations: 3
## n= 927

##
## Call:
## survreg(formula = SurvObj ~ race + smoke + yschool + ybirth,
##     data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 146.8739    38.1150  3.85 0.00012
## raceblack    -0.1485     0.1175 -1.26 0.20620
## raceother    -0.2717     0.1095 -2.48 0.01306
## smokeyes     -0.2252     0.0883 -2.55 0.01077
```

```
## yschool       0.0884     0.0222  3.99 6.7e-05
## ybirth       -0.0735     0.0193 -3.81 0.00014
## Log(scale)    0.1391     0.0236  5.88 4.0e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3382.7   Loglik(intercept only)= -3402.8
##  Chisq= 40.18 on 5 degrees of freedom, p= 1.4e-07
## Number of Newton-Raphson Iterations: 3
## n= 927


##
## Call:
## survreg(formula = SurvObj ~ smoke + yschool + ybirth, data = bfeed,
##     dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 149.7212    38.2083  3.92 8.9e-05
## smokeyes     -0.1664     0.0856 -1.94  0.0521
## yschool       0.0993     0.0218  4.56 5.0e-06
## ybirth       -0.0750     0.0193 -3.88  0.0001
## Log(scale)    0.1427     0.0236  6.04 1.6e-09
##
## Scale= 1.15
##
## Log Normal distribution
## Loglik(model)= -3386.1   Loglik(intercept only)= -3402.8
##  Chisq= 33.39 on 3 degrees of freedom, p= 2.7e-07
## Number of Newton-Raphson Iterations: 3
## n= 927


##
## Call:
## survreg(formula = SurvObj ~ yschool + ybirth, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 149.4927    38.2869  3.90 9.4e-05
## yschool       0.1072     0.0214  5.01 5.5e-07
## ybirth       -0.0750     0.0194 -3.87 0.00011
## Log(scale)    0.1448     0.0236  6.13 9.1e-10
##
## Scale= 1.16
##
## Log Normal distribution
## Loglik(model)= -3388   Loglik(intercept only)= -3402.8
##  Chisq= 29.63 on 2 degrees of freedom, p= 3.7e-07
## Number of Newton-Raphson Iterations: 3
## n= 927


##
## Call:
## survreg(formula = SurvObj ~ ybirth, data = bfeed, dist = "lognormal")
##                Value Std. Error     z       p
## (Intercept) 82.4388    36.3920  2.27   0.023
```

```
## ybirth      -0.0405     0.0184 -2.20   0.028
## Log(scale)   0.1580     0.0237  6.68 2.4e-11
##
## Scale= 1.17
##
## Log Normal distribution
## Loglik(model)= -3400.4   Loglik(intercept only)= -3402.8
##  Chisq= 4.83 on 1 degrees of freedom, p= 0.028
## Number of Newton-Raphson Iterations: 2
## n= 927


## # ------------------------------------------------------------------------------------------------------------
## # Initial Model:
## Call:
## coxph(formula = formula, data = data, method = "efron")
##
##   n= 927, number of events= 892
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## smoke 0.22702   1.25486  0.07374 3.079  0.00208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## smoke     1.255     0.7969     1.086      1.45
##
## Concordance= 0.526  (se = 0.009 )
## Likelihood ratio test= 9.18  on 1 df,   p=0.002
## Wald test            = 9.48  on 1 df,   p=0.002
## Score (logrank) test = 9.52  on 1 df,   p=0.002
##
## # ------------------------------------------------------------------------------------------------------------
## ### iter num = 1, Forward Selection by LR Test: + race
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race, data = data,
##     method = "efron")
##
##   n= 927, number of events= 892
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## smoke 0.27943   1.32237  0.07552 3.700 0.000216 ***
## race  0.15893   1.17226  0.04550 3.493 0.000478 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## smoke     1.322     0.7562     1.140     1.533
## race      1.172     0.8531     1.072     1.282
##
## Concordance= 0.551  (se = 0.011 )
## Likelihood ratio test= 20.82  on 2 df,   p=3e-05
## Wald test            = 21.3  on 2 df,   p=2e-05
## Score (logrank) test = 21.4  on 2 df,   p=2e-05
##
```

```
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##    smoke     race
## 1.022593 1.022593
## # ------------------------------------------------------------------------------------------------
## ### iter num = 2, Forward Selection by LR Test: + ybirth
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth,
##     data = data, method = "efron")
##
##   n= 927, number of events= 892
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## smoke  0.29709   1.34594  0.07574 3.922 8.77e-05 ***
## race   0.16719   1.18198  0.04558 3.668 0.000245 ***
## ybirth 0.05569   1.05727  0.01680 3.315 0.000917 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## smoke      1.346     0.7430     1.160     1.561
## race       1.182     0.8460     1.081     1.292
## ybirth     1.057     0.9458     1.023     1.093
##
## Concordance= 0.565  (se = 0.012 )
## Likelihood ratio test= 31.87  on 3 df,   p=6e-07
## Wald test            = 32.19  on 3 df,   p=5e-07
## Score (logrank) test = 32.33  on 3 df,   p=4e-07
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##    smoke     race   ybirth
## 1.044828 1.044811 1.001262
## # ------------------------------------------------------------------------------------------------
## ### iter num = 3, Forward Selection by LR Test: + yschool
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool, data = data, method = "efron")
##
##   n= 927, number of events= 892
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## smoke    0.24504   1.27768  0.07837  3.127  0.00177 **
## race     0.14201   1.15258  0.04664  3.045  0.00233 **
## ybirth   0.07172   1.07435  0.01790  4.007 6.14e-05 ***
## yschool -0.04985   0.95137  0.01908 -2.613  0.00898 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## smoke      1.2777     0.7827    1.0957    1.4898
## race       1.1526     0.8676    1.0519    1.2629
## ybirth     1.0744     0.9308    1.0373    1.1127
## yschool    0.9514     1.0511    0.9165    0.9876
```

```
##
## Concordance= 0.575  (se = 0.012 )
## Likelihood ratio test= 38.7  on 4 df,   p=8e-08
## Wald test            = 39.05  on 4 df,   p=7e-08
## Score (logrank) test = 39.19  on 4 df,   p=6e-08
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##    smoke     race   ybirth  yschool
## 1.182042 1.059091 1.002535 1.131705
## # ------------------------------------------------------------------------------------------------
## ### iter num = 4, Forward Selection by LR Test: + poverty
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool + poverty, data = data, method = "efron")
##
##   n= 927, number of events= 892
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## smoke    0.25749   1.29368  0.07844  3.282  0.00103 **
## race     0.15221   1.16440  0.04675  3.256  0.00113 **
## ybirth   0.07068   1.07324  0.01789  3.951 7.78e-05 ***
## yschool -0.06232   0.93958  0.02002 -3.113  0.00185 **
## poverty -0.19924   0.81935  0.09228 -2.159  0.03085 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## smoke      1.2937     0.7730    1.1093    1.5087
## race       1.1644     0.8588    1.0625    1.2761
## ybirth     1.0732     0.9318    1.0363    1.1115
## yschool    0.9396     1.0643    0.9034    0.9772
## poverty    0.8194     1.2205    0.6838    0.9818
##
## Concordance= 0.576  (se = 0.012 )
## Likelihood ratio test= 43.51  on 5 df,   p=3e-08
## Wald test            = 43.75  on 5 df,   p=3e-08
## Score (logrank) test = 43.83  on 5 df,   p=3e-08
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##    smoke     race   ybirth  yschool  poverty
## 1.206682 1.072935 1.005629 1.129337 1.042396
## # ================================================================================================
## *** Stepwise Final Model (in.lr.test: sle = 0.15; out.lr.test: sls = 0.15; variable selection restri
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool + poverty, data = data, method = "efron")
##
##   n= 927, number of events= 892
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## smoke    0.25749   1.29368  0.07844  3.282  0.00103 **
## race     0.15221   1.16440  0.04675  3.256  0.00113 **
```

