Survival Analysis Project Report

Maher SEBAI (A18)

May 1, 2019

Contents

1 Introduction

The current report attempts to analyse two Birth spacing datasets in the context of `Event History Analysis`, a field that borrows a lot of the Clinical Survival Analysis techniques to study sociological and historical phenomena. Both datasets are provided by the Medical Birth Registry of Norway. The first dataset describes the **first to second birth spacing**. The second dataset, in the same spirit, describes the **second to third birth spacing**. The two dataset does not share exactly the same covariates. We will try to study and highlight what factor is influential for the observed birth spacing. In the first section we will import, clean the two datasets and introduce the embedded covariates. In the second section we perform some exploratory analysis that will guide us in the remaining report. In the third section logrank test for groups will helps us discover the first influential categorical factors. the next section, Cox Proportional hazard modeling will help us select the significant covariate and quantify their effect size. The conclusion will be the occasion to list our findings

2 Data Acquisition and environement setup

Let's start by loading our toolkit packages. `tidyverse` package for data wrangling, and `survival` being the backbone of our inferential survival analysis.

```
library(tidyverse)
library(survival)
```

Next we will retreive our two datasets. The source is The Medical Birth Registry of Norway which was established in 1967 and contains information on all births in Norway since that time. In the first dataset we use data from the registry on the time **between first and second births** for a selection of **53 558 women**, and look at how this is affected if the first child dies within one year of its birth. The first dataset covariates are the following:

- age: age of mother at first birth (in years)
- sex: sex of first child (1=boy, 2=girl)
- death: first child died within one year (0=no, 1=yes)
- time: time from first birth to second birth or censoring (in days)
- status: censoring indicator (0=censored, 1=birth)

```
raw.second.birth <- read.table("http://folk.uio.no/borgan/abg-2008/data/second_births.txt",
                               header = T)
str(raw.second.birth)
```

```
## 'data.frame':    53558 obs. of  5 variables:
##  $ age   : num  29.3 21.9 24 29.6 22.8 29.7 23.2 27.2 22.1 21.2 ...
##  $ sex   : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ death : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ time  : int  192 424 1241 280 570 65 1344 296 1002 797 ...
##  $ status: int  0 1 0 0 1 0 0 0 1 1 ...
```

The dataset needs a bit a preprocessing to specify that sex and death covariates are to be treated as categorical covariates. We change also the scale of time from days to months for easier interpretations

```
df.second.birth <- mutate(raw.second.birth,
  sex = factor(sex, levels = c('1', '2'), labels = c('Boy', 'Girl')),
  death = factor(death, levels = c('0', '1'), labels = c('Alive', 'Dead')),
  time = round(time / 30.5, digits = 2)  # birth spacing in months instead of days
  )
head(df.second.birth)
```

```
##    age  sex death  time status
## 1 29.3 Girl Alive  6.30      0
## 2 21.9 Girl Alive 13.90      1
## 3 24.0  Boy Alive 40.69      0
## 4 29.6  Boy Alive  9.18      0
## 5 22.8 Girl Alive 18.69      1
## 6 29.7 Girl Alive  2.13      0
```

In the second dataset we use data from the same registry and listing the time **between second and third births** for a selection of **16 116 women**, and we look at how the gender of the two older children affect the likelihood of a woman having a third birth. The second dataset covariates are the following:

- age: age of mother at first birth (in years)
- spacing: time between first and second births (in days)
- sibs: genders of the first two children (1 = boy, boy; 2 = girl, girl; 3 = boy, girl; 4 = girl, boy)
- time: time from second birth to third birth or censoring (in days)
- status: censoring indicator (0=censored, 1=birth)

```
raw.third.birth  <- read.table("http://folk.uio.no/borgan/abg-2008/data/third_births.txt",
                               header = T)
str(raw.third.birth)
```

```
## 'data.frame':    16116 obs. of  5 variables:
##  $ age    : num  21.9 22.8 22.1 21.2 22.4 20.4 26.9 24.8 26.7 26.1 ...
##  $ spacing: int  424 570 1002 797 1003 1578 1029 1060 528 617 ...
##  $ sibs   : int  4 4 4 4 2 4 4 3 4 4 ...
##  $ time   : int  1826 1502 1988 652 1868 1589 142 707 316 825 ...
##  $ status : int  1 0 1 1 0 0 0 0 0 0 ...
```

The same preprocessing is needed for this dataset too.

```
df.third.birth <- mutate(raw.third.birth,
  sibs = factor(sibs, levels = c('1', '2', '3', '4'),
                labels = c('Boy-Boy', 'Girl-Girl', 'Boy-Girl', 'Girl-Boy')),
  spacing = round(spacing / 30.5, digits = 2),# birth spacing in months instead of days
  time = round(time / 30.5, digits = 2) # birth spacing in months instead of days
  )
head(df.third.birth)
```

```
##    age spacing      sibs  time status
## 1 21.9   13.90  Girl-Boy 59.87      1
## 2 22.8   18.69  Girl-Boy 49.25      0
## 3 22.1   32.85  Girl-Boy 65.18      1
## 4 21.2   26.13  Girl-Boy 21.38      1
## 5 22.4   32.89 Girl-Girl 61.25      0
## 6 20.4   51.74  Girl-Boy 52.10      0
```

3 Exploratory Analysis

We start by exploring the two data set we have collected and cleaned. A brief summary of the first one:

```
summary(df.second.birth)
```

```
##       age          sex          death            time
##  Min.   :14.60   Boy :27842   Alive:53296   Min.   :  0.00
##  1st Qu.:22.30   Girl:25716   Dead :  262   1st Qu.: 14.23
##  Median :24.30                              Median : 25.15
##  Mean   :24.25                              Mean   : 29.57
##  3rd Qu.:26.30                              3rd Qu.: 38.98
##  Max.   :30.50                              Max.   :166.23
##      status
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3051
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

let's check the relative frequencies of sex, death and status

