Exploring Factors Related to Duration of Breastfeeding

Introduction

In this study, we are going to explore factors related to duration of breastfeeding through survival analysis. We are using a data set consisting of 927 mother-infant pairs for which the mothers chose to breastfeed their children, the children were born after 1978, at gestational ages between 20-45 weeks. The data set include: minimum of duration of breastfeeding and time on study in weeks (length), indicator of whether the breastfeeding was completed (complete), race of mother (race), if mother in poverty (poverty), if mother smoked at time of birth of child (smoke), if mother drank at time of birth of child (alcohol), mother's age at birth of child (age), year of child's birth (birthyr), years of education of mother (educ), and if mother sought prenatal care after third month of pregnancy or never (prenatal3).

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.6.2
```

```
library(EnvStats)
```

```
## Warning: package 'EnvStats' was built under R version 3.6.3
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
```

```
##     extract
```

```
# read in data
dat <- read_csv('E:/NYU/2020SPRING/SurvivalAnalysis/final/breastfeed.csv')
```

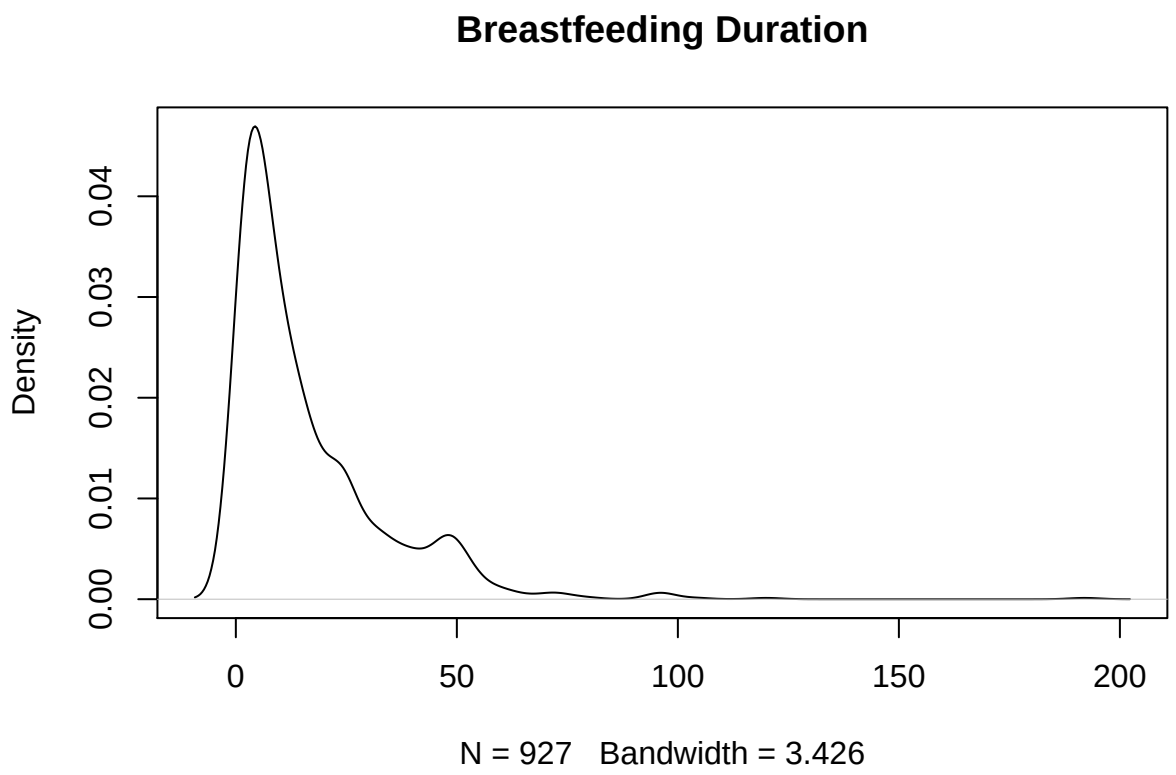
```
## Parsed with column specification:
## cols(
##   length = col_double(),
##   complete = col_double(),
##   race = col_double(),
##   poverty = col_double(),
##   smoke = col_double(),
##   alcohol = col_double(),
##   age = col_double(),
##   birthyr = col_double(),
##   educ = col_double(),
##   prenatal3 = col_double()
## )
```

```
attach(dat)
```

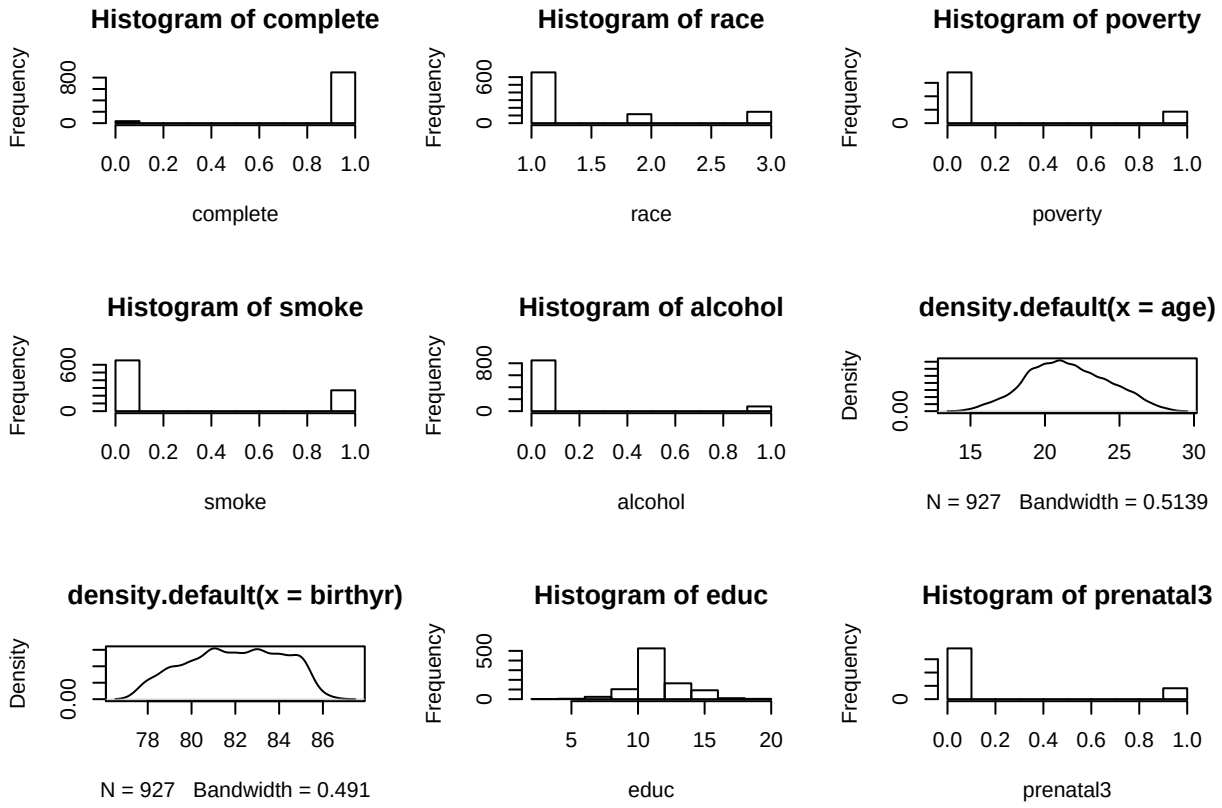
Descriptive Analysis

First we explore the variables in the data set.

