Survival Analysis of Breastfeeding Cessation

STAT 639V Survival Analysis Final Project

Carson Stacy

December 17, 2021

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.

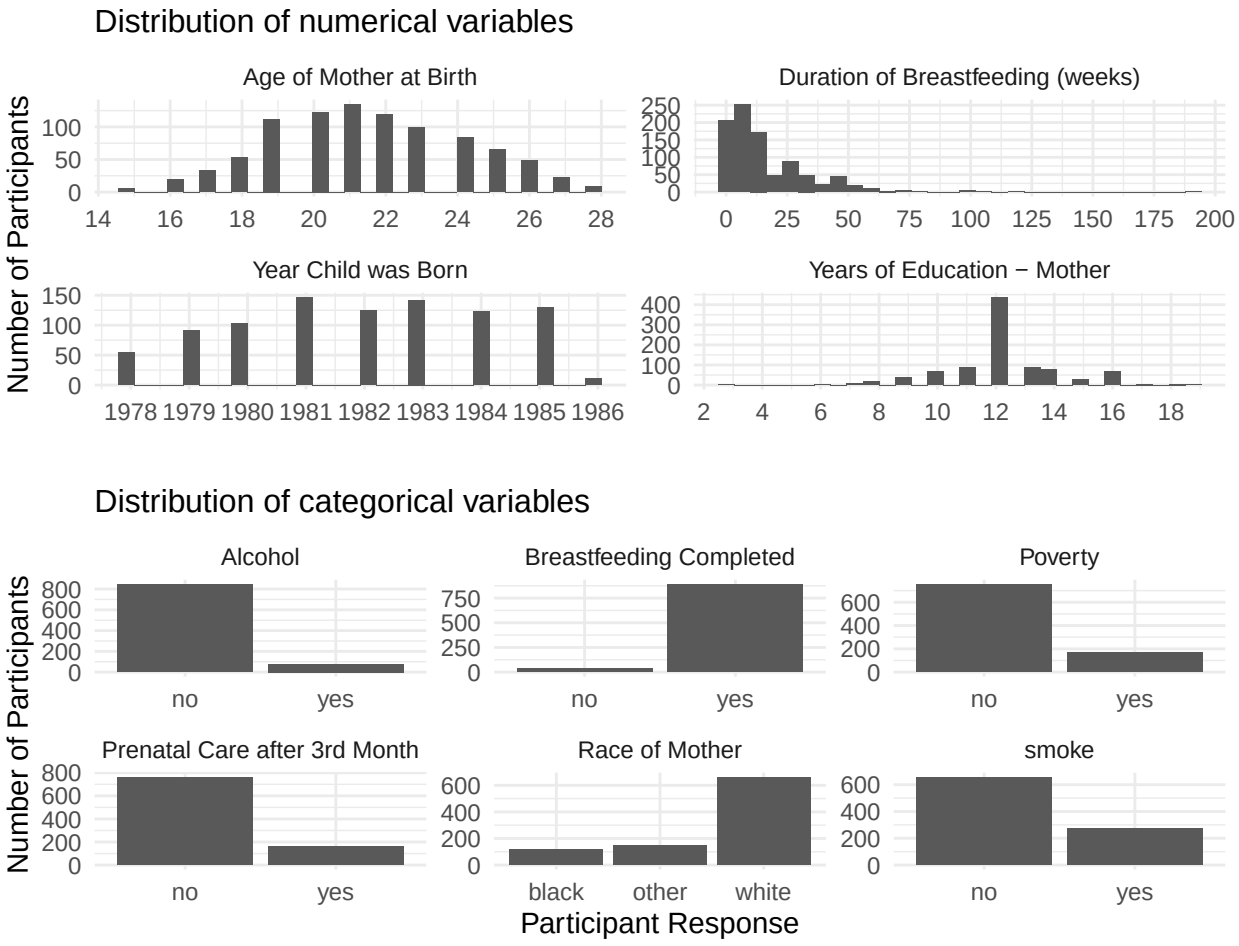

Figure 1: Distributions of categorical and numerical variables comprising the data set.

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. The predefined significance level α will be set at 0.05 for determining significance for this analysis unless otherwise specified. 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.

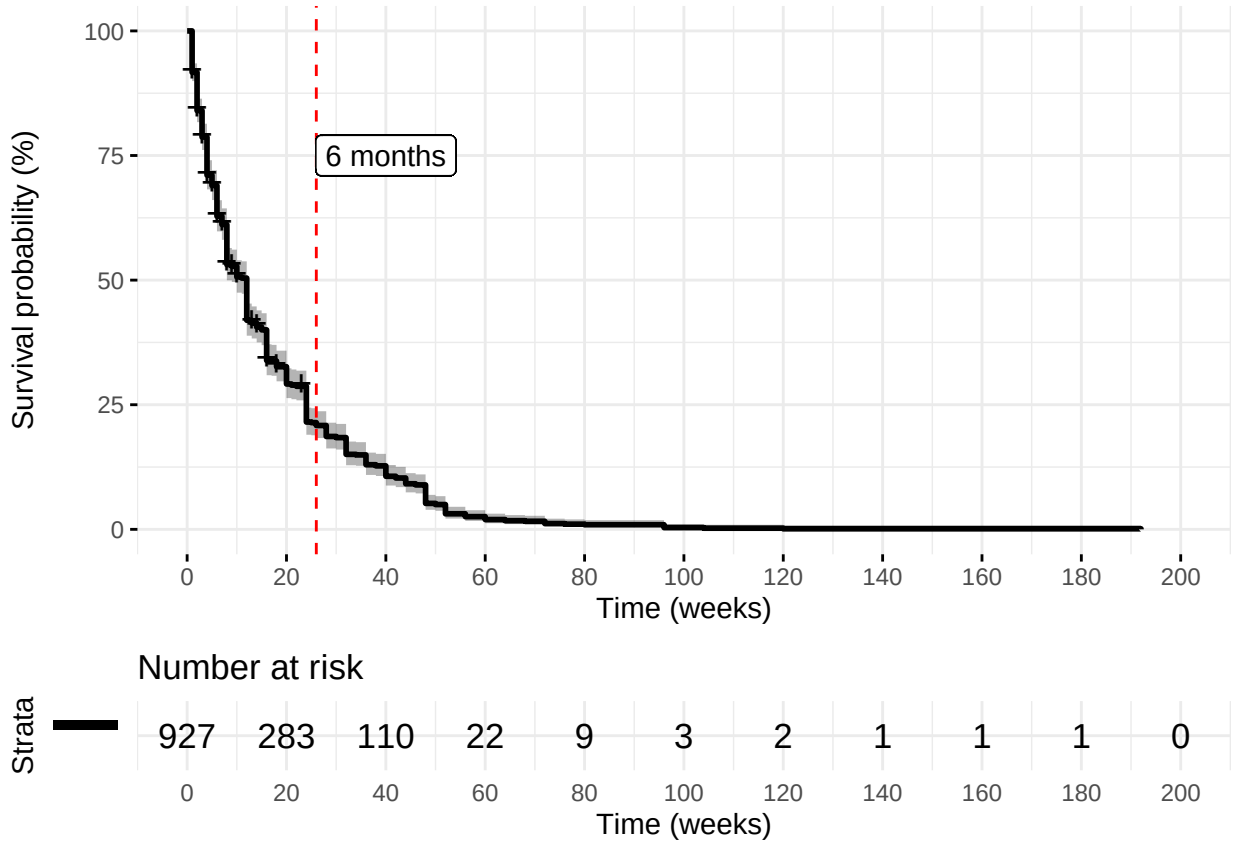

Figure 2: KM curve for all participants for duration of breastfeeding. Gray shaded region represents a pointwise 95% confidence interval for the survival curve. The symbol + corresponds to a censoring time. The number of participants still breastfeeding at a given number of weeks is shown below the survival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

