Survival Analysis of breast feeding times

Inês Fortes

last update: 14/07/2022

1. Introduction

In the past decade research has indicated that it is optimal for babies to be exclusively breast fed until 6 months-old (Kramer & Kakuma, 2012). However, not all mothers breast feed their babies, and among those who do it, not all of them continue until the sixth month. The reasons to stop breast feeding can be very diverse: they can be physical, psychological or even cultural (Wright & Schanler, 2001).

The present study investigates the major factors involved in stopping breastfeeding. For that, I analyzed the ‘bfeed’ database from the R package “KMsurv” (Klein & Moeschberger, 1997). This database results from the National Longitudinal Survey of Youth which started in 1979 in the USA. This particular database was obtained from 1983 to 1988, when women in the survey were asked about their previous pregnancies (since 1978). This data set contains reports from 927 first-born children to mothers who chose to breast feed their baby. For this study I used survival analysis methods because this database includes some censored observations, that is, not all subjects reported the event (stop breast feeding) at the end of the survey.

2. Exploratory data analysis

The response variable is the breastfeeding duration. Thus, the time origin is the birth of the child and the event of interest is stopping breastfeeding, signaled by the indicator variable (‘delta’). Table 1 shows the variables included in the database.

Table 1: Description of variables in bfeed

Variable	Description
duration	Duration of breast feeding in weeks
delta	Indicator of completed breast feeding (1=yes, 0=no)
race	Race of mother (1=white, 2=black, 3=other)
poverty	Mother in poverty (1=yes, 0=no)
smoke	Mother smoked at birth of child (1=yes, 0=no)
alcohol	Mother used alcohol at birth of child (1=yes, 0=no)
agemth	Age of mother at birth of child in years
ybirth	Year of birth
yschool	Education level of mother (years of school)
pc3mth	Prenatal care after third month (1=yes, 0=no)

Note that agemth, ybirth and yschool are continuous variables while race, poverty, smoke, alcohol,and pc3mth are categorical variables (the former with 3 levels and the others with 2 levels). The ‘bfeed’ database has 927 observations and no NA values. Only 35 observations were censored.

Tables 2 and 3 show statistics for the categorical and continuous variables, respectively.

Table 2: Statistics of the categorical variables

variable		N (%)	Censored (%)
race	White	662 (71.4)	28 (4.2)
	Black	117 (12.6)	4 (3.4)
	Other	148 (16.0)	3 (2)
poverty	No	756 (81.6)	32 (4.2)
	Yes	171 (18.4)	3 (1.8)
smoke	No	657 (70.9)	28 (4.3)
	Yes	270 (29.1)	7 (2.6)
alcohol	No	848 (91.5)	32 (3.8)
	Yes	79 (8.5)	3 (3.8)
pc3mth	No	763 (82.3)	27 (3.5)
	Yes	164 (17.7)	8 (4.9)

Table 3: Statistics of the continuous variables

variable	Mean	SD	Min	Max
agemth	21.539	2.686	15	28
ybirth	81.974	2.139	78	86
yschool	12.207	1.928	3	19

From Table 2 we can see that most mothers were white (71.4%), were not in poverty (81.6%), did not smoke (70.9%) or drank alcohol during pregnancy (91.5%), and did not have prenatal care after the third month (82.3%). As said before, only a small percentage of observations (4%) were censored. From Table 3 we can see that mothers aged from 15 to 28 years-old (mean = 21.54), births occurred between 1978 and 1986 and mothers attended school for an average of 12.21 years (min = 3, max = 19).

3. Estimation of Survival

Even though the sample has a low number of censored observations, to obtain an empirical estimation of the survival function I used the nonparametric Kaplan-Meier estimator (Kaplan and Meier, 1958), which was specifically developed to deal with right-censored data. For that, the estimator uses a redistribution to the right algorithm, by assuming that a censored time is equally likely to be observed in any point in the future. This estimator indicates $P(T > t)$, that is, the probability of observing an event after time t. This function starts at 1 and it is a decreasing step function with steps only at the observed event times. Figure 1 shows the Kaplan-Meier estimator curve for the breast feeding data.