```
plot(density(length),main='Breastfeeding Duration')
```



```
par(mfrow=c(3,3))
hist(complete)
hist(race)
hist(poverty)
hist(smoke)
hist(alcohol)
plot(density(age))
plot(density(birthyr))
hist(educ)
hist(prenatal3)
```



We can see that there is not much censoring in the data. We also notice that there are three levels of races, with the majority of white. Thus, we create indicators for white and non-white.

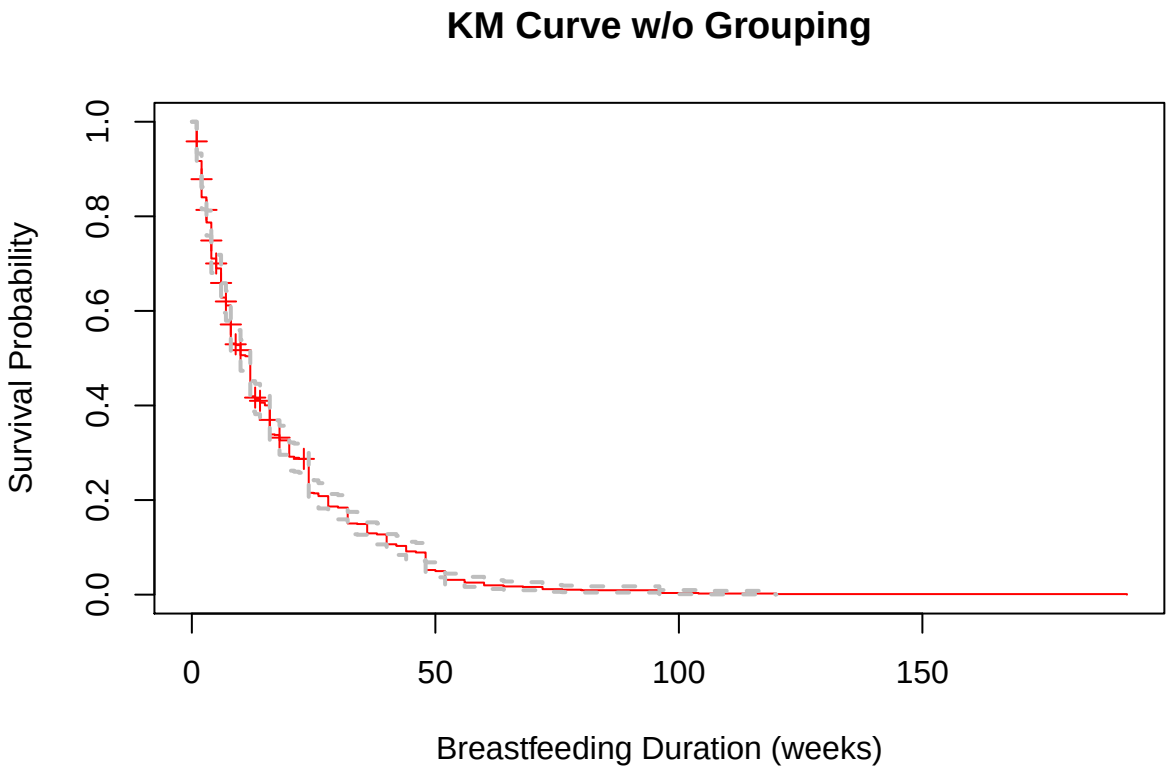
```
dat <- dat %>% mutate(white=if_else(race==1,1,0))
attach(dat)
```

```
## The following objects are masked from dat (pos = 3):
##
##     age, alcohol, birthyr, complete, educ, length, poverty, prenatal3,
##     race, smoke
```

Overall Survival Distribution

We estimate the overall survival distribution without grouping through the Kaplan-Meier estimator. We plot the KM curves and confidence intervals through the log-log approach.

