Survival analysis on breastfeeding completion with bfeed data frame with R

Introduction

The bfeed data frame (Klein and Moeschberger , 1997) from KMSurv library contains 927 rows and 10 columns of data about breastfeeding completion. The principal parameters of bfeed for survival analysis are "delta" and the temporal variable "duration". Other parameters are shown in Table 1.

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

Table 1: Variation in bfeed data frame in library KMSurv

In this report, the general prediction model is, firstly, created using every feature provided in the data frame with the Kaplan Meier (KM) estimator to determine the survival function. Then, the hypotheses testing is carried to enhance the model, by dropping and merging statistically insignificant features with testing methods including log-rank test on KM model, and Wald test and likelihood ratio test on cox proportional hazard model (Coxph) to address two Types of errors. Type-I error is when the null hypothesis cannot be rejected when the feature has no impact, whereas the Type-II error the null hypothesis is rejected when the feature has an impact. Finally, the model's quality is tested with the residual analysis.

Keywords: survival analysis, bfeed, R, breast feeding

General statistical information about bfeed

bfeed contains 10 columns of features which will be explored here in order to provide a general outlook of the data frame in Table 2 and Table 3.

Features (unit)	min	1st quartile	median	mean	3rd quartile	max
duration (week)	1.00	4.00	10.00	16.18	24.00	192.00
agemth (years old)	15.00	20.00	21.00	21.54	23.00	28.00
ybirth (years)	78.00	80.00	82.00	81.97	84.00	86.00
yschool (years)	3.00	12.00	12.00	12.21	13.00	19.00

Table 2: General statistical information of numerical data in bfeed data frame in library KMSurv

Features (unit)	Category: number of samples		
delta	0: 35	1: 892	
race	1: 662	2: 117	3: 148
poverty	0: 756	1: 171	
smoke	0: 657	1: 270	
alcohol	0: 848	1: 79	
pc3mth	0: 763	1: 164	

Table 3: General statistical information of factor data in bfeed data frame in library KMSurv

Density histogram charts of the "agemth" and the "yschool" features show that the two features are generally normally distributed. The ybirth-duration dot plot shows that there is an outlier in the data frame with duration of 192 weeks of breastfeeding. They are shown in Figure 1.

