Survival Analysis Final Exam

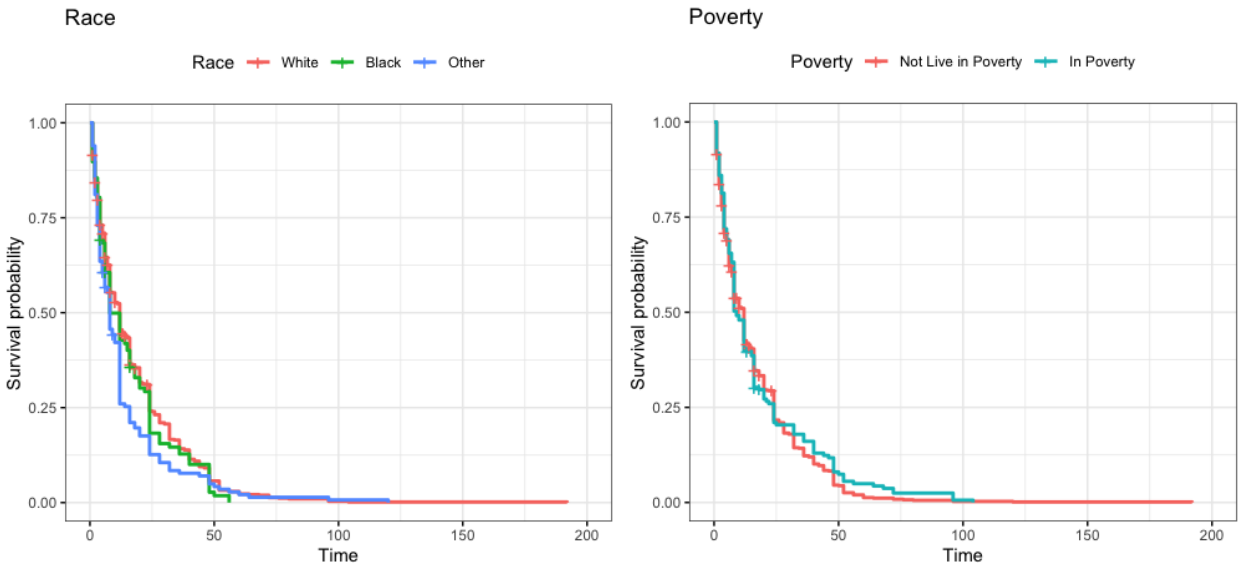
The analysis is aimed to study what factors are related to the duration of breastfeeding.