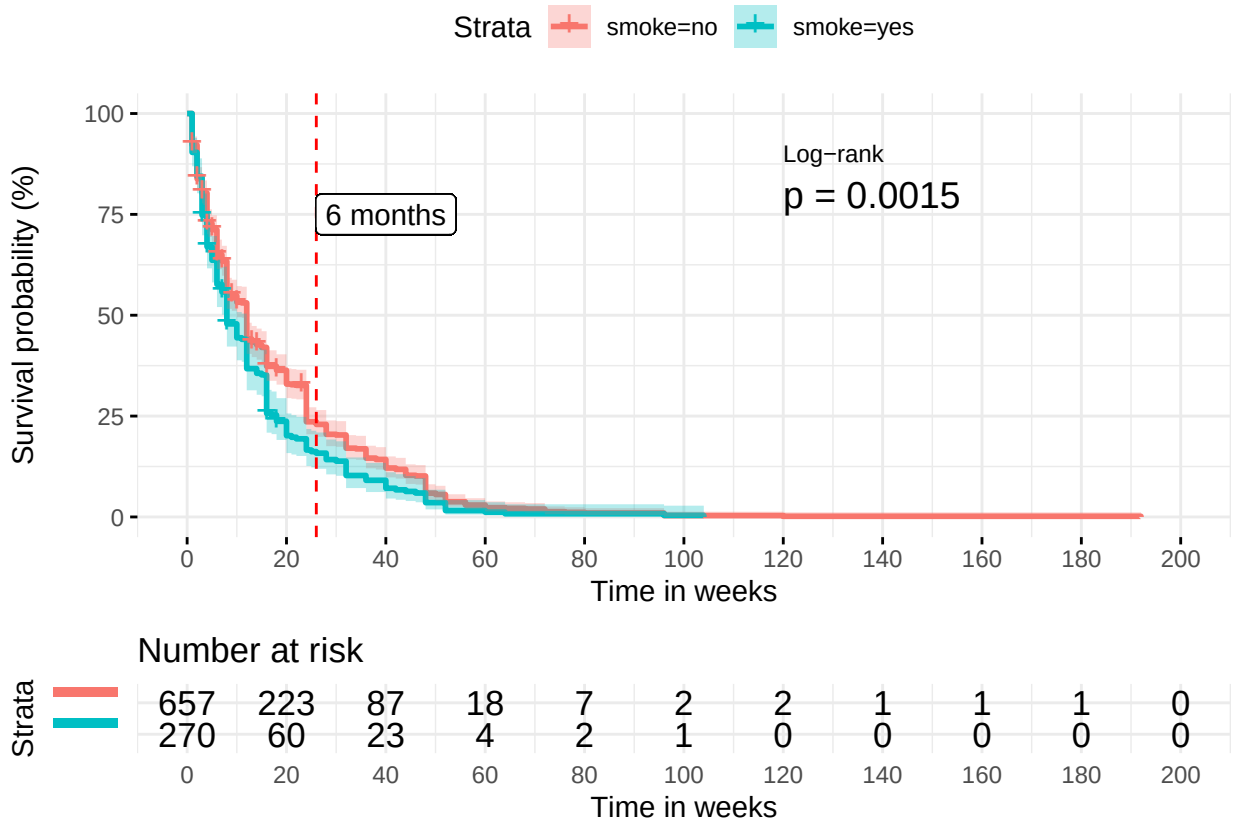

Figure 3: KM curve for duration of breastfeeding according to whether mother smoked when child was born. Shaded regions represents a pointwise 95% confidence interval for the survival curves. The symbol + corresponds to a censoring time. The number of participants still breastfeeding by group at a given 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 Figure S1.

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 λ.

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.

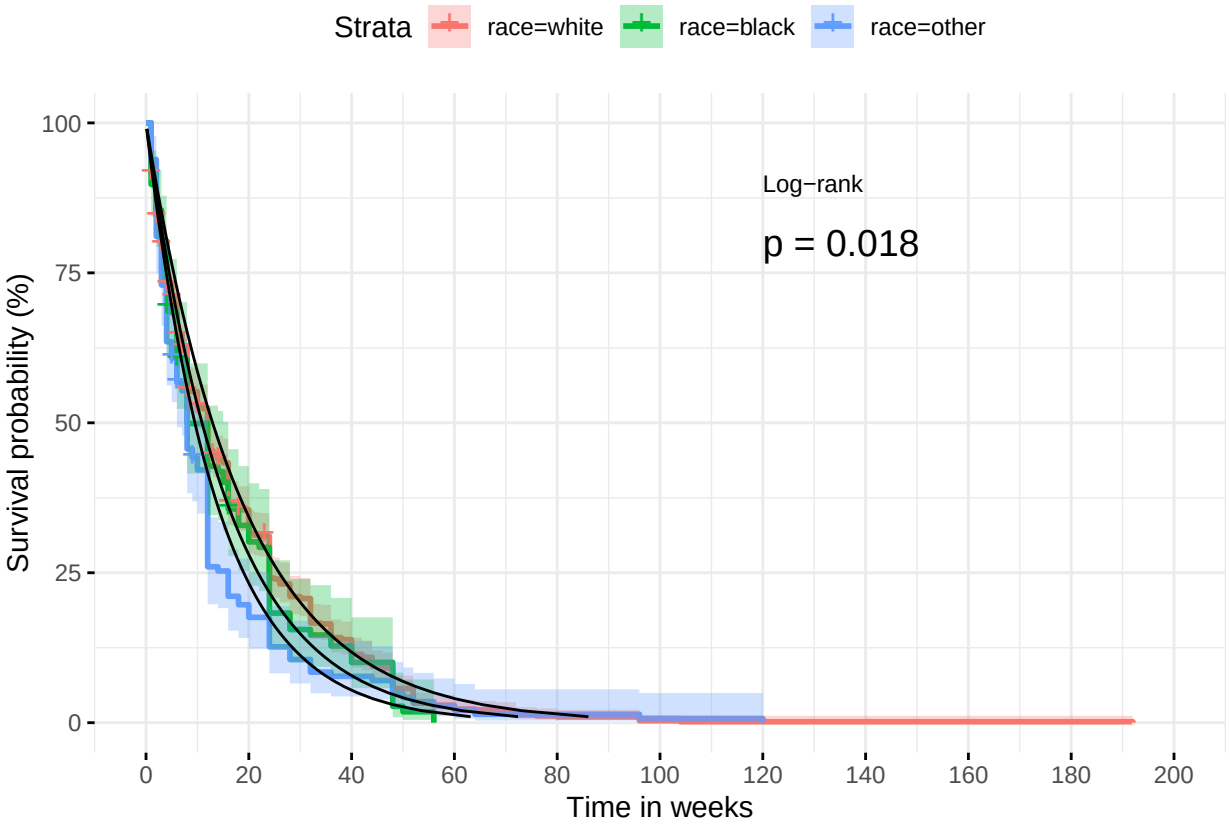

Figure 4: Comparing KM curve for duration of breastfeeding according to mothers' race with exponential regression model.

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, holding all other parameters constant. 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 α can be thought of as a baseline log-hazard, while λ 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.

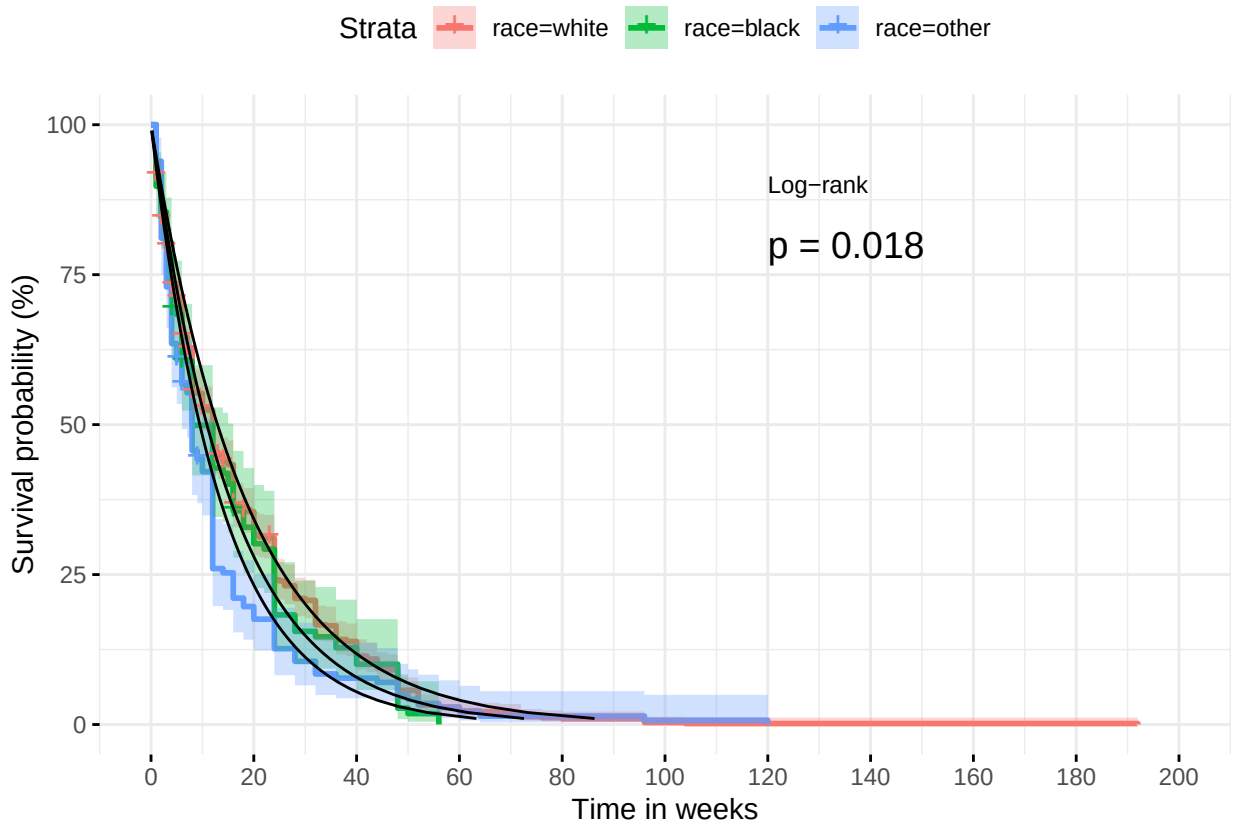

Figure 5: Comparing KM curve for duration of breastfeeding according to mothers' race with the Weibull regression model.

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 α 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 ϕ represents the probability density function of the normal distribution while Φ is the cumulative distribution function of the normal distribution. The distribution of $ln(X) \sim N(\mu, \sigma)$ with mean μ and standard deviation σ.

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 S2). As an AFT model, differences between groups should be able to be modeled as accelerated 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.