```
## ybirth   0.07068   1.07324  0.01789  3.951 7.78e-05 ***
## yschool -0.06232   0.93958  0.02002 -3.113  0.00185 **
## poverty -0.19924   0.81935  0.09228 -2.159  0.03085 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## smoke      1.2937     0.7730    1.1093    1.5087
## race       1.1644     0.8588    1.0625    1.2761
## ybirth     1.0732     0.9318    1.0363    1.1115
## yschool    0.9396     1.0643    0.9034    0.9772
## poverty    0.8194     1.2205    0.6838    0.9818
##
## Concordance= 0.576  (se = 0.012 )
## Likelihood ratio test= 43.51  on 5 df,   p=3e-08
## Wald test            = 43.75  on 5 df,   p=3e-08
## Score (logrank) test = 43.83  on 5 df,   p=3e-08
##
## --------------- Variance Inflating Factor (VIF) ---------------
## Multicollinearity Problem: Variance Inflating Factor (VIF) is bigger than 10 (Continuous Variable) o
##    smoke     race   ybirth  yschool  poverty
## 1.206682 1.072935 1.005629 1.129337 1.042396
```

## Cox Proportional Hazards model of time-independent variables

Cox model approach is a semi-parametric model useful with fixed-time covariates

here is the cox model of the entire bfeed dataset:

```
m1 <- coxph(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth +
              yschool, data = bfeed)
summary(m1)
```

```
## Call:
## coxph(formula = SurvObj ~ race + poverty + smoke + alcohol +
##     agemth + ybirth + yschool, data = bfeed)
##
##   n= 927, number of events= 892
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## raceblack   0.18000   1.19721  0.10503  1.714 0.086561 .
## raceother   0.29350   1.34112  0.09709  3.023 0.002504 **
## povertyyes -0.22222   0.80074  0.09364 -2.373 0.017638 *
## smokeyes    0.24435   1.27679  0.07948  3.074 0.002109 **
## alcoholyes  0.15937   1.17277  0.12297  1.296 0.194979
## agemth     -0.01551   0.98461  0.01881 -0.824 0.409676
## ybirth      0.07864   1.08182  0.02036  3.863 0.000112 ***
## yschool    -0.05668   0.94490  0.02309 -2.455 0.014103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
```

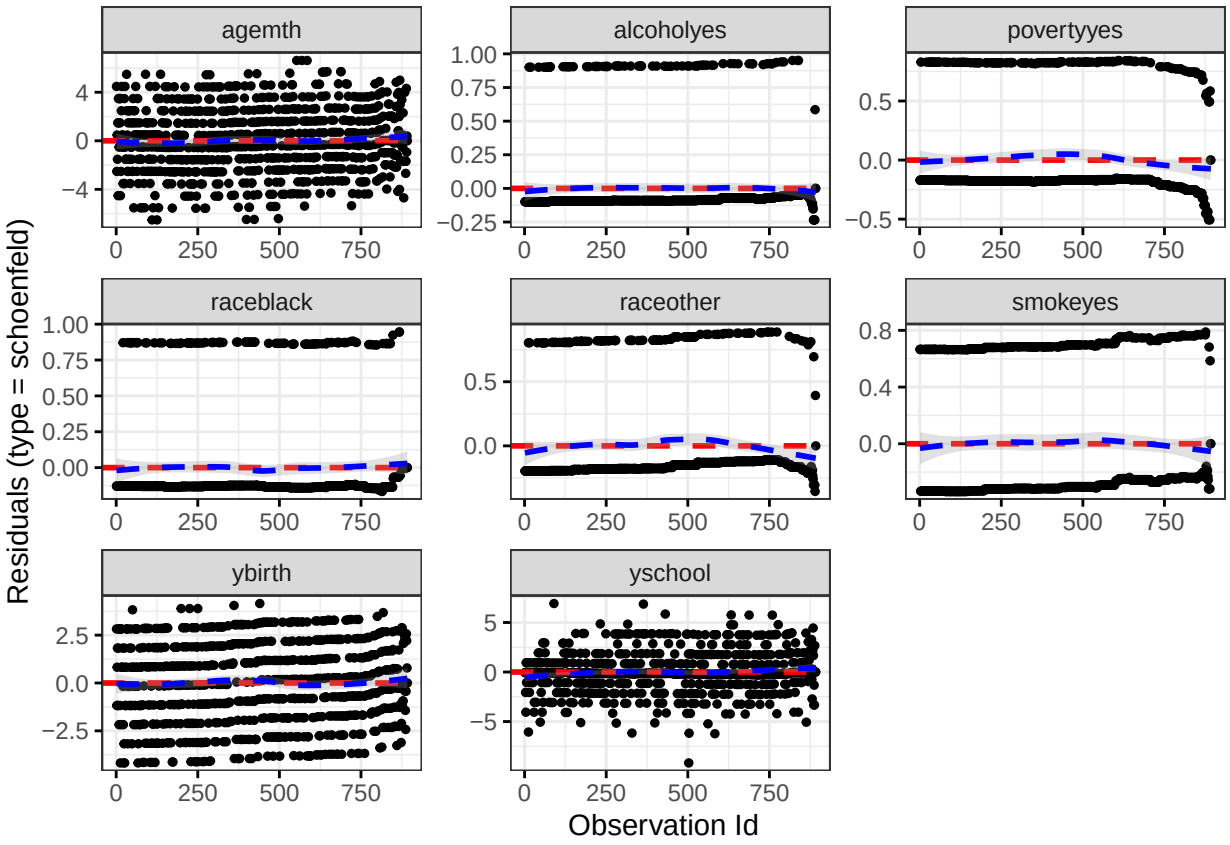
```
## raceblack     1.1972     0.8353    0.9745    1.4709
## raceother     1.3411     0.7456    1.1087    1.6222
## povertyyes    0.8007     1.2488    0.6665    0.9620
## smokeyes      1.2768     0.7832    1.0926    1.4920
## alcoholyes    1.1728     0.8527    0.9216    1.4924
## agemth        0.9846     1.0156    0.9490    1.0216
## ybirth        1.0818     0.9244    1.0395    1.1259
## yschool       0.9449     1.0583    0.9031    0.9886
##
## Concordance= 0.577  (se = 0.012 )
## Likelihood ratio test= 45.97  on 8 df,   p=2e-07
## Wald test            = 46.29  on 8 df,   p=2e-07
## Score (logrank) test = 46.33  on 8 df,   p=2e-07
```

