Survival Case Study on Breastfeeding

Wanrun Yang & Leo Liu

Contents

```
library(KMsurv)
library(survival)
library(survminer)
library(dplyr)
library(MASS)
library(survivalmodels)
```

Background & Objective

Breastfeeding may have an impact on the mental health of children and adolescents, with shorter periods of breastfeeding associated with higher rates of behavioural problems [9]. [7] indicates that smoking by mothers with breast feeding may not only reduce the protective properties of milk but also affect the health of the infant, thereby causing adverse changes in the composition of milk. And contrary to common wisdom, drinking alcohol may not have a significant effect on breastfeeding mothers [8]. Exploring simple and dependable predictors, early identification of mothers who are unable to complete breast feeding and appropriate intervention in the early stages of nursing will provide a higher chance for mothers who complete breast feeding [2,4,5].

Our research goal is to explore the relationship between different factors and the cessation of breast feeding by using the existing breast feeding data set, and find the factors to model and build the final model distribution. In our research, we use `duration` as the time variable and `delta` as the state variable. If the status is 1, which means completed breast feeding, and 0, which means incomplete breast feeding The following is a basic introduction to the data set.

Description of data

Our research data set is from mothers with breast feeding through survival paths mapping based on time series data [1]. The data included duration of breast feeding,the indicator of completed breast feeding, and some covariates such as race of mother, whether the mother was in poverty, whether the mother smoked at birth of child, whether the mother used alcohol at birth of child, age of mother at birth of child, year of birth, education level of mother, and whether to receive prenatal care after 3 months.

```
data(bfeed)
head(bfeed)
```

```
##   duration delta race poverty smoke alcohol agemth ybirth yschool pc3mth
## 1       16     1    1       0     0       1     24     82      14      0
## 2        1     1    1       0     1       0     26     85      12      0
## 3        4     0    1       0     0       0     25     85      12      0
## 4        3     1    1       0     1       1     21     85       9      0
## 5       36     1    1       0     1       0     22     82      12      0
## 6       36     1    1       0     0       0     18     82      11      0
```

Theory

Single Factor Analysis

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

n_i : number of subjects at risk at event time t_i. d_i : number of events at event time t_i.

Log-rank Test

A 2×2 table conditional on information at t_i:

	Group 1	Group 2	Totals
Death	d_{1i}	d_{2i}	d_i
Alive	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
At risk	n_{1i}	n_{2i}	n_i

d_{1i} follows a hypergeometric distribution with mean $\hat{e}_{1i}$ and variance $\hat{v}_{1i}$:

$$\hat{e}_{1i} = d_i \left(\frac{n_{1i}}{n_i} \right)$$
$$\hat{v}_{1i} = d_i \left(\frac{n_i - d_i}{n_i - 1} \right) \left(\frac{n_{1i}}{n_i} \right) \left(1 - \frac{n_{1i}}{n_i} \right)$$

Test statistic for two-sample comparison:

$$Q_1 = \frac{\left[\sum_{i=1}^{m} (d_{i1} - \hat{e}_{1i}) \right]^2}{\sum_{i=1}^{m} \hat{v}_{1i}}$$

Q_1 follows χ^2-distribution with 1 degree of freedom under H_0.

Proportional Hazards Model

Proportional hazards model [6]:

$$h(t; x_i) = h_0(t) exp(\beta x_i)$$

When $x = 0$ or 1, a group indicator, the hazard ratio between two groups is:

$$HR = exp(\beta)$$

When HR < 1, interpretation of 1-HR% reduction in hazard rate in group 1 as compared to group 0.

Partial likelihood function:

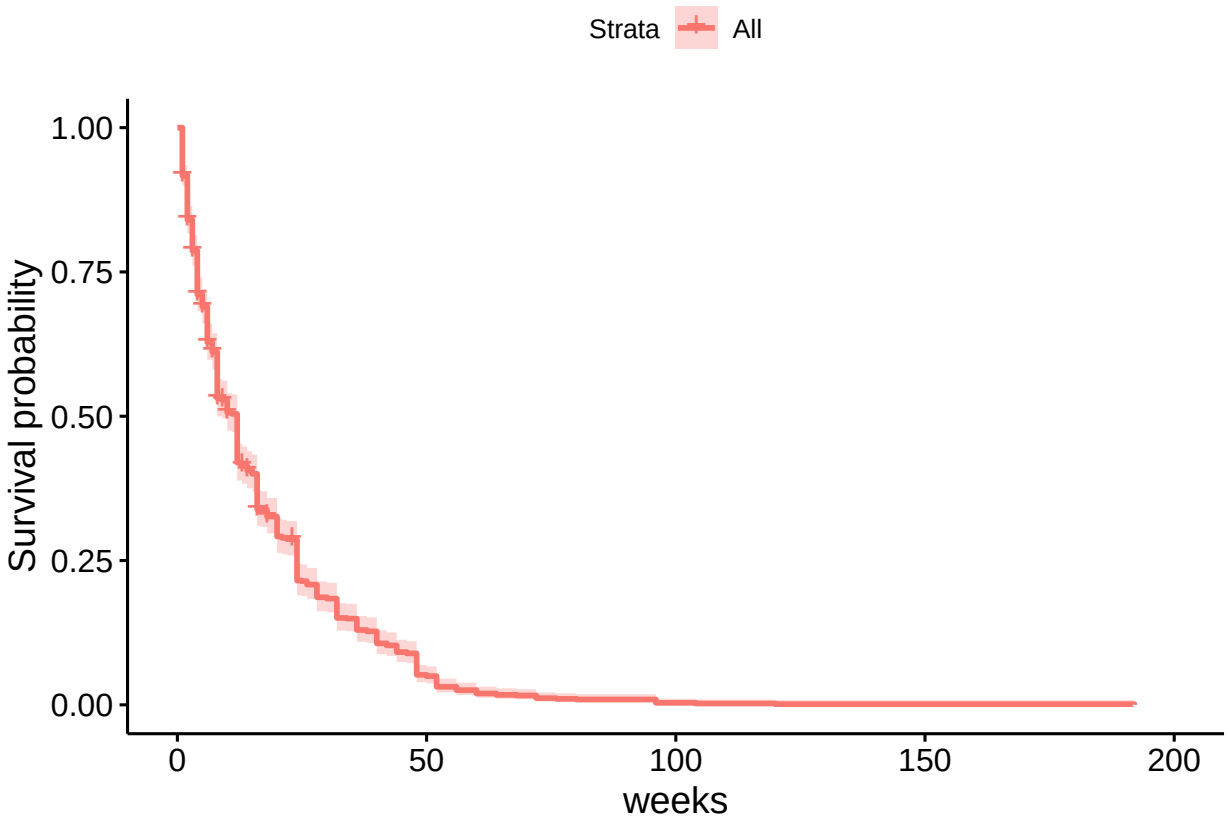
$$l_p(\beta) = \prod_{i=1}^{m} [\frac{exp\{\beta x_i\}}{\sum_{j \in R(t_i)} \{\beta x_j\}}]$$

Data Analysis

Overview

First of all, the Kaplan-Meier estimator is used to estimate survival probabilities of the breastfeeding mothers. When the time is more than 25 weeks, the survival probability is less than 0.25.

```
KM.devariation<-survfit(Surv(duration,delta)~1,data=bfeed)
ggsurvplot(KM.devariation,xlab="weeks", ylab="Survival probability")
```



We converted some continuous variables into categorical variables for KM plotting.

```
bfeed = bfeed %>%
  mutate(
    agemth = case_when(between(agemth, 15, 21) ~ "<=21",
                       between(agemth, 22, 30) ~ ">21"),
    ybirth = case_when(between(ybirth, 78, 82) ~ "<=82",
                       between(ybirth, 83, 86) ~ ">82"),
    yschool = case_when(between(yschool, 3, 11) ~ "[3,12]",
                       between(yschool, 12, 12) ~ "12",
                       between(yschool, 13, 20) ~ "(12,20]",))
```

Single Factor Analysis

```
KM.drrace<-survfit(Surv(duration,delta)~race,data=bfeed)
ggsurvplot(KM.drrace,xlab="weeks", ylab="Survival probability", legend.title='race',legend.labs=c('white
```