Method

Descriptive Statistics

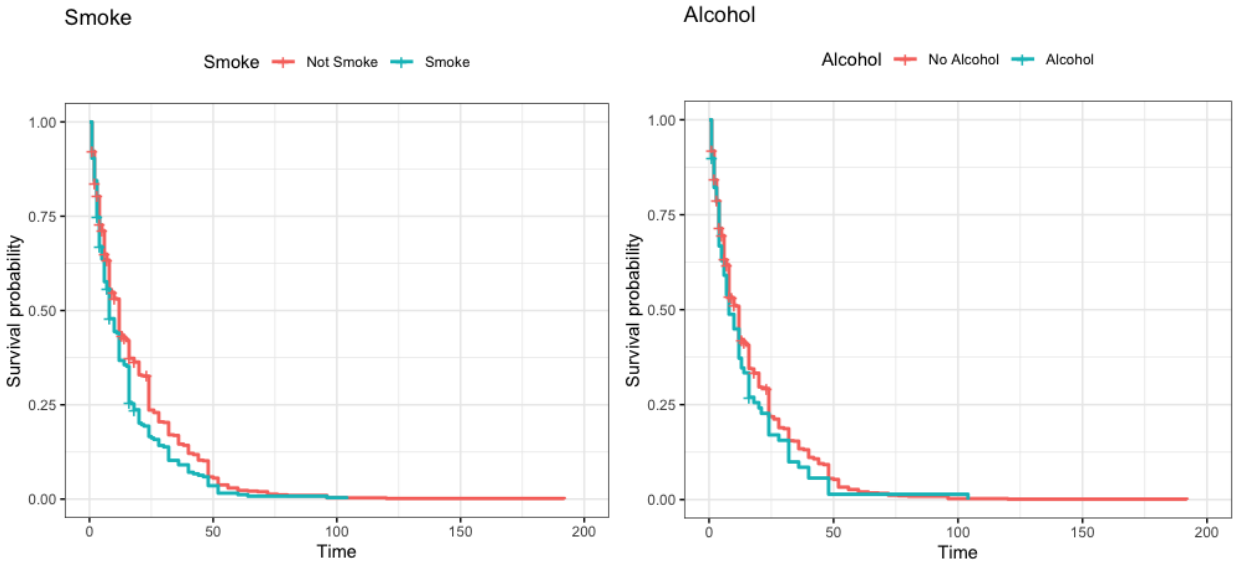
In order to understand the the dataset, descriptive statistics is needed to gain understanding of the distributions. The dataset contains information of breastfeeding length and indicator whether it was completed (1 yes, 0 no) for 927 subjects. In addition, there are also information about subject characterestiscs including race, poverty, smoke, alcohol which are binary variables: 1 if yes, 0 if no. Moreover, we also have age, year of birth and years of education for all subjects. I analyzed the distributions of each variable within the dataset.

The dataset contains 3 race categories: white, blackc and others. Among all subjects, 71.4% of them are White, 12.6% of them are black, and 16% of them are in other races. Kaplan-Meier Estimator was conducted. According to the plot, all races had similar breastfeeding probability However, as time goes by, white had longer breasfeeding length compared to other races.