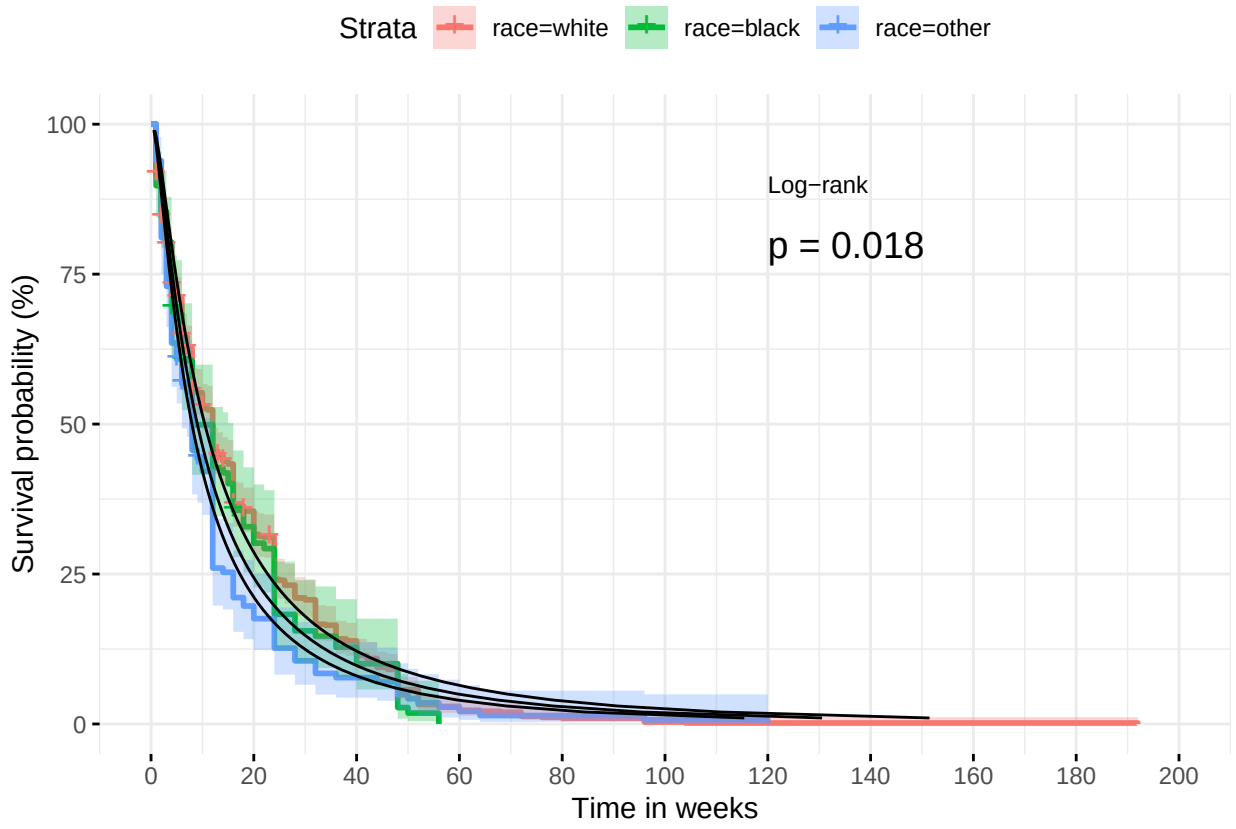

Figure 6: Comparing KM curve for duration of breastfeeding according to mothers' race with the Weibull regression model.

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 β 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 occurrence. An HR value close to zero means the factor increases time until the event.

Goodness of Fit

In order to optimize the Cox model fit, an iterative forward-backwards stepwise selection approach was utilized via the `My.stepwise.coxph` function of the `My.stepwise` package. The significance level for entry and significance level for stay were both set to 0.15 for the fitting procedure. The code used for the resulting best model from this process is shown below. This approach selected a model with five variables: race, smoking, years of education, year child was born, and poverty. This model has one more variable than the parametric models above. The newly included variable is whether the mother is considered living in poverty.

Code for stepwise selection of Cox PH model:

```
# create variable list for stepwise selection of variables in models:
my.variable.list <- c("race", "poverty", "smoke", "alcohol",
                      "agemth", "ybirth", "yschool", "pc3mth")

# run the stepwise selection
My.stepwise::My.stepwise.coxph(Time = "duration",
                               Status = "delta",
                               variable.list = my.variable.list,
                               data = bfeed,
                               sle = 0.15, sls = 0.15)
# best model:
"duration ~ smoke + race + ybirth + yschool + poverty"
```

The output of the `coxph` function from the `survival` package for the step-wise selected CoxPH model is shown below:

```
## Call:
## coxph(formula = SurvObj ~ race + poverty + smoke + ybirth + yschool,
##     data = bfeed)
##
##   n= 927, number of events= 892
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## raceblack   0.19218   1.21189  0.10433  1.842 0.065461 .
## raceother   0.29369   1.34137  0.09726  3.020 0.002530 **
## povertyyes -0.20286   0.81639  0.09265 -2.189 0.028562 *
## smokeyes    0.25855   1.29505  0.07848  3.295 0.000985 ***
## ybirth      0.07104   1.07362  0.01791  3.967 7.29e-05 ***
## yschool    -0.06316   0.93879  0.02014 -3.136 0.001711 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## raceblack     1.2119     0.8252    0.9878    1.4868
## raceother     1.3414     0.7455    1.1086    1.6231
## povertyyes    0.8164     1.2249    0.6808    0.9790
## smokeyes      1.2950     0.7722    1.1104    1.5104
## ybirth        1.0736     0.9314    1.0366    1.1120
## yschool       0.9388     1.0652    0.9025    0.9766
##
## Concordance= 0.575  (se = 0.012 )
## Likelihood ratio test= 43.69  on 6 df,   p=9e-08
## Wald test            = 43.8  on 6 df,   p=8e-08
## Score (logrank) test = 43.9  on 6 df,   p=8e-08
```

From this output, we see that the p-values `Pr(>|z|)` is significantly less than the threshold significance level of $\alpha = 0.05$ except for the indicator variable of race being black, which has a p-value of 0.065. The global statistical significance tests of the model show significant p-values ($p \leq 8 \times 10^{-8}$), indicating the overall test is a reasonable fit for the data.

A plot comparing the KM curve of the data to that predicted by this Cox PH model are shown below in Figure 7:

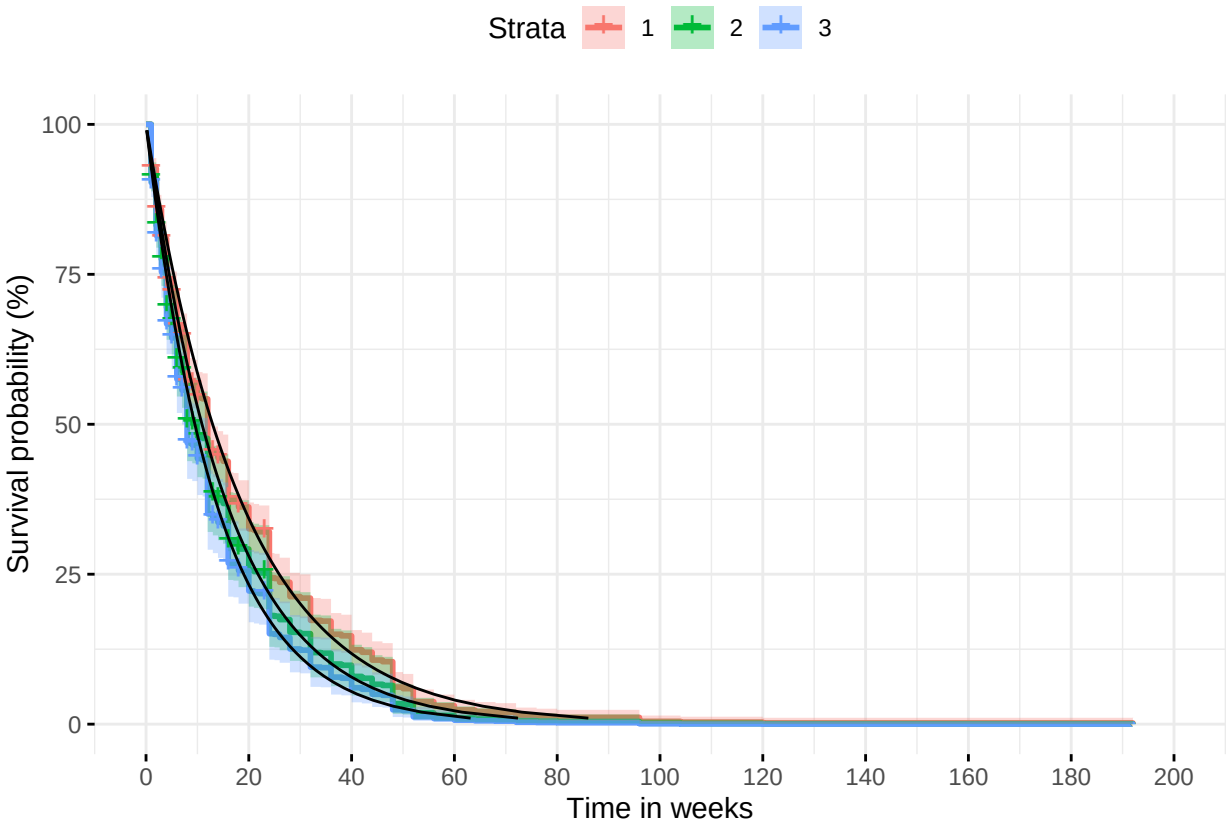

Figure 7: Comparing KM curve for duration of breastfeeding according to mothers' race with Cox PH model predictions.