```
km.all <- survfit(Surv(length,complete)~1,conf.type='log-log')
plot(km.all,main='KM Curve w/o Grouping',xlab='Breastfeeding Duration (weeks)',ylab='Survival Probabili
```



The red solid line is the overall survival curve, while the two grey dashed lines are the 95% log-log confidence interval. Censoring is also marked in the plot.

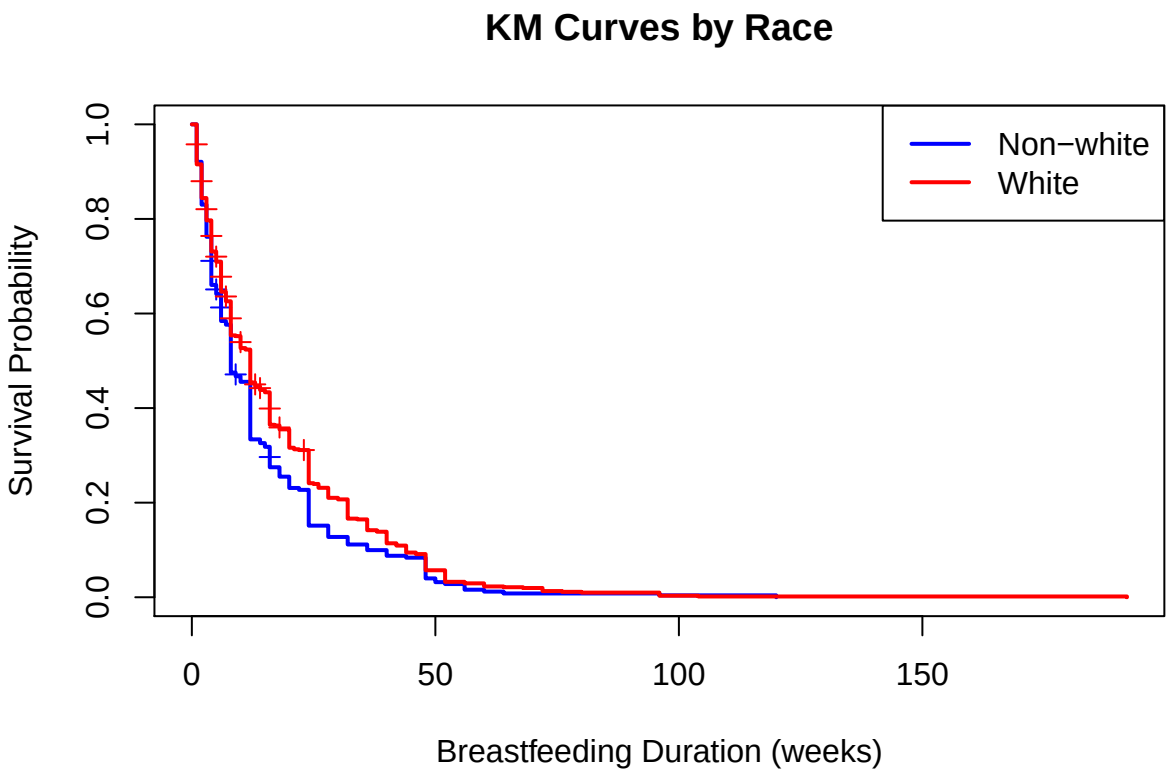
Two-Group Comparisons

We compare survival distributions grouping by white, poverty, smoke, alcohol, and prenatal3.

Race

First we compare the survival distributions of white and non-white by eyeballing their KM curves.

```
km.race <- survfit(Surv(length,complete)~white)
plot(km.race,main='KM Curves by Race',xlab='Breastfeeding Duration (weeks)',ylab='Survival Probability'
legend("topright",col=c('blue','red'),lwd=rep(2,2),c('Non-white','White'))
```



From the KM curves we can see the survival probabilities for white mother is higher than that of non-white mother. Then we use the CMH logrank test to compare between white and non-white groups.

```
survdiff(Surv(length,complete)~white)
```

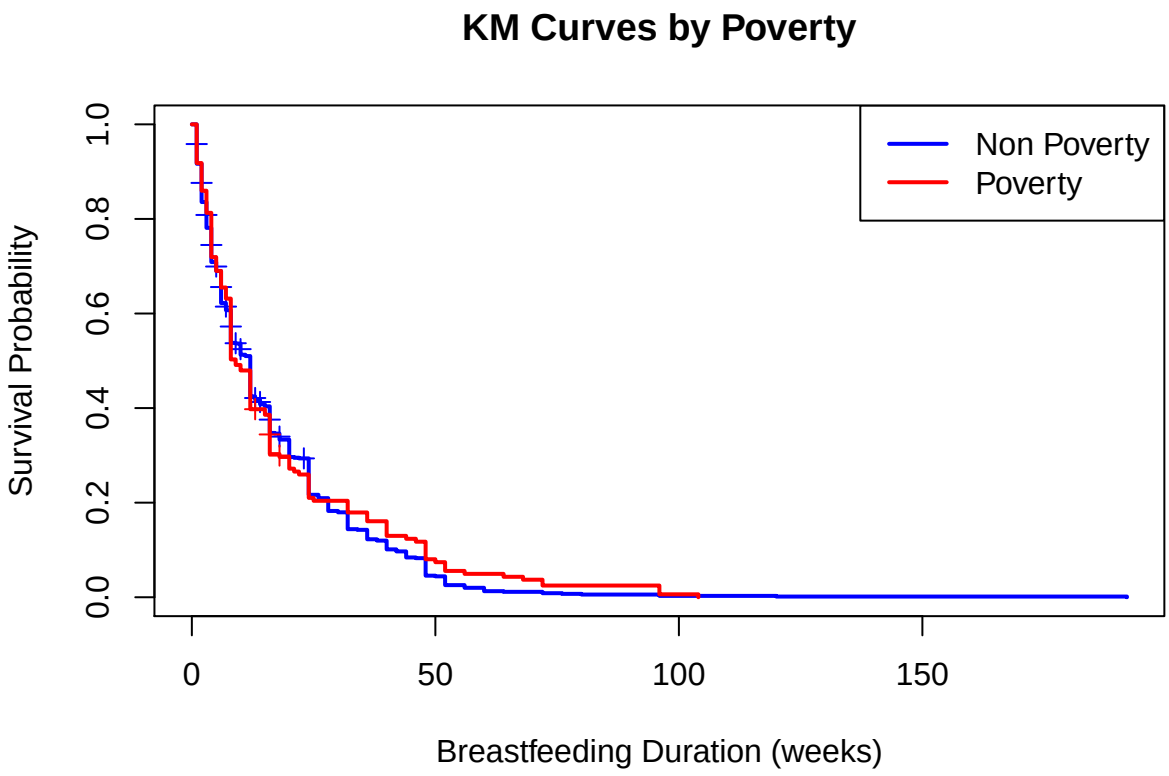
```
## Call:
## survdiff(formula = Surv(length, complete) ~ white)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## white=0 265      258      228      4.08      6.31
## white=1 662      634      664      1.40      6.31
##
##  Chisq= 6.3  on 1 degrees of freedom, p= 0.01
```

The significant result indicates a difference in breastfeeding duration for white group and non-white group. The breastfeeding duration for white mother is significantly longer than that of non-white mother.

Poverty

First we compare the survival distributions of mother in poverty with mother not in poverty by eyeballing their KM curves.