```
# termplot(m1) # useful but long in rmd file

cox.zph(m1)
```

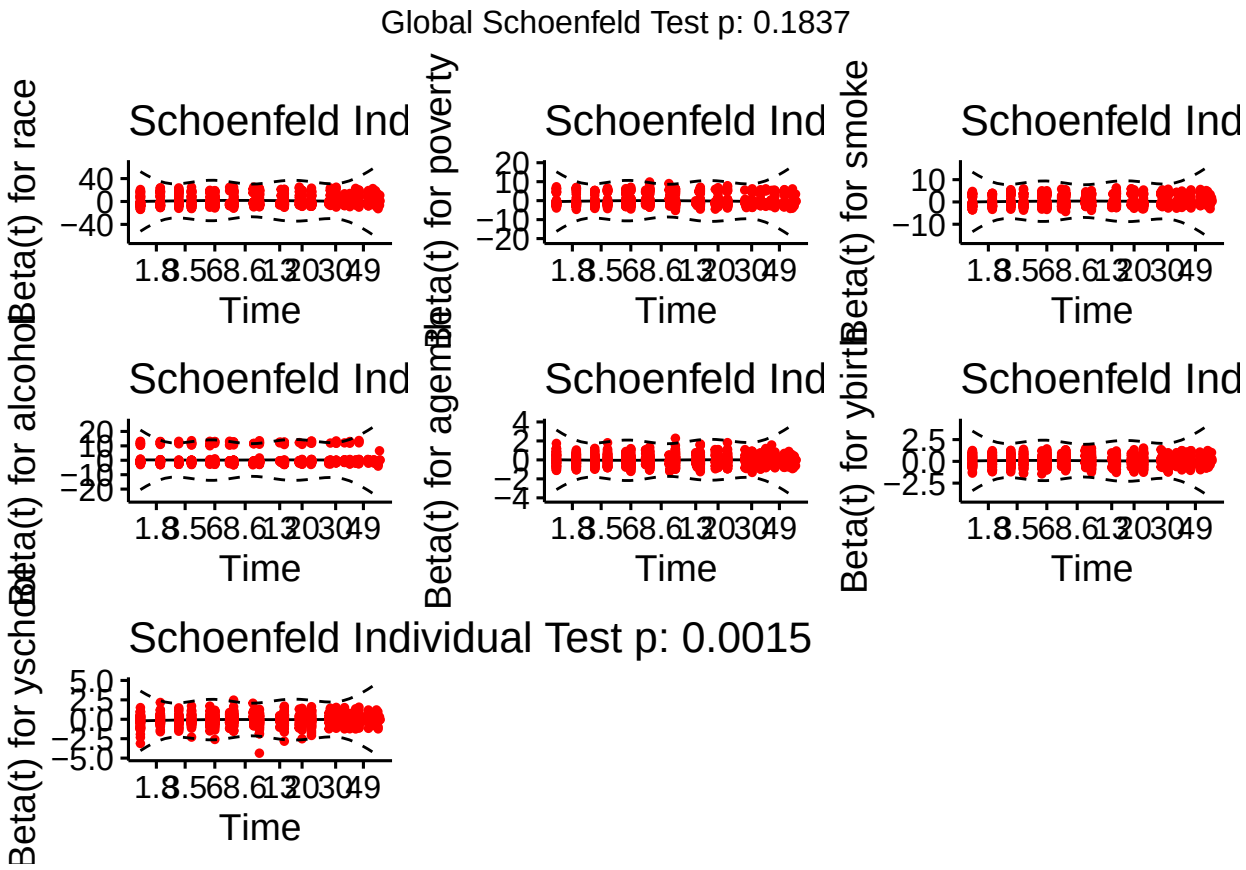
```
##          chisq df      p
## race     1.947  2 0.3778
## poverty  2.731  1 0.0984
## smoke    0.165  1 0.6849
## alcohol  0.049  1 0.8249
## agemth   3.692  1 0.0547
## ybirth   0.758  1 0.3838
## yschool 10.103  1 0.0015
## GLOBAL  11.329  8 0.1837
```

```
ggcoxdiagnostics(m1, type = "schoenfeld")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#alternative code to above line  (I think)
ggcoxzph(cox.zph(m1))
```

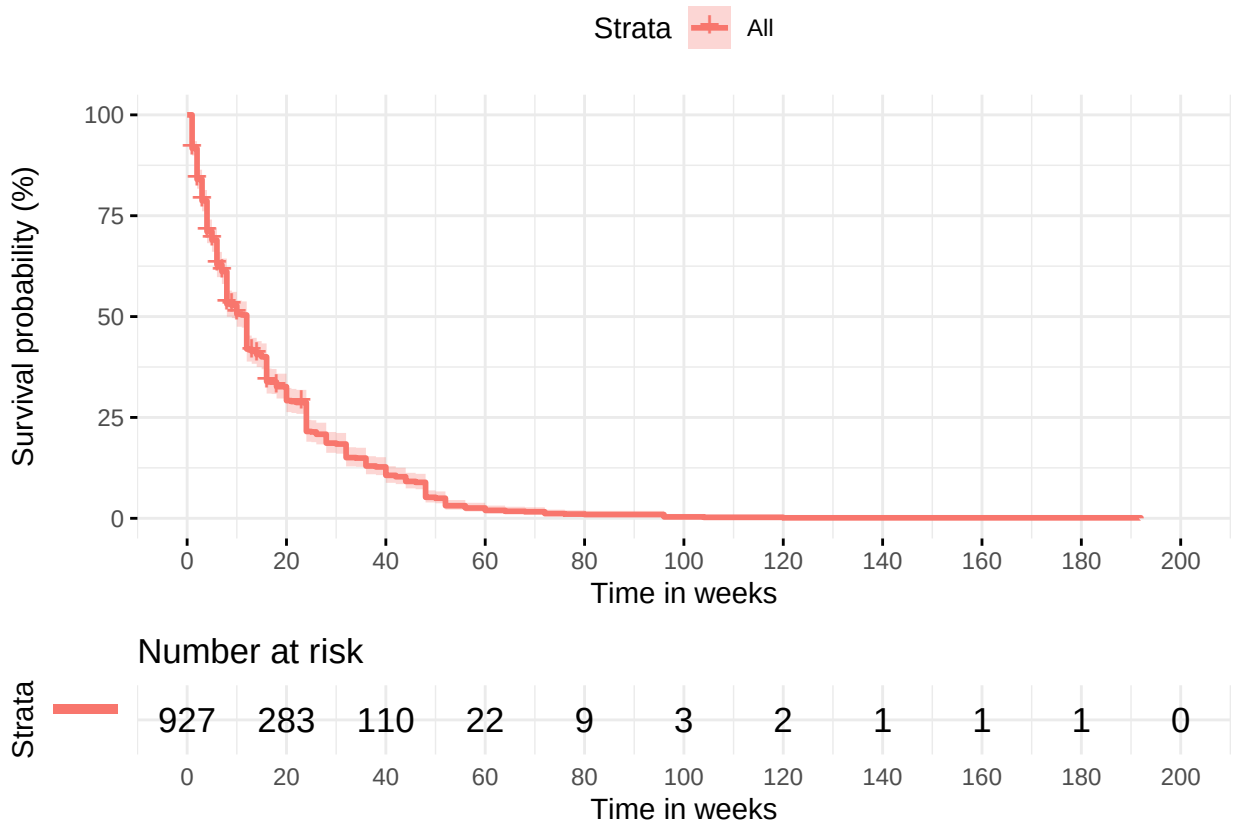



### Cross-Group Comparisons

Kaplan Meier Curve of entire dataset combined:

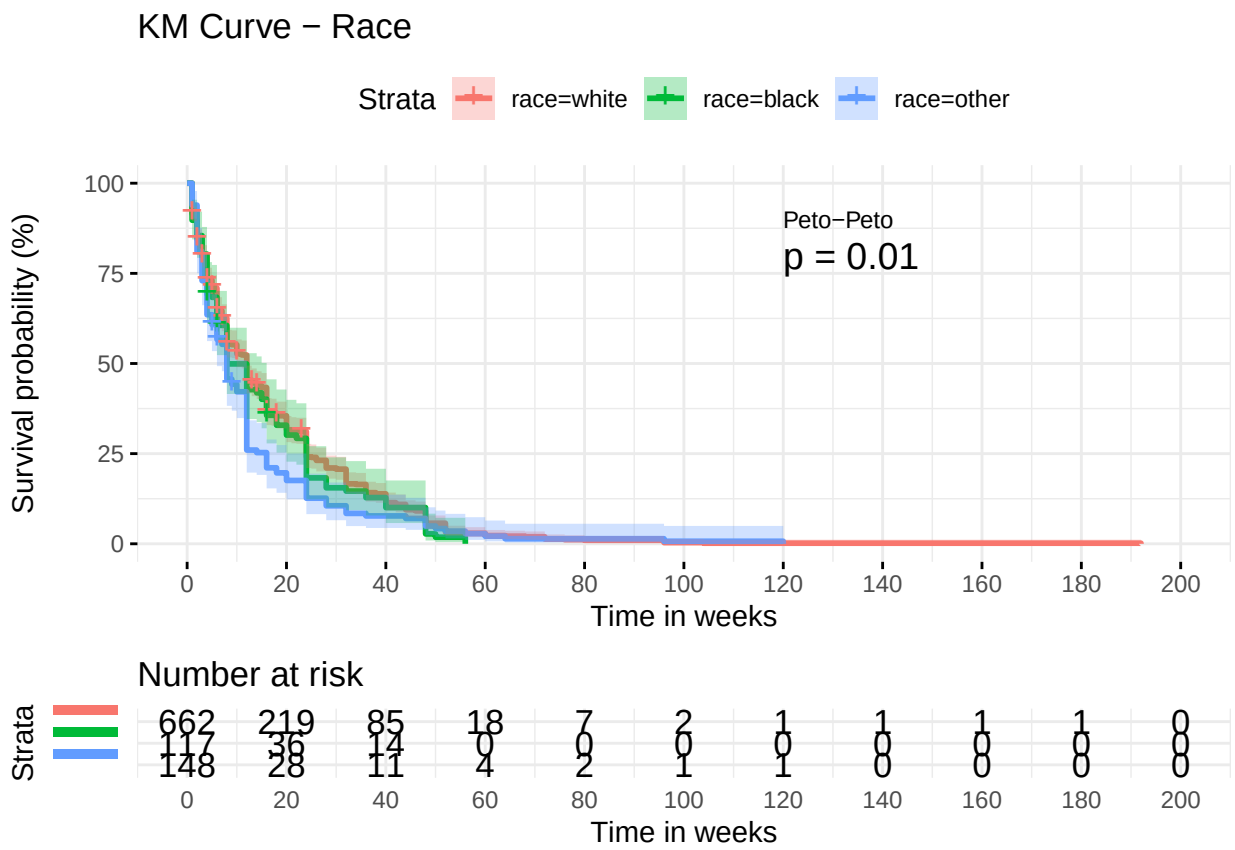
```
# create survival object:
km.as.one <- survfit(SurvObj ~ 1, data = bfeed)
# summary(km.as.one)