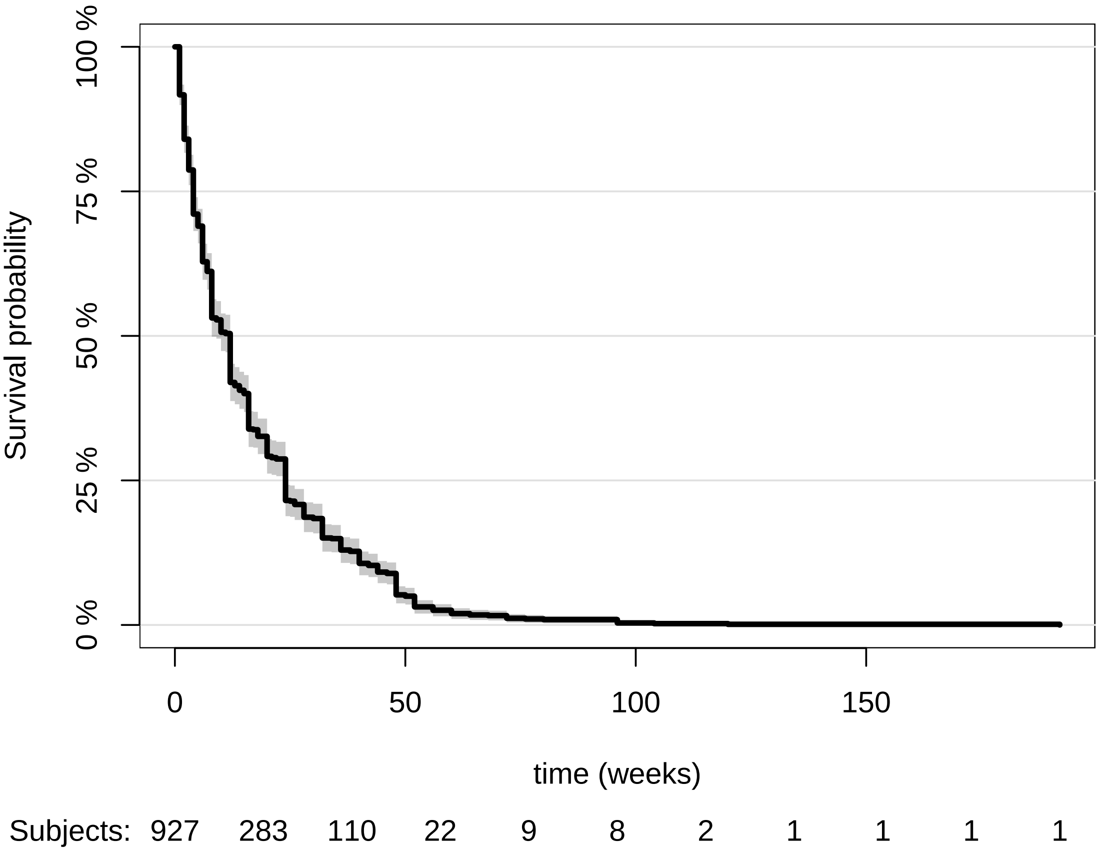


Figure 1: Kaplan-Meier estimator of survival.

Figure 2 shows that because there are only 4% of censored data, the empirical survival curve if there were no censored observations is very similar to the Kaplan-Meier estimator of survival. Also note that because of the distribution to the right algorithm, the Kaplan-Meier estimator is always equal or above the empirical curve.