```
prop.table(table(df.second.birth$sex))
```

```
##
##       Boy      Girl
## 0.5198476 0.4801524
```

```
prop.table(table(df.second.birth$death))
```

```
##
##       Alive        Dead
## 0.995108107 0.004891893
```

```
prop.table(table(df.second.birth$status))
```

```
##
##         0         1
## 0.6948915 0.3051085
```

Boys and Girls are evenly distributed in the second birth as expected. Death of first Baby is a rare event (less than 0.5%). It will be interesting to study the effect of such event of the second birth spacing. We notice also that around 69% of the observations are **censored (status = 0)**. The cause of censoring are not explained in the dataset documentation, so we can not infer that 69% of woman in Norway perfer to have a single child and not giving birth afterwards.

Let's have a look at the women's age distribution:

```
hist(df.second.birth$age, xlab = 'Mother age', main = 'Histogram of Mother age at second birth')
```

Mother's age looks nearly normally distributed with mean age at 24.25 years-old. Let's explore the second dataset for the second to third birth.

```
summary(df.third.birth)
```

```
##       age           spacing             sibs          time
##  Min.   :16.00   Min.   :  0.00   Boy-Boy  :4334   Min.   :  0.00
##  1st Qu.:21.70   1st Qu.: 22.38   Girl-Girl:3759   1st Qu.:  9.90
##  Median :23.40   Median : 29.93   Boy-Girl :4067   Median : 20.82
##  Mean   :23.32   Mean   : 32.61   Girl-Boy :3956   Mean   : 24.93
##  3rd Qu.:25.00   3rd Qu.: 39.34                    3rd Qu.: 35.58
##  Max.   :29.10   Max.   :120.56                    Max.   :117.74
##      status
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1093
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

How is the previous birth's sex distribution and status distribution ?

```
prop.table(table(df.third.birth$sibs))
```

```
##
##   Boy-Boy Girl-Girl  Boy-Girl  Girl-Boy
## 0.2689253 0.2332465 0.2523579 0.2454703
```

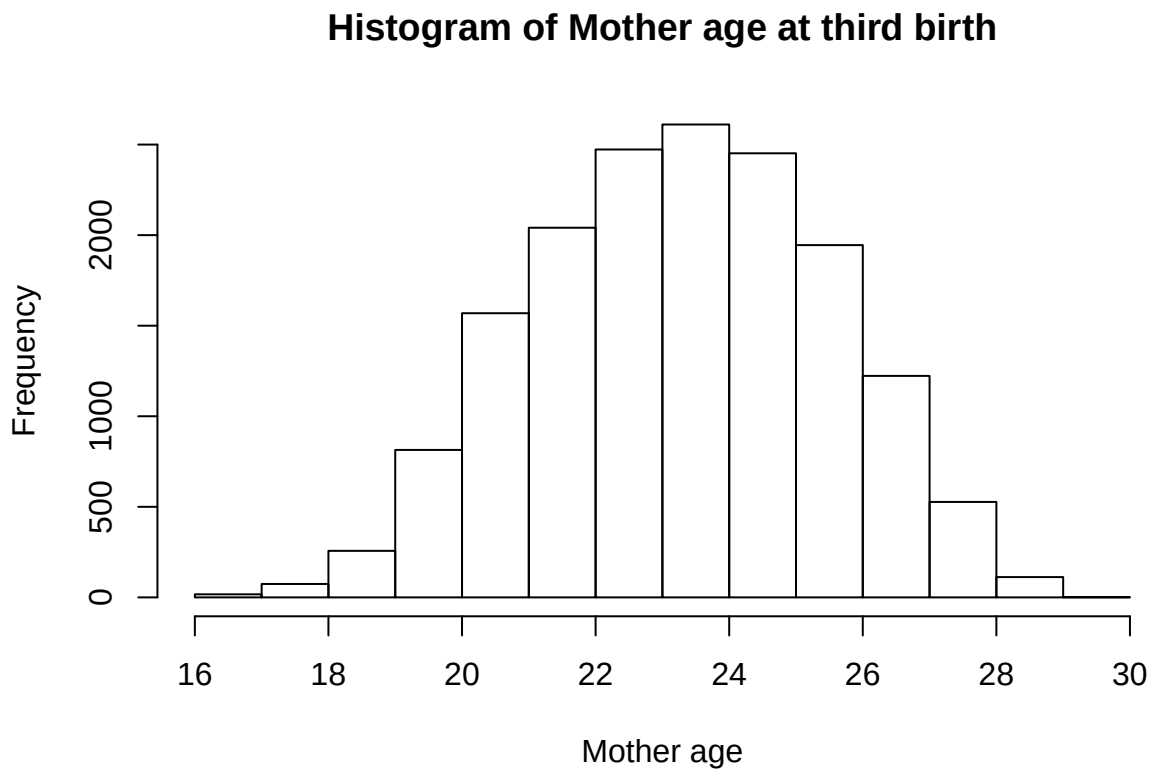
```
prop.table(table(df.third.birth$status))
```

```
##
##         0         1
## 0.8907297 0.1092703
```

Siblings is uniformally distributed among the 4 permutations (around 25% each). 89% of the observation are censored for the second to third birth spacing. Again we can not conclude as we don't have the reason of censorship.