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.by.ybirth <- survfit(SurvObj ~ ybirth, 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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  # title = "KM Curve - Duration of Breast Feeding",
  fun = "pct"                 #show survival function as percentage
)
```

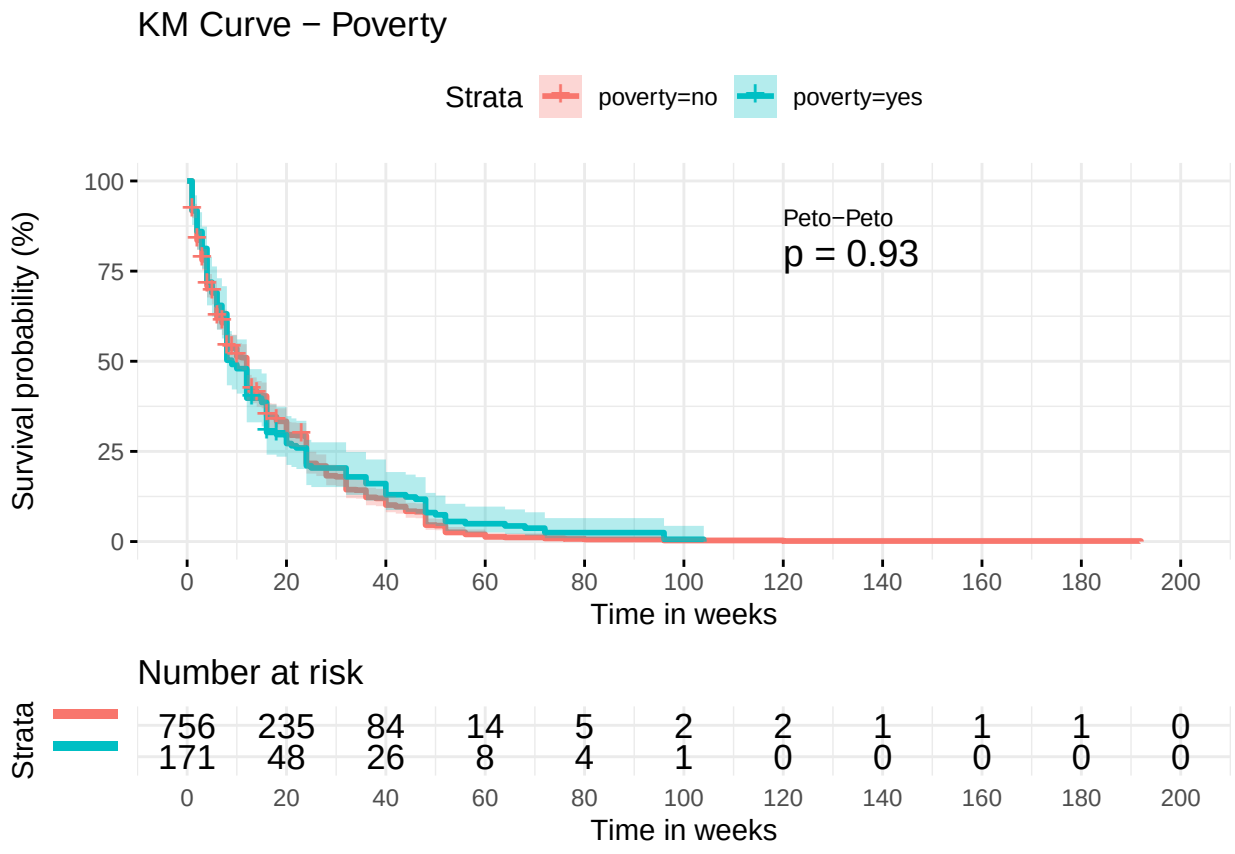