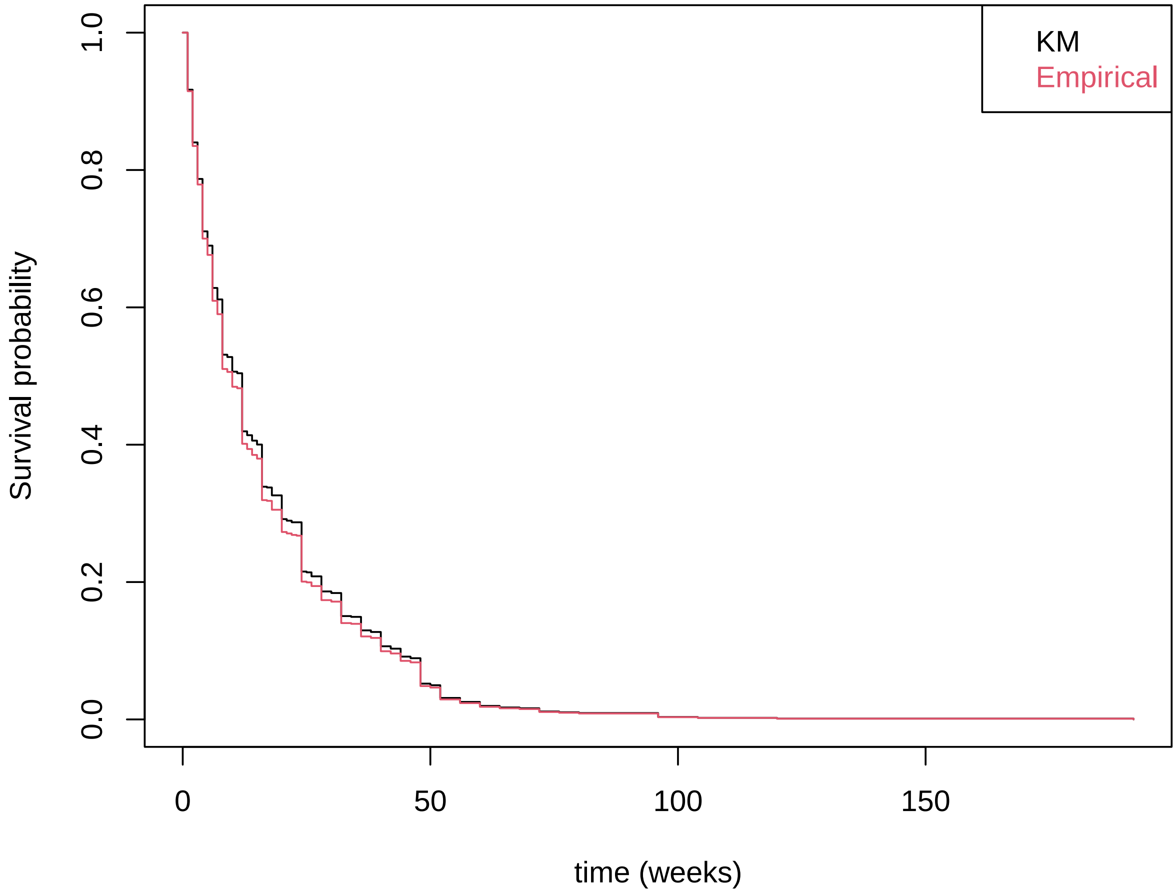


Figure 2: Kaplan-Meier estimator of survival vs Empirical estimator.

```
## Call: survfit(formula = Surv(duration, as.integer(delta)) ~ 1, data = bfeed)
##
##         n events rmean* se(rmean) median 0.95LCL 0.95UCL
## [1,] 927    892   16.9     0.614     12       8      12
##     * restricted mean with upper limit =  192
```

The median survival time, i.e., the time at which 50 % of the events have occurred can be estimated from the curve, and it is 12 weeks. The mean residual time is 16.9 weeks (SE = 0.614). Table 4 shows the survival estimates at different moments in time.

```
km <- survfit(Surv(duration,delta)~1,data=bfeed)
```

```
## List of 18
##  $ n           : int 927
##  $ time        : num [1:14] 1 4 8 12 16 20 24 32 40 48 ...
##  $ n.risk       : num [1:14] 927 722 547 447 352 283 248 159 110 77 ...
##  $ n.event      : num [1:14] 77 190 162 99 71 41 66 56 38 47 ...
##  $ n.censor     : num [1:14] 2 9 14 2 5 2 1 0 0 0 ...
##  $ surv         : num [1:14] 0.917 0.711 0.531 0.419 0.339 ...
##  $ std.err      : num [1:14] 0.00906 0.01493 0.01654 0.01643 0.01582 ...
##  $ cumhaz       : num [1:14] 0.0831 0.3271 0.6038 0.8229 1.0229 ...
##  $ std.chaz     : num [1:14] 0.00947 0.02011 0.02966 0.03694 0.04391 ...
##  $ type         : chr "right"
##  $ logse        : logi TRUE
##  $ conf.int     : num 0.95
##  $ conf.type    : chr "log"
##  $ lower        : num [1:14] 0.899 0.682 0.5 0.388 0.309 ...
##  $ upper        : num [1:14] 0.935 0.741 0.565 0.453 0.371 ...
##  $ call         : language survfit(formula = Surv(duration, delta) ~ 1, data = bfeed)
##  $ table        : Named num [1:9] 927 927 927 892 16.9 ...
##   ..- attr(*, "names")= chr [1:9] "records" "n.max" "n.start" "events" ...
##  $ rmean.endtime: num 192
##  - attr(*, "class")= chr "summary.survfit"
```

Table 4: Survival estimates at different weeks

Week	Survival	95% CI
1	0.917	0.899-0.935
4	0.711	0.682-0.741
8	0.531	0.500-0.565
12	0.419	0.388-0.453
16	0.339	0.309-0.371
20	0.292	0.263-0.323
24	0.215	0.190-0.244
32	0.150	0.129-0.176
40	0.106	0.088-0.129
48	0.052	0.039-0.069
72	0.012	0.006-0.021
96	0.003	0.001-0.011
144	0.001	0.000-0.008
192	0.000	NA-NA

From Table 4 we can see that after approximately 1 month (4 weeks) only 71% of mothers were still breast feeding. We can also use the Kaplan-Meier estimator to estimate survival in different groups. At the end of approximately 6 months (24 weeks), the recommended time to feed babies exclusively with milk, the percentage of mothers still breast feeding drops to about 22%. We can also see that even though they are rare, there are mothers breast feeding their babies for more than 3 years.

Kaplan-Meier estimates can also be computed for different groups. By plotting the survival functions for each categorical variable it was possible to observe that when curves differed, they did not cross. This might be an indication that the proportional hazards are held in this data. So, to check for differences between groups I ran log-rank tests (Mantel-cox tests) for each categorical variable. The results are shown in Table 5.

Table 5: Log-rank tests for the categorical variables.

Variable	Chi-Square	df	p-value
race	8.066	2	0.018
poverty	0.713	1	0.398
smoke	10.093	1	0.001
alcohol	2.01	1	0.156
pc3mth	0.162	1	0.687

Only the race and smoke variables showed differences in the survival curves. Figure 3 shows that non-smokers took more time to stop breast feeding.