Parameter Estimates

Hazard ratios are a useful tool with an intuitive interpretation. The estimated hazard ratios for each parameter either univariate or included in the final Cox PH model are shown in Table 11 below. These values are calculated by exponentiation of the coefficient estimates such as the `coef` column of the Cox model from above. The 95% confidence intervals and p-values for the coefficient estimates are also shown in parenthesis next to each HR estimate. Rows with no value correspond to either the baseline group for which the hazard ratio is compared for the given variable, or a variable that was exlcuded from the best Cox PH model. For continuous variables, instead the hazard ratio corresponding to an increase of one in the variable.

Table 11: Hazard Ratios (HR) for Cox PH model of time-independent variables

Duration	Group	Count (%)	HR (univariable)	HR (multivariable)
race	white	662 (71.4)	-	-
	black	117 (12.6)	1.12 (0.91-1.37, p=0.280)	1.21 (0.99-1.49, p=0.065)
	other	148 (16.0)	1.29 (1.08-1.55, p=0.006)	1.34 (1.11-1.62, p=0.003)
poverty	no	756 (81.6)	-	-
	yes	171 (18.4)	0.93 (0.78-1.10, p=0.379)	0.82 (0.68-0.98, p=0.029)
smoke	no	657 (70.9)	-	-
	yes	270 (29.1)	1.25 (1.09-1.45, p=0.002)	1.30 (1.11-1.51, p=0.001)
alcohol	no	848 (91.5)	-	-
	yes	79 (8.5)	1.18 (0.93-1.49, p=0.168)	-
agemth	Mean (SD)	21.5 (2.7)	0.99 (0.97-1.02, p=0.632)	-
ybirth	Mean (SD)	82.0 (2.1)	1.05 (1.02-1.09, p=0.003)	1.07 (1.04-1.11, p<0.001)
yschool	Mean (SD)	12.2 (1.9)	0.96 (0.92-0.99, p=0.009)	0.94 (0.90-0.98, p=0.002)
pc3mth	no	763 (82.3)	-	-
	yes	164 (17.7)	1.04 (0.87-1.23, p=0.690)	-

Among categorical predictor variables, the highest hazard ratio of the complete Cox PH model is the category of "other" in the `race` variable with a hazard ratio of $HR = 1.34$, $CI = 1.11, 1.52$. This hazard ratio means that, holding all other variables constant, a mother whose race is "other" has a 34% higher hazard rate than a white mother with a 95% confidence interval between 11% to 52% increased hazard.

For numerical variables included in the model, significant effects are predicted in response to changes in both years of education of the mother as well as the year the child was born. The hazard ratio for years of education was $HR = 0.94$, $CI = 0.90, 0.98$ which corresponds to approximately a 6% less likely to stop breastfeeding at a given time for each additional year of education the mother completes. Surprisingly, the year in which the child was born increased the hazard for early cessation of breastfeeding, with $HR = 1.07$, $CI = 1.04, 1.11$. This means that children born each year are 7% more likely to stop breastfeeding at a given time than a child born the year before.

Potential predictor variables excluded from the final model were: alcohol use, age of the mother, and prenatal care visit after the third month.

Assumptions of the Cox PH survival model

The Cox proportional hazards model has three important assumptions: (1) that survival times are independent between individuals in the study, (2) that the relationship between predictors and the hazard is multiplicative, and (3) that the hazard ratio is constant over time.

For the first assumption, it is reasonable to expect that the breastfeeding duration of random survey participants in this study are independent.

In order to test the assumption of proportional hazards, a test of proportional hazards was performed via the `cox.zph` function of the `survival` package in r. The results of this test are shown in Table 12. The H_0 of the p-value test is that the hazard for the variable is proprotional, so large p-values indicate the proportional hazards assumption is appropriate.

Table 12: Test of Proprtional Hazards Assumption of Cox Regression Model

	chisq	df	p
race	2.0635	2	0.3564
poverty	2.5446	1	0.1107
smoke	0.1131	1	0.7367
ybirth	0.9255	1	0.3360
yschool	10.0075	1	0.0016
GLOBAL	11.1034	6	0.0852

We see that the p-value for years of education is small, indicating the proportional hazards assumption for this variable is likely not appropriate. The global p-value for the Cox model is smaller than would be desired. The p-values for the remaining variables are appropriate for a PH assumption.

A graphical test of proportional hazards is shown below in Figure 8.

Discussion

Analysis of time-to-cessation of breastfeeding data is an approach with potential utility for characterizing patterns of breastfeeding behavior. Utilization of model approaches to exploring breastfeeding patterns from nonparametric, to semi-parametric and parametric approaches resulted in many different lenses with which to view a similar pattern: that race, smoking behaviors, and education of a mother correlate with differences in duration of breastfeeding. Furthermore, the downward trend in breastfeeding durations from year to year is important knowledge for public health officials (in the 1980's) to be aware of. The use of survival analysis techniques with this data allowed for the inclusion of participants who had not yet stopped breastfeeding when data was collected, while other approaches would likely not include this information as part of the analysis.

This analysis was an interesting exercise in the use of survival analysis; however, it is not perfect. Although residual plots of the Cox model appear reasonable, the test for proportional hazards in table 12 indicate the assumption of the Cox model appears to be violated by the variable years of education. Therefore, this model could be improved by fitting a Cox model with `yschool` as a time-dependent variable.

As the backward parameter 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. Additionally, he 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.

Finally, A significant improvement to this analysis would be first splitting the data into a training and testing dataset, in order to better validate the accuracy of the models and correct for potential overfitting of the data.

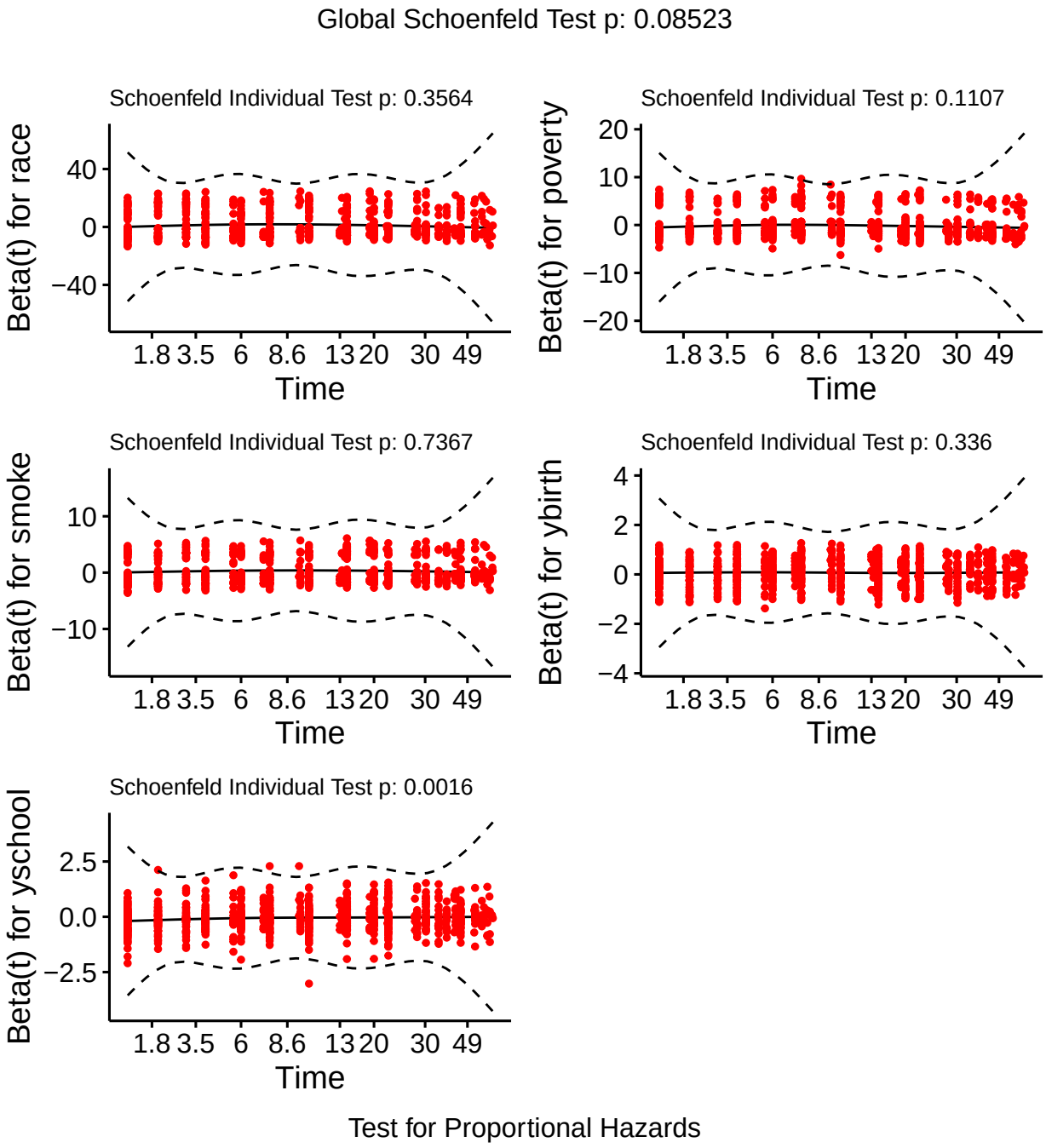

Figure 8: Schoenfield plot of residuals for Cox proportional hazards assumption

Conclusion

The data analyzed here from the 1970s and 1980s is by no means adequate evidence to challenge any public health advice. Rather, this analysis shows the utility of time-to-event analysis for considering breastfeeding behaviors and supports the hypothesis that breastfeeding duration appears to be impacted by environmental and social factors. The findings that prenatal care appointments following 3 months had no impact on breastfeeding duration was a surprising finding. It has been widely reported that breastfeeding behavior for infants has been on the decline in the United States over the past few decades; however, the magnitude of effect of the year the child was born on duration breast feeding was striking if these data are representative of the larger population. This analysis suggests that a mother's decisions about breastfeeding appear to correlate with some of the predictor variables collected as part of the NLYS study. Perhaps this information could be useful for public health officials hoping to increase breastfeeding rates in the United States.