```
#KM curve according to race
ggsurvplot(
```

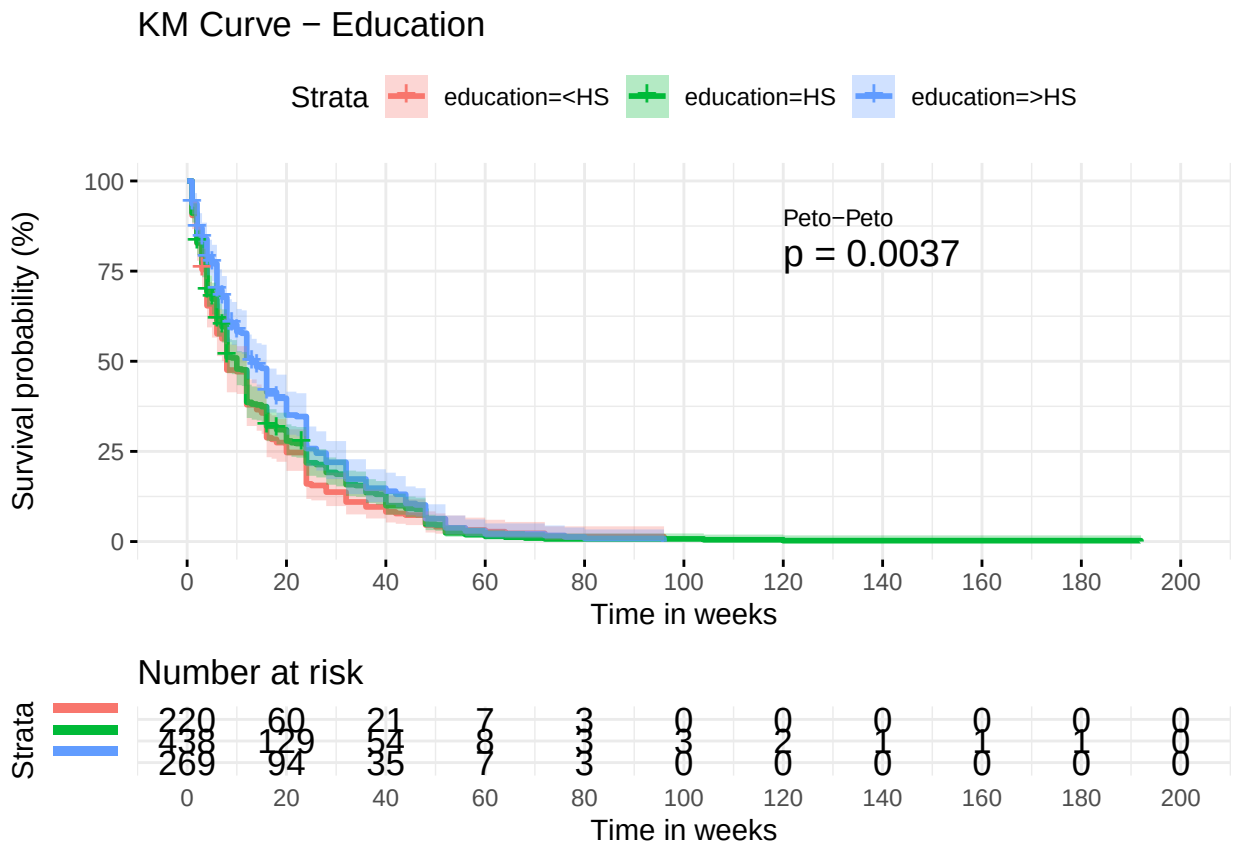
```
  km.by.race,                 # survfit object with calculated statistics.
  data = bfeed,               # data used to fit survival curves.
  risk.table = TRUE,          # show risk table.
  pval = TRUE,                # show p-value of log-rank test.
  pval.coord = c(120,80),     # location of pval
  pval.method = TRUE,          # show type of pval shown
  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(120,90), # location of p-value text
 pval.method.size = 3,       # p-val text size
 title = "KM Curve - Race",
 fun = "pct"                 #show survival function as percentage
)
```



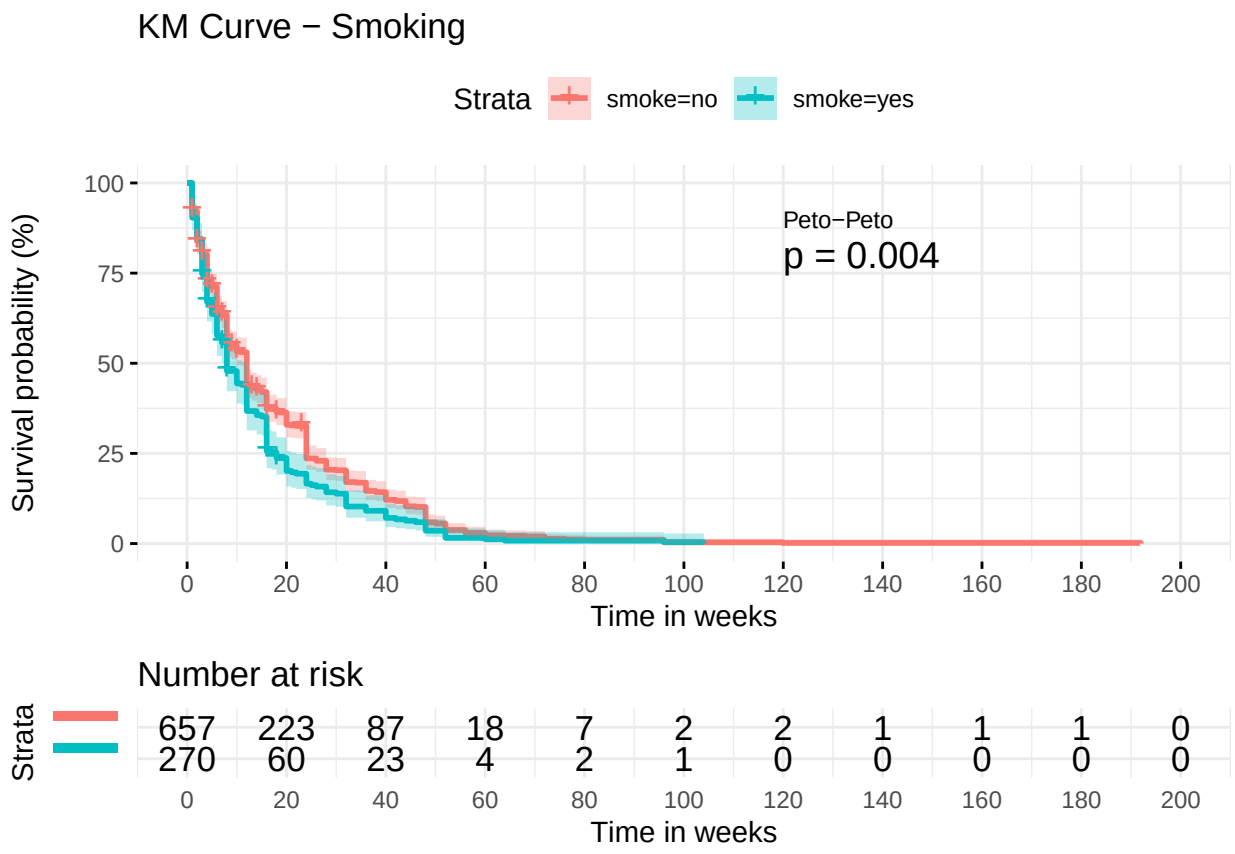
```
#KM curve according to poverty
ggsurvplot(
   km.by.poverty,                  # survfit object with calculated statistics.
   data = bfeed,                 # data used to fit survival curves.
   risk.table = TRUE,            # show risk table.
   pval = TRUE,                  # show p-value of log-rank test.
   pval.coord = c(120,80),       # location of pval
   pval.method = TRUE,           # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,         # p-val text size
  title = "KM Curve - Poverty",
  fun = "pct"                   #show survival function as percentage
)
```



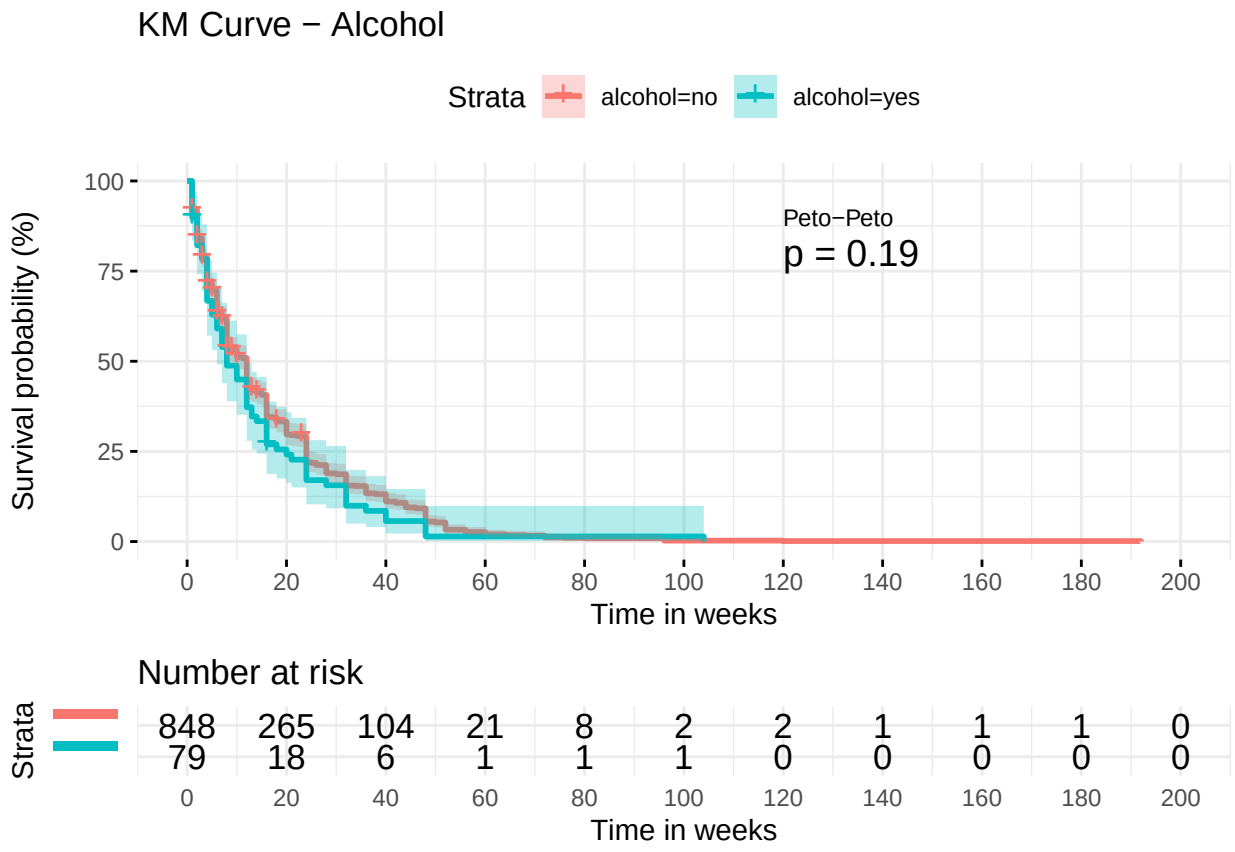
```r
#KM curve according to education
ggsurvplot(
   km.by.education,                  # survfit object with calculated statistics.
   data = bfeed,                  # data used to fit survival curves.
   risk.table = TRUE,             # show risk table.
   pval = TRUE,                   # show p-value of log-rank test.
   pval.coord = c(120,80),        # location of pval
   pval.method = TRUE,             # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,          # p-val text size
  title = "KM Curve - Education",