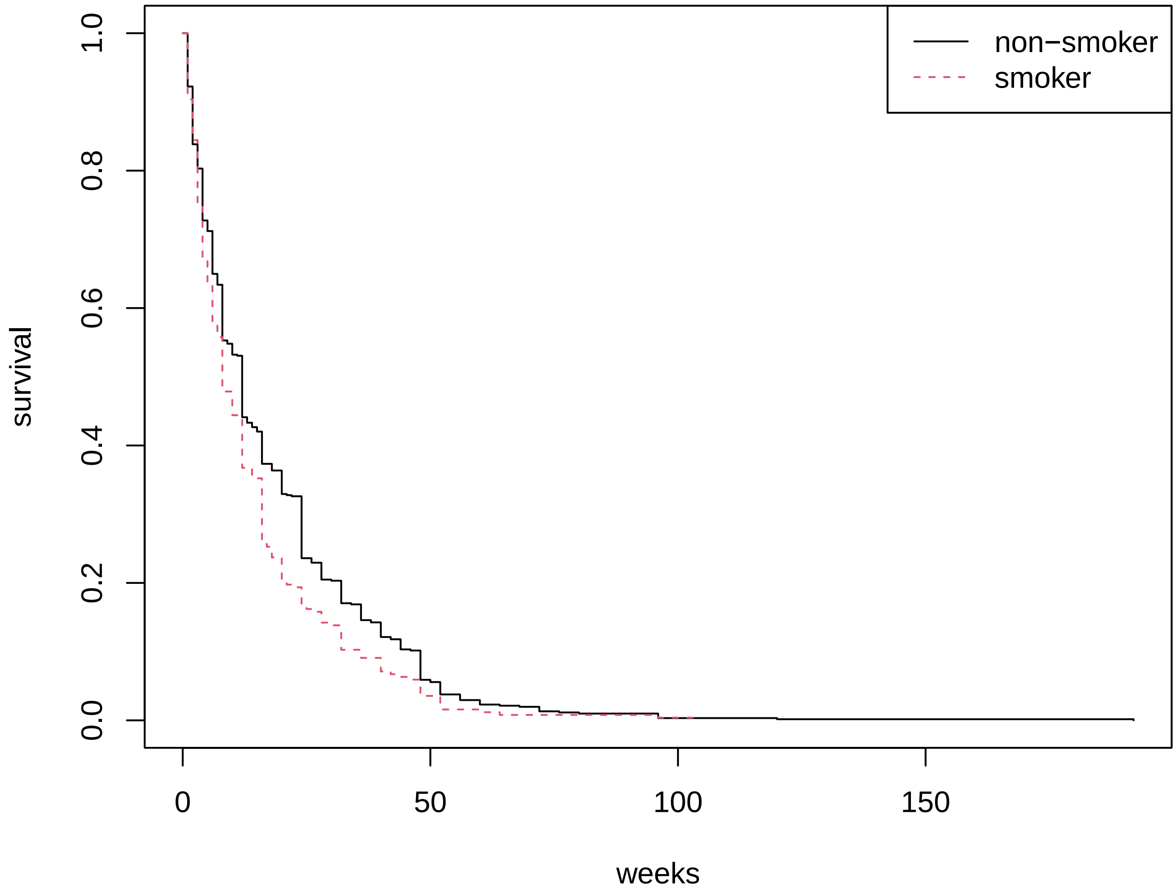


Figure 3: Kaplan-Meier estimator of survival for smokers and non-smokers.

Regarding race, Figure 4 shows that there seems to be no difference in the survival curves between white and black mothers, but the mothers in the ‘other’ category curve seems to be below the other two. In fact, a post hoc pairwise comparison with BH correction (Benjamini & Hochberg, 1995) between races showed that there is only a difference between white mothers and mothers from other races (p = 0.016).

R Code for pairwise comparisons:

```
res <- pairwise_survdiff(Surv(duration,delta)~race,data=bfeed,p.adjust.method = "BH",rho=0)
```

