Survival Data analysis: Principles

Survival Analysis of Breastfeeding Duration

Andrew Kamya

November, 2023

Introduction

Survival analysis is a statistical approach used to analyze time-to-event data. In this project, we examine the duration of breastfeeding using the `bfeed` dataset from the `KMsurv` package. We apply non-parametric, semi-parametric, and parametric survival models to understand factors influencing breastfeeding duration.

Data Preparation

We begin by loading the necessary packages and dataset.

```
library(survival)
library(KMsurv)
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
```

The key variables of interest are: - `duration`: Breastfeeding duration in weeks. - `delta`: Indicator for completed breastfeeding (1=yes, 0=no). - `smoke`: Whether the mother smoked at the child's birth (1=yes, 0=no).

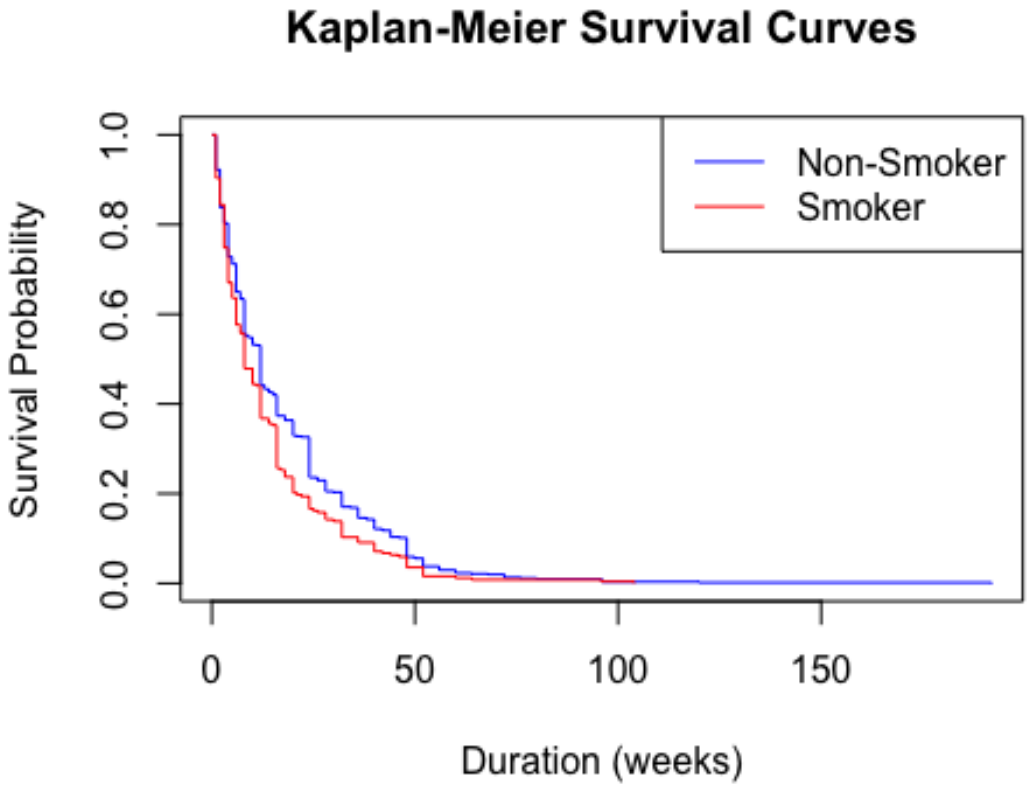
Kaplan-Meier Estimation

We estimate and plot the survival functions for smoking and non-smoking mothers.

```
data1 <- subset(bfeed, smoke == 1)
data0 <- subset(bfeed, smoke == 0)
```

```r
fit1 <- survfit(Surv(duration, delta) ~ 1, data = data1)
fit0 <- survfit(Surv(duration, delta) ~ 1, data = data0)

plot(fit0, col = "blue", conf.int = FALSE, main = "Kaplan-Meier Survival
Curves",
     xlab = "Duration (weeks)", ylab = "Survival Probability")
lines(fit1, col = "red", conf.int = FALSE)
legend("topright", legend = c("Non-Smoker", "Smoker"), col = c("blue",
"red"), lty = 1)
```



Interpretation: The survival curves suggest that mothers who smoke tend to stop breastfeeding earlier than non-smokers.