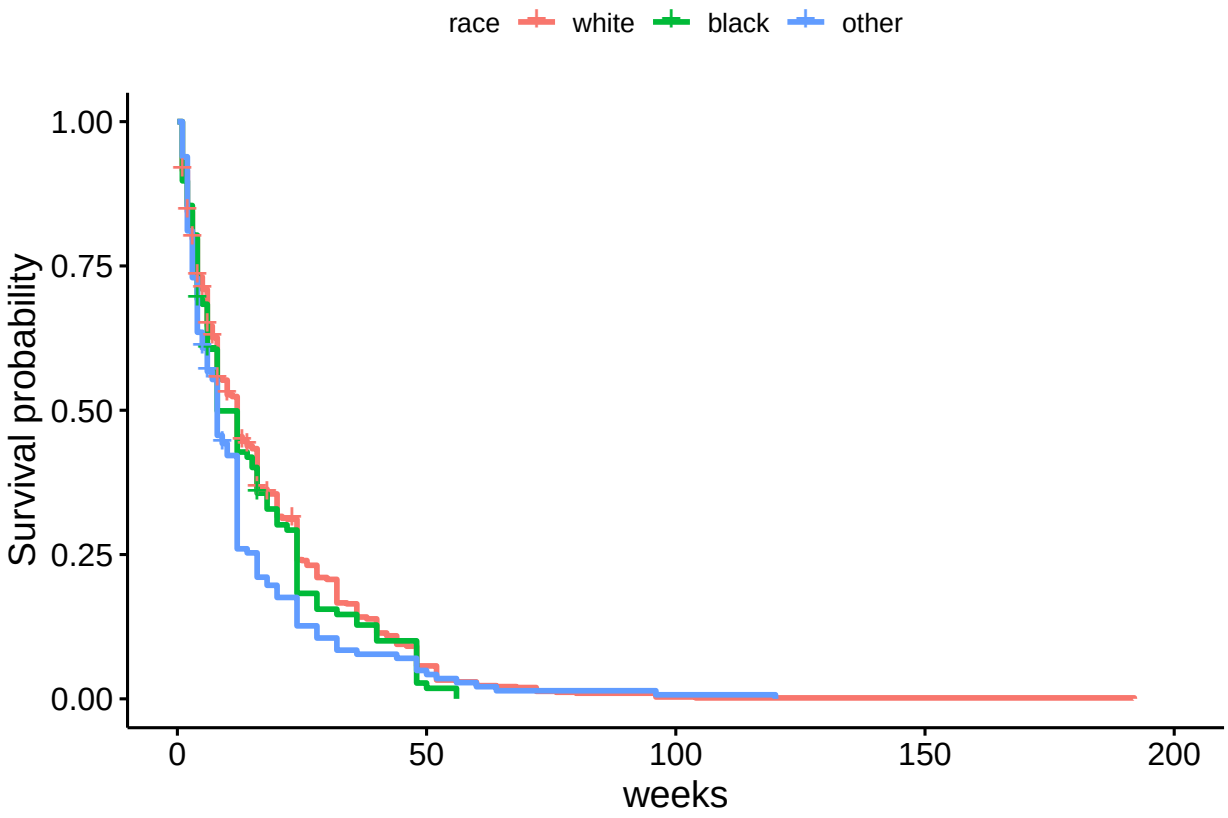
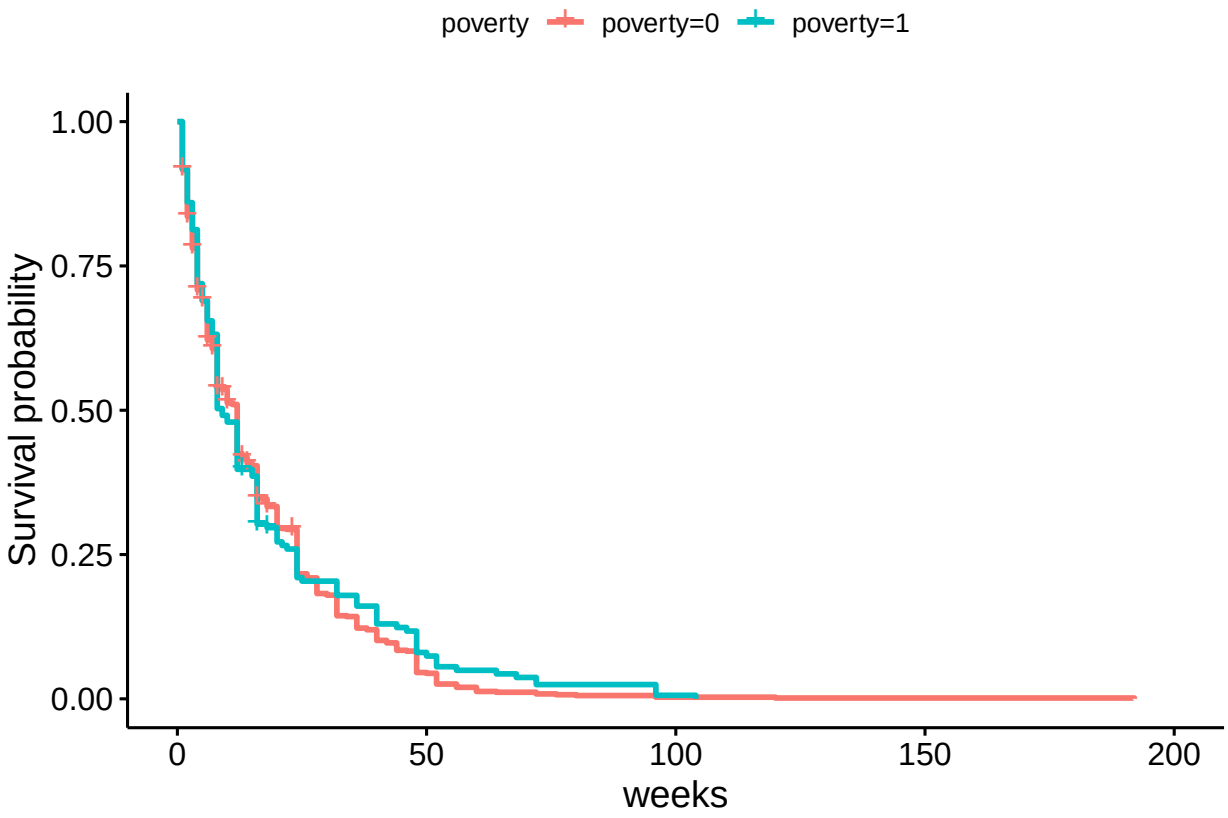


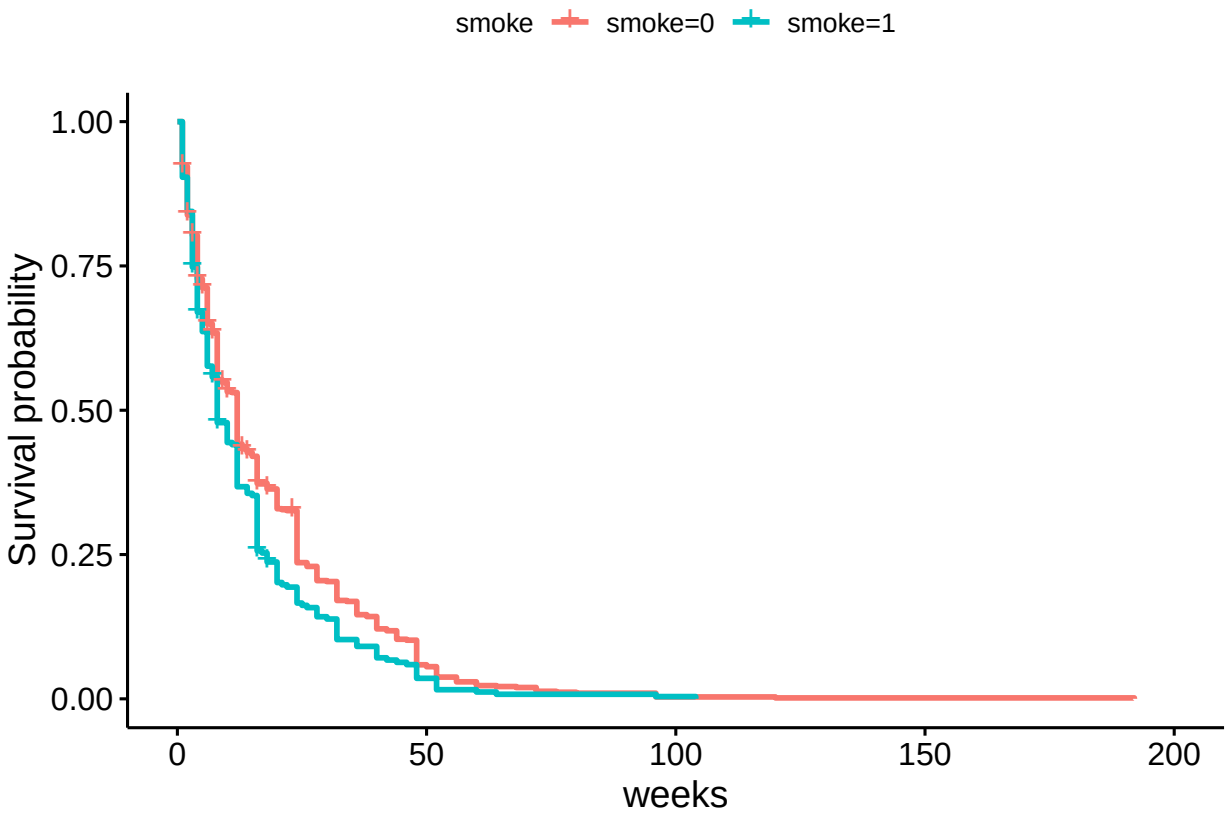
Figure below shows that the survival probability of mother completing breast feeding of mother in poverty is higher than that not in poverty in general, especially when the duration is more than 25 weeks.

```
KM.drpoverty<-survfit(Surv(duration,delta)~poverty,data=bfeed)
ggsurvplot(KM.drpoverty,xlab="weeks", ylab="Survival probability", legend.title='poverty')
```



The survival probability of mother completing breast feeding of mother smoking is lower than that not smoke.

```
KM.drsmoke<-survfit(Surv(duration,delta)~smoke,data=bfeed)
ggsurvplot(KM.drsmoke,xlab="weeks", ylab="Survival probability", legend.title='smoke')
```



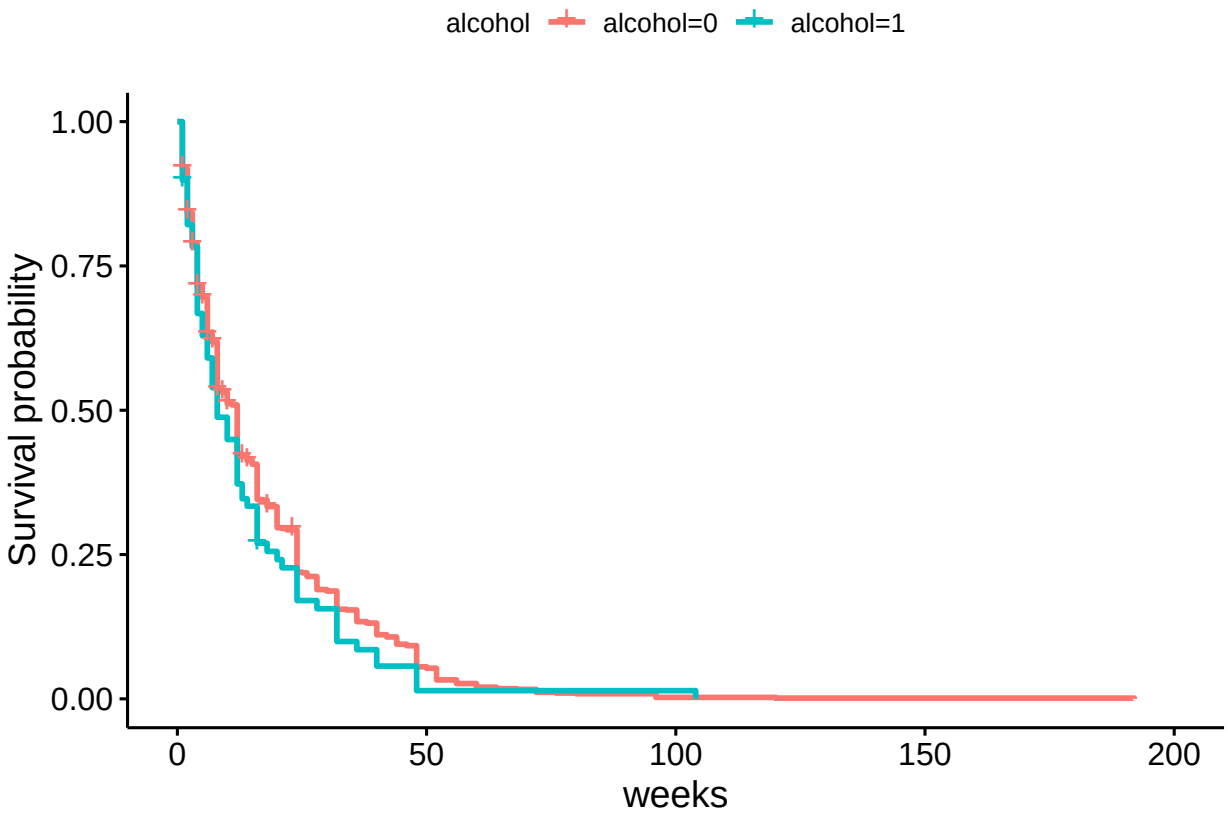
According to the figure, though there is barely any overlap between those curves, that could not completely attribute to the effect of the various treatments. So we use k sample log-rank test to compare those survival distributions of different smoke level. When the significance level is 0.5, P value of the test is 0.001 and the survival distribution of different smoke level group have significance difference.

```
survdiff(Surv(duration,delta)~smoke,data=bfeed)
```