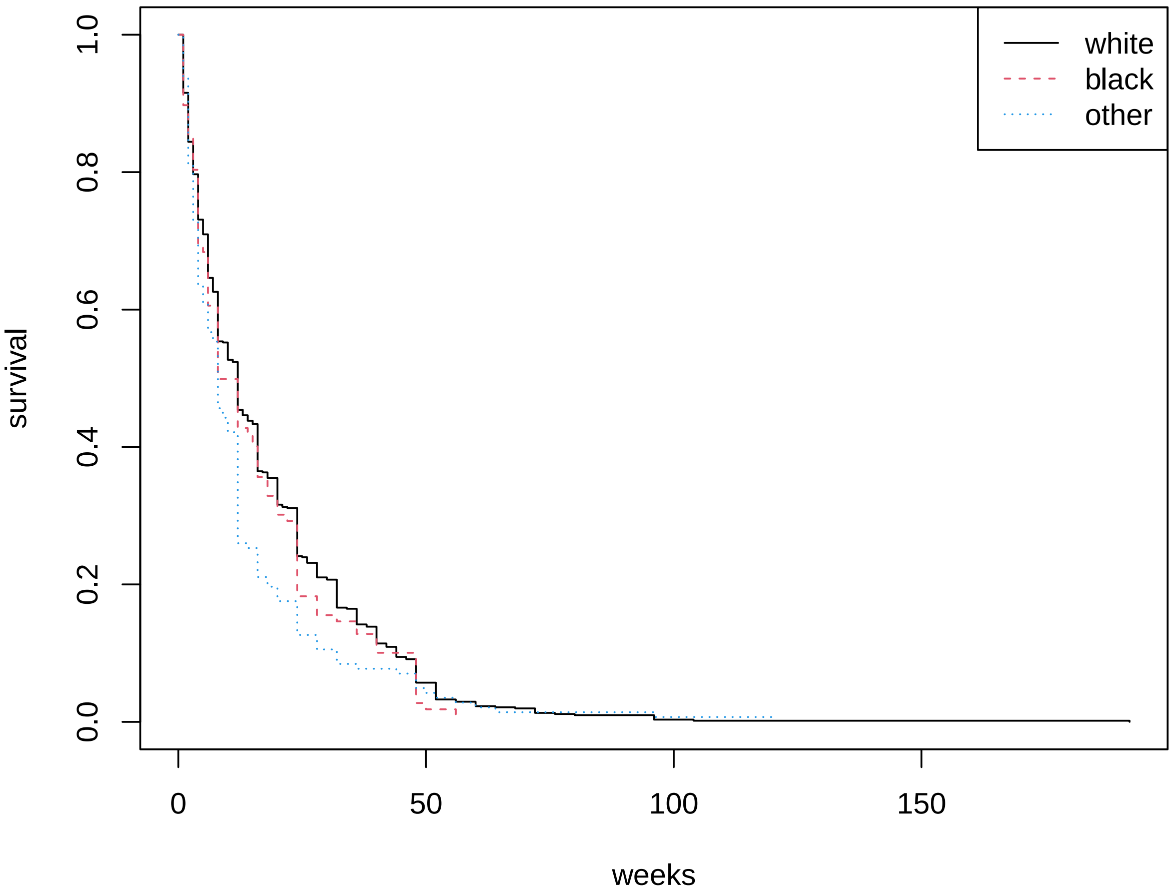


Figure 4: Kaplan-Meier estimator of survival for race.

To check if there were differences in the survival curves according to the mother's age, mother's education or baby's birth year, these variables were categorized such that each category holds approximately the same number of observations (or at least it results in a symmetrical distribution). The resulting distributions are presented in Tables 6, 7 and 8.

Table 6: Mother Age categories

Mother Age categories	N
15-18	110
18-21	370
21-24	303
24-28	144

Table 7: Year of birth categories

Year of birth categories	N
78-80	249
81-83	414
84-86	264

Table 8: Years of school categories

Years of school categories	N
3-11	220
12	438
12-19	269

Table 9 shows that there are differences between the categories in the year of birth and the mother's years of school. Post hoc pairwise comparison with BH corrections showed that in both variables the differences occur between the two extreme categories. There are differences in the survival curves between births before 1980 and those after 1984 (p = 0.0082), and there are differences between attending school for less than 12 years and more than 12 years (p = 0.033). The other comparisons were not statistically significant.

Table 9: Log-rank tests for the continuous variables (after categorization).

Variable	Chi-Square	df	p-value
agemth	3.852	2	0.146
ybirth	9.295	1	0.002
yschool	6.965	1	0.008

4. Univariate Cox analysis

To analyze the effect of the covariates in the breast feeding times I first fitted a Cox regression model to each variable, and then fitted a multiple regression with all the variables. One important assumption of this model is that there should be proportional hazards, that is, the hazard curves should not cross. In the analysis of the Kaplan-Meier survival curves that seemed to be the case for all variables, so I proceeded with the Cox model. It is important that the reference category has a sufficient number of observations to be able to detect differences between groups. Therefore, for the categorical variables the reference category always contained more observations than the other categories.

Table 10 shows the results for the simple Cox regression.

Table 10: Univariate Cox regression.

Variable		HR	95% CI	p-value
agemth		0.994	0.968-1.020	0.632
ybirth		1.051	1.017-1.086	0.003
yschool		0.956	0.924-0.989	0.009
race	white	1		
	black	1.117	0.914-1.365	0.28
	other	1.29	1.076-1.546	0.006
poverty	no	1		
	yes	0.927	0.783-1.098	0.379
smoke	no	1		
	yes	1.255	1.086-1.45	0.002
alcohol	no	1		
	yes	1.18	0.932-1.494	0.168
pc3mth	no	1		
	yes	1.036	0.871-1.231	0.69

Congruent with the Kaplan-Meier survival curves analysis, the variables that seem to significantly affect breast feeding times are: year of birth (ybirth), mother's years of education (yschool), race (between whites and other races) and smoking. The strongest predictor of stopping breast feeding seems to be being of other race, where the hazard ratio is 1.29, meaning that a mother of other race is 29% more likely to stop breast feeding. Similarly, smoking increases the hazard in 26%. Concerning the year of birth, the results indicate that for each year the hazard increases by 5%. Finally, per each year the mother studies, the hazard of stop breast feeding decreases by 5%.

4.1 Assumption of proportional hazard ratio

Before proceeding to a multivariate cox regression, and even though a visual inspection was already done, it is important to test for the proportional hazard ratio assumption. For that, the Schoenfeld residuals are tested, as these should be independent of time when the assumption of proportional hazard ratio is held. The cox.zph() function in R tests the independence of Schoenfeld residuals across time of each covariate as well as the global model.

```
fit<-coxph(Surv(duration,delta)~
             factor(race)+
             factor(poverty)+
             factor(smoke)+
             factor(alcohol)+
             factor(pc3mth)+
             agemth+
             ybirth+
             yschool,data=bfeed)
cox.zph(fit)
```

```
##                    chisq df      p
## factor(race)      1.8973  2 0.3873
## factor(poverty)   2.8255  1 0.0928
## factor(smoke)     0.1554  1 0.6934
## factor(alcohol)   0.0481  1 0.8263
## factor(pc3mth)    0.8394  1 0.3596
## agemth           3.7256  1 0.0536
## ybirth           0.7681  1 0.3808
## yschool         10.0658  1 0.0015
## GLOBAL          11.4124  9 0.2485
```