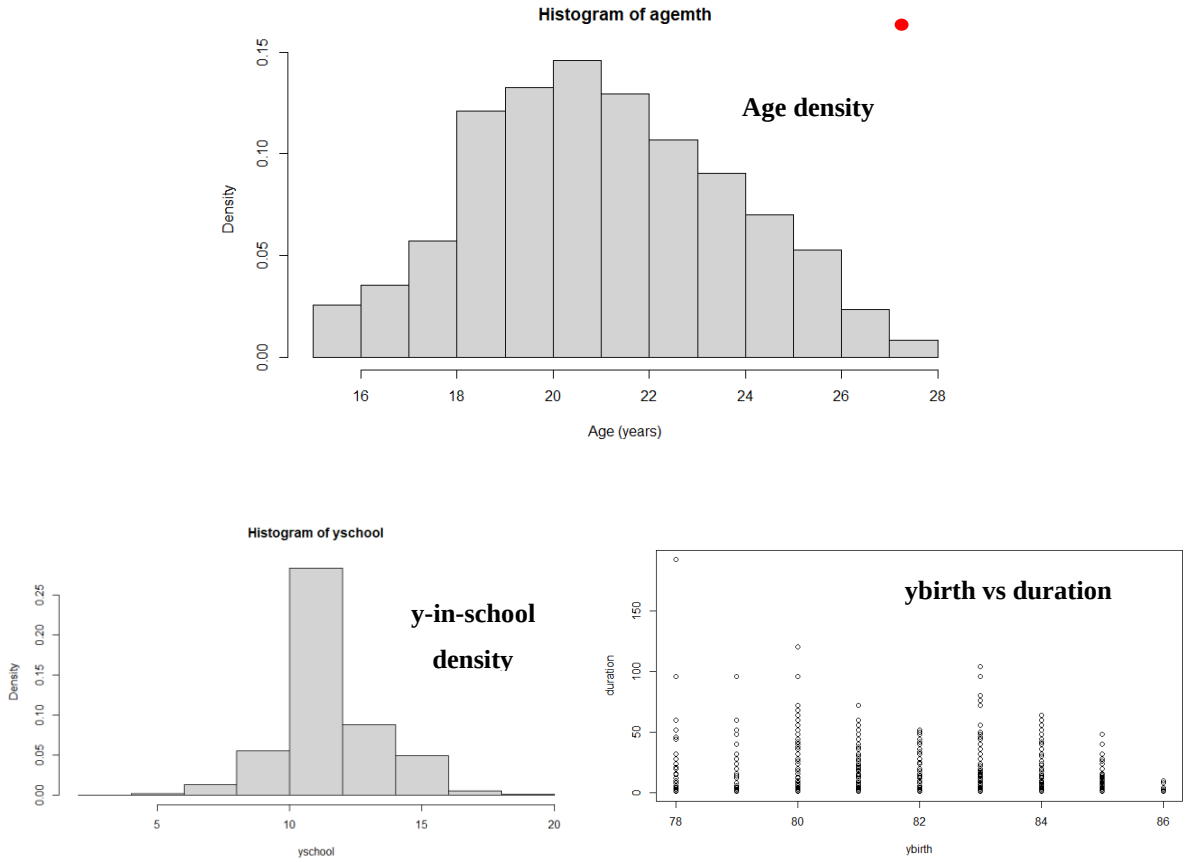

Figure1. Density histogram charts of agemth, yschool, and dot plot of ybirth-duration

Estimation of the survival function with KM estimator and Type-I error hypothesis testing

The KM method estimate breastfeeding completion survival function with 95% confidential interval (CI) of type "log" – as none of upper CI is above 1.00. The result is as shown on Figure 2. More charts were plotted using the KM estimator to see the impact of each factor feature on the survival time as in Figure 3.

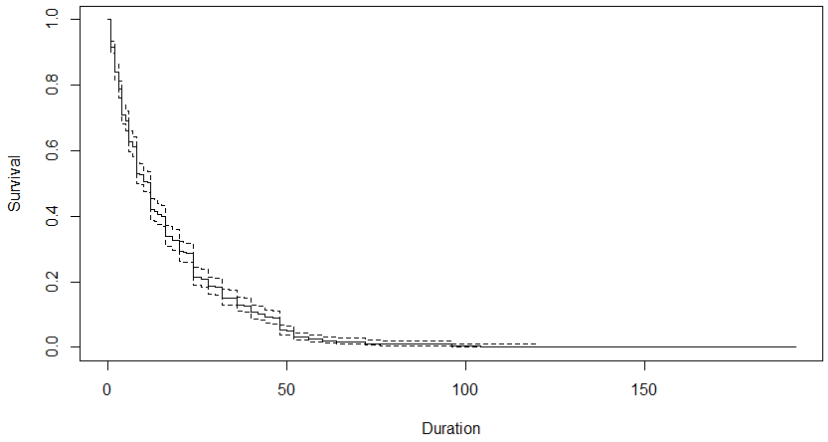

Figure 2. KM estimation of survival function of bfeed data frame with 95% CI, type "log"