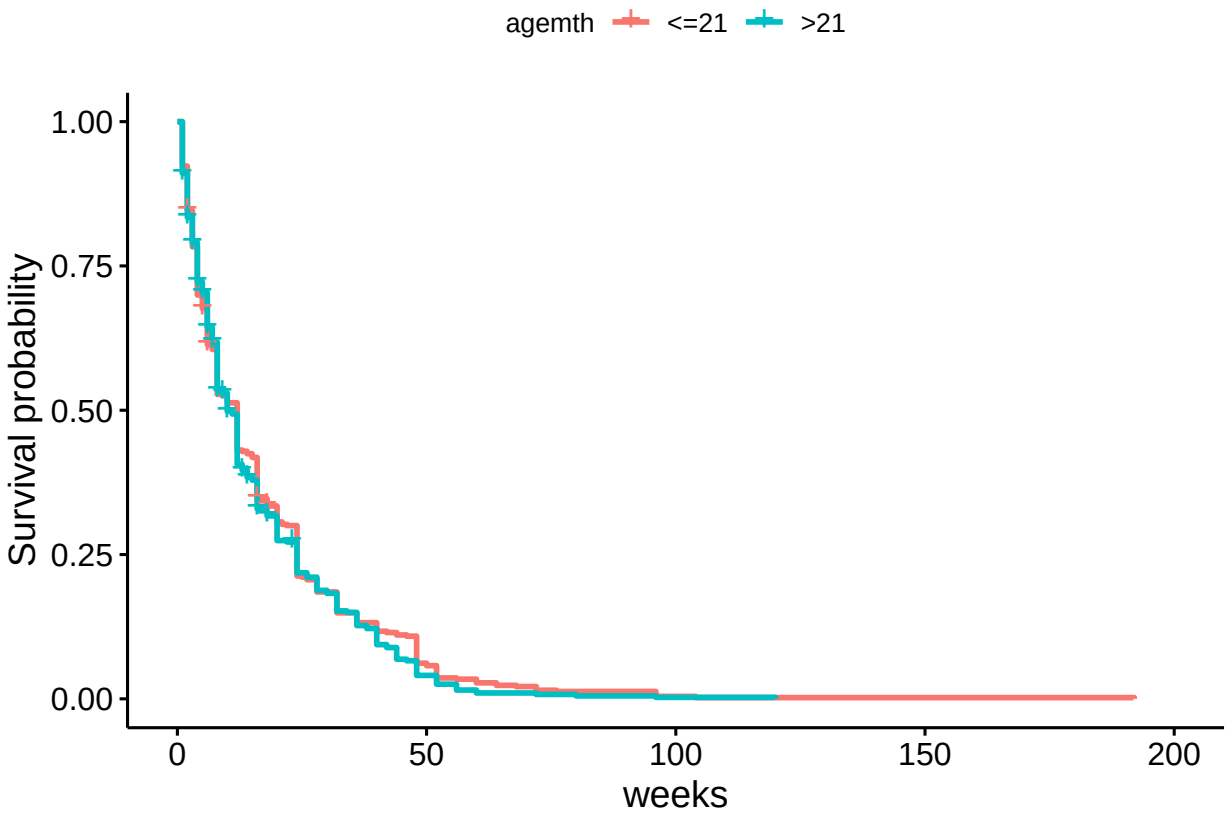
```
## Call:
## survdiff(formula = Surv(duration, delta) ~ smoke, data = bfeed)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## smoke=0 657      629      667      2.21      10.1
## smoke=1 270      263      225      6.56      10.1
##
##  Chisq= 10.1  on 1 degrees of freedom, p= 0.001
```

The survival probability of mother completing breast feeding of mother using alcohol is lower than that not use alcohol.

```
KM.dralcohol<-survfit(Surv(duration,delta)~alcohol,data=bfeed)
ggsurvplot(KM.dralcohol,xlab="weeks", ylab="Survival probability", legend.title='alcohol')
```

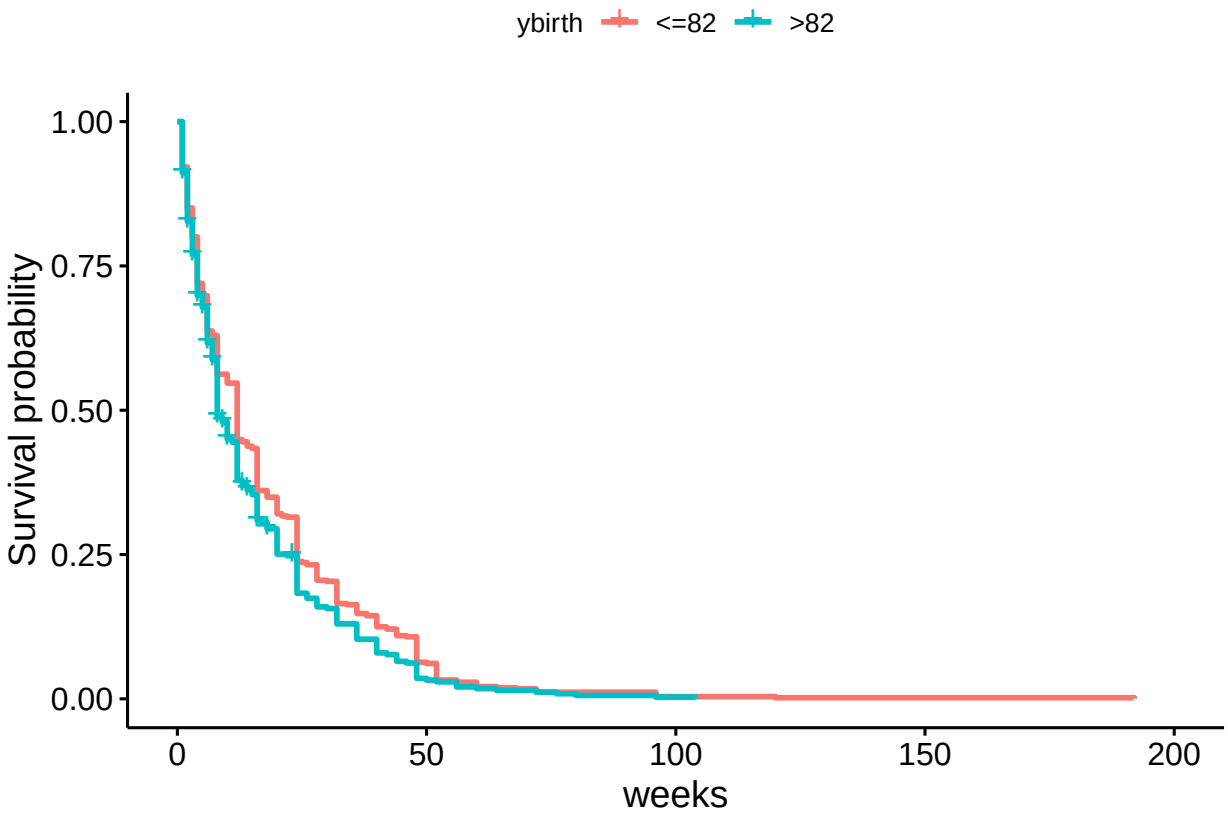



```
KM.dragemth<-survfit(Surv(duration,delta)~agemth,data=bfeed)
ggsurvplot(KM.dragemth,xlab="weeks", ylab="Survival probability", legend.title='agemth',legend.labs=c('
```



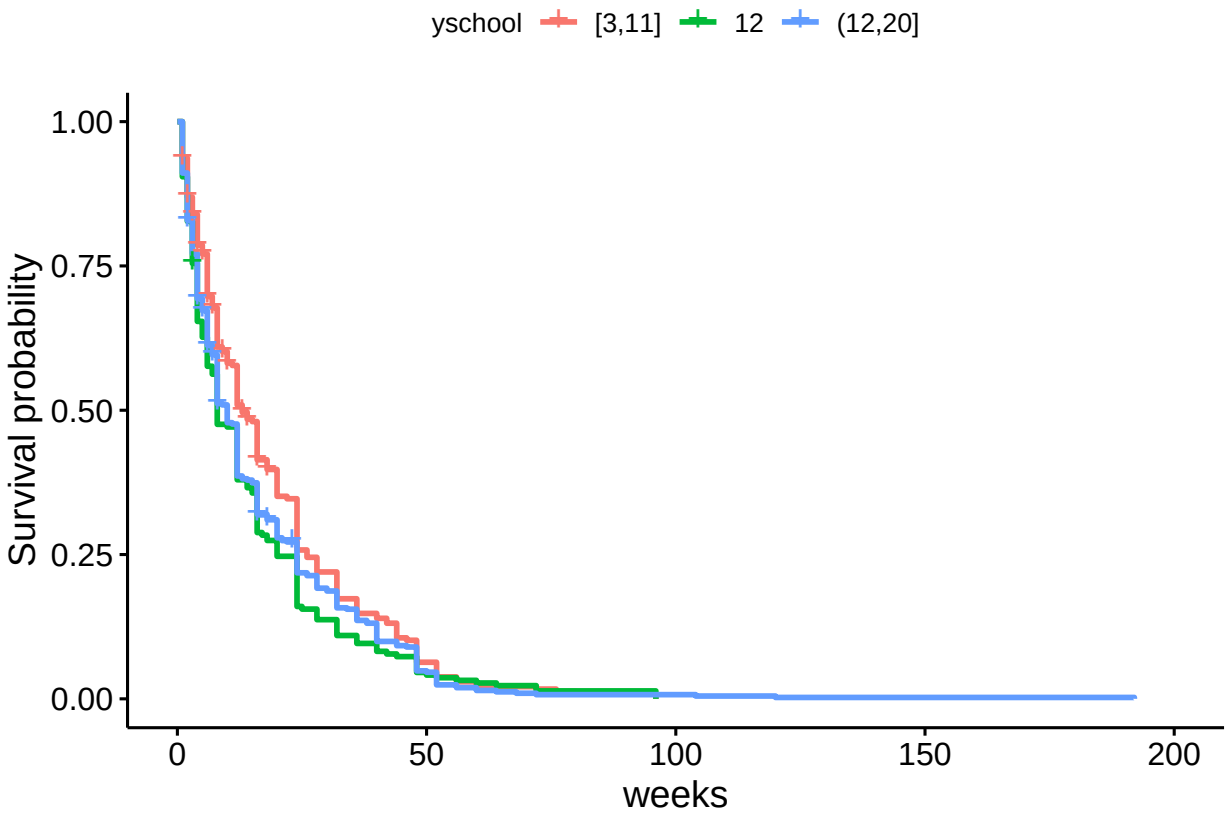
The survival probability of mother completing breast feeding of mother who birth lower than 1982 is higher than that who birth greater than 1982.

```
KM.drybirth <-survfit(Surv(duration,delta)~ybirth,data=bfeed)
ggsurvplot(KM.drybirth,xlab="weeks", ylab="Survival probability", legend.title='ybirth',legend.labs=c('
```



There may be a negative correlation between maternal education and survival probability.