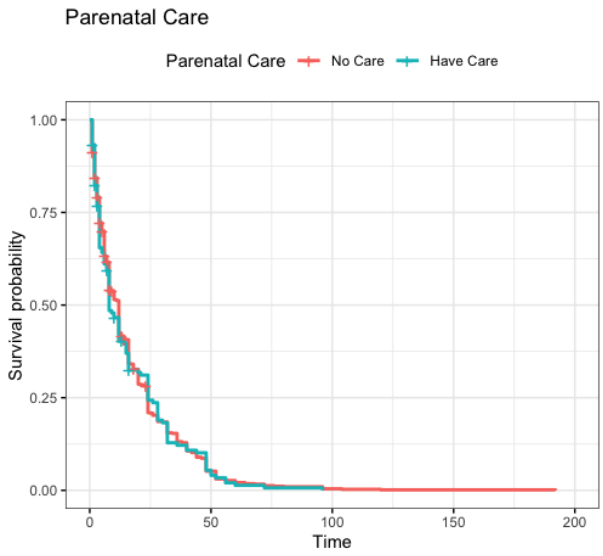


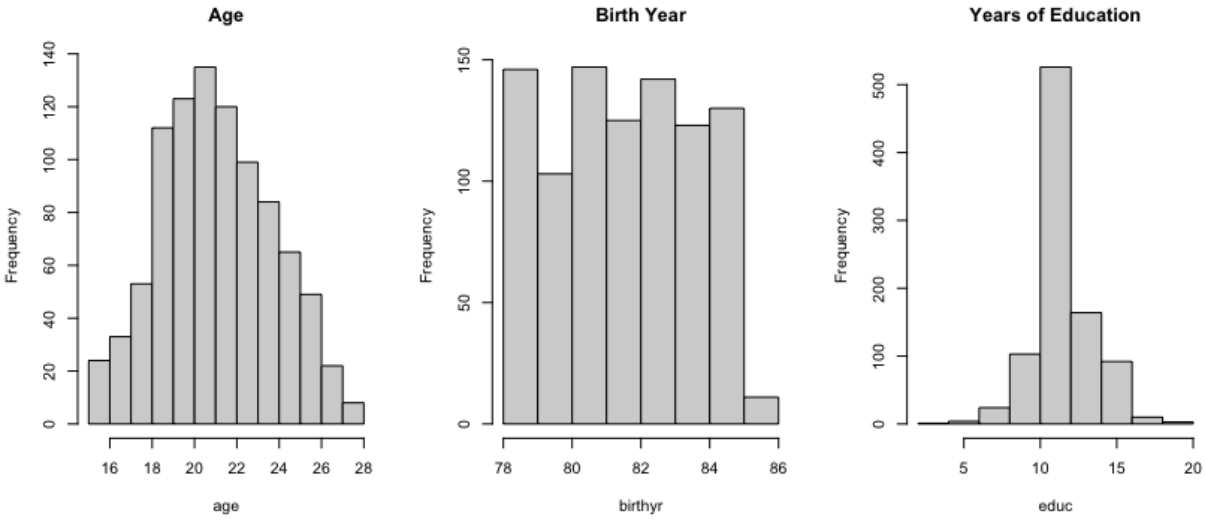
Poverty is a feature to investigate subjects' living condition. 82.5% of the mothers live in poverty, and 18.4% of them are not. According to the Kaplan Meier survival plot, mothers from both group had similar breastfeeding probabilities; however, more affulent mothers had longer breastfeeding length than poor mothers.

In smoking status, 70.9% of the mothers do not smoke, and 29.1% of them smoke. As for alcohol consumption, 91.4% of the mothers did not have alcohol, and 8.5% of them had alcohol. According to the Kaplan Meier survival plot, mothers from smoking group had similar breastfeeding probabilities at the beginning; however, non-smoking and non-drinking mothers had higher breastfeeding probabilities and longer breastfeeding length as time passes.



Among all subjects, 82.3% of the mothers did not seek or never saught prenatal care after third month of pregnancy, and only 17.7% of them did. According to the Kaplan Meier survival plot, mothers from both group had similar breastfeeding probabilities. Differ from previous covariates, they did not have very obvious different breastfeeding probabilities, yet, mothers without prenatal care had longer breastfeeding time length.





Among all mothers participated in this study, their average age is 21 years old and most of them are or below 23 years old, with youngest mothers at 15 years old and oldest mothers age 28. There are 222(24%) of them are below 20 years old, 626 (67.5%) of them are between 20 to 25 years old, and only 79 (8.5%) of them are above 25 years old. Most of them were born in the 80s, between 1981 to 1985. They had 12 years of education on average and mostly, with least years of education at 3 years and longest at 19 years of education.

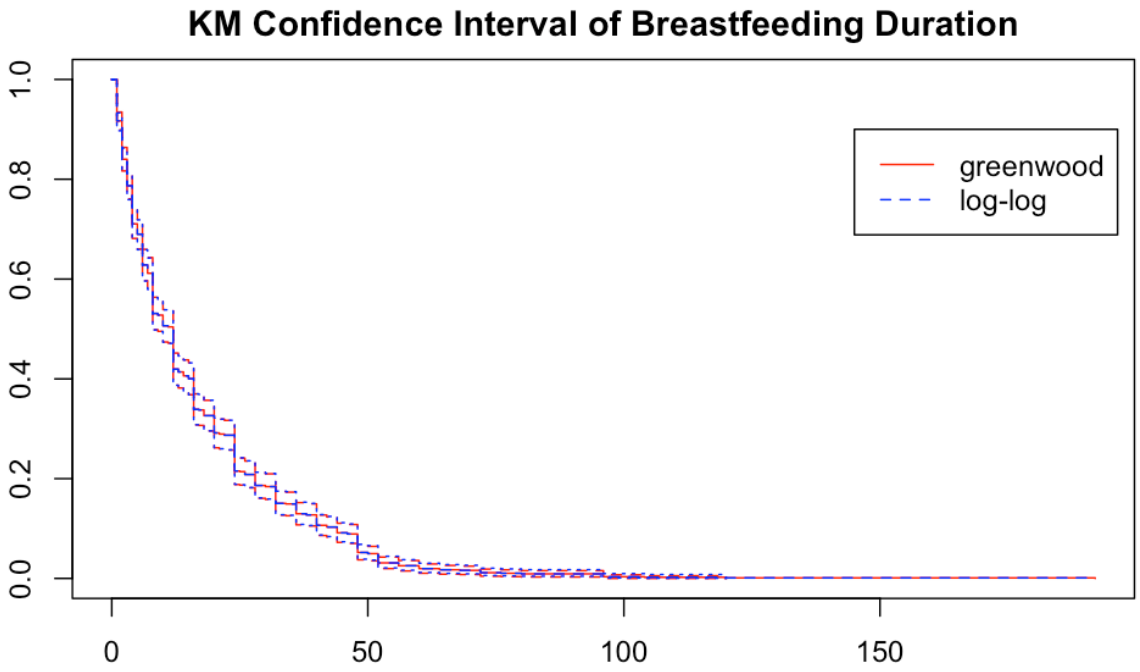
The above descriptive analysis helped me to gain basic knowledge of the sample. These subjects are young moms with 12 years of education on average. Majority of them are white, did not drink alcohol or smoke at time of birth, and did not live in poverty.