  fun = "pct"                    #show survival function as percentage
)
```



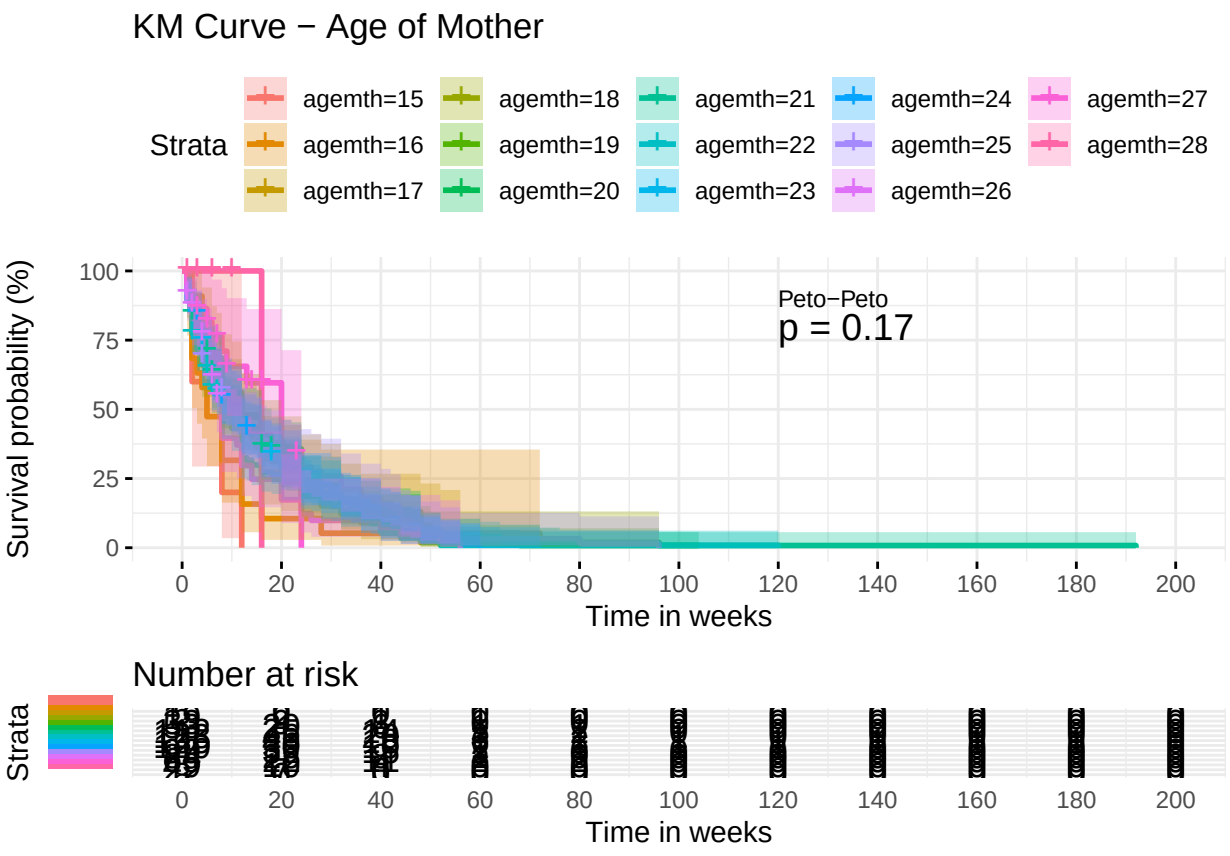
```
#KM curve according to smoking
ggsurvplot(
   km.by.smoke,                 # survfit object with calculated statistics.
   data = bfeed,               # data used to fit survival curves.
   risk.table = TRUE,          # show risk table.
   pval = TRUE,                # show p-value of log-rank test.
   pval.coord = c(120,80),     # location of pval
   pval.method = TRUE,          # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Smoking",
  fun = "pct"                 #show survival function as percentage
)
```



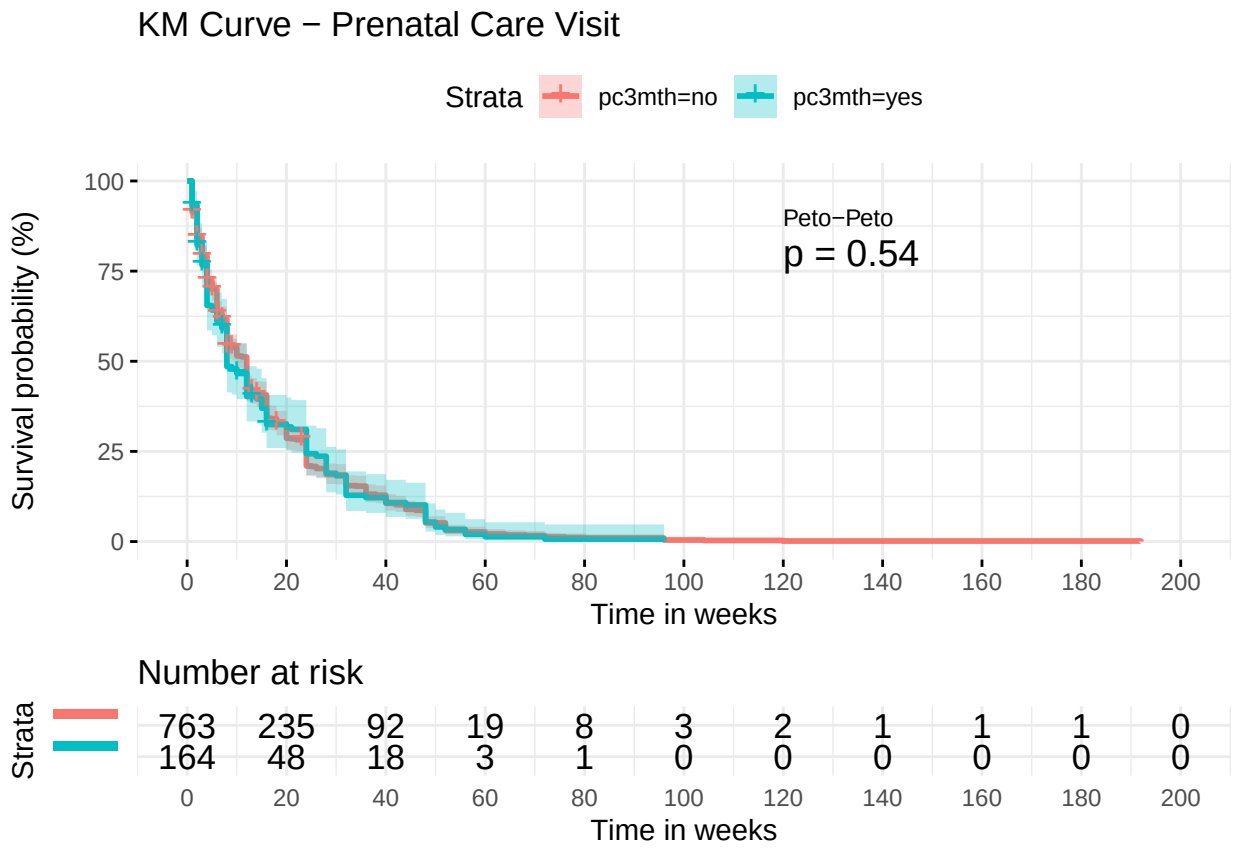
```
#KM curve according to alcohol
ggsurvplot(
   km.by.alcohol,                  # survfit object with calculated statistics.
   data = bfeed,                 # data used to fit survival curves.
   risk.table = TRUE,            # show risk table.
   pval = TRUE,                  # show p-value of log-rank test.
   pval.coord = c(120,80),       # location of pval
   pval.method = TRUE,           # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,         # p-val text size
  title = "KM Curve - Alcohol",
  fun = "pct"                   #show survival function as percentage
)
```



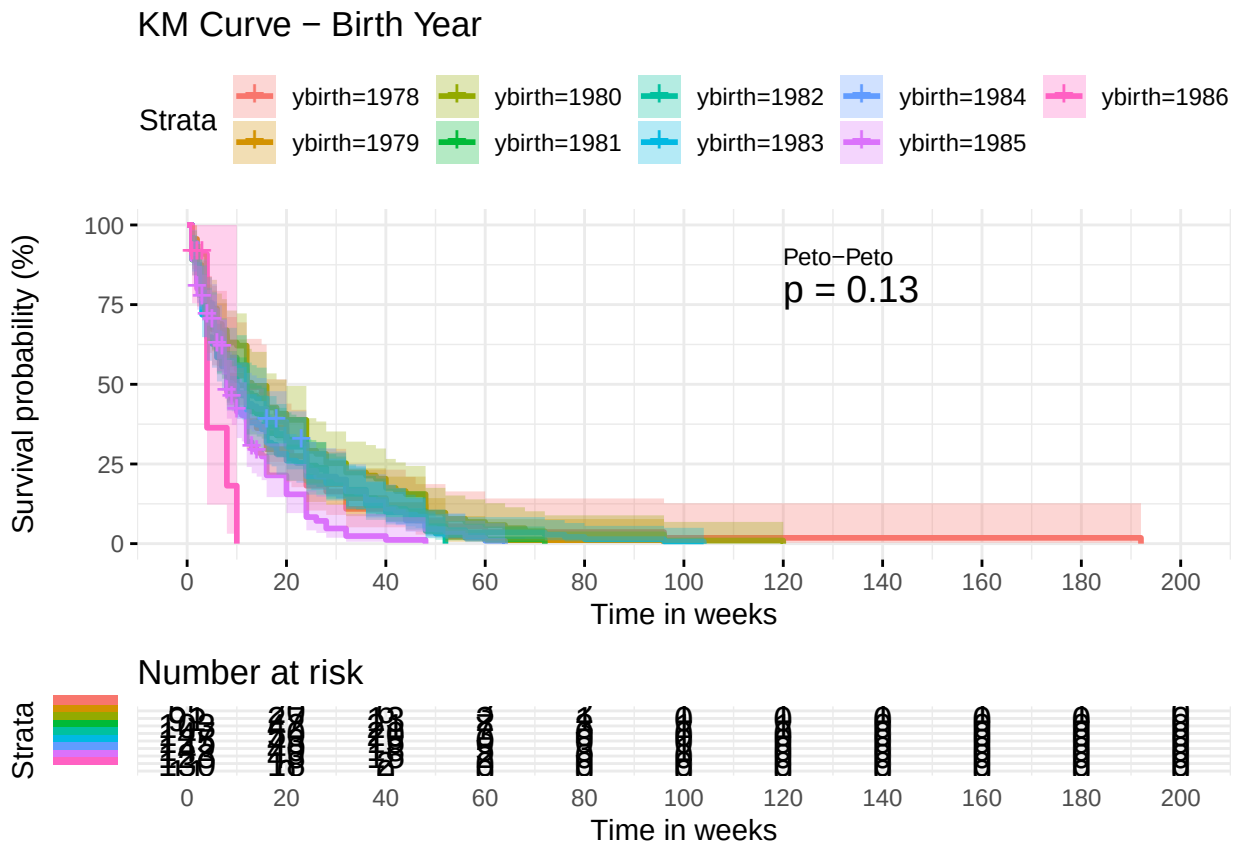
```r
#KM curve according to age of mother at birth of child
ggsurvplot(
   km.by.agemth,                 # survfit object with calculated statistics.
   data = bfeed,                # data used to fit survival curves.
   risk.table = TRUE,           # show risk table.
   pval = TRUE,                 # show p-value of log-rank test.
   pval.coord = c(120,80),      # location of pval
   pval.method = TRUE,           # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Age of Mother",
  fun = "pct"                 #show survival function as percentage
)
```