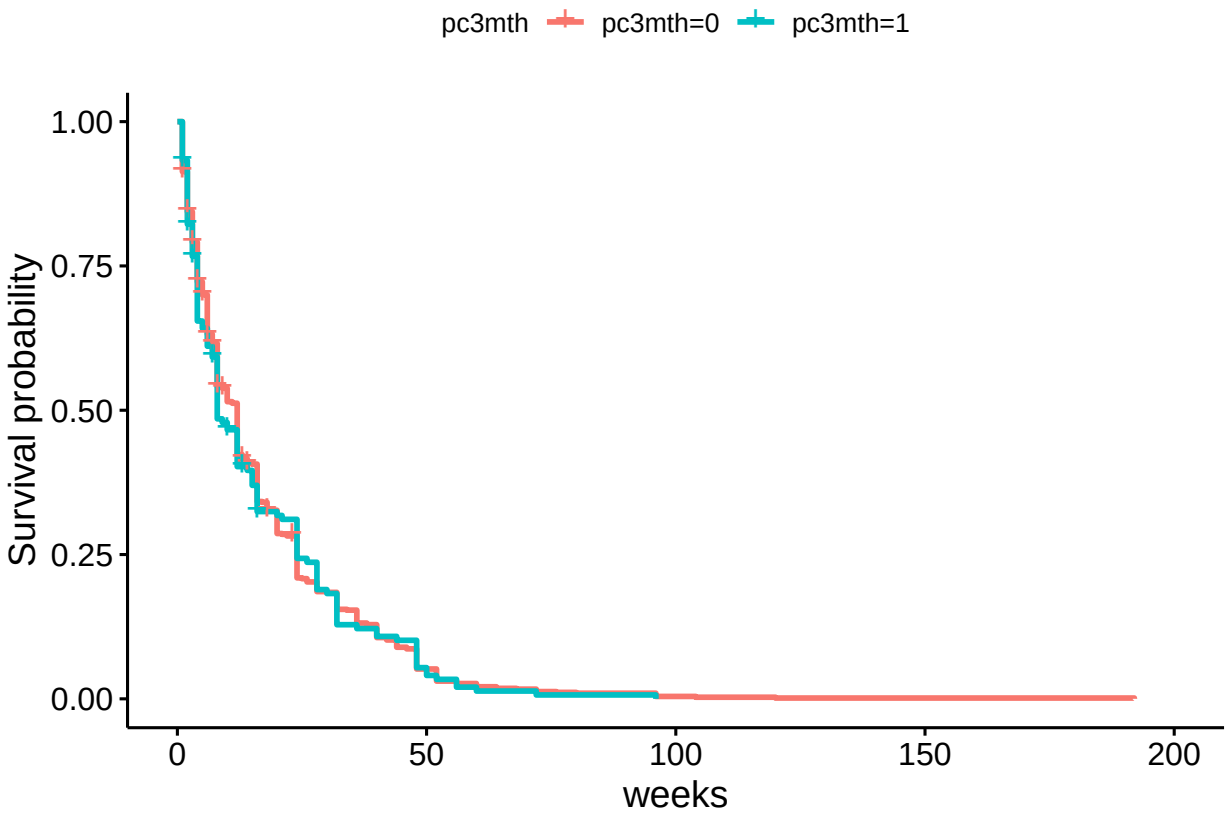
```
KM.dryschool <-survfit(Surv(duration,delta)~yschool,data=bfeed)
ggsurvplot(KM.dryschool,xlab="weeks", ylab="Survival probability", legend.title='yschool',legend.labs=c
```




```
survdiff(Surv(duration,delta)~yschool,data=bfeed)
```

```
## Call:
## survdiff(formula = Surv(duration, delta) ~ yschool, data = bfeed)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## yschool=(12,20] 269      248      280      3.68     6.174
## yschool=[3,12]  220      219      199      2.02     3.002
## yschool=12      438      425      413      0.35     0.754
##
##  Chisq= 7  on 2 degrees of freedom, p= 0.03
```

```
KM.drpc3mth <-survfit(Surv(duration,delta)~pc3mth,data=bfeed)
ggsurvplot(KM.drpc3mth,xlab="weeks", ylab="Survival probability", legend.title='pc3mth')
```



Model

In this section, Cox Proportional Hazard Model with careful variable selection is applied step by step.

Variable Selection

The model with all covariates is redundant and inefficient. Forward selection method are applied in this part to get the best covariate subset and possible interaction terms. First, we converted some categorical variables to factor types then used AIC for variable selection.

```
data(bfeed)
bfeed = bfeed %>%
  mutate(
    race = as.factor(bfeed$race),
    poverty = as.factor(bfeed$poverty),
    smoke = as.factor(bfeed$smoke),
    alcohol = as.factor(bfeed$alcohol),
    pc3mth = as.factor(bfeed$pc3mth),
    agemth = as.factor(case_when(between(agemth, 15, 21) ~ "<=21",
                                 between(agemth, 22, 30) ~ ">21")),
    ybirth = as.factor(case_when(between(ybirth, 78, 82) ~ "<=82",
                                 between(ybirth, 83, 86) ~ ">82")),
    yschool = as.factor(case_when(between(yschool, 3, 11) ~ "[3,12]",
                        between(yschool, 12, 12) ~ "12",
                        between(yschool, 13, 20) ~ "(12,20]",)))
```

```
stepAIC(coxph(Surv(duration,delta)~1, data=bfeed), direction = "forward", scope = list(upper=coxph(Surv
```

```
## Start:  AIC=10382.23
## Surv(duration, delta) ~ 1
##
##           Df   AIC
## + smoke    1 10375
## + race     2 10379
## + ybirth   1 10379
## + yschool  2 10379
## <none>       10382
## + alcohol  1 10382
## + agemth   1 10383
## + poverty  1 10383
## + pc3mth   1 10384
##
## Step:  AIC=10375.05
## Surv(duration, delta) ~ smoke
##
##           Df   AIC
## + race     2 10367
## + ybirth   1 10371
## <none>       10375
## + yschool  2 10375
## + agemth   1 10375
## + poverty  1 10376
## + alcohol  1 10376
## + pc3mth   1 10377
##
## Step:  AIC=10367.4
## Surv(duration, delta) ~ smoke + race
##
##           Df   AIC
## + ybirth   1 10362
## + poverty  1 10366
## + agemth   1 10367
## <none>       10367
## + alcohol  1 10368
## + pc3mth   1 10369
## + yschool  2 10369
##
## Step:  AIC=10362.01
## Surv(duration, delta) ~ smoke + race + ybirth
##
##           Df   AIC
## + yschool  2 10362
## + poverty  1 10362
## <none>       10362
## + alcohol  1 10363
## + agemth   1 10363
## + pc3mth   1 10364
##
## Step:  AIC=10361.78
## Surv(duration, delta) ~ smoke + race + ybirth + yschool
```

```
##
##           Df   AIC
## + poverty  1 10360
## + agemth   1 10361
## <none>       10362
## + alcohol  1 10363
## + pc3mth   1 10363
##
## Step:  AIC=10359.85
## Surv(duration, delta) ~ smoke + race + ybirth + yschool + poverty
##
##           Df   AIC
## + agemth   1 10360
## <none>       10360
## + alcohol  1 10360
## + pc3mth   1 10362
##
## Step:  AIC=10359.49
## Surv(duration, delta) ~ smoke + race + ybirth + yschool + poverty +
##     agemth
##
##           Df   AIC
## <none>       10360
## + alcohol  1 10360
## + pc3mth   1 10361
```

```
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool + poverty + agemth, data = bfeed)
##
##                    coef exp(coef) se(coef)      z       p
## smoke1          0.25760   1.29382  0.07869  3.274 0.00106
## race2           0.19409   1.21421  0.10461  1.855 0.06354
## race3           0.31643   1.37222  0.09754  3.244 0.00118
## ybirth>82       0.17860   1.19554  0.07326  2.438 0.01478
## yschool[3,12]   0.30659   1.35879  0.11261  2.723 0.00648
## yschool12       0.19359   1.21360  0.08531  2.269 0.02324
## poverty1       -0.17370   0.84055  0.09319 -1.864 0.06233
## agemth>21       0.12197   1.12972  0.07941  1.536 0.12454
##
## Likelihood ratio test=38.74  on 8 df, p=5.502e-06
## n= 927, number of events= 892
```

```
cph = coxph(Surv(duration, delta) ~ smoke + race + ybirth +
    yschool + poverty + agemth, data=bfeed)
```

We used the `stepAIC` function to automatically screen variables according to the AIC value. It can be seen that variables selected at each step all reduce AIC, smoke is the first variable selected, race is the second variable selected, and so on. After forward selection method, since all variables expect `agemth` were significant at the 0.1 level and AIC was small, we derived the preliminary main effects model. `Smoke`, `race`, `ybirth`, `yschool`, `poverty`, `agemth` are selected variables to be added to the model. In addition, we examine the interactions in the model. It can be seen that the original model works better.

```
model.inter.1 <- coxph(Surv(duration,delta)~smoke* race+ ybirth+yschool+poverty+agemth, data=bfeed)
model.inter.2 <- coxph(Surv(duration,delta)~race+ ybirth*smoke+yschool+poverty+agemth, data=bfeed)
model.inter.3 <- coxph(Surv(duration,delta)~race+ ybirth+yschool*smoke+poverty+agemth, data=bfeed)
```

```
model.inter.4 <- coxph(Surv(duration,delta)~race+ ybirth+yschool+poverty*smoke+agemth, data=bfeed)
model.inter.5 <- coxph(Surv(duration,delta)~race+ ybirth+yschool+poverty+agemth*smoke, data=bfeed)
```

```
AIC(cph, model.inter.1, model.inter.2,model.inter.3,model.inter.4,model.inter.5)
```

```
##                df      AIC
## cph             8 10359.49
## model.inter.1  10 10362.55
## model.inter.2   9 10360.58
## model.inter.3  10 10363.39
## model.inter.4   9 10361.39
## model.inter.5   9 10361.19
```