```
km.poverty <- survfit(Surv(length,complete)~poverty)
plot(km.poverty,main='KM Curves by Poverty',xlab='Breastfeeding Duration (weeks)',ylab='Survival Probabi
legend("topright",col=c('blue','red'),lwd=rep(2,2),c('Non Poverty','Poverty'))
```



From the KM curves we don't see much difference between the two groups. Then we use the CMH logrank test to compare between poverty and non-poverty groups.

```
survdiff(Surv(length,complete)~poverty)
```

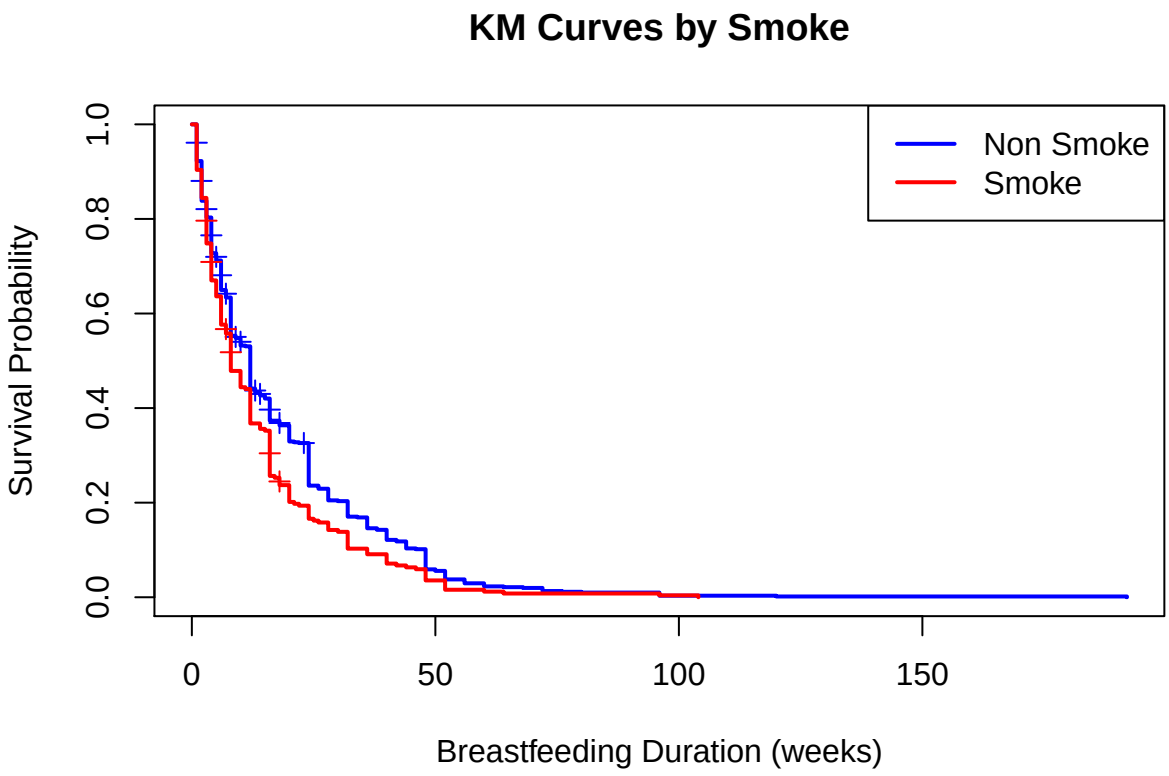
```
## Call:
## survdiff(formula = Surv(length, complete) ~ poverty)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## poverty=0 756      724      715     0.121     0.713
## poverty=1 171      168      177     0.489     0.713
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.4
```

The non-significant result indicates no difference in breastfeeding duration for poverty and non-poverty mothers.

Smoke

First we compare the survival distributions of mother smoked with mother not smoked by eyeballing their KM curves.

```
km.smoke <- survfit(Surv(length,complete)~smoke)
plot(km.smoke,main='KM Curves by Smoke',xlab='Breastfeeding Duration (weeks)',ylab='Survival Probabilit
legend("topright",col=c('blue','red'),lwd=rep(2,2),c('Non Smoke','Smoke'))
```



From the KM curves we can see the survival probabilities for mother who smoke is lower than that of mother not smoked. Then we use the CMH logrank test to compare between smoke and non-smoke groups.