The mother's age distribution:

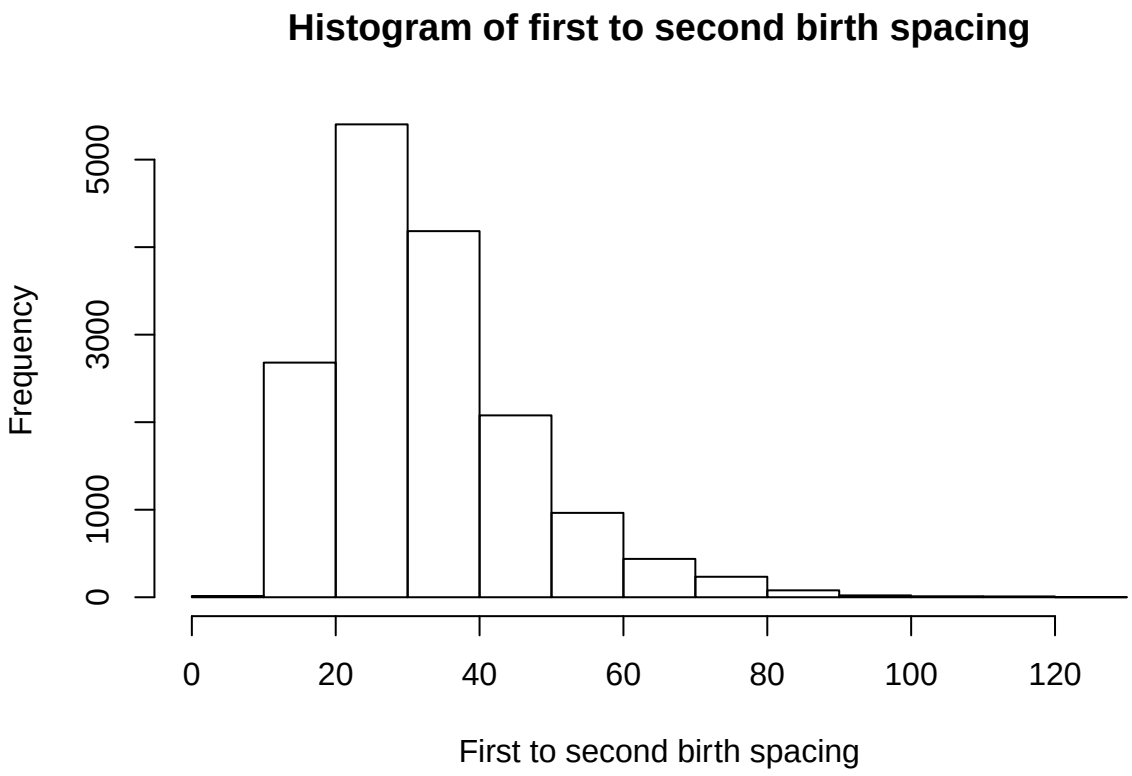
```
hist(df.third.birth$age, xlab = 'Mother age', main = 'Histogram of Mother age at third birth')
```



Again the age is nearly normally distributed with, **interestingly**, a mean at 23.32 years-old (**younger** than the previous dataset subjects)

The spacing distribution:

```
hist(df.third.birth$spacing, xlab = 'First to second birth spacing',
     main = 'Histogram of first to second birth spacing')
```



The spacing distribution for previous birth is right skewed. We'll see if measures are needed to removed this skewness when modeling the data.

4 Inferential Statistics

4.1 Kaplan-Meir Estimator

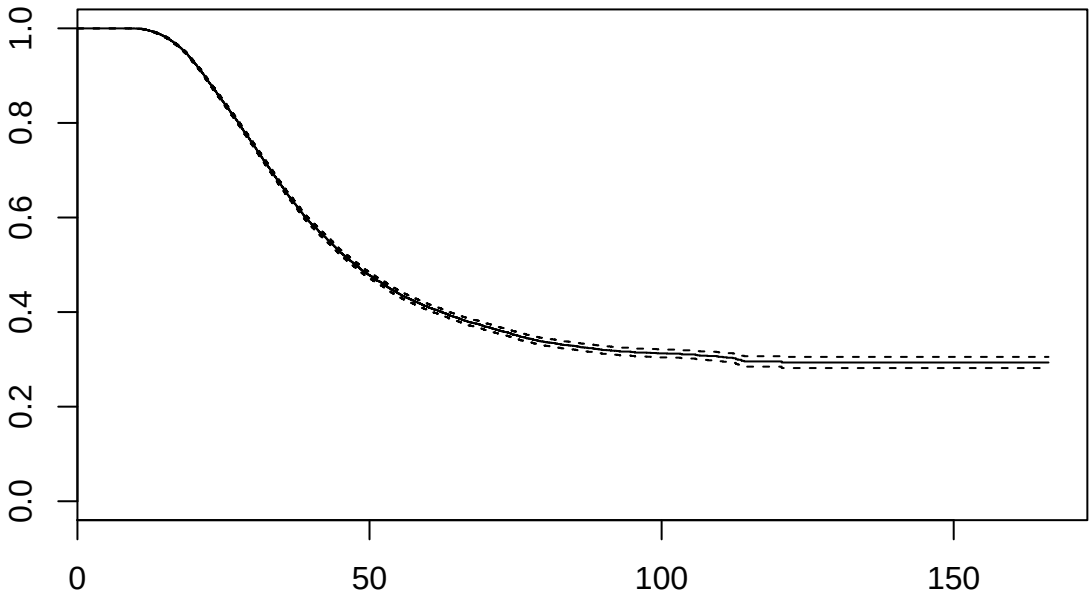
Let's process the first dataset:

```
second.birth.fit.KM <- survfit(Surv(time, status) ~ 1, data = df.second.birth, conf.type = "log-log")
second.birth.fit.KM
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = df.second.birth,
##     conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
## 53558.0 16341.0    47.7    47.0    48.2
```

the median spacing between first and second birth is 47.7 months (~ 4 years, Nowegians prefer to dedicate a good amount of time to their first child before conceiving the next).

```
plot(second.birth.fit.KM)
```