Cox Proportional Hazards Model

Next, we fit a Cox proportional hazards model to estimate relative risks.

```r
fit_cox <- coxph(Surv(duration, delta) ~ as.factor(race) + poverty + smoke +
                 alcohol + agemth + ybirth + yschool + pc3mth, data = bfeed)
summary(fit_cox)
```

```
## Call:
## coxph(formula = Surv(duration, delta) ~ as.factor(race) + poverty +
##     smoke + alcohol + agemth + ybirth + yschool + pc3mth, data = bfeed)
##
##   n= 927, number of events= 892
##
##                       coef exp(coef) se(coef)      z Pr(>|z|)
## as.factor(race)2  0.18601   1.20444  0.10545  1.764  0.07775 .
## as.factor(race)3  0.29583   1.34424  0.09719  3.044  0.00234 **
## poverty          -0.21837   0.80383  0.09384 -2.327  0.01996 *
## smoke             0.24701   1.28019  0.07957  3.104  0.00191 **
## alcohol           0.16153   1.17531  0.12304  1.313  0.18922
## agemth           -0.01570   0.98442  0.01880 -0.835  0.40363
## ybirth            0.07977   1.08303  0.02041  3.908 9.32e-05 ***
## yschool          -0.05802   0.94363  0.02315 -2.506  0.01222 *
## pc3mth           -0.05789   0.94376  0.09015 -0.642  0.52080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## as.factor(race)2    1.2044     0.8303    0.9795    1.4810
## as.factor(race)3    1.3442     0.7439    1.1111    1.6263
## poverty             0.8038     1.2440    0.6688    0.9661
## smoke               1.2802     0.7811    1.0953    1.4963
## alcohol             1.1753     0.8508    0.9235    1.4958
## agemth              0.9844     1.0158    0.9488    1.0214
## ybirth              1.0830     0.9233    1.0406    1.1272
## yschool             0.9436     1.0597    0.9018    0.9874
## pc3mth              0.9438     1.0596    0.7909    1.1262
##
## Concordance= 0.577  (se = 0.012 )
## Likelihood ratio test= 46.39  on 9 df,   p=5e-07
## Wald test            = 46.87  on 9 df,   p=4e-07
## Score (logrank) test = 46.92  on 9 df,   p=4e-07
```

Interpretation: - Smoking increases the hazard of stopping breastfeeding (`HR = 1.28`, $p < 0.01$). - Higher maternal education (`yschool`) is associated with longer breastfeeding duration.

Parametric Accelerated Failure Time (AFT) Models

We compare Weibull, log-logistic, and log-normal AFT models.

```
fit_weibull <- survreg(Surv(duration, delta) ~ as.factor(race) + poverty +
smoke +
                       alcohol + agemth + ybirth + yschool + pc3mth,
                       data = bfeed, dist = "weibull")
fit_loglogistic <- survreg(Surv(duration, delta) ~ as.factor(race) + poverty
+ smoke +
```

```
                            alcohol + agemth + ybirth + yschool + pc3mth,
                            data = bfeed, dist = "loglogistic")
fit_lognormal <- survreg(Surv(duration, delta) ~ as.factor(race) + poverty +
smoke +
                           alcohol + agemth + ybirth + yschool + pc3mth,
                           data = bfeed, dist = "lognormal")

list(weibull = logLik(fit_weibull), loglogistic = logLik(fit_loglogistic),
lognormal = logLik(fit_lognormal))
```

```
## $weibull
## 'log Lik.' -3382.031 (df=11)
##
## $loglogistic
## 'log Lik.' -3403.114 (df=11)
##
## $lognormal
## 'log Lik.' -3380.193 (df=11)
```

Interpretation: - The log-normal model provides the best fit based on log-likelihood values. - Results are consistent with the Cox model, confirming smoking negatively impacts breastfeeding duration.

Conclusion

Survival analysis of breastfeeding duration shows that maternal smoking significantly shortens breastfeeding duration. Higher education is associated with prolonged breastfeeding. The log-normal AFT model best fits the data, supporting findings from the Cox model.

This analysis highlights the importance of maternal characteristics in influencing breastfeeding behavior and can inform public health policies.