The global test is not statistically significant, meaning that globally the assumption of proportional hazards is held. In any case, the two variables that could possibly pose some problems are agemth and yschool. As these are continuous variables, one possibility is to include nonlinear effects and recheck the assumption of the proportional hazard ratios.

```
fit2<-coxph(Surv(duration,delta)~
             factor(race)+
             factor(poverty)+
             factor(smoke)+
             factor(alcohol)+
             factor(pc3mth)+
             pspline(agemth)+
             ybirth+
             pspline(yschool),data=bfeed)
cox.zph(fit2)
```

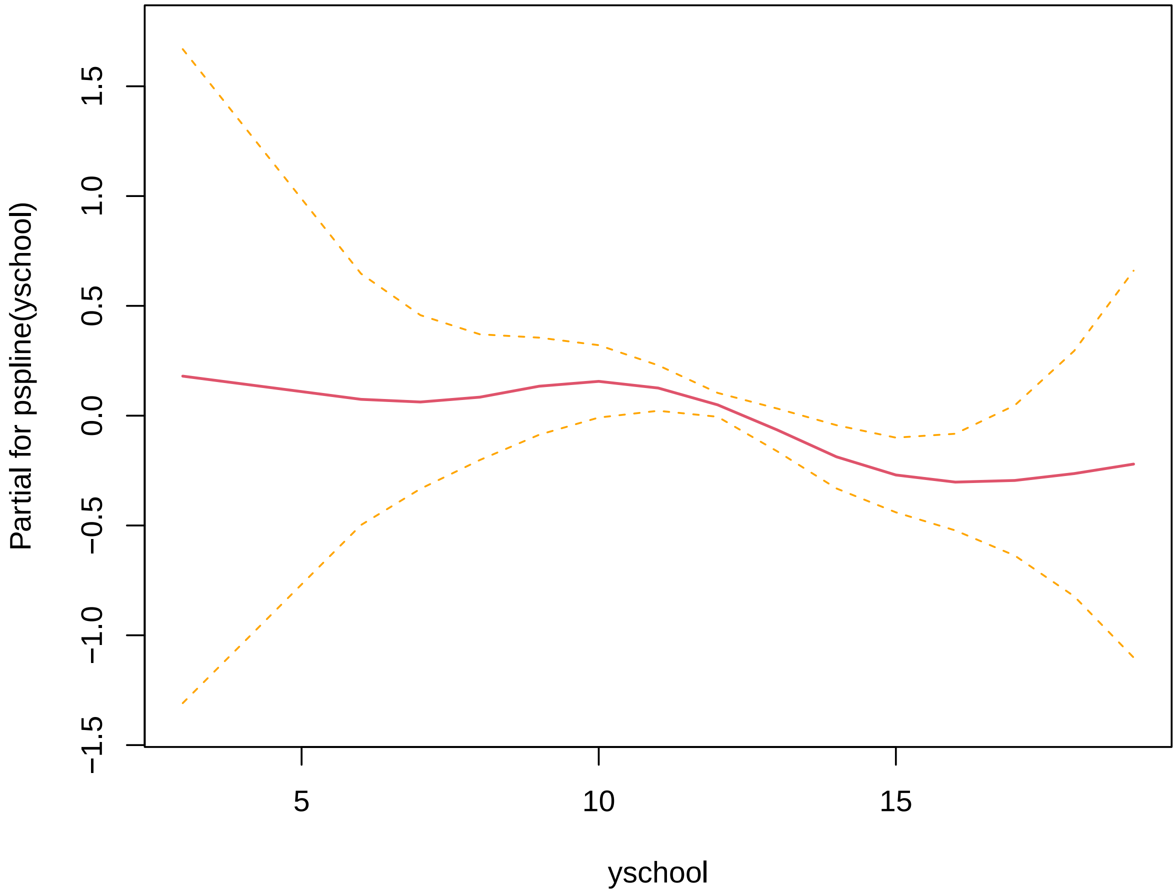
```
##                     chisq    df      p
## factor(race)      1.8743  1.99 0.3894
## factor(poverty)   2.5787  1.00 0.1076
## factor(smoke)     0.0666  1.00 0.7952
## factor(alcohol)   0.0339  1.00 0.8529
## factor(pc3mth)    1.0381  0.99 0.3059
## pspline(agemth)   4.5414  4.08 0.3484
## ybirth            0.6107  1.00 0.4336
## pspline(yschool) 13.4302  4.03 0.0095
## GLOBAL           17.2300 15.07 0.3100
```