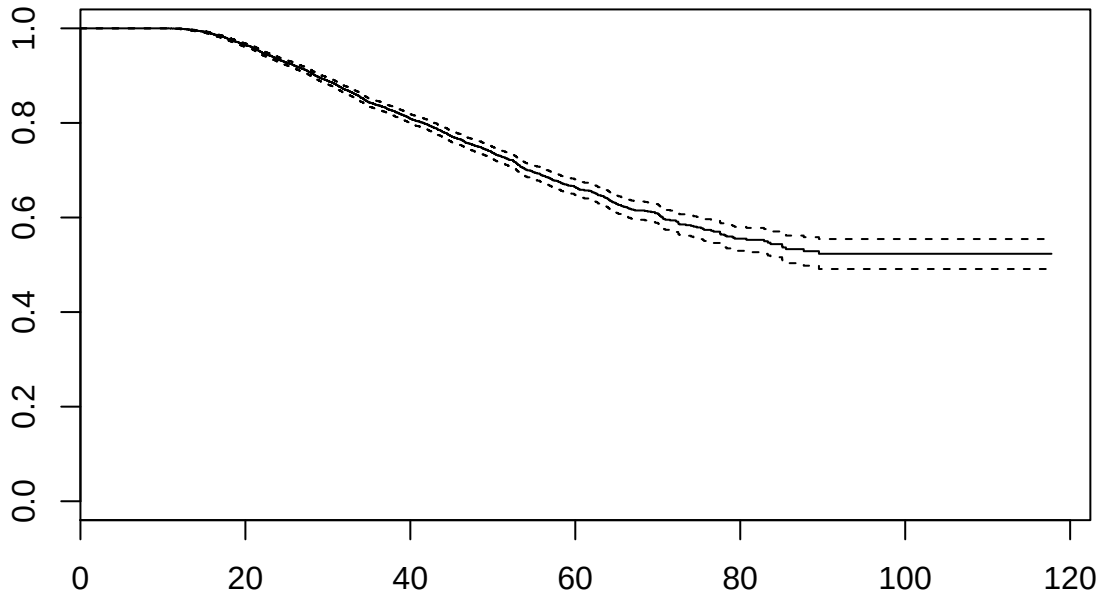
Let's analyse the second dataset:

```
third.birth.fit.KM <- survfit(Surv(time, status) ~ 1, data = df.third.birth, conf.type = "log-log")
third.birth.fit.KM
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = df.third.birth,
##     conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
## 16116.0  1761.0      NA    87.7      NA
```

The median in this case is `NA` meaning that the Survival function never reached the probability value of 0.5 . the 95% lower confidence level is 87.7 months(more than 7 years)

```
plot(third.birth.fit.KM)
```



4.2 Logrank test: Sex covariate effect

Now we get serious about our analysis. The first dataset exhibits two categorical covariates each with two levels(sex and death). This is perfect to comparing both strata using logrank tests. Let's start with the first covariate: sex of the first child

```
survfit(Surv(time, status) ~ sex, data = df.second.birth, conf.type = "log-log")
```

```
## Call: survfit(formula = Surv(time, status) ~ sex, data = df.second.birth,
##     conf.type = "log-log")
##
##              n events median 0.95LCL 0.95UCL
## sex=Boy  27842   8544   47.4    46.4    48.2
## sex=Girl 25716   7797   47.8    47.0    48.8
```

The median spacing is pretty close for the two groups. Let's perform a logrank test to compare the two groups

```
survdiff(Surv(time, status) ~ sex, data = df.second.birth)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = df.second.birth)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Boy  27842     8544     8461     0.811      1.68
## sex=Girl 25716     7797     7880     0.871      1.68
##
##  Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

the logrank test output a p value of 0.2, given a significance threshold at 0.05 it is pretty poor. We conclude that the sex of the first baby is **not affecting** the spacing between the first to second birth. If that was the case, populations would exhibit a dangerous imbalance between woman and men proportions.

4.3 Logrank test: Death covariate effect

Let's now study the tragic event where the first child is dead within a year of its birth. How would this event influence the next baby birth spacing:

```
survfit(Surv(time, status) ~ death, data = df.second.birth, conf.type = "log-log")
```

```
## Call: survfit(formula = Surv(time, status) ~ death, data = df.second.birth,
##     conf.type = "log-log")
##
##                 n events median 0.95LCL 0.95UCL
## death=Alive 53296  16188   47.7    47.0    48.3
## death=Dead    262    153   22.3    19.6    26.4
```

The median spacing duration is notably different (22.3 months when first baby is dead vs 47 months otherwise). A logrank test to help us conclude:

```
survdiff(Surv(time, status) ~ death, data = df.second.birth)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ death, data = df.second.birth)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## death=Alive 53296    16188  16274.5      0.46       113
## death=Dead    262      153     66.5    112.63       113
##
##  Chisq= 113  on 1 degrees of freedom, p= <2e-16
```

p value is **significantly low** confirming the effect of the first baby death on shortening the time spacing to give a second child. Naturally parents affected by such tragedy are inclined to make a second baby faster. Later, a proportional Cox Model will help us quantify the effect coefficient.

4.4 Logrank test: Parity covariate effect

We have two datasets that can be treated as two groups if combined so that we can also study if the parity (order of delivery) can contribute to the birth spacing. Essentially, we want to know if parents change the spacing of birth for the second to third birth compared to the first to second birth.

For this study we need to combine the two datasets attaching a label mentioning the origin dataset. We choose also to remove from the first dataset the observations where first child is dead to be closer to second dataset spirit.

```
df.part1 <- df.second.birth %>% filter(death == 'Alive') %>%
  transmute(age = age, time = time, status = status, group = as.factor('second'))
summary(df.part1)
```

```
##       age             time            status          group
##  Min.   :14.60   Min.   :  0.00   Min.   :0.0000   second:53296
##  1st Qu.:22.30   1st Qu.: 14.26   1st Qu.:0.0000
##  Median :24.30   Median : 25.18   Median :0.0000
##  Mean   :24.25   Mean   : 29.58   Mean   :0.3037
##  3rd Qu.:26.30   3rd Qu.: 39.02   3rd Qu.:1.0000
##  Max.   :30.50   Max.   :166.23   Max.   :1.0000
```

We preprocess the second dataframe and combine the two.