```r
#KM curve according to prenatal care after 3rd month
ggsurvplot(
   km.by.pc3mth,                 # surufit 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.
   pval.coord = c(120,80),      # location of pval
   pval.method = TRUE,           # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Prenatal Care Visit",
  fun = "pct"                 #show survival function as percentage
)
```



```r
#KM curve according to birth year
ggsurvplot(
   km.by.ybirth,                 # 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.
   pval.coord = c(120,80),      # location of pval
   pval.method = TRUE,           # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Birth Year",
  fun = "pct"                 #show survival function as percentage
)
```



<!–# 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 –>

## Supplemental Figures

Table S1: coefficients for PH exponential model

| | value |
|---|---|
| (Intercept) | -159.7446 |
| raceblack | 0.1733 |
| raceother | 0.3096 |
| smokeyes | 0.2669 |
| yschool | -0.0503 |
| ybirth | 0.0794 |

Table S2: coefficients for PH Weibull model

| | value |
|---|---|
| (Intercept) | -159.2647 |
| raceblack | 0.1728 |
| raceother | 0.3090 |
| smokeyes | 0.2663 |
| yschool | -0.0503 |
| ybirth | 0.0792 |

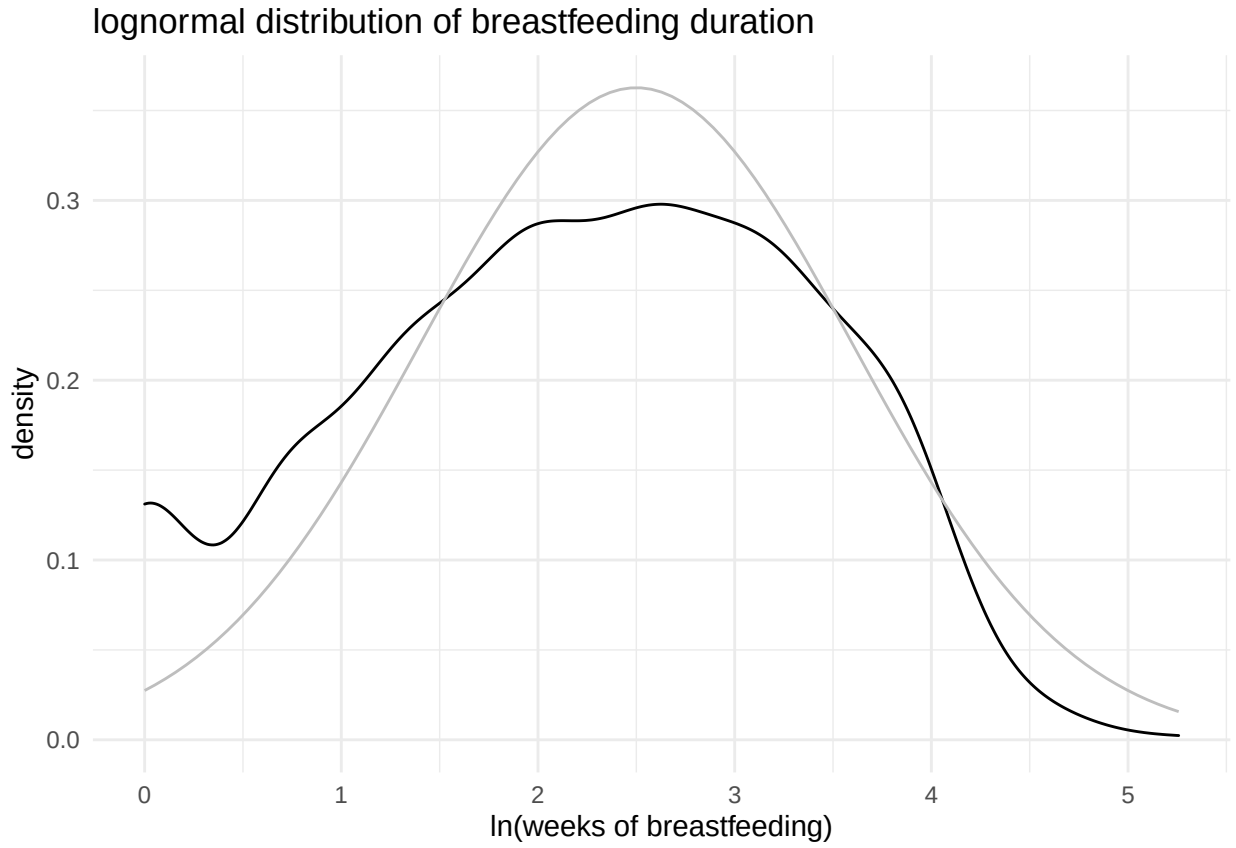



Figure S1: Distribution of natural log tranformed breastfeeding duration alongside normal distribution $N(2.5, 1.1)$ (grey)

### code used to create this report