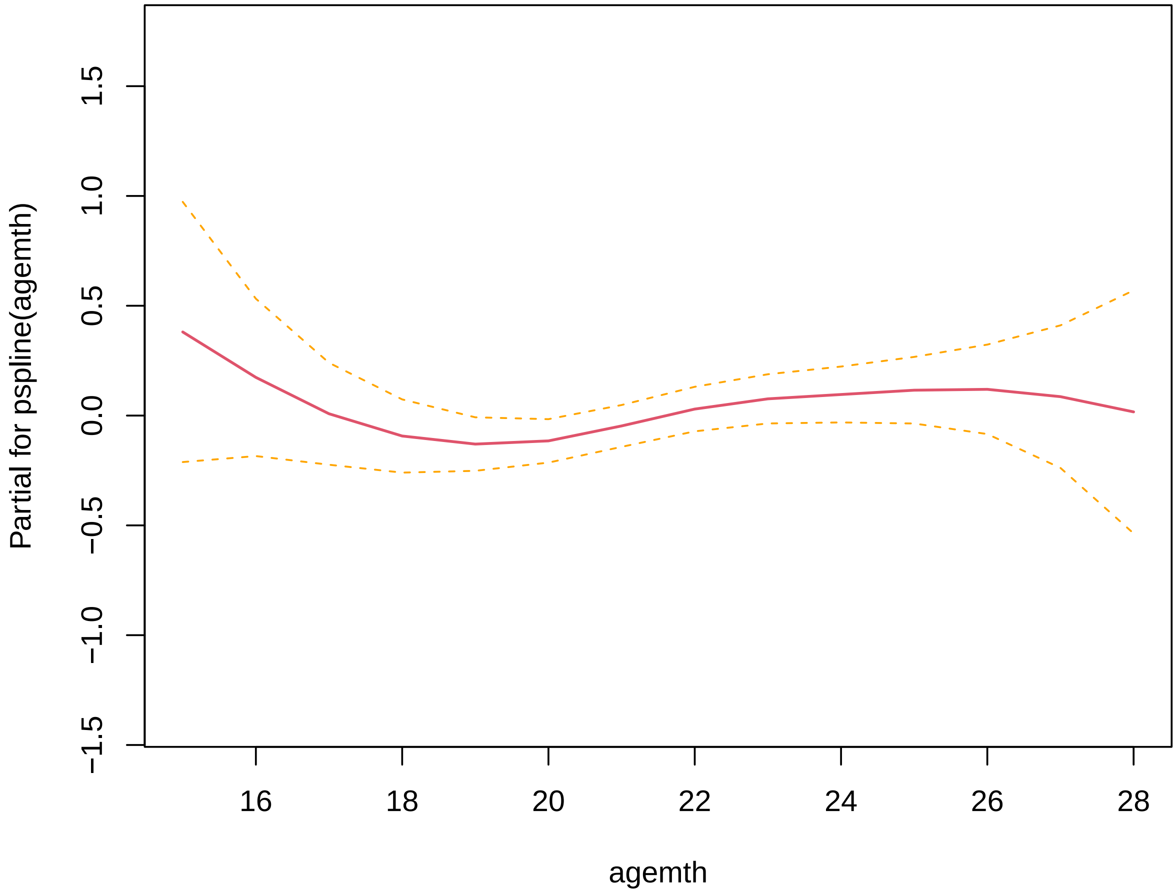
```
fit2
```

```
## Call:
## coxph(formula = Surv(duration, delta) ~ factor(race) + factor(poverty) +
##     factor(smoke) + factor(alcohol) + factor(pc3mth) + pspline(agemth) +
##     ybirth + pspline(yschool), data = bfeed)
##
##                              coef se(coef)     se2   Chisq   DF       p
## factor(race)2              0.1718   0.1064  0.1062  2.6073 1.00  0.1064
## factor(race)3              0.3256   0.0980  0.0977 11.0258 1.00  0.0009
## factor(poverty)1         -0.2014   0.0955  0.0953  4.4488 1.00  0.0349
## factor(smoke)1             0.2406   0.0798  0.0797  9.0823 1.00  0.0026
## factor(alcohol)1           0.1823   0.1239  0.1237  2.1654 1.00  0.1411
## factor(pc3mth)1          -0.0751   0.0919  0.0916  0.6681 1.00  0.4137
## pspline(agemth), linear  -0.0136   0.0190  0.0190  0.5079 1.00  0.4761
## pspline(agemth), nonlin                             7.1611 3.08  0.0709
## ybirth                     0.0810   0.0206  0.0206 15.3886 1.00 8.8e-05
## pspline(yschool), linear -0.0548   0.0239  0.0238  5.2724 1.00  0.0217
## pspline(yschool), nonlin                            4.0010 3.03  0.2649
##
## Iterations: 6 outer, 16 Newton-Raphson
##      Theta= 0.958
##      Theta= 0.902
## Degrees of freedom for terms= 2.0 1.0 1.0 1.0 1.0 4.1 1.0 4.0
## Likelihood ratio test=58.3  on 15.1 df, p=5e-07
## n= 927, number of events= 892
```

Adding a nonlinear effect on agemth seems to improve the model, because the p-value increases from 0.054 to 0.348. In fact, the Cox regression results indicate that this variable should be added with a nonlinear effect (p = 0.071). For the yschool adding a nonlinear effect does not seem to improve the model, and this variable should be added with a linear effect (p = 0.022).

In the next two plots we can see the regression terms against their predictors (and standard errors) for the agemth and yschool variables with nonlinear effects. We can easily fit a straight line within the confidence intervals bands for yschool (i.e., a linear effect is plausible). The reason for wider bands on the extreme values may be due to the low number of observations on the extremes. On the other hand, for the mother's age (agemth), it seems that young mothers have an increased hazard of stopping breast feeding, which declines at about 19 years-old and slightly increases again for ages above 19.