Dummy Variables

Since birth year and age are continuous numerical variables, their values will influence the outcome greatly. Therefore, it is better to change it to categorical binary variables, and this transformation is easier for later model selection and building. Years of Education stays as continuous numerical variables, as it indicat one additional year of education relative to breastfeeding hazard ratio.

Birth year is seperated to 2 levels: people born before the 80s and during the 80s. Age is divided into 3 categories: below20 years old, in mid 20s (between 20 to 25 years old) and late 20s (older than 25 years old). Originally, race has 3 levels indicating White, Black and Other race; here, race is seperated into 3 binary variables, each represent white, black and other race.

The purpose of the study is to investigate what factors are and how much they are reated to breastfeed duration. Fist of all, I would like to examine the survival probability of breastfeeding duration and events without any covariates, by looking at the confidence interval plot (below). I used Greenwood's and log-log approximation to find the 95% confidence interval for Kaplan Meier estimates. For the entire dataset, the 2 confidence intervals are very similar.



It is useful to model as a function of a possible set of covariates to study this association. All ties are dealt using “breslow”. In the beginning, I created a null cox model, without any covariates in it. The -2*log liklihood of the null model is 10503.22.

Collett’s model selction approach is chosen for this study. This approach assumes that all variables are thought to be on an equal footing. The first step is to fit multiple univariate models, for each of the covariate; the identify and select significant predictors at level 0.2. After creating dummy variales, there are 13 variabels; therefore, 13 univariate cox proportional hazard models are created. Their coefficients(betas), hazard ratio, p-values, and -2LogLikelihood values are shown in table below:

	beta	HR	p.value	neg.2LogL
white	-0.17400	0.84100	0.01910	10499.51000
black	0.05370	1.05500	0.59500	10504.58700
other	0.22700	1.25500	0.01280	10498.97500
poverty	-0.06770	0.93500	0.43200	10504.24000
smoke	0.21900	1.24500	0.00301	10496.33000
alcohol	0.15900	1.17300	0.18500	10503.18100
prenatal3	0.03310	1.03400	0.70700	10504.72600
below20	0.05250	1.05400	0.49900	10504.41200
mid20s	-0.03330	0.96700	0.62900	10504.63400
late20s	-0.01530	0.98500	0.87500	10504.84100
before80	-0.05170	0.95000	0.57000	10504.54000
in80s	0.05170	1.05300	0.57000	10504.54000
educ	-0.04210	0.95900	0.01500	10498.94400

Based on the Cox Proportional Hazards models, different race and birth year have varying influence on breastfeeding completion probability. For example, white mothers are 0.841 times the hazard ratio for completing breastfeeding than non-white mothers, holding other covariates constant. On the other hand, other race mothers are 1.255 times the hazard for completing breastfeeding,holding other covariates constant. None of the birth year group is significant.The hazard ratio for smoking mothers is 1.245 times than non-smoking mothers,holding other covariates constant; and mothers who consume alcohol are 1.173 times the hazard ratio than non-drinking mothers.