From the result of ANOVA, expect `yschool` and `agemth`, each variable is significant.

```
anova(cph)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(duration, delta)
## Terms added sequentially (first to last)
##
##          loglik   Chisq Df Pr(>|Chi|)
## NULL    -5191.1
## smoke   -5186.5  9.1801  1   0.002447 **
## race    -5180.7 11.6457  2   0.002959 **
## ybirth  -5177.0  7.3944  1   0.006543 **
## yschool -5174.9  4.2339  2   0.120398
## poverty -5172.9  3.9240  1   0.047601 *
## agemth  -5171.7  2.3576  1   0.124675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Expect `agemth`, each variable in our model is significant.

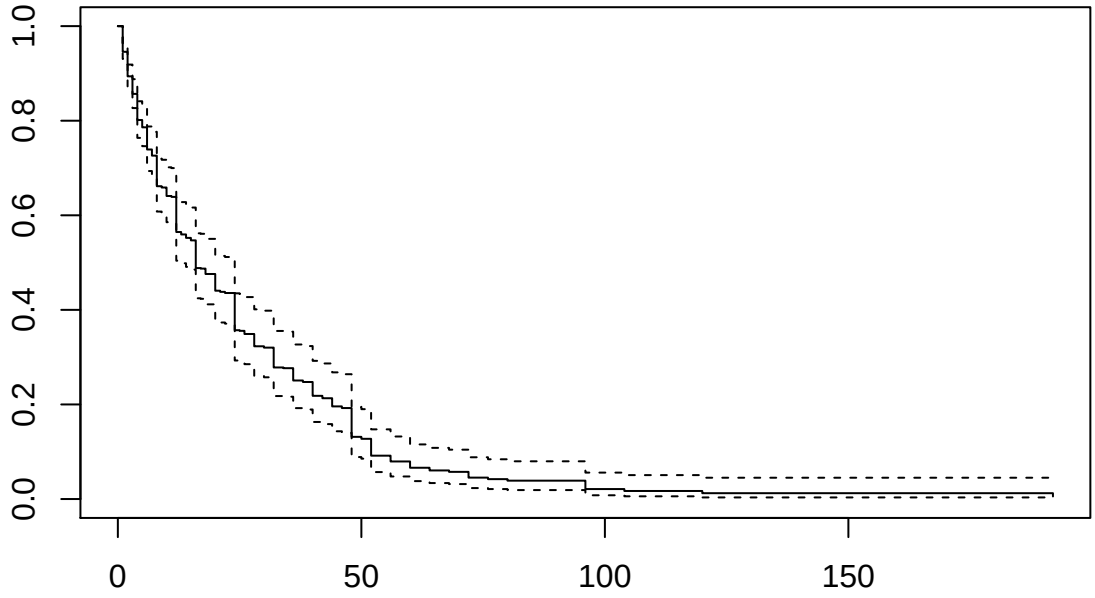
```
summary(cph)
```

```
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool + poverty + agemth, data = bfeed)
##
##   n= 927, number of events= 892
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## smoke1         0.25760   1.29382  0.07869  3.274  0.00106 **
## race2          0.19409   1.21421  0.10461  1.855  0.06354 .
## race3          0.31643   1.37222  0.09754  3.244  0.00118 **
## ybirth>82      0.17860   1.19554  0.07326  2.438  0.01478 *
## yschool[3,12]  0.30659   1.35879  0.11261  2.723  0.00648 **
## yschool12      0.19359   1.21360  0.08531  2.269  0.02324 *
## poverty1      -0.17370   0.84055  0.09319 -1.864  0.06233 .
## agemth>21      0.12197   1.12972  0.07941  1.536  0.12454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## smoke1           1.2938     0.7729    1.1089     1.510
```

```
## race2            1.2142     0.8236    0.9891     1.491
## race3            1.3722     0.7287    1.1334     1.661
## ybirth>82        1.1955     0.8364    1.0356     1.380
## yschool[3,12]    1.3588     0.7360    1.0897     1.694
## yschool12        1.2136     0.8240    1.0267     1.434
## poverty1         0.8405     1.1897    0.7002     1.009
## agemth>21        1.1297     0.8852    0.9669     1.320
##
## Concordance= 0.573  (se = 0.012 )
## Likelihood ratio test= 38.74  on 8 df,   p=6e-06
## Wald test            = 39.09  on 8 df,   p=5e-06
## Score (logrank) test = 39.22  on 8 df,   p=4e-06
```

It can be seen that the smoke coefficient is 0.25760, and the 95% confidence interval for its exp(coefficient) is [1.1089,1.510]. The rest of the results are in the summary above.

```
plot(survfit(cph))
```

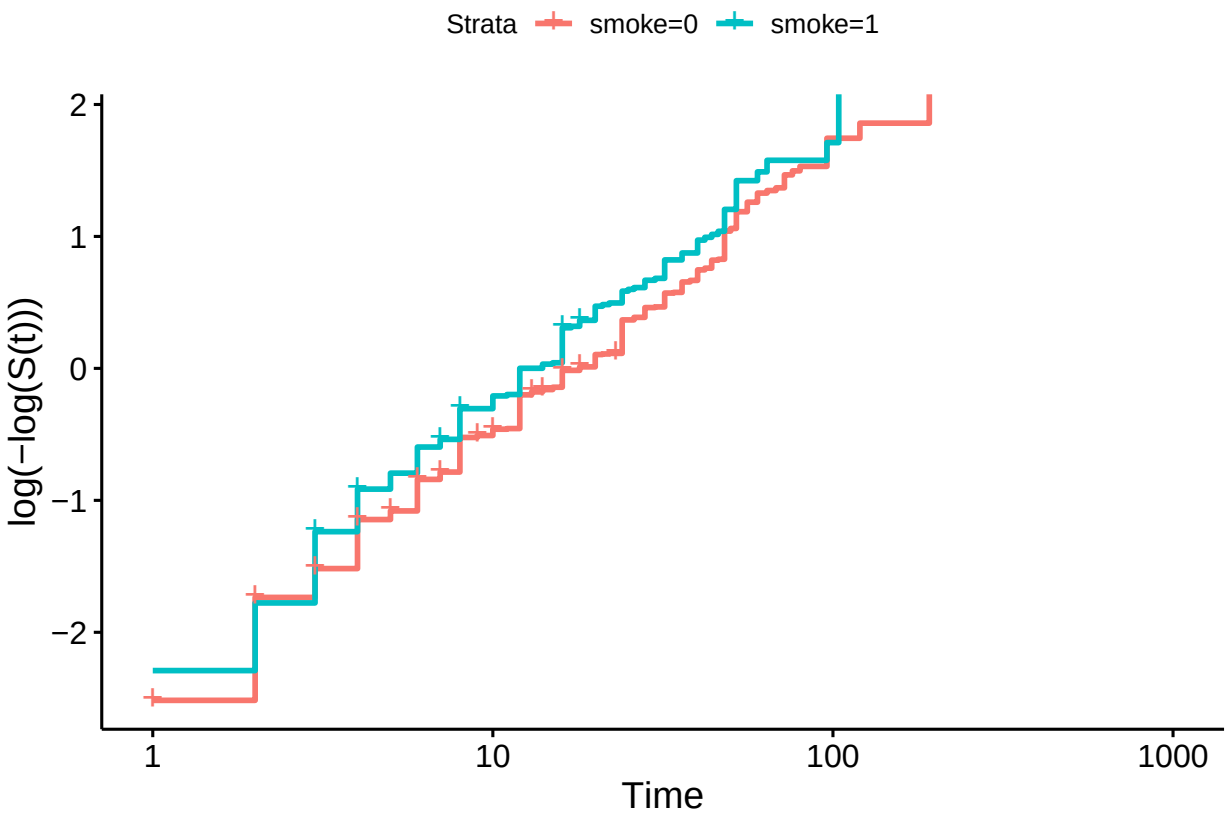


Evaluation

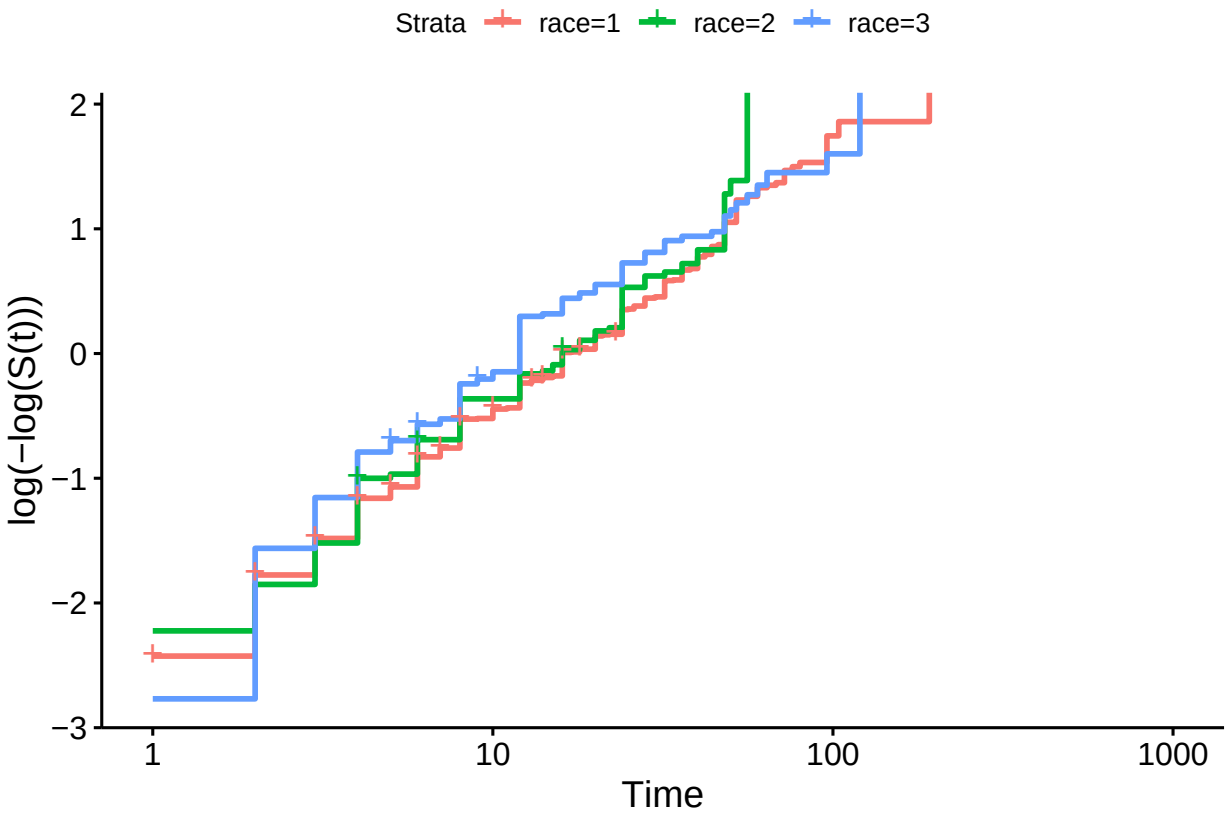
Testing Proportional Hazards Assumption

The Cox model relies on the proportional hazards (PH) assumption, implying that the factors investigated have a constant impact on the hazard over time. Cox Proportional Hazard Model works well only when proportional hazards assumption is satisfied. We draw the log-log plot of covariates. We can find that except for `yschool`, there is strong parallelism, so the ph assignment is satisfied. For `yschool` and other covariates, we use the `cox.zph` function to further test.

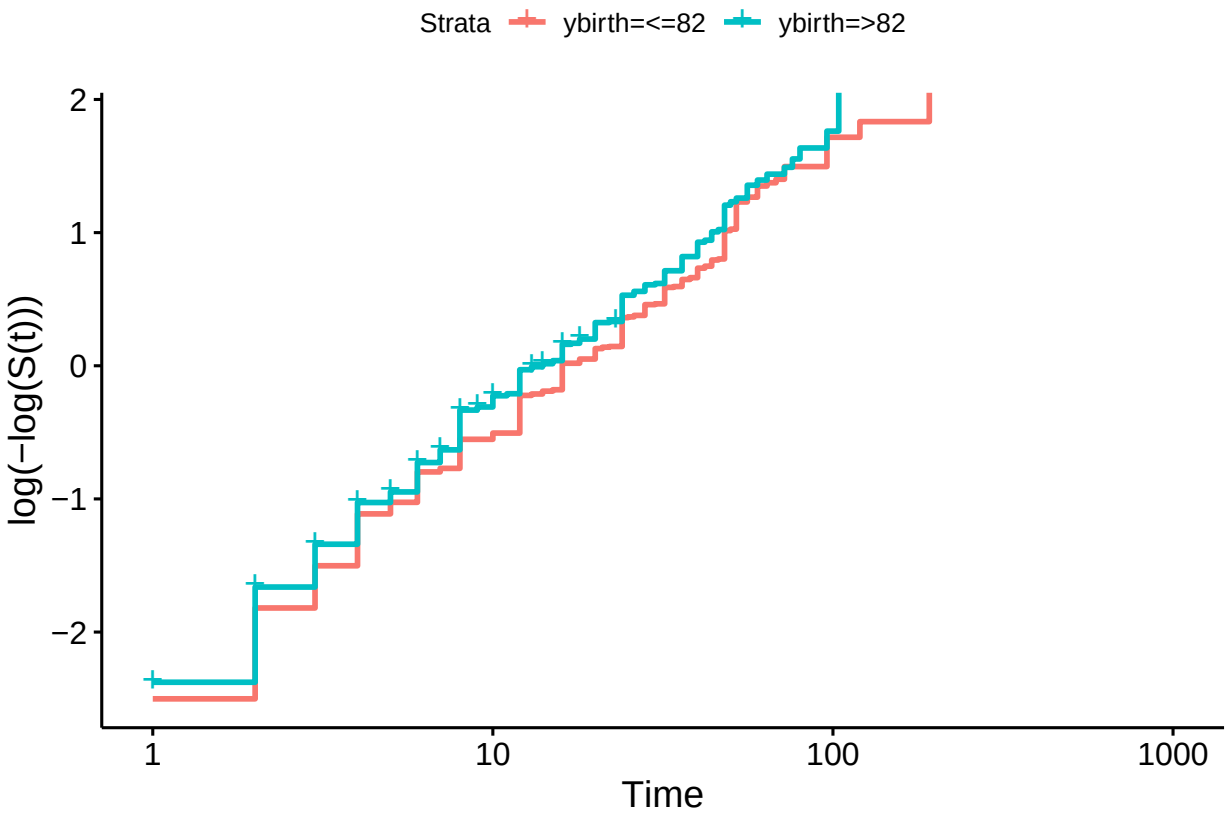
```
ggsurvplot(survfit(Surv(duration,delta)~smoke, data=bfeed),fun = "cloglog", main="Smoke Log-log Plot")
```



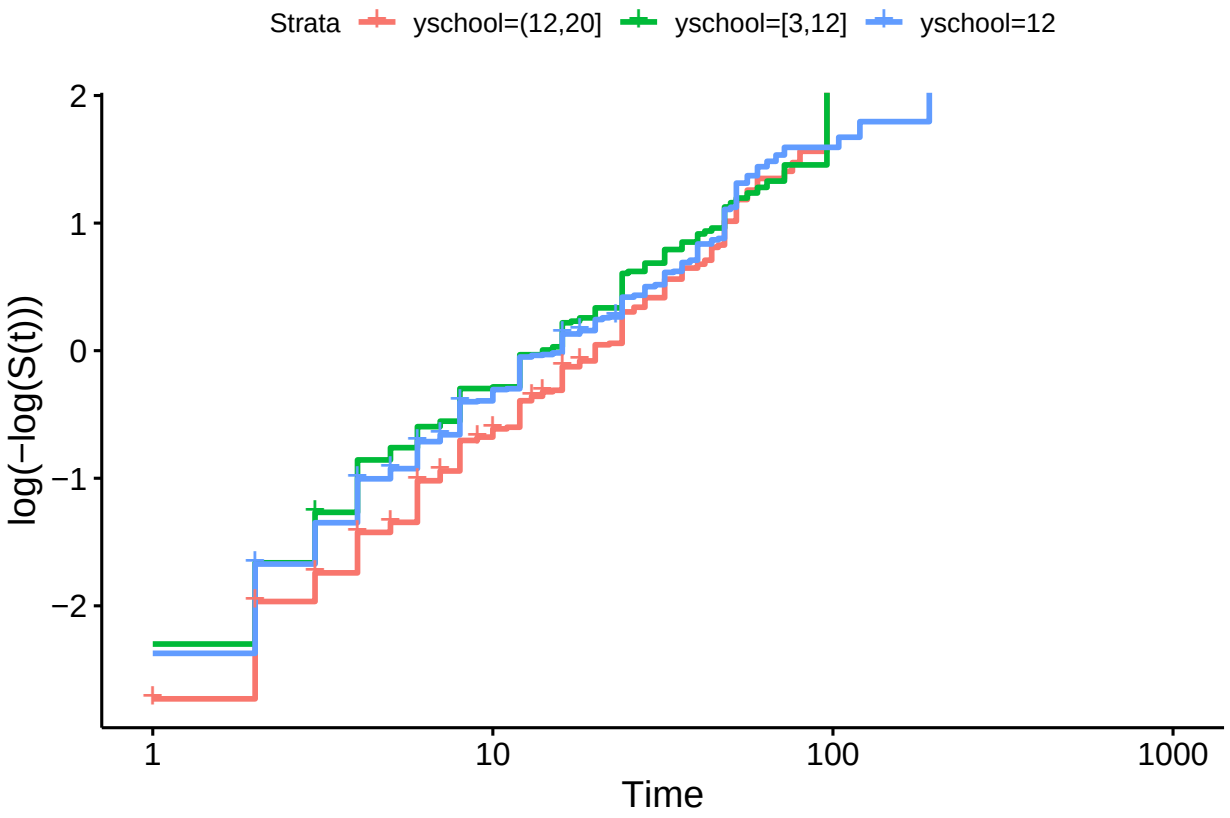