```
survdiff(Surv(length,complete)~smoke)
```

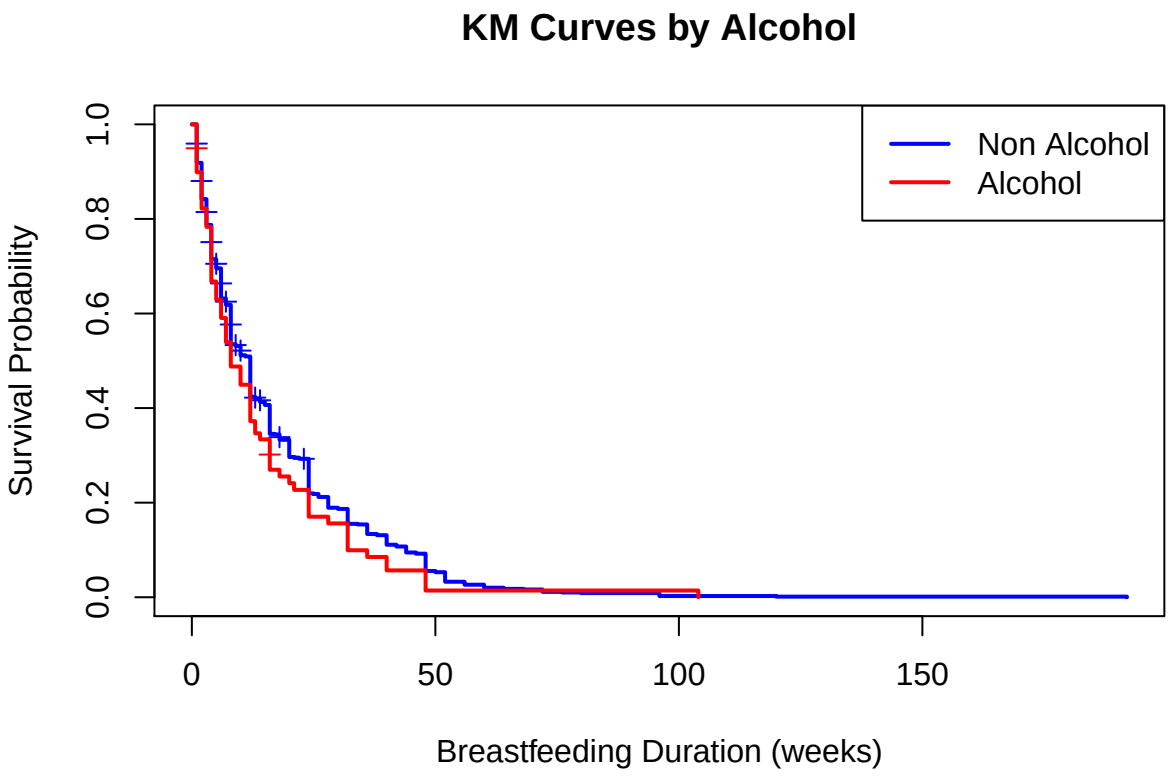
```
## Call:
## survdiff(formula = Surv(length, complete) ~ smoke)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## smoke=0 657      629      667      2.21      10.1
## smoke=1 270      263      225      6.56      10.1
##
##  Chisq= 10.1  on 1 degrees of freedom, p= 0.001
```

The significant result indicates a difference in breastfeeding duration for mother smoked and mother not smoke. The breastfeeding duration for smoked mother is significantly lower than that of non-smoked mother.

Alcohol

First we compare the survival distributions of mother drank with mother not drank by eyeballing their KM curves.

```
km.alcohol <- survfit(Surv(length,complete)~alcohol)
plot(km.alcohol,main='KM Curves by Alcohol',xlab='Breastfeeding Duration (weeks)',ylab='Survival Probabi
legend("topright",col=c('blue','red'),lwd=rep(2,2),c('Non Alcohol','Alcohol'))
```



From the KM curves we can see the survival probabilities for mother who drank is lower than that of mother not drank. Then we use the CMH logrank test to compare between drank and non-drank groups.

```
survdiff(Surv(length,complete)~alcohol)
```

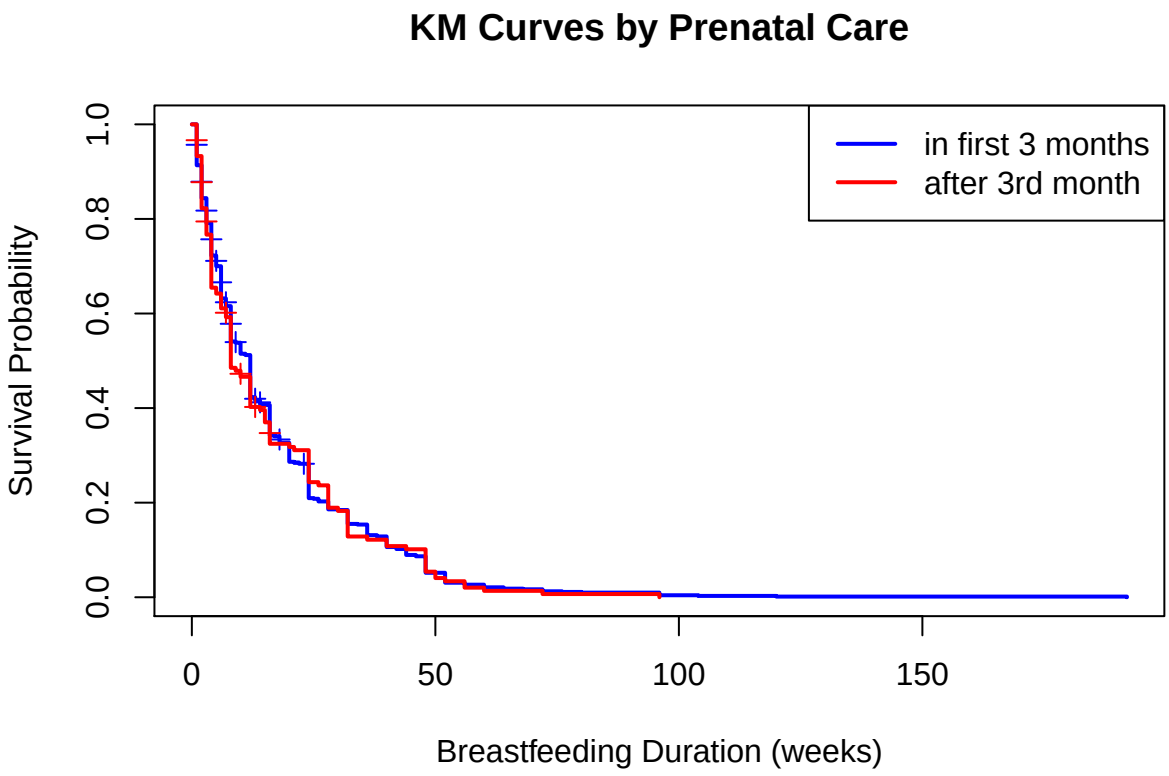
```
## Call:
## survdiff(formula = Surv(length, complete) ~ alcohol)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## alcohol=0 848      816    826.3     0.129      2.01
## alcohol=1  79       76     65.7     1.628      2.01
##
##  Chisq= 2  on 1 degrees of freedom, p= 0.2
```

The non-significant result indicates no difference in breastfeeding duration for alcohol and non-alcohol mothers.

Prenatal Care

First we compare the survival distributions of mother sought prenatal care after third month of pregnancy or never with mother sought prenatal care in first three months of pregnancy by eyeballing their KM curves.