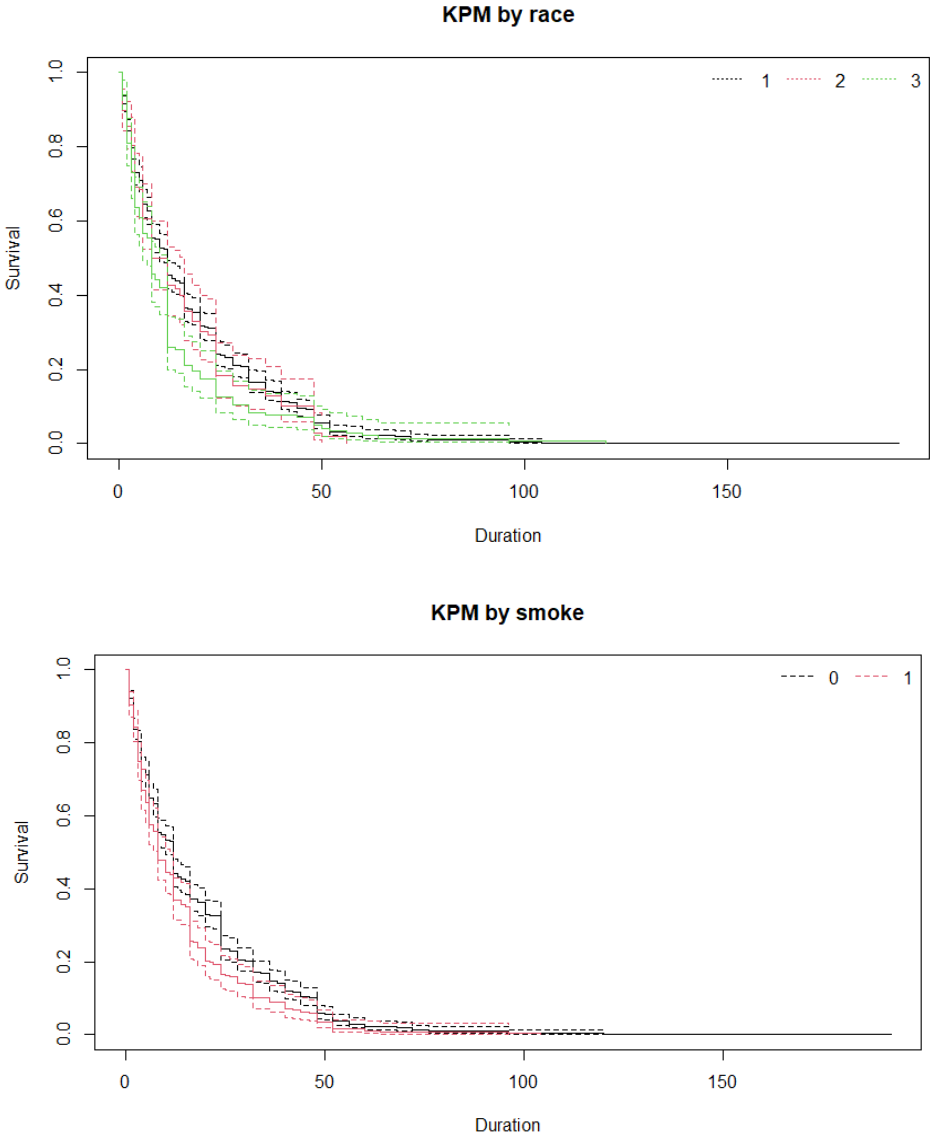

Figure 3. Examples of KM estimation of survival function by group of features of bfeed data frame with 95% CI, type "log"; race (top), smoking status (bottom)

The log-rank test has also been carried out to determine the impact of each factor to test the Type-I error. The combined result of the KM estimator and the log-tank test is presented in Table 4 and the given **level of significance (α) is 5%**, with the following hypotheses:

- H0 (null) hypothesis: the factor has an impact to the duration of breastfeeding
- H1 hypothesis: the factor **does not** have the impact

Model	**Median* (weeks)**	**S = 0.25 in CI 95%**	**chi square**	**p-value**
general	8 - 11	22 - 24	-	-
grouped by race	1: 10 - 11	1: 24 - 26	8.1	0.02

Model	Median* (weeks)	S = 0.25 in CI 95%	chi square	p-value
(type: weeks)	2: 8	2: 24		
	3: 6 - 10	3: 12 - 20		
grouped by poverty	0: 10 - 11	0: 22 - 24	0.7	0.4
(type: weeks)	1: 8 - 10	1: 16 - 25		
grouped by smoke	0: 10 - 11	0: 24 - 26	10.1	0.001
(type: weeks)	1: 8 - 11	1: 20 - 21		
grouped by alcohol	0: 9 - 11	0: 24	2	0.2
(type: weeks)	1: 6 - 10	1: 16 - 28		
grouped by pc3mth	0: 10 - 11	0: 22 - 24	0.2	0.7
(type: weeks)	1: 8 - 10	1: 21 - 30		

Table 3: Survival function of KM estimator and the log-rank test result of bfeed;
*Median is when Survival (S) = 0.5 is within 95% interval

It can be interpreted that half and 75% of the samples completed breastfeeding after 8–11 and after 22-24 weeks respectively from the KM estimator' result of the general model. The log-rank test suggests that, among factors, only the "race" and "smoke" are statistically significant to the prediction of duration of breastfeeding of the sample groups as the null hypothesis can be rejected when Chi square > 3.841 and p-value < 0.05. Their plots are shown in Figure 3.

