Data ScienceTech Institute

The effect of gender of first two children on the likelihood of a woman having a third birth

- Survival analysis assignment -

Students

Katarina Milivojević A19

Amine El Farssi S19

Teacher

Antonio, Fabio Di Narzo

Biot, 20th May 2020

Time to third birth data

The Time to third birth data of the Medical Birth Registry of Norway contains records of all births in Norway since 1967. For the purpose of this assignment, we used a sample of 16 116 women[1], along with the following information:

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

There was right-censoring in this data, as some women possibly had not yet gave the third birth until the moment this data was collected, or they, due to different reasons were lost to follow-up after the second birth (for example, moving abroad, death, etc).

Data preprocessing

Data preprocessing in this analysis included adding an ID to each case, as well as adding variables of time between first and second birth in months (*spacing_months*), and time from second birth to third birth or censoring in months (*time_months*). These two variables were computed by dividing variables *spacing* and *time* with 30.5, as the approximately average number of days in a month. The variable *age over 20* was created as well, with codes 1 = at least 20, and 0 = less than 20.

[1] The data was downloaded from the selection of 16 116 cases from the initial data set (http://folk.uio.no/borgan/abg-2008/data/third_births.txt), along with the original description (http://folk.uio.no/borgan/abg-2008/data/third_births_description.html).

Research question and objectives

The research question of this study was how the gender of the first two children affects the likelihood of a woman having a third birth over time (*risk*, in terms of survival analysis). Accordingly, the following objectives were set:

1. To visualize and compare the probability over time for having the third birth after the second birth later or not having it at all (*survival probability*, in terms of survival analysis) for each combination of the first two babies' genders.
2. To estimate a model for a survival time (from the second to the third birth), with respect to age of the mother at first birth, time between first and second birth and genders of the first two children. To evaluate the quality of this model.
3. To identify, using this model, the first eighty women from the new unlabeled dataset who have the highest risk rates for the third birth.

Methods used

The data was analyzed in Rstudio software, using programming language R.

The first objective, regarding comparisons of survival probability, was addressed by plotting the Kaplan-Meier estimator for the whole sample and for each of the four groups of women by genders of their children. Survival curves of the groups were compared using the log-rank test.

The second objective, regarding the model estimation and evaluation, was addressed with multivariate Cox regression. The assumptions of proportionality of hazards, absence of outliers and linearity were tested by plotting Schoenfeld, deviance, and Martingale residuals respectively.

The third objective, related to predictions of the risk rates using the Cox model, was addressed by implementing the obtained model in a selected unlabeled dataset, and sorting the cases, to identify the top ten.

Time between the second and the third birth was used both in days and in months: in months mainly for easier interpretation of the plots, and in days for more precision in the Cox regression model.

Descriptive characteristics of the data

The data for modelling included N = 16 116 observations.

The average age of mother at first birth was M = 23.32, with the median of 23.40 years (SD = 2.20, min = 16.00, max = 29.10). The average time between the first and second childbirth was M = 32.61 months, while the median time was 29.93 months (SD = 14.02, min = 0.00, max = 120.56). Table 1 presents frequencies and percentages of each combination of genders of the first two children.

Table 1. Frequencies and percentages of the genders of the first two children

Genders	**Frequency**	**Percent (%)**
Boy, boy	4 334	26.9
Girl, girl	3759	23.3
Boy, girl	4067	25.2
Girl, boy	3956	24.5

N = 16 116.

Out of the whole sample, 10.9% (n = 1 761) women have given the third birth. For them, the average time between the second and the third birth was M = 33.03 months, with the median of 29.80 months (SD = 14.70, min = 10.59, max = 89.54).

Results

Survival probability for having the third birth depending on the group

Survival probabilities for having the third birth after the second one are presented in Figures 1 and 2 in the form of Kaplan-Meier plots. Horizontal axes represent time in months (used for easier interpretability), and the vertical axes display the probability of surviving, i.e., in this study, probability of not having the third birth. Each vertical drop in the curves represent an event, i.e. a childbirth. At the zero month, on the both figures the survival probabilities were 1.0, i.e. none of the women had the third baby at the same time at having the second baby. After eight years (months), on the both figures, percentage of women was zero, which means that no woman was likely to have the third baby after this time from the second birth.