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.

Supplemental Materials

Supplementary figures and tables are available here. Code used to generate this report is available at the bottom of this document.

Figures

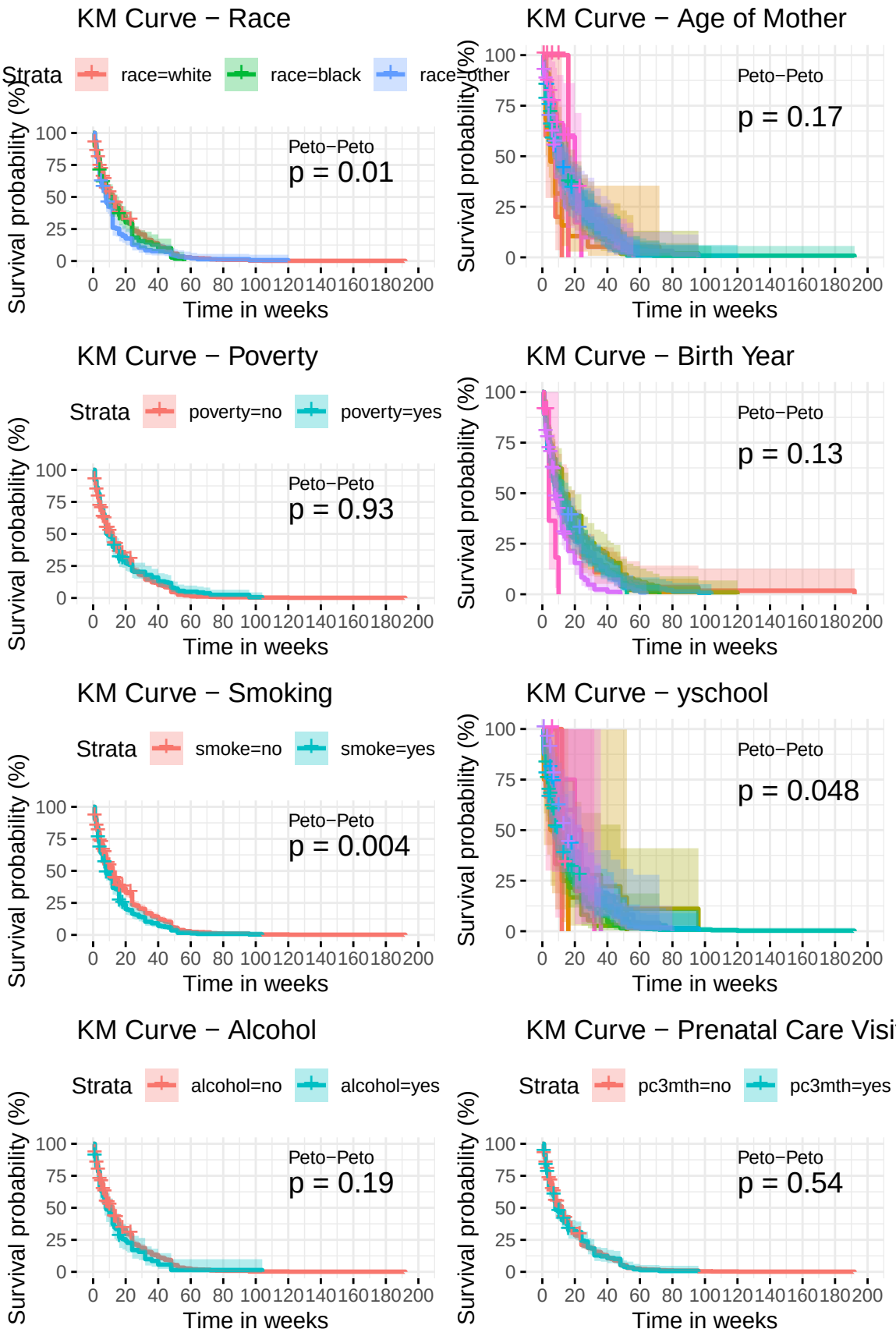

Figure S1: KM curves for all predictor variables

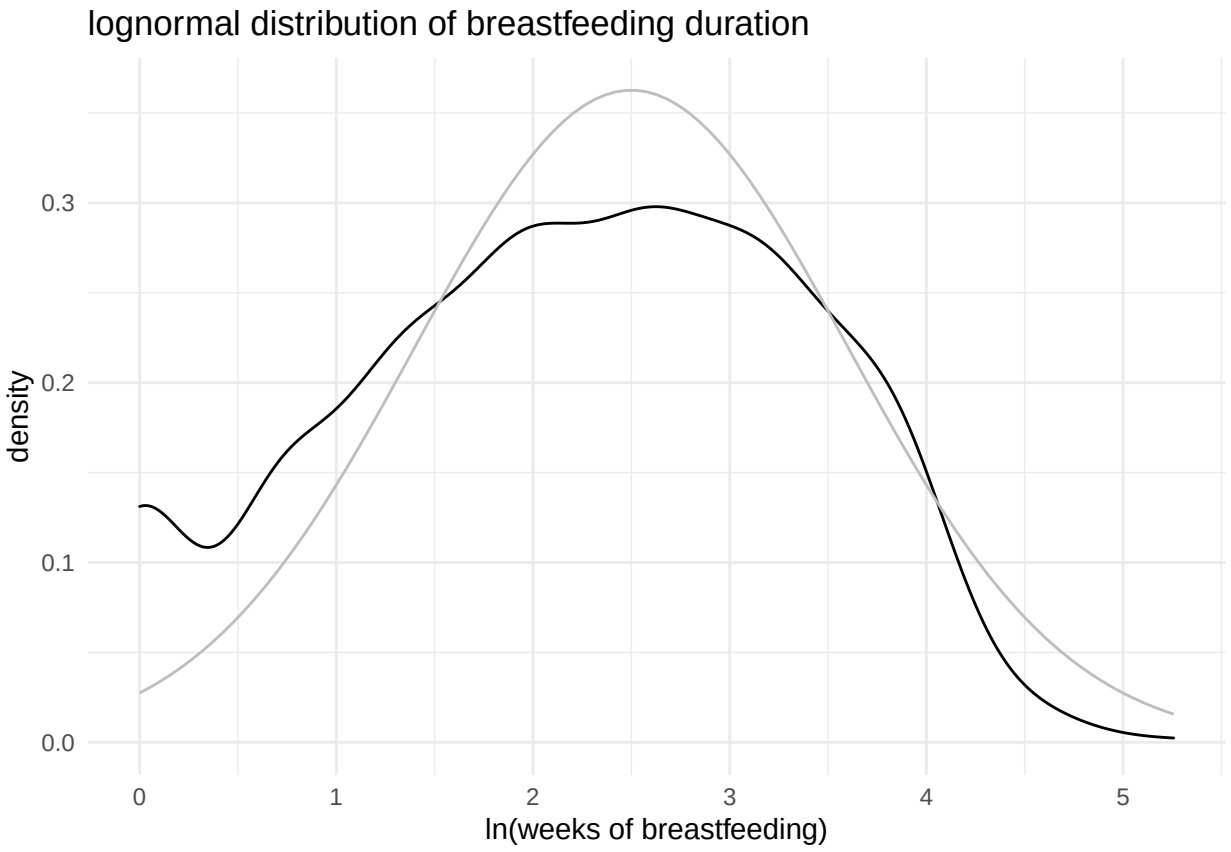

Figure S2: Distribution of natural log tranformed breastfeeding duration alongside normal distribution $N(2.5, 1.1)$ (grey)

Tables

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

Table S3: Comparison of parameter coefficients for parametric AFT models

	Exponential	Weibull	Lognormal
(Intercept)	159.7446	159.6332	146.8739
raceblack	-0.1733	-0.1732	-0.1485
raceother	-0.3096	-0.3097	-0.2717
smokeyes	-0.2669	-0.2669	-0.2252
yschool	0.0503	0.0504	0.0884
ybirth	-0.0794	-0.0794	-0.0735

Code used to create this report

```
library(tidyverse) # package for data analysis
library(ggsci) # for graph colors
library(KMsurv) # package with data set
library(survival) # package for survival analysis
library(survminer) # for ggsurvplot viz
library(rstatix) # package for piped stat tests

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

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

```
# data processing

bfeed_og <- bfeed %>%
  mutate(SurvObj = Surv(duration, delta == 1) )