Coxph and Type-II error hypothesis testing

With Coxph model and the Wald test, it is possible to confirm that "alcohol" and "pc3mth" can be dropped as their Wald's p-values are > 0.05 as well as the numerical "agemth", for the same reason. However, the Wald's p-value of "poverty" is <0.05 which suggests that it may be significant. LRT is performed on all of them before further decision. "race" is another interesting feature as it has two degrees of freedom (dof), and the p-value of "race = 2" is > 0.05, while being < 0.05, for "race = 3". To improve the model, group 2 and 3 of "race" are merged which means the "new race" only consider if the sample's race is white (1) or others (2). "smoke", "ybirth", and "yschool" are kept as their Wald test's p-values < 0.05. The LRT is applied on nested models to check Type-II error of each feature that could potential be dropped, except for the "new race" feature where Akaike's Information Criterion (AIC) is applied as it is not a nested model. The result of the hypothesis testing is shown in Table 4.

Model (- drop)	Wald test's p-value	LRT or AIC's result
The general model => race + poverty + smoke + alcohol + agemth + ybirth + yschool + pc3mth		
(- alcohol)	0.1892	p-value = 0.1982
(- pc3mth)	0.5208	p-value = 0.5184
The reduced general model => race + poverty + smoke + agemth + ybirth + yschool		
(- poverty)	0.0241	p-value = 0.02199
(- agemth)	0.4142	p-value = 0.4134
(- race) (+ new race)	2: 0.0836	df = 6 / 5
	3: 0.0025	AIC 10350.54 / 10349.17
		*p-value = 0.00155 (Wald test)

Table 4: The Wald test's p-value and the result of the LRT and the AIC on features in the question.

From Table 4, it can be concluded that "alcohol", "pc3mth", and "agemth" can be dropped as their p-values from the Wald test and the LRT concur to be > 0.05, in contrary to "poverty" which should not be dropped to avoid Type-II error as both the Wald test and LRT suggest that it is not possible to reject the null hypothesis. Regarding "race" feature, the AIC test and the new Wald test's p-value suggest that transforming it to "new race" improves the model; although very slightly.

The best fit model and the Residual analysis

The best fit model of bfeed's survival analysis is the reduced model containing five features out of original eight as presented in Table 5 whereas the interpretation of Coxph of the model is in Table 6.

general model	race + poverty + smoke + alcohol + agemth + ybirth + yschool + pc3mth
best fit model	new race + poverty + smoke + ybirth + yschool

Table 5: Comparison between the general and the best fit models after KM and Coxph analysis

feature	Coefficient (coef)	Standard error se(coef)	Interpretation (when other features are the same including the duration, we can estimate that ….)
poverty	-0.20666	0.09253	samples in poverty are ~20% ± 9% less likely to have already completed breastfeeding
smoke	0.25418	0.07826	samples who smoke are ~25% ± 8% more likely to have already completed breastfeeding
ybirth	0.07173	0.01788	the likelihood of completing breastfeeding is up ~7% ± 2% with one unit increase in ybirth
yschool	-0.06569	0.01995	the likelihood of completing breastfeeding is up ~7% ± 2% with one unit increase in yschool
new race	0.24617	0.07779	samples in poverty are ~24% ± 8% less likely to have already completed breastfeeding

Table 5: Comparison between the general and the best fit models after KM and Coxph analysis

The residual analysis is conducted to evaluate the quality via graphical observation. Some selected results presented in Figure 4. The straight flat line in (A) and (C) are the evidence that "yschool" and "smoke" **conform linearity check** and the same can be said to other features **except in (B)** where a clear trend line is formed; indicating that the model performs poorly with increase in "duration". In (D), it is an example of a Schoenfeld analysis which is used to check that **the proportional hazard assumption in not violated** over time. It is true for all significant features