From above result table, white,other race,smoke,alcohol,Educ have p.values less than 0.2. -2LogLiklihood values also confirms this variable selection. Therefore, these covariates are included for multivariate model.

In step 2, I fitted a multivariate model with all the significant predictors chosen before, and use backward selection to eliminate non-significant variables at level p2=0.10. Aftering fitting a multivariate cox model with white, other race, smoke, alcohol and educ, and this model has -2 Loglikelihood of 10358.46. Only smoke has p-value less than 0.1.

```
Call:
coxph(formula = Surv(length, complete) ~ white + other + smoke +
    alcohol + educ, data = dat)

            coef exp(coef) se(coef)      z       p
white   -0.14438   0.86556  0.10323 -1.399 0.16191
other    0.14726   1.15866  0.12746  1.155 0.24796
smoke    0.23937   1.27045  0.07921  3.022 0.00251
alcohol  0.13826   1.14827  0.12210  1.132 0.25750
educ    -0.02512   0.97519  0.01810 -1.388 0.16520

Likelihood ratio test=23.77  on 5 df, p=0.0002401
n= 927, number of events= 892
```

Step 3 starts with the final step 2 model (smoke only), I added each of the non-signifcant varable from Step 1 individually to the model, using forward selection, and selected those variables at 0.1 significant level. The non-significant variables were black, poverty, prenatal3, age, and born before the 80s and in the 80s. The p-values for each of the original non signifcant predictors are shown below:

	Neg2LogLike <dbl>	p.value <dbl>
Smoke	10373.05	0.002079805
Smoke+Black	10372.18	0.345744559
Smoke+Poverty	10371.50	0.218678110
Smoke+Prenatal Care	10373.03	0.894669820
Smoke+Born before 80s	10372.45	0.442407442
Smoke+Born in the 80s	10372.45	0.442407442
Smoke+Below20s	10372.80	0.614745686
Smoke+Mid20s	10372.92	0.718289773
Smoke+late20s	10373.04	0.908787697

Results

After considering non-signifcant covariates, none of these variables have p-value less than 0.1. In addition, chi square of 2.706 is for p=0.1. None of the additional variables leed to reductions in the -2* Log Likelihood of 2.706 or more. Therefore, the final model includes only smoke as covaraites.

$$Breastfeed\ length = \lambda_0(t) * \exp\ (0.227 * Smoke)$$

	Coefficients	Exp(Coef)	Lower. 95% CI	Upper. 95% CI	Std.Err	p-value
Smoke	0.227	1.255	1.086	1.45	0.074	0.002

Only smoking is significantly (p-value<0.05) positively associated with breastfeeding duration proportional hazards. Smoke has coefficients of 0.227 and the model has hazard ratio of 1.255. Within 95% confidence interval, the hazard ratio is between 1.086 and 1.45. The final model has AIC score of 10375.05, and -2*log likelihood value of 10373.05.