5. Multivariate Cox Model

Next, the step() function is applied to the multivariate cox model with all variables. This function computes the AIC for the model without each variable, and on each step it eliminates the variable if it increases the AIC until this metric can not be further minimized.

```
fit4<-coxph(Surv(duration,delta)~
             factor(race)+
             factor(poverty)+
             factor(smoke)+
             factor(alcohol)+
             factor(pc3mth)+
             pspline(agemth)+
             ybirth+
             yschool,data=bfeed)
step(fit4)
```

```
## Start:  AIC=10353.41
## Surv(duration, delta) ~ factor(race) + factor(poverty) + factor(smoke) +
##     factor(alcohol) + factor(pc3mth) + pspline(agemth) + ybirth +
##     yschool
##
##                        Df   AIC
## - factor(pc3mth)  0.99572 10352
## - pspline(agemth) 4.07160 10352
## - factor(alcohol) 0.99866 10353
## <none>                    10353
## - yschool         0.99854 10357
## - factor(poverty) 0.99932 10357
## - factor(race)    1.99358 10361
## - factor(smoke)   1.00001 10361
## - ybirth          0.99919 10367
##
## Step:  AIC=10351.91
## Surv(duration, delta) ~ factor(race) + factor(poverty) + factor(smoke) +
##     factor(alcohol) + pspline(agemth) + ybirth + yschool
##
##                        Df   AIC
## - pspline(agemth) 4.07588 10351
## - factor(alcohol) 0.99865 10352
## <none>                    10352
## - yschool         0.99885 10355
## - factor(poverty) 0.99929 10356
## - factor(race)    1.99393 10359
## - factor(smoke)   1.00007 10359
## - ybirth          0.99872 10365
##
## Step:  AIC=10350.94
## Surv(duration, delta) ~ factor(race) + factor(poverty) + factor(smoke) +
##     factor(alcohol) + ybirth + yschool
##
##                   Df   AIC
## - factor(alcohol)  1 10350
## <none>               10351
## - factor(poverty)  1 10354
## - factor(race)     2 10357
## - factor(smoke)    1 10358
## - yschool          1 10360
## - ybirth           1 10364
##
## Step:  AIC=10350.54
## Surv(duration, delta) ~ factor(race) + factor(poverty) + factor(smoke) +
##     ybirth + yschool
##
##                   Df   AIC
## <none>               10350
## - factor(poverty)  1 10354
## - factor(race)     2 10357
## - yschool          1 10358
## - factor(smoke)    1 10359
## - ybirth           1 10364
## Call:
## coxph(formula = Surv(duration, delta) ~ factor(race) + factor(poverty) +
##     factor(smoke) + ybirth + yschool, data = bfeed)
##
##                      coef exp(coef) se(coef)      z        p
## factor(race)2     0.19218   1.21189  0.10433  1.842 0.065461
## factor(race)3     0.29369   1.34137  0.09726  3.020 0.002530
## factor(poverty)1 -0.20286   0.81639  0.09265 -2.189 0.028562
## factor(smoke)1    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
##
## Likelihood ratio test=43.69  on 6 df, p=8.517e-08
## n= 927, number of events= 892
```

As expected, the final model includes race, poverty, smoke, year of birth and mother's education as factors that significantly increase the hazard of stopping breast feeding. Compared to being white, being black increases the hazard by 21%, and being of other race increases the hazard by 34%. Being in poverty decreases the hazard by 22%. Mothers who smoke are 30% more likely of stopping breast feeding. Moreover, each increment in the year of birth (between 1978 and 1986) increases the hazard by 7%. At last, each increment in the number of years that mothers studied decreases the hazard by 7%.

6. Conclusion

In this study a series of survival analysis methods were applied to understand which are the major risk factors to stop breast feeding. Because this data set has a low number of censored observations, probably a common regression would wild similar results (the author tested a multiple linear regression and the conclusions would be very similar). However, by using these methods we are able to use all information contained in the database.

These results need to be interpreted within their context: they refer to mothers in the USA from 1978 to 1986. The factors that increased the hazard of stopping breast feeding were being not white in race, smoking, and the year of birth. It is possible that there was a white privilege, in which mothers of other races, for the nature of their jobs, had to start working earlier or more hours, making it more difficult to breast feed their babies. At the same time, as years passed by, from 1978 to 1986, more women became independent and started working, so it seems reasonable to expect them to be less available to breast feed. At last, although it is not clear why smoking (but not drinking) influences breast feeding, in can be for health or cultural reasons (e.g., mothers can get sick more easily, mothers don't like to breast feed after smoking). Contrarily, the two variables that decreased the hazard was being poor, and having more years of education. Breast feeding is the cheapest way to feed a baby, so it makes sense that poor mothers resort to this feeding more frequently. Concerning the years of school, it is possible that mothers who studied more have better jobs that allow them to better conciliate work and family life. It would be interesting to study this topic in a more recent database, comparing different countries.

7. References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

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

Kramer, M. S., & Kakuma, R. (2012). Optimal duration of exclusive breastfeeding. Cochrane database of systematic reviews, (8).

Wright, A. L., & Schanler, R. J. (2001). The resurgence of breastfeeding at the end of the second millennium. The Journal of nutrition, 131(2), 421S-425S.