```
km.prenatal3 <- survfit(Surv(length,complete)~prenatal3)
plot(km.prenatal3,main='KM Curves by Prenatal Care',xlab='Breastfeeding Duration (weeks)',ylab='Survival
legend("topright",col=c('blue','red'),lwd=rep(2,2),c('in first 3 months','after 3rd month'))
```



From the KM curves we don't see much difference between the two groups. Then we use the CMH logrank test to compare between groups.

```
survdiff(Surv(length,complete)~prenatal3)
```

```
## Call:
## survdiff(formula = Surv(length, complete) ~ prenatal3)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## prenatal3=0 763      736      740     0.024     0.162
## prenatal3=1 164      156      152     0.117     0.162
##
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.7
```

The non-significant result indicates no difference in breastfeeding duration for mother sought prenatal care after third month of pregnancy or never and mother sought prenatal care in first three months of pregnancy.

By comparing between groups, we found that there are significant group differences between white mother and non-white mother as well as between smoked mother and non-smoked mother.

Fitting a Cox PH Regression Model

We try to find a model including all possible important covariates through the Cox Proportional Hazards regression, assuming proportional hazards. The Peto-Breslow method is used to adjust for ties. Then we do regression diagnostics afterwards.

Variable Selection

We conduct variable selection through Collett's model selection approach. First, fit a univariate model for each covariate, and identify the predictors significant at 0.20 level through Wald tests:

```
covariates <- c('white','poverty','smoke','alcohol','age','birthyr','educ','prenatal3')
univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('Surv(length,complete)~',x)))
# fit a univariate model for each covariate
univ_models <- lapply(univ_formulas, function(x){coxph(x,data=dat,ties="breslow")})
# extract data
univ_results <- lapply(univ_models,
                       function(x){
                          x <- summary(x)
                          p.value<-signif(x$wald["pvalue"], digits=2)
                          wald.test<-signif(x$wald["test"], digits=2)
                          beta<-signif(x$coef[1], digits=2) #coeficient beta
                          HR <-signif(x$coef[2], digits=2) #exp(beta)
                          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                          HR <- paste0(HR, " (",
                                       HR.confint.lower, "-", HR.confint.upper, ")")
                          res<-c(beta, HR, wald.test, p.value)
                          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",
                                        "p.value")
                          return(res)
                         })
# print results table
as.data.frame(t(as.data.frame(univ_results,check.names=FALSE)))
```

```
##              beta HR (95% CI for HR) wald.test p.value
## white       -0.17   0.84 (0.73-0.97)       5.5   0.019
## poverty    -0.068    0.93 (0.79-1.1)      0.62    0.43
## smoke        0.22      1.2 (1.1-1.4)       8.8   0.003
## alcohol      0.16     1.2 (0.93-1.5)       1.8    0.18
## age       -0.0048         1 (0.97-1)      0.13    0.72
## birthyr     0.048          1 (1-1.1)       8.2  0.0043
## educ       -0.042   0.96 (0.93-0.99)       5.9   0.015
## prenatal3   0.033       1 (0.87-1.2)      0.14    0.71
```

Based on the Wald tests, white, smoke, alcohol, birthyr, and educ are significant at 0.2 level, and thus they are included into the multivariate model in the next step.

Second, fit a multivariate model with all significant univariate predictors, and use backward selection to eliminate non-significant variables at 0.1 level:

```
fit1 <- coxph(Surv(length,complete)~white+smoke+alcohol+birthyr+educ,ties="breslow")
fit1
```

```
## Call:
## coxph(formula = Surv(length, complete) ~ white + smoke + alcohol +
##     birthyr + educ, ties = "breslow")
##
##             coef exp(coef) se(coef)      z        p
## white   -0.20640   0.81350  0.07715 -2.675 0.007462
## smoke    0.22009   1.24619  0.07914  2.781 0.005419
## alcohol  0.12314   1.13104  0.12205  1.009 0.313015
```

```
## birthyr  0.06939   1.07186  0.01785  3.887 0.000102
## educ    -0.05085   0.95042  0.01902 -2.674 0.007506
##
## Likelihood ratio test=35.26  on 5 df, p=1.336e-06
## n= 927, number of events= 892
```

Alcohol is removed from this step because its Wald test suggests non-significant result at 0.1 level. Thus, we fit a Cox PH model without alcohol:

```
fit2 <- coxph(Surv(length,complete)~white+smoke+birthyr+educ,ties="breslow")
fit2
```

```
## Call:
## coxph(formula = Surv(length, complete) ~ white + smoke + birthyr +
##     educ, ties = "breslow")
##
##             coef exp(coef) se(coef)      z        p
## white   -0.20781   0.81236  0.07718 -2.693  0.00709
## smoke    0.23246   1.26171  0.07810  2.977  0.00292
## birthyr  0.06960   1.07208  0.01784  3.902 9.53e-05
## educ    -0.04947   0.95173  0.01898 -2.607  0.00913
##
## Likelihood ratio test=34.27  on 4 df, p=6.549e-07
## n= 927, number of events= 892
```

Then start with the model in step 2 (i.e. model includes white, smoke, birthyr, and educ), we re-consider each of the non-significant variables from step 1 (i.e. poverty, age, and prenatal3) using forward selection, with significance level of 0.1:

```
add1(fit2,scope=~white+smoke+birthyr+educ+poverty+age+prenatal3,test="Chisq")
```

```
## Single term additions
##
## Model:
## Surv(length, complete) ~ white + smoke + birthyr + educ
##           Df   AIC    LRT Pr(>Chi)
## <none>       10479
## poverty    1 10476 4.2729  0.03872 *
## age        1 10480 0.2082  0.64816
## prenatal3  1 10480 0.5151  0.47294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on results of likelihood ratio tests, poverty is significant at 0.1 level, which means that adding poverty results in a significant reduction in the -2logL. Thus, the final model should include white, smoke, birthyr, educ, and poverty, not considering quadratic and interaction terms.