```
ggsurvplot(survfit(Surv(duration,delta)~race, data=bfeed),fun = "cloglog", main="Race Log-log Plot")
```



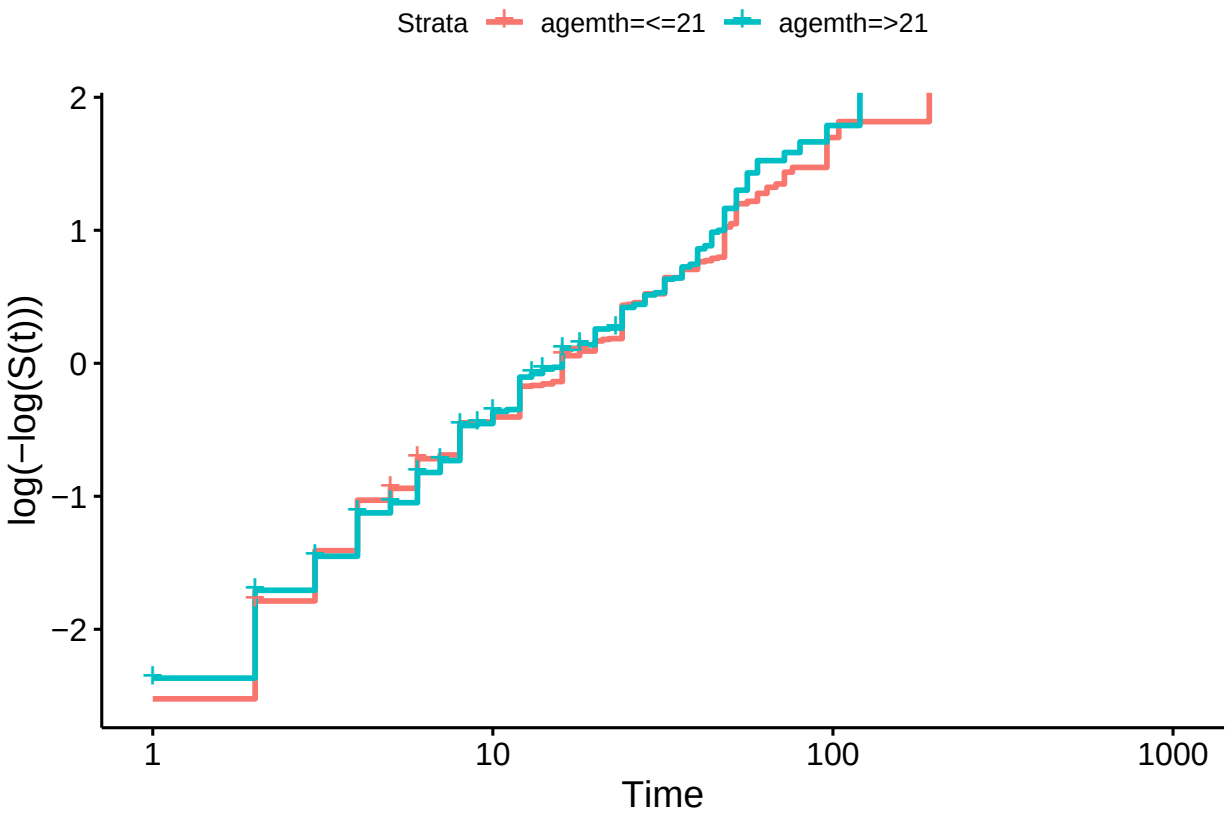

```
ggsurvplot(survfit(Surv(duration,delta)~ybirth, data=bfeed),fun = "cloglog", main="Ybirth Log-log Plot"
```




```
ggsurvplot(survfit(Surv(duration,delta)~yschool, data=bfeed),fun = "cloglog", main="Yschool Log-log Plo
```