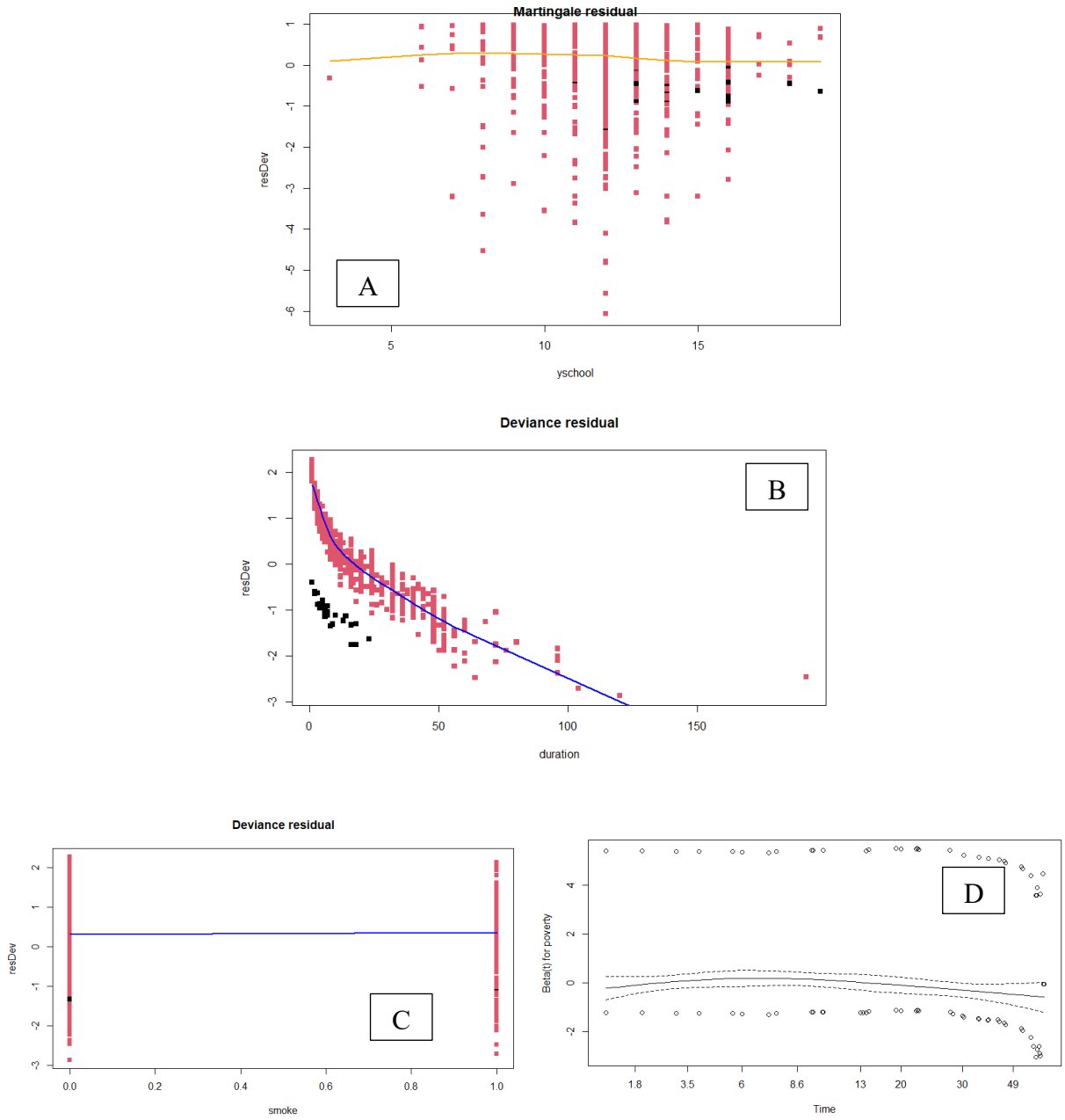

Figure 4. Examples of residual analysis of the best fit models; martingale yschool (A), and deviance duration (B) and smoke (C), and Schoenfeld residual of poverty (D)

Conclusion

To conclude, the life table generated by KM estimator of the general model show that 50% of samples completed breastfeeding after 8-11. The number rises to 75% after 22-24 weeks. The result of hypotheses tests, namely log-rank, Wald, and LRT test, suggest that five out eight features in the data frame, namely “poverty”, “smoke”, “ybirth”, “yshool” and “race”, are statistically significant; and for race, the samples only need to be considered if they are white or others. Despite the satisfaction of proportional hazard assumption and linearity is conform for almost all feature, the model has a poor performance with increase in “duration”, as shown by non-linearity in the residual analysis. The log transformation was performed but cannot improve the situation. The finding suggests that the “duration” feature may lack some of its components and cannot provide reliable prediction.

Reference

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