Final Model

From all of the above steps we finally get a Cox PH model, with Breslow method to adjust for ties:

```
fit.coxph <- coxph(Surv(length,complete)~white+smoke+birthyr+educ+poverty,ties="breslow")
fit.coxph
```

```
## Call:
## coxph(formula = Surv(length, complete) ~ white + smoke + birthyr +
##     educ + poverty, ties = "breslow")
##
```

```
##             coef exp(coef) se(coef)      z        p
## white   -0.22890   0.79541  0.07780 -2.942 0.003258
## smoke    0.24549   1.27825  0.07827  3.136 0.001710
## birthyr  0.06870   1.07112  0.01783  3.853 0.000117
## educ    -0.06100   0.94083  0.01988 -3.068 0.002154
## poverty -0.18813   0.82851  0.09235 -2.037 0.041632
##
## Likelihood ratio test=38.55  on 5 df, p=2.93e-07
## n= 927, number of events= 892
```

The final model includes white, smoke, birthyr, educ, and poverty as predictors, indicating that those are important factors associated with breastfeeding duration.

Results

According to the results of the final Cox PH model, whether the mother is white, whether the mother smoked at time of birth of child, year of child's birth, years of mother's education, and whether mother is in poverty are important factors related to duration of breastfeeding. Mother being white, mother with more years of education, and mother in poverty are associated with less risk of shorter breastfeeding duration, while mother smoked at time of birth of child and larger year of child's birth are associated with higher risk of shorter breastfeeding duration.

As for marginal effects of each predictor, holding other factors constant, being white is associated with 20% less risk of shorter breastfeeding duration; holding other factors constant, mother smoked at time of birth of child is associated with 28% more risk of shorter breastfeeding duration; holding other factors constant, a one year increase in year of child's birth is associated with 7% more risk of shorter breastfeeding duration; holding other factors constant, a one year increase in years of education of mother is associated with 6% less risk of shorter breastfeeding duration; holding other factors constant, being mother in poverty is associated with 17% less risk of shorter breastfeeding duration.

Diagnostics

Deviance Residuals

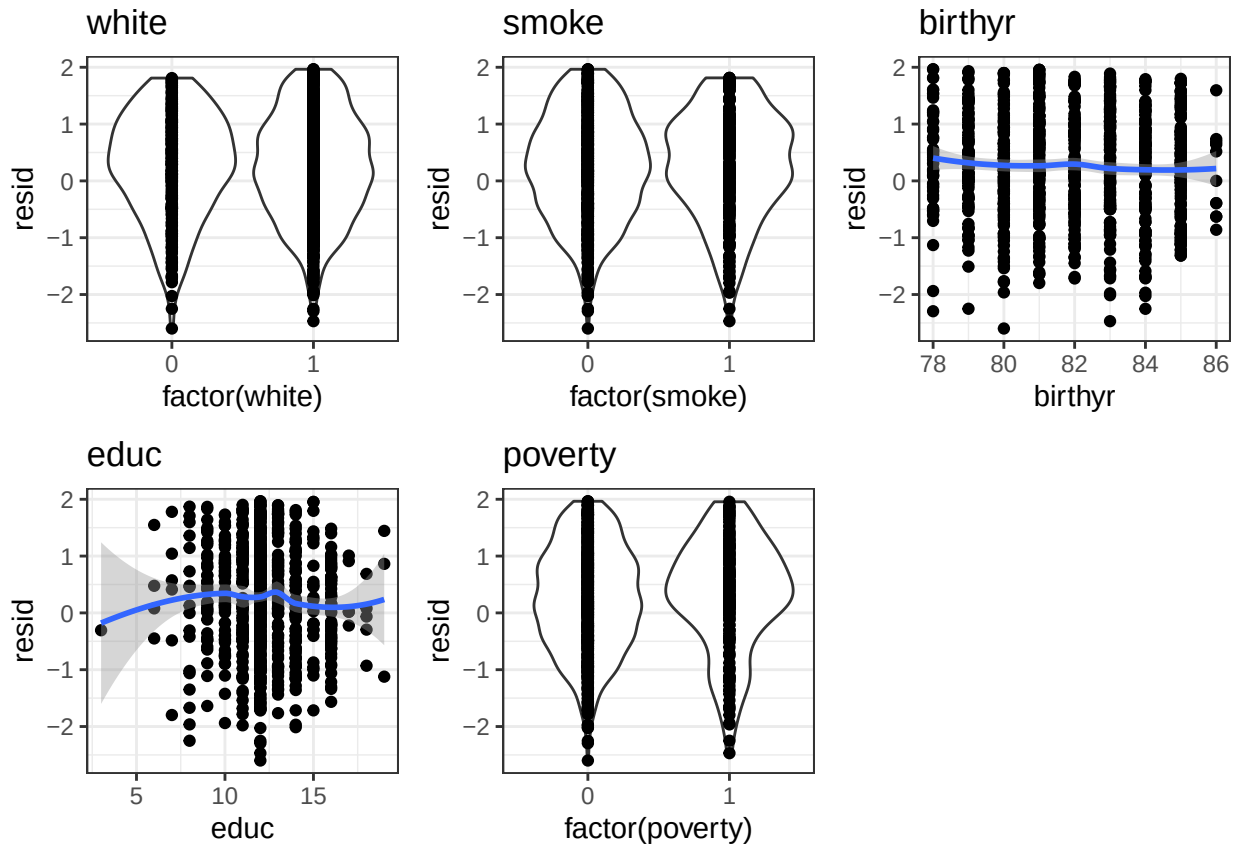
After fitting the Cox PH model, we use residual plots for regression diagnostics. Here we are calculating the deviance residuals.

```
# Martingale residuals
dat$resid <- residuals(fit.coxph,type='deviance')
# white
res1 <- dat %>% ggplot(aes(factor(white),resid)) +
    geom_violin() +
    geom_point() +
    labs(title='white') +
    theme_bw() +
    theme(legend.key = element_blank())
# smoke
res2 <- dat %>% ggplot(aes(factor(smoke),resid)) +
    geom_violin() +
    geom_point() +
    labs(title='smoke') +
    theme_bw() +
    theme(legend.key = element_blank())
# birthyr
res3 <- dat %>% ggplot(aes(birthyr,resid)) +
```

```
    geom_point() +
    geom_smooth() +
    labs(title='birthyr') +
    theme_bw() +
    theme(legend.key = element_blank())
# educ
res4 <- dat %>% ggplot(aes(educ,resid)) +
    geom_point() +
    geom_smooth() +
    labs(title='educ') +
    theme_bw() +
    theme(legend.key = element_blank())
# poverty
res5 <- dat %>% ggplot(aes(factor(poverty),resid)) +
    geom_violin() +
    geom_point() +
    labs(title='poverty') +
    theme_bw() +
    theme(legend.key = element_blank())

ggarrange(res1,res2,res3,res4,res5,ncol=3,nrow=2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



All residuals plots seem to be fine, without structure or trend detected.

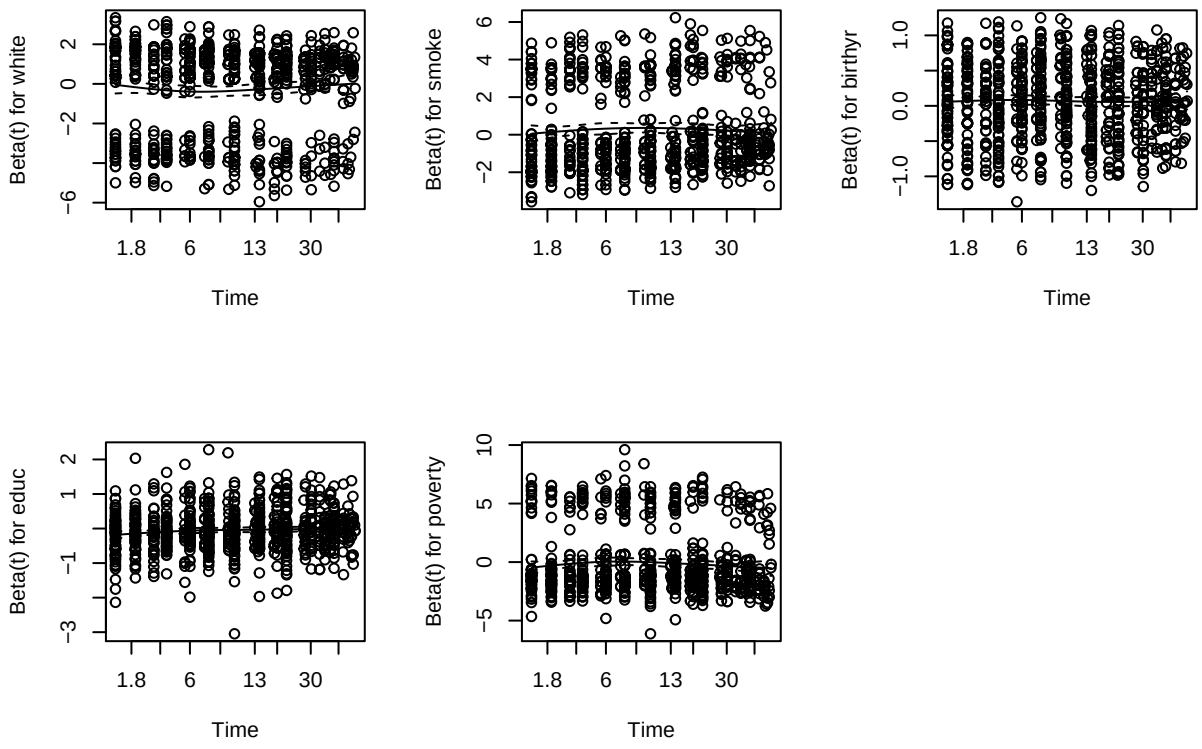
Proportional Hazards Assumption

Since all of the analyses so far are under proportional hazards assumption, we want to assess the PH assumption for the Cox PH model. Here we use the cox.zph() function to check proportionality.