```
ggsurvplot(survfit(Surv(duration,delta)~agemth, data=bfeed),fun = "cloglog", main="Agemth Log-log Plot")
```

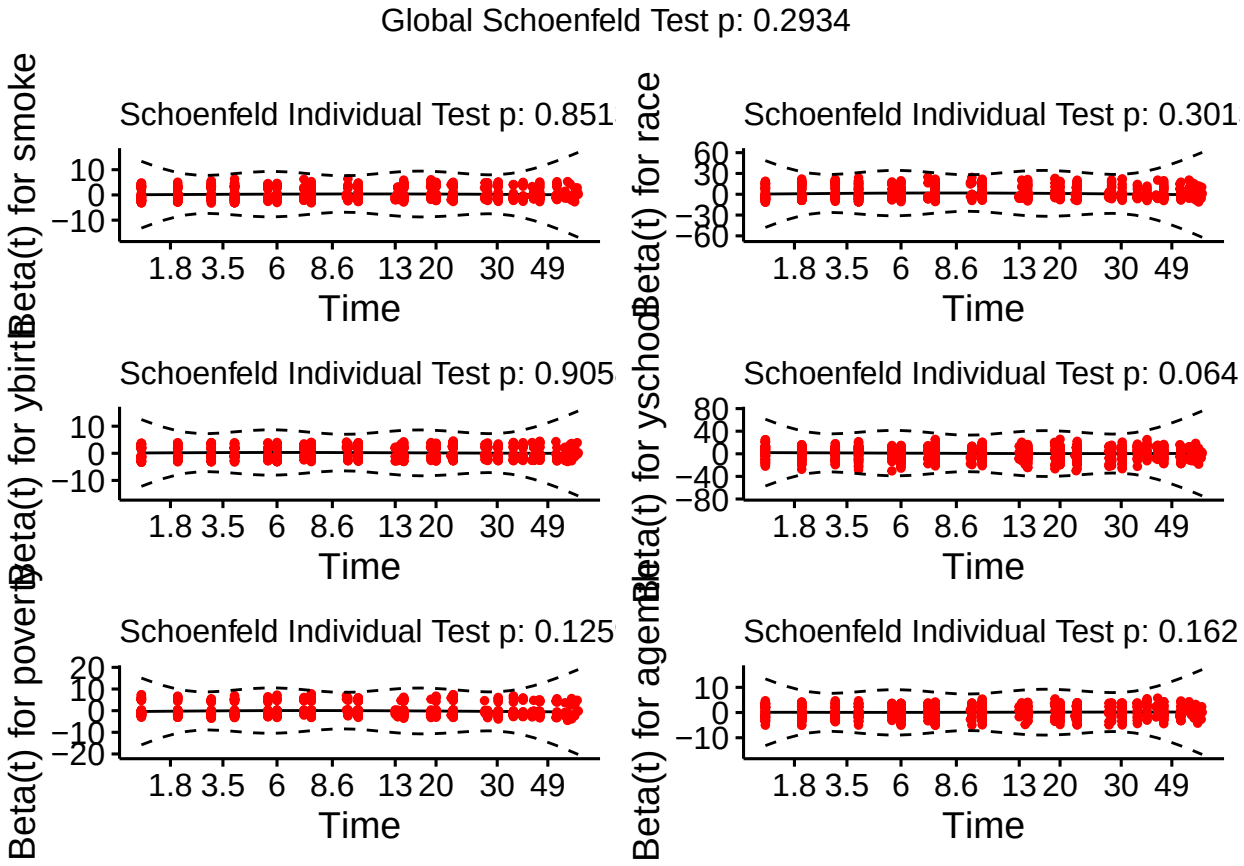


Correlation coefficient between transformed survival time and the scaled Schoenfeld residuals is tested for the assumption. If the correlation is significantly non-zero, then PH assumption is not satisfied. The result shows that there is significant deviation from the proportional hazards assumption for all variables.

```
test.ph <- cox.zph(cph)
test.ph
```

```
##          chisq df     p
## smoke   0.0351  1 0.851
## race    2.3994  2 0.301
## ybirth  0.0140  1 0.906
## yschool 5.4985  2 0.064
## poverty 2.3428  1 0.126
## agemth  1.9557  1 0.162
## GLOBAL  9.6107  8 0.293
```

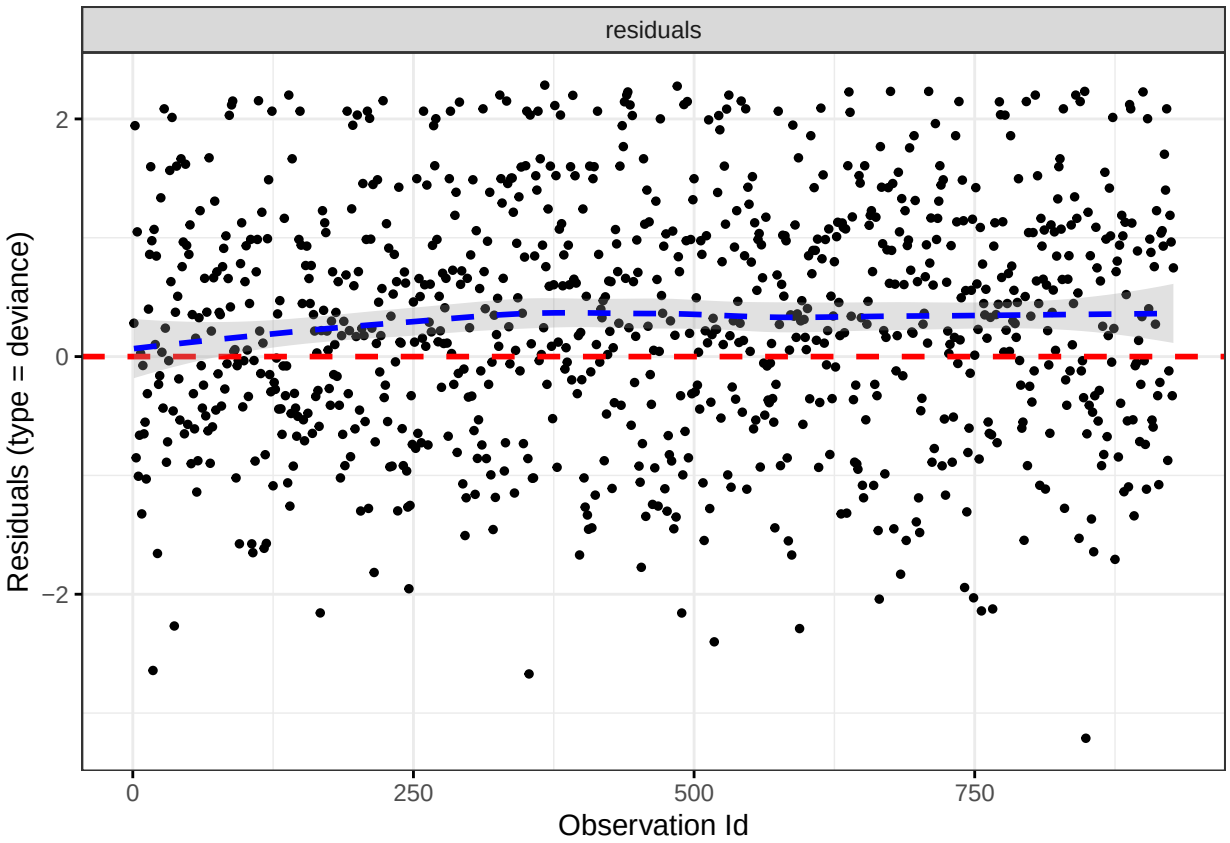
```
ggcoxzph(test.ph,font.main = 12)
```



Testing Influential Observations

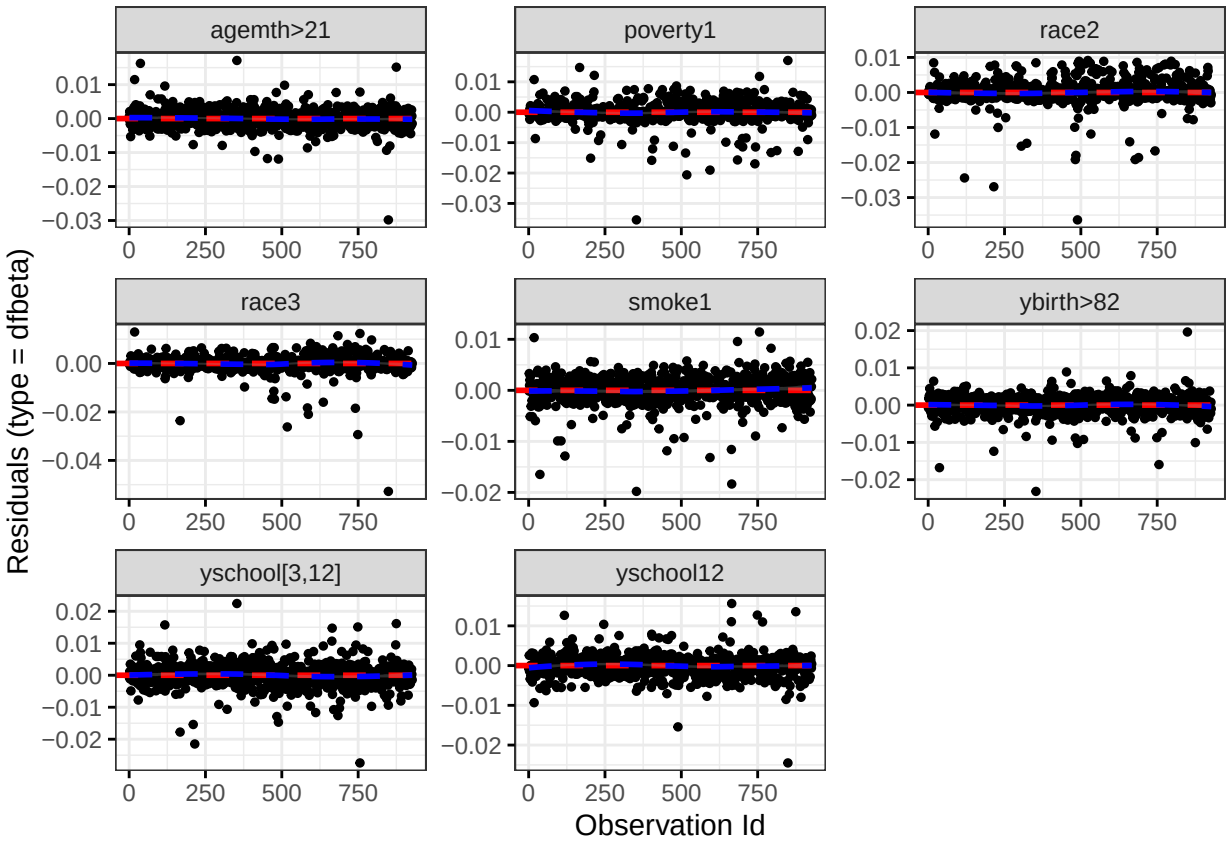
In this part, deviance residuals are visualized to check outliers who may be influential. The deviance residual is a normalized transform of the martingale residual. These residuals should be roughly symmetrically distributed about zero with a standard deviation of 1. Plot looks symmetric between 0 and 0.5 and there is no apparent outliers.

```
ggcoxdiagnostics(cph, type = 'deviance', linear.predictions = FALSE)
```



We examine the estimated change in the regression coefficient plotted after each observation is removed in turn. The figure shows that a comparison of the amplitude of the maximum dfbeta value with the regression coefficient shows that these observations are not significantly different from each of the exclusive results and are evenly distributed on both sides of the reference line of $x = 0$.