# raw data processing
bfeed <- bfeed %>%
  mutate(race = factor(
              recode(race, "1" = "white", "2" = "black", "3"="other"),
              levels = c("white", "black", "other")
              ),
         status = delta,
         delta = factor(
              recode(delta, "1" = "yes", "0" = "no"),
              levels = c("no", "yes")
              ),
         poverty = factor(
              recode(poverty, "1" = "yes", "0" = "no"),
              levels = c("no", "yes")
              ),
         smoke = factor(
```

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

          ),
        ybirth = ybirth + 1900 , # making data more exact for visualization
        SurvObj = Surv(duration, delta == "yes"), #add Survival Obj variable
        race_o = case_when(
         race == "other" ~ 1,
         race != "other" ~ 0),
        race_b = case_when(
         race == "black" ~ 1,
         race != "black" ~ 0)
  )
```

```
#table 2
bfeed %>%
  select(-c(race_o, race_b, education, status)) %>%
  head(n=5) %>%
  knitr::kable(caption = "Time to cessation of breastfeeding data set")
```

```
n_censored <- bfeed %>%
  filter(delta=="no") %>%
  count() %>%
  pull()
```

```
Fig1a
fig1 <- bfeed %>%
  select(-c(status, SurvObj, race_o, race_b)) %>% # artificial duplicate variables
  rename("Age of Mother at Birth" = agemth, # change names
         "Duration of Breastfeeding (weeks)" = duration,
         "Breastfeeding Completed" = delta,
         "Year Child was Born" = ybirth,
         "Years of Schooling - Mother" = yschool,
         "Race of Mother" = race,
```

```
         "Poverty" = poverty,
         "Alcohol" = alcohol,
         "Prenatal Care after 3rd Month" = pc3mth,
         "Years of Education - Mother" = yschool
         ) %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value))+
  geom_histogram(bins = 30)+
  facet_wrap(~key, scales = "free") +
  theme_minimal() +
  scale_x_continuous(n.breaks = 10)+
  labs(x=" ", y="Number of Participants") +
  theme(plot.title = element_text(size = 12)) +
  ggtitle("Distribution of numerical variables")
```

```
#fig1b
fig2 <- bfeed %>%
  select(-education) %>%
  rename("Age of Mother at Birth" = agemth, # change names
         "Duration of Breastfeeding (weeks)" = duration,
         "Breastfeeding Completed" = delta,
         "Year Child was Born" = ybirth,
         "Years of Schooling - Mother" = yschool,
         "Race of Mother" = race,
         "Poverty" = poverty,
         "Alcohol" = alcohol,
         "Prenatal Care after 3rd Month" = pc3mth,
         "Years of Education - Mother" = yschool
         ) %>%
  select_if(negate(is.numeric)) %>%
  gather() %>%
  ggplot(aes((value)))+
  geom_bar()+
  facet_wrap(~key, scales = "free")+
  theme_minimal() +
  # scale_fill_grey() +
  theme(legend.position = "none") +
  ggtitle("Distribution of categorical variables") +
  theme(plot.title = element_text(size = 12)) +
  labs(x = "Participant Response", y = "Number of Participants")

ggarrange(fig1, fig2, nrow =2)
```

```
#fig2
# create survival object:
km.as.one <- survfit(SurvObj ~ 1, data = bfeed)

#KM plot combining all participants
gg.km.as.one <- 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.
   palette = "black",           # make combined curve black
   xlab = "Time (weeks)",    # customize X axis label.
   break.time.by = 20,         # break X axis in time intervals by 20.
   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",
  legend = "none",
  fun = "pct"                 #show survival function as percentage
)
gg.km.as.one$plot <- gg.km.as.one$plot +
  geom_vline(xintercept=26, color="red", linetype = "dashed") + # add 6mo marker (WHO recommendation)
  geom_label(label="6 months", x=40, y = 75)
gg.km.as.one

# get 6 month survival time estimate
pct6mo.all <- gg.km.as.one$data.survplot %>%
  filter(time == 26) %>%
  pull(surv)

#upperbound
u6mo.all <- gg.km.as.one$data.survplot %>%
  filter(time == 26) %>%
  pull(upper)

# lowerbound for est
l6mo.all <- gg.km.as.one$data.survplot %>%
  filter(time == 26) %>%
  pull(lower)
```

```
#prep fig3
km.by.smoke <- survfit(SurvObj ~ smoke, data = bfeed)

#KM curve according to smoking
gg.km.by.smoke <- 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 = "1",   # normal 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
)

gg.km.by.smoke$plot <- gg.km.by.smoke$plot +
  geom_vline(xintercept=26, color="red", linetype = "dashed") + # add 6mo marker (WHO recommendation)
  geom_label(label="6 months", x=40, y = 75)

# get 6 month survival time estimate smoke
pct6mo.smoke <- gg.km.by.smoke$data.survplot %>%
  filter(time == 26) %>%
  pull(surv)

#upperbound
u6mo.smoke <- gg.km.by.smoke$data.survplot %>%
  filter(time == 26) %>%
  pull(upper)

# lowerbound for est
l6mo.smoke <- gg.km.by.smoke$data.survplot %>%
  filter(time == 26) %>%
  pull(lower)
```

```
#figure3
gg.km.by.smoke
```

```
#table 3

# calculate log-rank test p-values for each categorical variable

# race
lr_p_race <- pchisq(
  survdiff(SurvObj ~ race, bfeed)$chisq,
  length(survdiff(SurvObj ~ race, bfeed)$n)-1, lower.tail=F
)

# poverty
lr_p_poverty <- pchisq(
  survdiff(SurvObj ~ poverty, bfeed)$chisq,
  length(survdiff(SurvObj ~ poverty, bfeed)$n)-1, lower.tail=F
)
```

```r
# smoke
lr_p_smoke <- pchisq(
  survdiff(SurvObj ~ smoke, bfeed)$chisq,
  length(survdiff(SurvObj ~ smoke, bfeed)$n)-1, lower.tail=F
)

# alcohol
lr_p_alcohol <- pchisq(
  survdiff(SurvObj ~ alcohol, bfeed)$chisq,
  length(survdiff(SurvObj ~ alcohol, bfeed)$n)-1, lower.tail=F
)

# prenatal care after 3mo
lr_p_pc3mth <- pchisq(
  survdiff(SurvObj ~ pc3mth, bfeed)$chisq,
  length(survdiff(SurvObj ~ pc3mth, bfeed)$n)-1, lower.tail=F
)

log_rank_pvals <- tibble(
  "Variable ID" = c("race", "poverty", "smoke", "alcohol", "pc3mth"),
  "p-value" = c(lr_p_race, lr_p_poverty, lr_p_smoke, lr_p_alcohol, lr_p_pc3mth)
)

log_rank_pvals %>%
  add_significance("p-value") %>%
  knitr::kable(caption = "p-values of log-rank test for difference in breastfeeding duration according
               label = "ns corresponds to a p-value greater than 0.05, * a p-value less than 0.05, and
```

```r
# prep for expo
# exponential parametric model

# create vector of AIC values for expo
expAIC <- rep(0.0, 9)

## include all parameters (diy backselection)

expreg_all <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth + yschool + pc3mth,
expAIC[1] <- AIC(expreg_all)
# summary(expreg_all) # has highest p-val: pc3mth

expreg_less1 <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth + yschool, bfeed,
expAIC[2] <- AIC(expreg_less1)
# summary(expreg_less1) # has highest p-val: agemth

expreg_less2 <- survreg(SurvObj ~ race +
                          poverty + smoke +
                          alcohol + yschool +
                          ybirth, bfeed, dist="exponential")

expAIC[3] <- AIC(expreg_less2)
# summary(expreg_less2) # has highest p-val: alcohol

expreg_less3 <- survreg(SurvObj ~ race  +
```

```
                          poverty + smoke + yschool + ybirth,
                          bfeed, dist="exponential")
expAIC[4] <- AIC(expreg_less3)
# summary(expreg_less3) # has highest p-val: poverty

expreg_less4 <- survreg(SurvObj ~ race + smoke + yschool + ybirth,
                        bfeed, dist="exponential")
expAIC[5] <- AIC(expreg_less4)
# summary(expreg_less4) # has highest p-val: race

expreg_less5 <- survreg(SurvObj ~  smoke + yschool + ybirth,
                        bfeed, dist="exponential")
expAIC[6] <- AIC(expreg_less5)
# summary(expreg_less5) # has highest p-val: smoke


expreg_less6 <- survreg(SurvObj ~  yschool + ybirth,
                        bfeed, dist="exponential")
expAIC[7] <- AIC(expreg_less6)
# summary(expreg_less6) # has highest p-val: yschool

expreg_less7 <- survreg(SurvObj ~ ybirth,
                        bfeed, dist="exponential")
expAIC[8] <- AIC(expreg_less7)
# summary(expreg_less7)

expreg_less8 <- survreg(SurvObj ~ 1,
                        bfeed, dist="exponential")
expAIC[9] <- AIC(expreg_less8)
# summary(expreg_less8)
```

```
#figure 4
# predicted curves based on exponential best model
pred.exp1 = predict(expreg_less4, newdata=list(race="white", smoke = "no", yschool=12, ybirth=1982),typ
pred.exp2 = predict(expreg_less4, newdata=list(race="black", smoke = "no", yschool=12, ybirth=1982),typ
pred.exp3 = predict(expreg_less4, newdata=list(race="other", smoke = "no", yschool=12, ybirth=1982),typ

# put preds together in tidy df
df = data.frame(y=seq(99,1,by=-1), race_white=pred.exp1, race_black=pred.exp2, race_other = pred.exp3)
df_long = gather(df, key= "race", value="time", -y)


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

#KM curve according to race
gg.km.by.race.exp <- 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 = "1",   # Pval using 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 plot for race adding survreg predictions