```
df.part2 <- df.third.birth %>%
  transmute(age = age, time = time, status = status, group = as.factor('third'))
summary(df.part2)
```

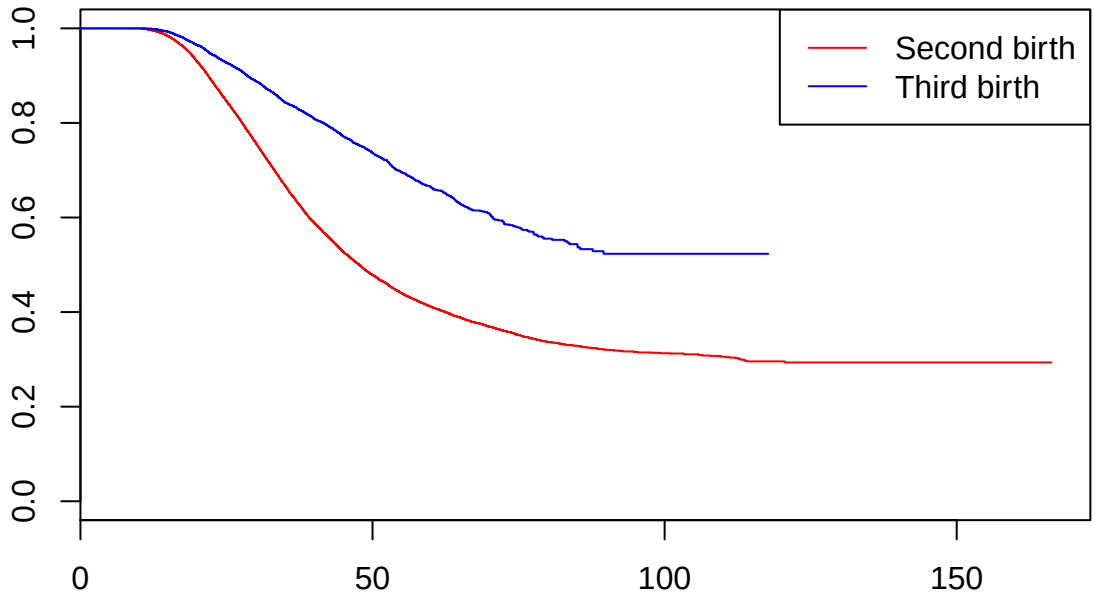
```
##       age             time            status         group
##  Min.   :16.00   Min.   :  0.00   Min.   :0.0000   third:16116
##  1st Qu.:21.70   1st Qu.:  9.90   1st Qu.:0.0000
##  Median :23.40   Median : 20.82   Median :0.0000
##  Mean   :23.32   Mean   : 24.93   Mean   :0.1093
##  3rd Qu.:25.00   3rd Qu.: 35.58   3rd Qu.:0.0000
##  Max.   :29.10   Max.   :117.74   Max.   :1.0000
```

```
df.combined.birth = rbind(df.part1, df.part2)
summary(df.combined.birth)
```

```
##       age             time            status          group
##  Min.   :14.60   Min.   :  0.00   Min.   :0.0000   second:53296
##  1st Qu.:22.10   1st Qu.: 13.15   1st Qu.:0.0000   third :16116
##  Median :24.00   Median : 24.30   Median :0.0000
##  Mean   :24.04   Mean   : 28.50   Mean   :0.2586
##  3rd Qu.:26.00   3rd Qu.: 38.26   3rd Qu.:1.0000
##  Max.   :30.50   Max.   :166.23   Max.   :1.0000
```

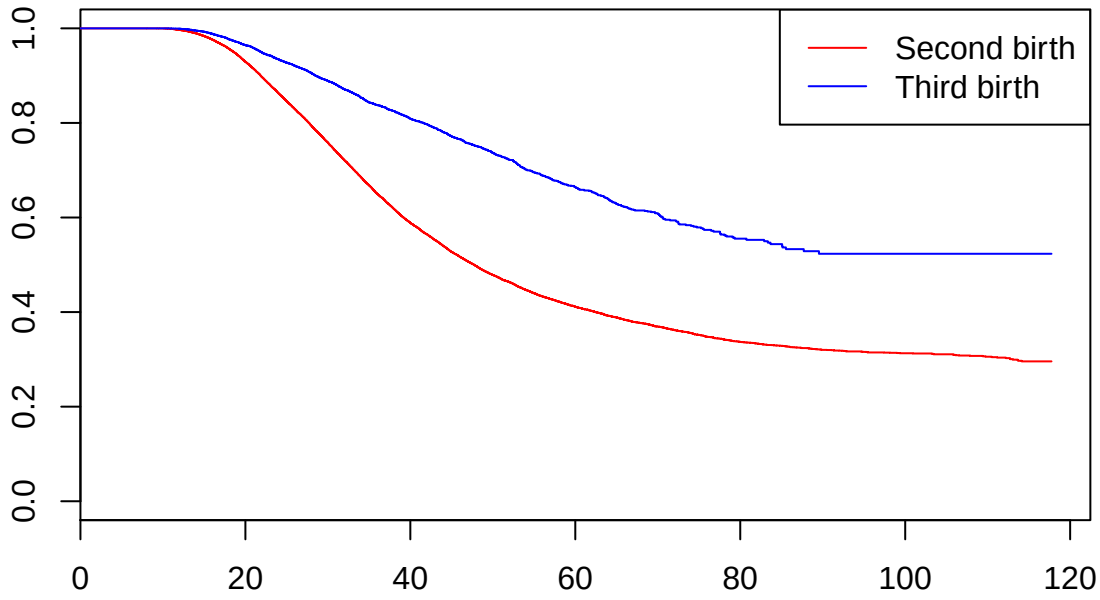
Let's study the effect of the parity group. We plot first the Kaplan-Meir curve

```
fit.comb <- survfit(Surv(time, status) ~ group, data = df.combined.birth)
plot(fit.comb, col=c("red", "blue"))
legend("topright", legend=c("Second birth", "Third birth"), col=c("red", "blue"), lty=1)
```



We adjust the second birth spacing curve by truncating what is exceeding the compared curve.