```
ggcoxdiagnostics(cph, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



Testing Non-Linearity

In the model, there is an assumption that continuous covariates have a linear form. Since all covariables are categorical variables, this step is skipped.

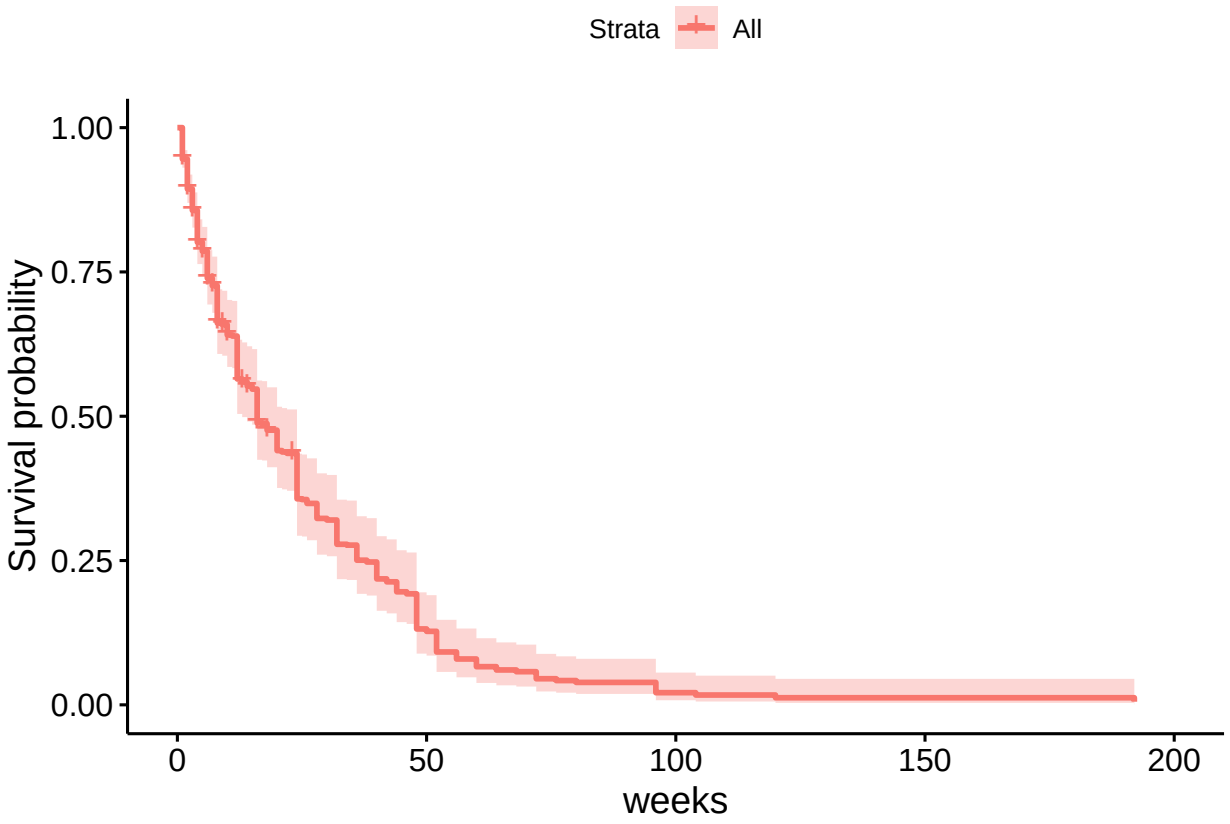
Model Interpretation

Model with covariates and coefficients is adopted after variable selection steps. Under Cox Proportional Hazard Model, exp(coef) means hazard ratio of the variable compared to the baseline hazard function. For instance, HR for smoke is 1.2938. It means that mothers in smoking is 0.7729 times more likely to finish breast feeding than people no smoking, when other variables are the same. HR larger than 1 indicate the danger factor, while HR smaller than 1 means positive effects.

Estimated Survival Curve

Through our final model, we estimate the curve.

```
ggsurvplot(survfit(cph),bfeed,xlab="weeks", ylab="Survival probability")
```




```
summary(cph)
```

```
## Call:
## coxph(formula = Surv(duration, delta) ~ smoke + race + ybirth +
##     yschool + poverty + agemth, data = bfeed)
##
##   n= 927, number of events= 892
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## smoke1         0.25760   1.29382  0.07869  3.274  0.00106 **
## race2          0.19409   1.21421  0.10461  1.855  0.06354 .
## race3          0.31643   1.37222  0.09754  3.244  0.00118 **
## ybirth>82      0.17860   1.19554  0.07326  2.438  0.01478 *
## yschool[3,12]  0.30659   1.35879  0.11261  2.723  0.00648 **
## yschool12      0.19359   1.21360  0.08531  2.269  0.02324 *
## poverty1      -0.17370   0.84055  0.09319 -1.864  0.06233 .
## agemth>21      0.12197   1.12972  0.07941  1.536  0.12454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## smoke1           1.2938     0.7729    1.1089     1.510
## race2            1.2142     0.8236    0.9891     1.491
## race3            1.3722     0.7287    1.1334     1.661
## ybirth>82        1.1955     0.8364    1.0356     1.380
## yschool[3,12]    1.3588     0.7360    1.0897     1.694
```

```
## yschool12        1.2136     0.8240    1.0267     1.434
## poverty1         0.8405     1.1897    0.7002     1.009
## agemth>21        1.1297     0.8852    0.9669     1.320
##
## Concordance= 0.573  (se = 0.012 )
## Likelihood ratio test= 38.74  on 8 df,   p=6e-06
## Wald test            = 39.09  on 8 df,   p=5e-06
## Score (logrank) test = 39.22  on 8 df,   p=4e-06
```

DeepSurv

Based on the above test results, we believe that there are no Time-varying conditions for model variables. In addition, recurring events and other states do not exist in the data set. We try to use neural networks to explore whether there are nonlinear conditions.

We can try to use neural networks to model survival analysis data, which means through the survival model of Cox proportional hazard deep neural network DeepSurv for model construction [10]. Compared with the random survival forest model, the prediction performance of the model was more stable, but the interpretability was worse than that of the cox proportional hazards model. We randomly sampled the original data set to create a training set and a test set, and used the training set for model training. For the trained model, We use The concordance statistic function in the survival package to calculate the concordance statistic.The concordance statistic compute the agreement between an observed response and a predictor.

```
set.seed(1)
sub <- sample(1:nrow(bfeed),round(nrow(bfeed)*2/3))
train <- bfeed[sub,]
test <- bfeed[-sub,]


deep <- deepsurv(Surv(duration,delta) ~ ., data = train, frac = 0.3, activation = "relu",
    num_nodes = c(4L, 8L, 4L, 2L), dropout = 0.1, early_stopping = TRUE, epochs = 100L,
    batch_size = 32L)
```

```
concordance(Surv(duration, delta) ~ predict(deep,test,type = "risk"), test)
```

```
## Call:
## concordance.formula(object = Surv(duration, delta) ~ predict(deep,
##     test, type = "risk"), data = test)
##
## n= 309
## Concordance= 0.5117 se= 0.02073
## concordant discordant     tied.x     tied.y    tied.xy
##      18064      17069       7444       1872        400
```

```
concordance(Surv(duration, delta) ~ predict(cph,test,type = "risk"), test)
```

```
## Call:
## concordance.formula(object = Surv(duration, delta) ~ predict(cph,
##     test, type = "risk"), data = test)
##
## n= 309
## Concordance= 0.4532 se= 0.02013
## concordant discordant     tied.x     tied.y    tied.xy
##      18710      22693       1174       2219         53
```

As can be seen from the size of concordance statistic, Deepsurv model is better than CoxPH.

Conclusion

In this report, we focus on the completion of the mother's lactation and perform a basic survival analysis. This dataset comes from [1] and is preprocessed into the dataset we use. First, the KM estimator is used to estimate the survival function. The Log-rank test was used to examine whether there were significant differences in survival function among the groups. A Cox proportional hazard model is then built. After forward variable selection, we got the model and explained part of HR. In addition, we tested the model and get the final model. We also tried to construct a Cox proportional hazard deep neural network model, and segmented data sets to verify the model effect. It can be seen that Deepsurv works better.

The hazard ratio of a variable of CoxPH has a reference value. HR for smoke is 1.2938. It means that mothers in smoking is 0.7729 times more likely to finish breast feeding than people no smoking, when other variables are the same. This is actually consistent with the conclusion [7]. In addition, the 95% confidence interval for the hazard ratio of smoking was [1.1089,1.510]. Furthermore, race also has an effect on the ability to complete lactation. The model shows that HR for blacks and others is greater than one, with 95% confidence intervals of respectively [0.9891, 1.491] and [1.1334, 1.661]. For the year of birth, we divided mothers into two categories, namely, the year of birth less than 1982 and greater than 1982. It can be seen that HR for birth is 1.1955 and its 95% confidence intervals for the hazard ratio of birth is [1.0356, 1.380]. As for the education level, we divided the education time into three categories: less than 12 years, equal to 12 years and more than 12 years. It can be found that the HR of the first two categories is greater than 1, 95% confidence intervals are [1.0897, 1.694] and [1.0267, 1.434]. However, for mothers in poverty, the model suggests an HR of less than 1, with a 95% confidence interval of [0.7002, 1.009], and we suspect that affluent mothers may provide their children with other supplements instead of breast feeding. For mothers older than 21, HR is greater than 1, and the 95% confidence interval is [0.9669, 1.320],

References

[1] Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data (Vol. 1230). New York: Springer.

[2] Kim, Y., Han, S., Choi, S., & Hwang, D. (2014). Inference of dynamic networks using time-course data. Briefings in bioinformatics, 15(2), 212-228.

[3] Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. (2018). Time-varying covariates and coefficients in Cox regression models. Annals of translational medicine, 6(7).

[4] Robert, E., Coppieters, Y., Swennen, B., & Dramaix, M. (2014). Breastfeeding duration: a survival analysis—data from a regional immunization survey. BioMed Research International, 2014.

[5] Abada, T. S., Trovato, F., & Lalu, N. (2001). Determinants of breastfeeding in the Philippines: a survival analysis. Social science & medicine, 52(1), 71-81.

[6] Kumar, D., & Klefsjö, B. (1994). Proportional hazards model: a review. Reliability Engineering & System Safety, 44(2), 177-188.

[7] Napierala, M., Mazela, J., Merritt, T. A., & Florek, E. (2016). Tobacco smoking and breastfeeding: effect on the lactation process, breast milk composition and infant development. A critical review. Environmental research, 151, 321-338.

[8] Haastrup, M. B., Pottegård, A., & Damkier, P. (2014). Alcohol and breastfeeding. Basic & clinical pharmacology & toxicology, 114(2), 168-173.

[9] Wendy, H., Oddy, and, Garth, & E., et al. (2010). The long-term effects of breastfeeding on child and adolescent mental health: a pregnancy cohort study followed for 14 years. Journal of Pediatrics.

[10] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology, 18(1), 1-12.