gg.km.by.race.exp <- gg.km.by.race.exp$plot + geom_line(data=df_long, aes(x=time, y=y, group=race))
gg.km.by.race.exp
```

```
#Table 4
summary_expAIC <- tibble(
  "# Parameters" = c(8:0),
  "AIC" = expAIC,
  "Equation" = c(
    "SurvObj ~ ybirth + yschool + ... + agemth + pc3mth",
    "SurvObj ~ ybirth + yschool + ... + alcohol + agemth",
    "SurvObj ~ ybirth + yschool + ... + poverty + alcohol",
    "SurvObj ~ ybirth + yschool + smoke + race + poverty",
    "SurvObj ~ ybirth + yschool + smoke + race", # lowest AIC
    "SurvObj ~ ybirth + yschool + smoke",
    "SurvObj ~ ybirth + yschool",
    "SurvObj ~ ybirth",
    "SurvObj ~ 1"
  )
)

summary_expAIC %>%
  knitr::kable(digits=5,caption="AIC of Backward selection of exponential survival model")
```

```
#Table 5
summary_expreg <- summary(expreg_less4)

best.exp.model <- expreg_less4

summary_expreg$table %>%
  knitr::kable(digits=4,caption = "Exponential Model: SurvObj ~ race + smoke + yschool + ybirth")#,capt
```

```
#prep Weibull
# Weibull parametric model

# create vector of AIC values for Weibull
WeibullAIC <- rep(0.0, 9)

## include all parameters (diy backselection)

Weibullreg_all <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth +
              yschool + pc3mth, bfeed, dist="weibull")
WeibullAIC[1] <- AIC(Weibullreg_all)
# summary(Weibullreg_all) # has highest p-val: pc3mth

Weibullreg_less1 <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth +
                yschool, bfeed, dist="weibull")
WeibullAIC[2] <- AIC(Weibullreg_less1)
# summary(Weibullreg_less1) # has highest p-val: agemth

Weibullreg_less2 <- survreg(SurvObj ~ race + poverty + smoke + alcohol + yschool + ybirth,
                        bfeed, dist="weibull")
WeibullAIC[3] <- AIC(Weibullreg_less2)
# summary(Weibullreg_less2) # has highest p-val: alcohol

Weibullreg_less3 <- survreg(SurvObj ~ race  + poverty + smoke + yschool + ybirth,
                        bfeed, dist="weibull")
WeibullAIC[4] <- AIC(Weibullreg_less3)
# summary(Weibullreg_less3) # has highest p-val: poverty

Weibullreg_less4 <- survreg(SurvObj ~ race + smoke + yschool + ybirth,
                        bfeed, dist="weibull")
WeibullAIC[5] <- AIC(Weibullreg_less4)
# summary(Weibullreg_less4) # has highest p-val: race

Weibullreg_less5 <- survreg(SurvObj ~  smoke + yschool + ybirth,
                        bfeed, dist="weibull")
WeibullAIC[6] <- AIC(Weibullreg_less5)
# summary(Weibullreg_less5) # has highest p-val: smoke


Weibullreg_less6 <- survreg(SurvObj ~  yschool + ybirth,
                        bfeed, dist="weibull")
WeibullAIC[7] <- AIC(Weibullreg_less6)
# summary(Weibullreg_less6) # has highest p-val: yschool

Weibullreg_less7 <- survreg(SurvObj ~ ybirth,
                        bfeed, dist="weibull")
WeibullAIC[8] <- AIC(Weibullreg_less7)
# summary(Weibullreg_less7)

Weibullreg_less8 <- survreg(SurvObj ~ 1,
                        bfeed, dist="weibull")
WeibullAIC[9] <- AIC(Weibullreg_less8)
# summary(Weibullreg_less8)
```

```r
# figure 5
# predicted curves based on exponential best model
pred.W1 = predict(Weibullreg_less4, newdata=list(race="white", smoke = "no", yschool=12, ybirth=1982),ty
pred.W2 = predict(Weibullreg_less4, newdata=list(race="black", smoke = "no", yschool=12, ybirth=1982),ty
pred.W3 = predict(Weibullreg_less4, newdata=list(race="other", smoke = "no", yschool=12, ybirth=1982),ty

# put preds together in tidy df
df = data.frame(y=seq(99,1,by=-1), race_white=pred.W1, race_black=pred.W2, race_other = pred.W3)
df_long = gather(df, key= "race", value="time", -y)


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

#KM curve according to race
gg.km.by.race.Weibull <- 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 = "1",    # Pval using 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 plot for race adding survreg predictions
gg.km.by.race.Weibull$plot <- gg.km.by.race.Weibull$plot + geom_line(data=df_long, aes(x=time, y=y, grou
gg.km.by.race.Weibull
```

```r
# Table 6
summary_WeibullAIC <- tibble(
  "# Parameters" = c(8:0),
  "AIC" = WeibullAIC,
  "Equation" = c(
    "SurvObj ~ ybirth + yschool + ... + agemth + pc3mth",
    "SurvObj ~ ybirth + yschool + ... + alcohol + agemth",
    "SurvObj ~ ybirth + yschool + ... + poverty + alcohol",
```

```
    "SurvObj ~ ybirth + yschool + smoke + race + poverty",
    "SurvObj ~ ybirth + yschool + smoke + race", # lowest AIC
    "SurvObj ~ ybirth + yschool + smoke",
    "SurvObj ~ ybirth + yschool",
    "SurvObj ~ ybirth",
    "SurvObj ~ 1"
  )
)

summary_WeibullAIC %>%
  knitr::kable(digits=5,caption="AIC of Backward selection of Weibull survival model")
```

```
# Table7
summary_Weibullreg <- summary(Weibullreg_less4)

best.Weibull.model <- Weibullreg_less4

summary_Weibullreg$table %>%
  knitr::kable(digits=4,caption = "Weibull Model: SurvObj ~ race + smoke + yschool + ybirth")#,caption=
```

```
#prep for lognorm
# lognormal parametric model

# create vector of AIC values for lognormal
lognormalAIC <- rep(0.0, 9)

## include all parameters (diy backselection)

lognormalreg_all <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth +
              yschool + pc3mth, bfeed, dist="lognormal")
lognormalAIC[1] <- AIC(lognormalreg_all)
# summary(lognormalreg_all) # has highest p-val: pc3mth


lognormalreg_less1 <- survreg(SurvObj ~ race + poverty + smoke + alcohol + agemth + ybirth +
                yschool, bfeed, dist="lognormal")
lognormalAIC[2] <- AIC(lognormalreg_less1)
# summary(lognormalreg_less1) # has highest p-val: agemth

lognormalreg_less2 <- survreg(SurvObj ~ race + poverty + smoke + alcohol + yschool + ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[3] <- AIC(lognormalreg_less2)
# summary(lognormalreg_less2) # has highest p-val: alcohol

lognormalreg_less3 <- survreg(SurvObj ~ race + poverty + smoke + yschool + ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[4] <- AIC(lognormalreg_less3)
# summary(lognormalreg_less3) # has highest p-val: poverty

lognormalreg_less4 <- survreg(SurvObj ~ race + smoke + yschool + ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[5] <- AIC(lognormalreg_less4)
# summary(lognormalreg_less4) # has highest p-val: race
```

```
lognormalreg_less5 <- survreg(SurvObj ~  smoke + yschool + ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[6] <- AIC(lognormalreg_less5)
# summary(lognormalreg_less5) # has highest p-val: smoke


lognormalreg_less6 <- survreg(SurvObj ~  yschool + ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[7] <- AIC(lognormalreg_less6)
# summary(lognormalreg_less6) # has highest p-val: yschool

lognormalreg_less7 <- survreg(SurvObj ~ ybirth,
                        bfeed, dist="lognormal")
lognormalAIC[8] <- AIC(lognormalreg_less7)
# summary(lognormalreg_less7)

lognormalreg_less8 <- survreg(SurvObj ~ 1,
                        bfeed, dist="lognormal")
lognormalAIC[9] <- AIC(lognormalreg_less8)
# summary(lognormalreg_less8)
```

```
#fig6
# predicted curves based on exponential best model
pred.lognormal1 = predict(lognormalreg_less4, newdata=list(race="white", smoke = "no", yschool=12, ybir
pred.lognormal2 = predict(lognormalreg_less4, newdata=list(race="black", smoke = "no", yschool=12, ybir
pred.lognormal3 = predict(lognormalreg_less4, newdata=list(race="other", smoke = "no", yschool=12, ybir

# put preds together in tidy df
df = data.frame(y=seq(99,1,by=-1), race_white=pred.lognormal1, race_black=pred.lognormal2, race_other =
df_long = gather(df, key= "race", value="time", -y)


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

#KM curve according to race
gg.km.by.race.lognormal <- 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 = "1",    # Pval using 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 plot for race adding survreg predictions
gg.km.by.race.lognormal$plot <- gg.km.by.race.lognormal$plot + geom_line(data=df_long, aes(x=time, y=y,
gg.km.by.race.lognormal
```

```
#table8
summary_lognormalAIC <- tibble(
  "# Parameters" = c(8:0),
  "AIC" = lognormalAIC,
  "Equation" = c(
    "SurvObj ~ ybirth + yschool + ... + agemth + pc3mth",
    "SurvObj ~ ybirth + yschool + ... + alcohol + agemth",
    "SurvObj ~ ybirth + yschool + ... + poverty + alcohol",
    "SurvObj ~ ybirth + yschool + smoke + race + poverty",
    "SurvObj ~ ybirth + yschool + smoke + race", # lowest AIC
    "SurvObj ~ ybirth + yschool + smoke",
    "SurvObj ~ ybirth + yschool",
    "SurvObj ~ ybirth",
    "SurvObj ~ 1"
  )
)

summary_lognormalAIC %>%
  knitr::kable(digits=5,caption="AIC of Backward selection of lognormal survival model")
```

```
#table9
summary_lognormalreg <- summary(lognormalreg_less4)

best.lognormal.model <- lognormalreg_less4

summary_lognormalreg$table %>%
  knitr::kable(digits=4,caption = "Lognormal Model: SurvObj ~ race + smoke + yschool + ybirth")#,captio
```

```
#table10
lognormalAIC <- round(lognormalAIC,0)
lognormalAIC[5] <- paste("\\color{red}{",round(lognormalAIC[5], 2), "}")

AIC_all <- tibble(