```
truncated.birth <- within(df.combined.birth, {
  status <- ifelse(time > max(df.part2$time), 0, status)
  time <- ifelse(time > max(df.part2$time), max(df.part2$time), time)
})
fit.comb <- survfit(Surv(time, status) ~ group, data = truncated.birth)
plot(fit.comb, col=c("red", "blue"))
legend("topright", legend=c("Second birth", "Third birth"), col=c("red", "blue"), lty=1)
```



Now we can compare the two groups by performing a logrank test.

```
survdiff(Surv(time, status) ~ group, data = truncated.birth)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ group, data = truncated.birth)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## group=second 53296    16187    14437       212      1087
## group=third  16116     1761     3511       872      1087
##
##  Chisq= 1087  on 1 degrees of freedom, p= <2e-16
```

The test shows a significant increase of Birth spacing for the third child compared to the second child. I can attest this observation, being a father of two childs, it is pretty fulfilling and the desire to have a third is quiet low.

5 Cox Proportional Hazard modeling

5.1 Second birth dataset modeling

The two datasets embed also some continuous covariates (age)and multilevel categorical covariate (siblings). let's model the first data set using a Cox PH model using all the covariates.

```
fit.second.ph <- coxph(Surv(time, status) ~ ., data = df.second.birth)
summary(fit.second.ph)
```

```
## Call:
```

```
## coxph(formula = Surv(time, status) ~ ., data = df.second.birth)
##
##   n= 53558, number of events= 16341
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age        0.087725  1.091687  0.003606 24.331   <2e-16 ***
## sexGirl   -0.018660  0.981513  0.015665 -1.191    0.234
## deathDead  0.888988  2.432667  0.081279 10.938   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0917     0.9160    1.0840     1.099
## sexGirl      0.9815     1.0188    0.9518     1.012
## deathDead    2.4327     0.4111    2.0744     2.853
##
## Concordance= 0.553  (se = 0.003 )
## Rsquare= 0.013   (max possible= 0.998 )
## Likelihood ratio test= 684.3  on 3 df,   p=<2e-16
## Wald test            = 701.1  on 3 df,   p=<2e-16
## Score (logrank) test = 709.5  on 3 df,   p=<2e-16
```

The model output tell us that `age` and `death` covariates are significant, in contrast of the `sex` which fail to be significant (already discovered in the previous logrank test).

We are going to let the software select automatically the best model covariate using the `step` funtion. Of course for such low dimensional dataset (only 3 covariates) this is overkill and ridiculous, but it is a good occasion to apply what we learned during the course :-)

```
best.second.ph <- step(fit.second.ph)
```

```
## Start:  AIC=320274.8
## Surv(time, status) ~ age + sex + death
##
##         Df    AIC
## - sex    1 320274
## <none>     320275
## - death  1 320364
## - age    1 320873
##
## Step:  AIC=320274.2
## Surv(time, status) ~ age + death
##
##         Df    AIC
## <none>     320274
## - death  1 320364
## - age    1 320873
```

No surprise, the sex covariate was exluded from the final model remaining with only age and death. Let's study their effect size:

```
summary(best.second.ph)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + death, data = df.second.birth)
##
##   n= 53558, number of events= 16341
```

```
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## age       0.087721  1.091683 0.003606 24.33   <2e-16 ***
## deathDead 0.890936  2.437411 0.081262 10.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age           1.092     0.9160     1.084     1.099
## deathDead     2.437     0.4103     2.079     2.858
##
## Concordance= 0.553  (se = 0.003 )
## Rsquare= 0.013   (max possible= 0.998 )
## Likelihood ratio test= 682.9  on 2 df,   p=<2e-16
## Wald test            = 699.7  on 2 df,   p=<2e-16
## Score (logrank) test = 708  on 2 df,   p=<2e-16
```

the likelihood ratio test is significant meaning that the model is sound. p-values are also remarkably significant for both covariates. The coefficient of both covariated is positive (respectively 0.08 and 0.89) leading to an amplifying effect.

- age: exp(coef) = 1.09 means that one year of increase of the mother's age amplify the chance/desire to have a second child
- death: exp(coef) = 2.43 means that woman experiencing the tragedy of death of first child are 2.43 times more willing to conceive a second child compared to woman who did not suffer such tragedy.

let's test if the Proportional Hazard Assumptions holds:

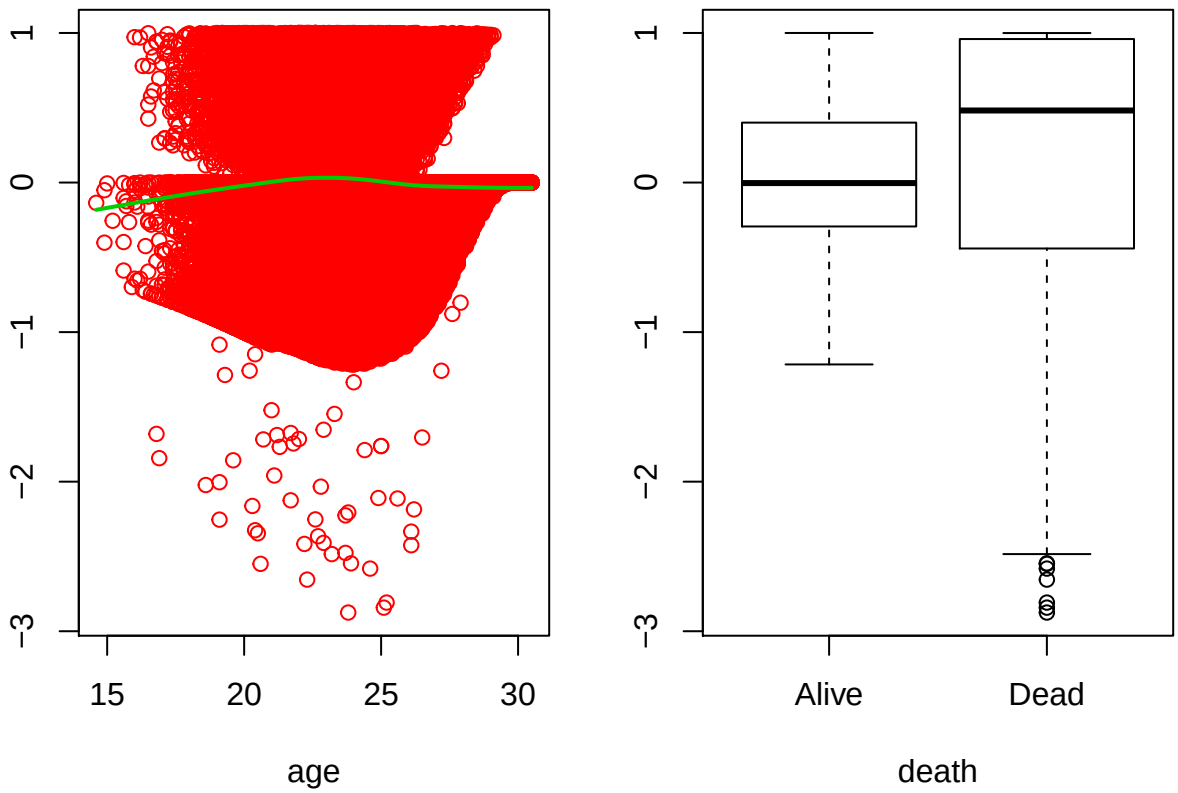
```
test.ph <- cox.zph(best.second.ph)
test.ph
```