```
# create survival object:
km.as.one <- survfit(SurvObj ~ 1, data = bfeed)
# 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.by.ybirth <- survfit(SurvObj ~ ybirth, 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(120,90), # location of p-value text
  pval.method.size = 3,      # p-val text size
  title = "KM Curve - Duration of Breast Feeding",
  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.
   pval.coord = c(120,80),   # location of pval
```

```
   pval.method = TRUE,         # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Race",
  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.
   pval.coord = c(120,80),    # location of pval
   pval.method = TRUE,         # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Poverty",
  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.
   pval.coord = c(120,80),    # location of pval
   pval.method = TRUE,         # show type of pval shown
```

```
   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(120,90), # location of p-value text
  pval.method.size = 3,      # p-val text size
  title = "KM Curve - Education",
  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.
   pval.coord = c(120,80),   # location of pval
   pval.method = TRUE,        # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,      # p-val text size
  title = "KM Curve - Smoking",
  fun = "pct"                #show survival function as percentage
)


#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.
   pval.coord = c(120,80),   # location of pval
```

```
   pval.method = TRUE,         # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Alcohol",
  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.
   pval.coord = c(120,80),    # location of pval
   pval.method = TRUE,         # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,       # p-val text size
  title = "KM Curve - Age of Mother",
  fun = "pct"                 #show survival function as percentage
)

#KM curve according to prenatal care after 3rd month
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.
   pval.coord = c(120,80),    # location of pval
   pval.method = TRUE,         # show type of pval shown
```

```
   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(120,90), # location of p-value text
  pval.method.size = 3,      # p-val text size
  title = "KM Curve - Prenatal Care Visit",
  fun = "pct"                #show survival function as percentage
)

#KM curve according to birth year
ggsurvplot(
   km.by.ybirth,               # 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.
   pval.coord = c(120,80),   # location of pval
   pval.method = TRUE,        # show type of pval shown
   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(120,90), # location of p-value text
  pval.method.size = 3,      # p-val text size
  title = "KM Curve - Birth Year",
  fun = "pct"                #show survival function as percentage
)
```

```
bfeed %>%
  ggplot(aes(x=ybirth, y=duration, color = delta)) +
  geom_point()+ geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

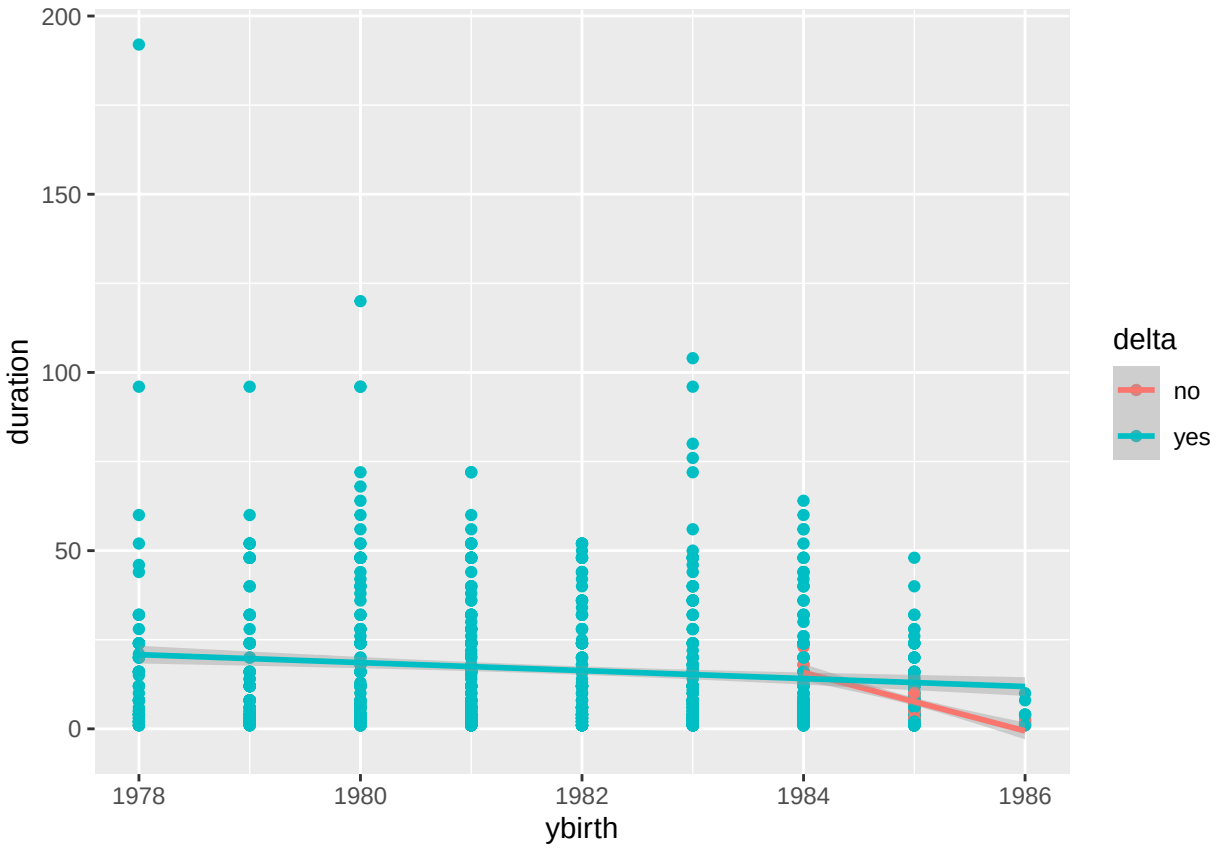



Figure S2: correlation between birth year and duration of breastfeeding