  "# Parameters" = c(8:0),
  "Exponential" = round(expAIC,0),
  "Weibull" = round(WeibullAIC,0),
  "Lognormal" = (lognormalAIC),
  "Equation" = c(
    "SurvObj ~ ybirth + yschool + ... + agemth + pc3mth",
    "SurvObj ~ ybirth + yschool + ... + alcohol + agemth",
```

```
    "SurvObj ~ ybirth + yschool + ... + poverty + alcohol",
    "SurvObj ~ ybirth + yschool + smoke + race + poverty",
    "SurvObj ~ ybirth + yschool + smoke + race", # lowest AIC
    "SurvObj ~ ybirth + yschool + smoke",
    "SurvObj ~ ybirth + yschool",
    "SurvObj ~ ybirth",
    "SurvObj ~ 1"
  )
)

AIC_all %>%
  knitr::kable(digits=4,
               caption = "AIC values across parametric models")
```

```
# create variable list for stepwise selection of variables in models:
my.variable.list <- c("race", "poverty", "smoke", "alcohol",
                      "agemth", "ybirth", "yschool", "pc3mth")

# run the stepwise selection
My.stepwise::My.stepwise.coxph(Time = "duration",
                               Status = "delta",
                               variable.list = my.variable.list,
                               data = bfeed,
                               sle = 0.15, sls = 0.15)
# best model:
"duration ~ smoke + race + ybirth + yschool + poverty"
```

```
# output in Rmd file
coxph.best.model <- coxph(SurvObj ~ race + poverty + smoke + ybirth + yschool, data = bfeed)

summary(coxph.best.model)
```

```
# predicted curves based on CoxPH best model
#fig7

race_df <- with(bfeed,
               data.frame(sex = rep("male",3),
                          race = c("white", "black", "other"),
                          ybirth = rep(mean(ybirth, na.rm = TRUE), 3),
                          yschool = rep(mean(yschool, na.rm = TRUE), 3),
                          smoke = rep("no",3),
                          poverty = rep("no",3)
                          )
               )
# race_df

km.cox.fit <- survfit(coxph.best.model, newdata = race_df)

# put preds together in tidy df
df = data.frame(y=seq(99,1,by=-1), race_white=pred.exp1, race_black=pred.exp2, race_other = pred.exp3)
df_long = gather(df, key= "race", value="time", -y)
```

```
#KM curve according to race
gg.km.by.race.cox <- ggsurvplot(
   km.cox.fit,                # 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 = "1",    # Pval using 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 plot for race adding survreg predictions
gg.km.by.race.cox <- gg.km.by.race.cox$plot + geom_line(data=df_long, aes(x=time, y=y, group=race))
gg.km.by.race.cox
```

```
#table 11 prep
explanatory = c("race", "poverty", "smoke", "alcohol",
                      "agemth", "ybirth", "yschool", "pc3mth")
dependent = "Surv(duration, delta)"
bfeed_og %>%
    mutate(
      race = as.factor(race),
      poverty = as.factor(poverty),
      smoke = as.factor(smoke),
      alcohol = as.factor(alcohol),
      pc3mth = as.factor(pc3mth)
    ) %>%
    finalfit::finalfit(dependent, explanatory,
                       explanatory_multi =
                         c("smoke", "race", "ybirth", "yschool", "poverty")
                       ) -> t

colnames(t)[3] <- "Count (%)"
colnames(t)[2] <- "Group"
colnames(t)[1] <- "Duration"

t <- tibble(t) %>%
```

```
  remove_rownames() %>%
  mutate(Group = c("white", "black", "other", "no", "yes","no", "yes","no", "yes", "Mean (SD)","Mean (SD
```

```
#table 11
t %>% knitr::kable(digits=4, caption = "Hazard Ratios (HR) for Cox PH model of time-independent variable
```

```
# create variable list for stepwise selection of variables in models:
my.variable.list <- c("race", "poverty", "smoke", "alcohol",
                      "agemth", "ybirth", "yschool", "pc3mth")

# run the stepwise selection
My.stepwise::My.stepwise.coxph(Time = "duration", Status = "delta", variable.list = my.variable.list, da
# best Cox PH model:
"duration ~ smoke + race + ybirth + yschool + poverty"
```

```
# termplot(m1) # useful but long in rmd file
#table12
cox.zph(coxph.best.model)$table%>%
  knitr::kable(digits=4, caption = "Test of Proprtional Hazards Assumption of Cox Regression Model")
```

```
# Figure 8

#alternative code to above line
ggcoxzph(cox.zph(coxph.best.model), font.main = 10, caption = "Test for Proportional Hazards")
```

```
#figS1prep
# 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 ~ yschool, 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
gg.all <- ggsurvplot(
   km.as.one,                # survfit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = F,        # 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
gg.race <- ggsurvplot(
   km.by.race,               # surufit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = F,        # show risk table.
   pval = TRUE,              # show p-value of log-rank test.
   pval.coord = c(120,70),   # 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
gg.poverty <- ggsurvplot(
   km.by.poverty,               # surufit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = F,        # show risk table.
   pval = TRUE,              # show p-value of log-rank test.
   pval.coord = c(120,70),   # 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
gg.yschool <- ggsurvplot(
   km.by.education,               # survfit object with calculated statistics.
   data = bfeed,              # data used to fit survival curves.
   risk.table = F,         # show risk table.
   pval = TRUE,               # show p-value of log-rank test.
   pval.coord = c(120,70),    # location of pval
   legend = "none",
   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 - yschool",
  fun = "pct"                 #show survival function as percentage
)


#KM curve according to smoking
gg.smoke <- ggsurvplot(
   km.by.smoke,                # survfit object with calculated statistics.
   data = bfeed,              # data used to fit survival curves.
   risk.table = F,         # show risk table.
   pval = TRUE,               # show p-value of log-rank test.
   pval.coord = c(120,70),    # 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
gg.alcohol <- ggsurvplot(
   km.by.alcohol,               # survfit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = F,        # show risk table.
   pval = TRUE,              # show p-value of log-rank test.
   pval.coord = c(120,70),   # 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
gg.agemth <- ggsurvplot(
   km.by.agemth,               # survfit object with calculated statistics.
   data = bfeed,             # data used to fit survival curves.
   risk.table = F,        # show risk table.
  legend = "none",
    pval = TRUE,              # show p-value of log-rank test.
   pval.coord = c(120,70),   # 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
gg.pc3mth <- ggsurvplot(
   km.by.pc3mth,                # survfit object with calculated statistics.
   data = bfeed,              # data used to fit survival curves.
   risk.table = F,         # show risk table.
   pval = TRUE,               # show p-value of log-rank test.
   pval.coord = c(120,70),    # 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
gg.ybirth <- ggsurvplot(
   km.by.ybirth,                # survfit object with calculated statistics.
   data = bfeed,              # data used to fit survival curves.
   risk.table = F,         # show risk table.
   pval = TRUE,               # show p-value of log-rank test.
   legend = "none",
   pval.coord = c(120,70),    # 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
)
```

```
#figS1
ggS1 <- list(
  # gg.all,
  gg.race,
  gg.poverty,
  gg.smoke,
  gg.alcohol,
  gg.agemth,
  gg.ybirth,
  gg.yschool,
  gg.pc3mth
)

arrange_ggsurvplots(ggS1, ncol = 2, nrow = 4)
```

```
bfeed %>%
  mutate(lognormduration = log(duration)) %>%
  ggplot(aes(x=lognormduration)) +
  geom_density() +
  theme_minimal() +
  labs(x="ln(weeks of breastfeeding)", y="density") +
  ggtitle("lognormal distribution of breastfeeding duration") +
  stat_function(fun = dnorm, args = list(2.5, 1.1), color = "gray")
```

```
#tableS1
exp.aft.ph <- -coef(best.exp.model)/best.exp.model$scale
#https://myweb.uiowa.edu/pbreheny/7210/f15/notes/10-15.pdf
exp.aft.ph %>%
  knitr::kable(digits=4,caption = "coefficients for PH exponential model", col.names = c( "value"))
```

```
#tableS2
Weibull.aft.ph <-
  -coef(best.Weibull.model)/best.Weibull.model$scale
#https://myweb.uiowa.edu/pbreheny/7210/f15/notes/10-15.pdf
```

```
Weibull.aft.ph %>%
  knitr::kable(digits=4,caption = "coefficients for PH Weibull model", col.names = c( "value"))
```

```
# comparison of coefficient estimates across the parametric models (table S3)
all.param.coef <- tibble(
  " " = names(best.exp.model$coefficients),
  "Exponential" = best.exp.model$coefficients,
  "Weibull" = best.Weibull.model$coefficients,
  "Lognormal" = best.lognormal.model$coefficients)

all.param.coef %>%
  knitr::kable(digits=4,caption = "Comparison of parameter coefficients for parametric AFT models")
```