```
##                rho chisq        p
## age        0.0657  66.7 3.09e-16
## deathDead -0.1040 176.6 2.65e-40
## GLOBAL         NA 248.7 9.66e-55
```

From the output above we notice that the test is strongly significant for each of the covariate and the global model is also strongly significant. Unfortunatly this means that the Cox Proportionality assumptions are not met and exhibit some time time dependency of the covariates.

```
df.second.birth$residual <- residuals(best.second.ph, type = "martingale")
par(mfrow = c(1, 2), mar = c(4.2, 2, 2, 2))
with(df.second.birth, {

  plot(age, residual,col = 2)
  lines(lowess(age, residual), lwd = 2, col = 3)
  plot(residual ~ death)
})
```



The residuals show indeed some non linearity. Maybe some advanced techniques are thus needed to dig deeper.

5.2 Third birth dataset modeling

We found during the exploratory phase that the spacing covariate is right skewed. We will add a modified version of the covariated (`log(1 + spacing)`) to counter the skewness effect.

```
fit.third.ph <- coxph(Surv(time, status) ~ spacing + log1p(spacing) + age + sibs, data = df.third.birth)
summary(fit.third.ph)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ spacing + log1p(spacing) +
##     age + sibs, data = df.third.birth)
##
##   n= 16116, number of events= 1761
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## spacing          0.007446  1.007473  0.005693  1.308  0.19091
## log1p(spacing) -0.493691  0.610370  0.166464 -2.966  0.00302 **
## age              0.006955  1.006980  0.014542  0.478  0.63244
## sibsGirl-Girl  -0.115231  0.891160  0.065063 -1.771  0.07655 .
## sibsBoy-Girl   -0.283112  0.753436  0.065812 -4.302 1.69e-05 ***
## sibsGirl-Boy   -0.266826  0.765806  0.065659 -4.064 4.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
```

```
## spacing           1.0075     0.9926    0.9963    1.0188
## log1p(spacing)    0.6104     1.6384    0.4405    0.8458
## age               1.0070     0.9931    0.9787    1.0361
## sibsGirl-Girl     0.8912     1.1221    0.7845    1.0124
## sibsBoy-Girl      0.7534     1.3273    0.6623    0.8572
## sibsGirl-Boy      0.7658     1.3058    0.6733    0.8710
##
## Concordance= 0.55  (se = 0.008 )
## Rsquare= 0.003   (max possible= 0.839 )
## Likelihood ratio test= 50.52  on 6 df,   p=4e-09
## Wald test            = 53.84  on 6 df,   p=8e-10
## Score (logrank) test = 53.64  on 6 df,   p=9e-10
```

Bingo! the model prefers much more the modified version of the covariate (p-value = 0.003) compared to the original covarated which p-value is pretty poor (0.19). Surprisingly, for the second to third birth spacing, the age covariated is not significant anymore! Let's automate the model building (overkill, i know :-))

```
best.third.ph <- step(fit.third.ph)
```

```
## Start:  AIC=29427.62
## Surv(time, status) ~ spacing + log1p(spacing) + age + sibs
##
##                  Df   AIC
## - age             1 29426
## - spacing         1 29427
## <none>              29428
## - log1p(spacing)  1 29433
## - sibs            3 29447
##
## Step:  AIC=29425.85
## Surv(time, status) ~ spacing + log1p(spacing) + sibs
##
##                  Df   AIC
## - spacing         1 29425
## <none>              29426
## - log1p(spacing)  1 29431
## - sibs            3 29445
##
## Step:  AIC=29425.29
## Surv(time, status) ~ log1p(spacing) + sibs
##
##                  Df   AIC
## <none>              29425
## - sibs            3 29445
## - log1p(spacing)  1 29446
```

The final model selected only log1p(spacing) and sibling covarate. Let's check their coefficient significance:

```
summary(best.third.ph)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ log1p(spacing) + sibs, data = df.third.birth)
##
##   n= 16116, number of events= 1761
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## log1p(spacing) -0.29286   0.74613  0.06073 -4.823 1.42e-06 ***
## sibsGirl-Girl  -0.11536   0.89104  0.06505 -1.773   0.0762 .
## sibsBoy-Girl   -0.28359   0.75307  0.06580 -4.310 1.63e-05 ***
## sibsGirl-Boy   -0.26548   0.76684  0.06565 -4.044 5.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## log1p(spacing)    0.7461      1.340    0.6624    0.8404
## sibsGirl-Girl     0.8910      1.122    0.7844    1.0122
## sibsBoy-Girl      0.7531      1.328    0.6620    0.8567
## sibsGirl-Boy      0.7668      1.304    0.6743    0.8721
##
## Concordance= 0.548  (se = 0.008 )
## Rsquare= 0.003   (max possible= 0.839 )
## Likelihood ratio test= 48.85  on 4 df,   p=6e-10
## Wald test            = 49.57  on 4 df,   p=4e-10
## Score (logrank) test = 49.48  on 4 df,   p=5e-10
```

All the coefficients are now negative, meaning that they have a shrinking/reducing effect size.

- for spacing: exp(coeff) = 0.74 means that 1 unit increase of log(1 + spacing) reduces the chance to have a third child. The more woman waites before having her second child the less the desire to have a third child, exactly the opposite effect of age for having a second child.
- for sibling: the reference level is 'Boy-Boy'. We discover that the levels 'Boy-Girl' and 'Girl-Boy' have a smaller coefficients (resp 0.75 and 0.76) thus less chance to have a third child compared to the level 'Girl-Girl' (which p-value is not significant by the way). This express the desire of parents whose first two childs are of the same sex their williness to have a third child hopefully of different sex, even though that desire is vanishing (negative coefficient).

let's test if the Proportional Hazard Assumptions holds:

```
test.ph <- cox.zph(best.third.ph)
test.ph
```

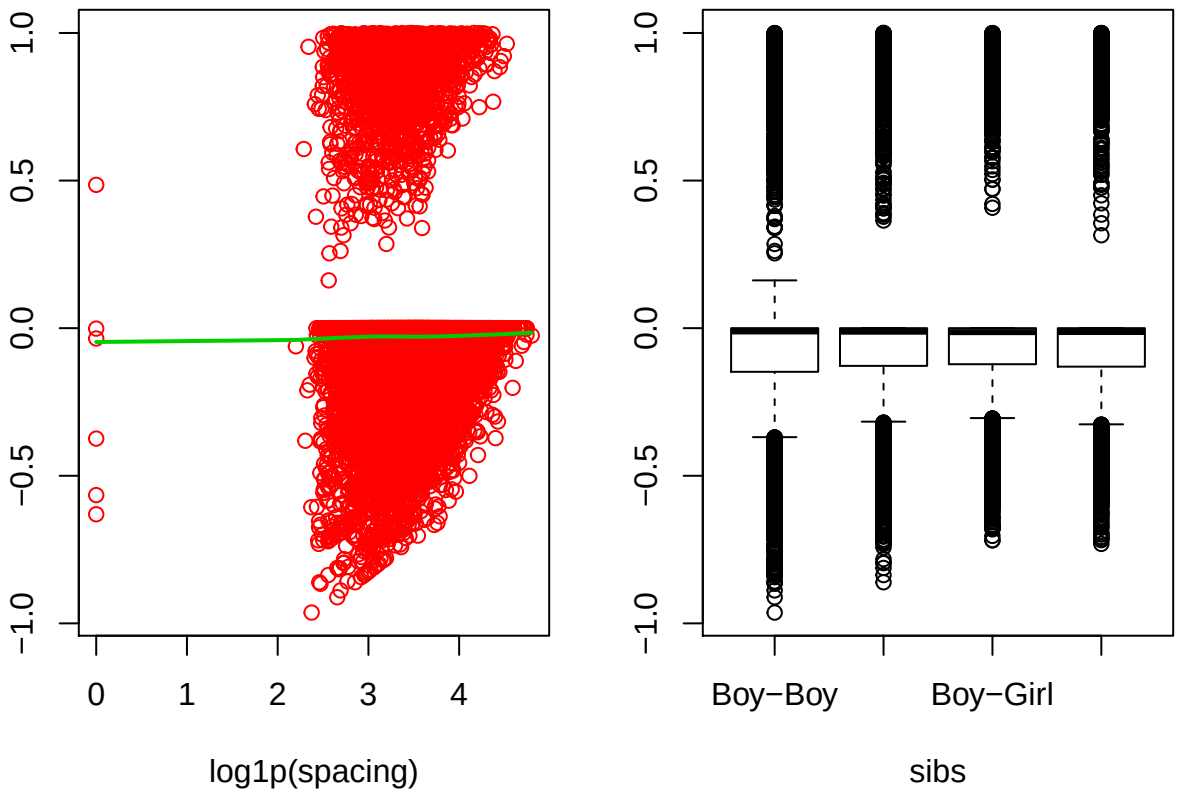
```
##                     rho  chisq     p
## log1p(spacing) -0.00456 0.0370 0.847
## sibsGirl-Girl   0.01837 0.5963 0.440
## sibsBoy-Girl    0.00373 0.0245 0.876
## sibsGirl-Boy   -0.00456 0.0366 0.848
## GLOBAL              NA 0.9435 0.918
```

The p-values of all covariates and the global model p-value are not significant, so we do not reject the Null Hypothisis. Meaning that the Model assumption holds and the model is sound.

Let's check the residuals:

```
df.third.birth$residual <- residuals(best.third.ph, type = "martingale")
par(mfrow = c(1, 2), mar = c(4.2, 2, 2, 2))
with(df.third.birth, {

  plot(log1p(spacing), residual,col = 2)
  lines(lowess(log1p(spacing), residual), lwd = 2, col = 3)
  plot(residual ~ sibs)
})
```



indeed, no non-linearity is shown by the residuals which explains why the model assumptions are validated.

6 Conclusion

We studied in this report two Birth spacing datasets from the Medical Birth Registry of Norway. the first dataset is first to second birth spacing, the other dataset is second to third birth dataset. We studied the factors that influence such spacing. We concluded the following items:

- the age and the more strongly the death of the first child amplify the chance for woman to have rapidely a second child. In contrast, the sex of the first child is not impacting this duration.
- the spacing between first to second birth is significantly lower than the second to third birth.
- for the second to third birth spacing, the age of the mother is no longer an significant factor, in contract with the birth spacing of the previous birth and the sex of the two previous kids. These factor have a negative coefficient leading to a vanishing chance to have third birth if the previous birth spacing increases or if the sex of the first two kids are different (a boy and a girl: royal combination as we say in France).

Hope you enjoyed reading.

kindly

Maher SEBAI