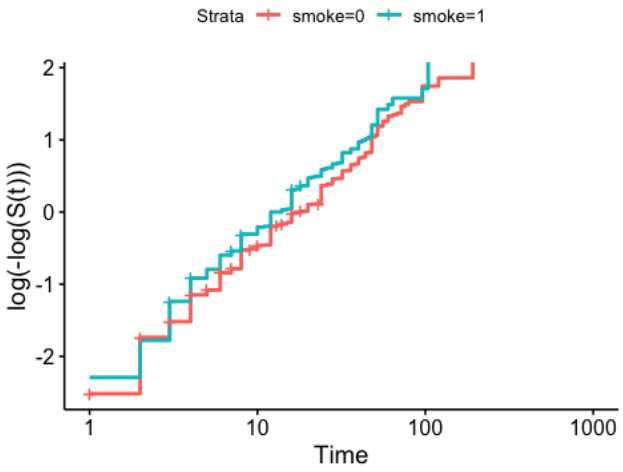
Conclusion

In conclusion, only smoking status is significantly associated with breastfeeding surivival probability. Smoking subjects have 1.255 times higher hazarad ratio than non-smoking subjects. Suprisingly, none of the age, race, poverty, alcohol status, and alcohol are significant to the breastfeeding duration cox proportion hazard. The age range is relatively small, and mostly of them are in their mid 20s. Although whtie, other race, alcohol and years of education was significant in step 1, but did not sustain in step 2 creteria.

Extra Credit

Assessing Proportional Hazard Assumption

The Cox PH model has 2 assumptions. First, the baseline hazard depends on time, but not on covariates. Second, the hazard ratio depends on the covariates, not on time. Assumption 2 is vital to establish a proportioanl hazard model and the hazard ratio is for a subject with covarite X comparing to the reference group. Since only smoke is chosen for the final model, I will check the PH assumption graphically, using plots of log(-log(S)) vs. log(time) for smoke group using KM estimates.



Kleinbaum suggests a strategy of assumotion the PH holds unless the estiamted log cumulative hazard curves cross, or look very unparallel over time. The above plot shows two lines cross and becomes unparallel and cross over time. This indicates that the final Cox model from above gives an average Hazard Ratio, averaged over event times.

Sample Size:

80% power to detect the observed hazard ratio for mothers who smoke versus mothers who do not smoke, with a two-sided significance level of 0.05

Alpha=0.05

Z=1.96

Beta=0.2

Z_beta=0.84

D=$\frac{4*(1.96+0.84)^2}{(\ln(1.255))^2}$=608.5=609

Different Approach on variable selection

Using the same set of variables, I performed a backward stepwise variable selection based on AIC scores. First I start with the full cox model with white, black, other, poverty, smoke, alcohol, below20s, mid20s, late20s, before 80s, in 80s, prenatal care and educ.

Model	AIC
FULL	10370.43
white + black + other + poverty + smoke + alcohol + prenatal3 + before80 + in80s + below20 + mid20s +educ	10370.43
white + black + other + poverty + smoke + alcohol + prenatal3 + before80 + below20 + mid20s + educ	10370.43
white + black + poverty + smoke + alcohol + prenatal3 + before80 + below20 + mid20s + educ	10370.43
white + black + poverty + smoke + alcohol + before80 + below20 + mid20s + educ	10368.59

white + black + poverty + smoke + alcohol + before80 + below20 + educ	10367.16
white + black + poverty + smoke + alcohol + before80 + educ	10365.28
white + poverty + smoke + alcohol + before80 + educ	10364.27
white + poverty + smoke + alcohol + educ	10363.58
white + poverty + smoke + educ	10363.37

Differ from Collett's method, if I were to choose backward AIC selection, my final model would have white, poverty, smoke and education as covariates.

	Coef	Exp(coef)	Lower. 95	Upp. 95	P-Value
white	-0.246	0.782	0.672	0.911	0.002
Poverty	-0.215	0.807	0.673	0.967	0.020
Smoke	0.258	1.295	1.111	1.509	0.000
educ	-0.040	0.961	0.926	0.997	0.034

All of the covariates are signficant. Holding other covaraites constant, being white is associated with 24.6% less risk of stopping breastfeeding. Living in poverty is associated with 0.807 times the hazard of stopping breastfeeding, holding other covariates constant. Education has similar negative association. The hazard ratio for one more year od education is 0.961, suggesting

education is protective with regards of stopping breastfeeding. On the contrary, smoking increases the hazard ratio by 1.295 times.