```
test.ph <- cox.zph(fit.coxph)
test.ph
```

```
##          chisq df      p
## white    0.961  1 0.3271
## smoke    0.122  1 0.7265
## birthyr  0.891  1 0.3452
## educ     9.956  1 0.0016
## poverty  2.042  1 0.1530
## GLOBAL  10.370  5 0.0654
```

```
par(mfrow=c(2,3))
plot(test.ph)
```



Based on the cox.zph() test, there is strong evidence of non-proportional hazards for educ, while the global test has a p-value slightly in excess of 0.05. Those graphs plot the scaled Schonenfeld residuals against KM transformed time, and no trend is detected from the plots.

Unfortunately, there is no time-varying covariate in this data set, so that we cannot fix non-proportionality of educ by including time-varying covariates or interactions with time. Since educ is regarded as a continuous variable, it's unrealistic to do a stratified analysis for each educ level.

Sample Size Calculation

We conduct a sample size calculation to have 90% power to detect the observed hazard ratio (1.2) for mothers who smoke versus mothers who do not smoke, with a two-sided significance level of 0.05.

```
4*(1.96+1.282)^2/log(1.2)^2
```

```
## [1] 1264.765
```

In order to have 90% power to detect the hazard ratio of 1.2 for mothers who smoke versus mothers who do not smoke with a two-sided significance level of 0.05, we need a sample size of 1265. However, this does not consider drop-offs.

Conclusion

In this study, we explore factors associated with duration of breastfeeding.

According to CMH Logrank tests, significant group differences are detected between white and non-white mothers as well as smoke and non-smoke mothers. Breastfeeding duration is significantly longer in white group than non-white group, and longer in non-smoke group than smoke group.

Through the Cox PH model, we found that whether the mother is white, whether the mother smoked at time of birth of child, year of child's birth, years of mother's education, and whether mother is in poverty are important factors related to duration of breastfeeding. Mother being white, mother with more years of education, and mother in poverty are associated with higher hazard ratio of longer breastfeeding duration, while mother smoked at time of birth of child and larger year of child's birth are associated with higher risk of shorter breastfeeding duration.

Overall, the Cox PH model fits well. No significant trend is detected from the residuals plots. The global proportional hazard assumption is satisfied, except for the educ covariate. However, due to the lack of time-varying covariate in the dataset, we cannot go further to improve this model by including time-varying covariates. Future work might focus on collecting time-varying covariates further improve the model.

Reference

https://bookdown.org/sestelo/sa_financial/how-to-evaluate-the-ph-assumption.html

http://www.sthda.com/english/wiki/cox-model-assumptions

https://stat.ethz.ch/R-manual/R-devel/library/survival/html/residuals.coxph.html