According to Figure 1, the survival proportion of the third birth of around 50.0 - 55.0% stabilizes after around 85 months, i.e. seven years. Therefore, seven years after the second birth, approximately 50.0 - 55.0% of women will not have the third baby yet.

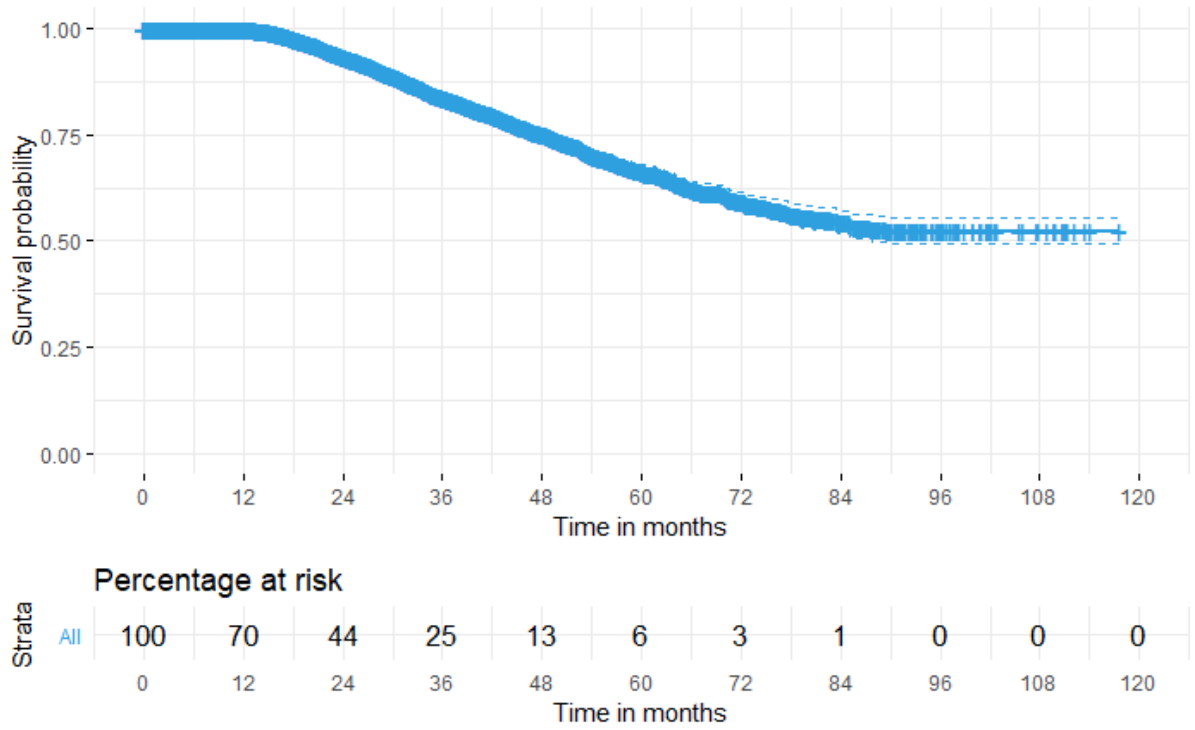

Figure 1. Survival probability of the third birth for all women

Figure 2 presents the survival probability for the third birth for the four groups of women, based on the genders of their first two children. These probability curves are very similar among each other, with the groups who have children of both genders having slightly higher overall survival probabilities, than the mothers of same-gender children. As shown in the figure, the result of the log-rank test was highly significant ($p < .0001$). There was a significant difference between the groups in survival probability.

Median survival times were not computed, as survival probabilities remained above 50.0% at the end of the measuring time. The exception was the group with two boys, with the median survival time of 85.1 month, which is around seven years. Therefore, women who had two boys have survival probability of 50.0% to have the third baby within around seven years from the second birth.

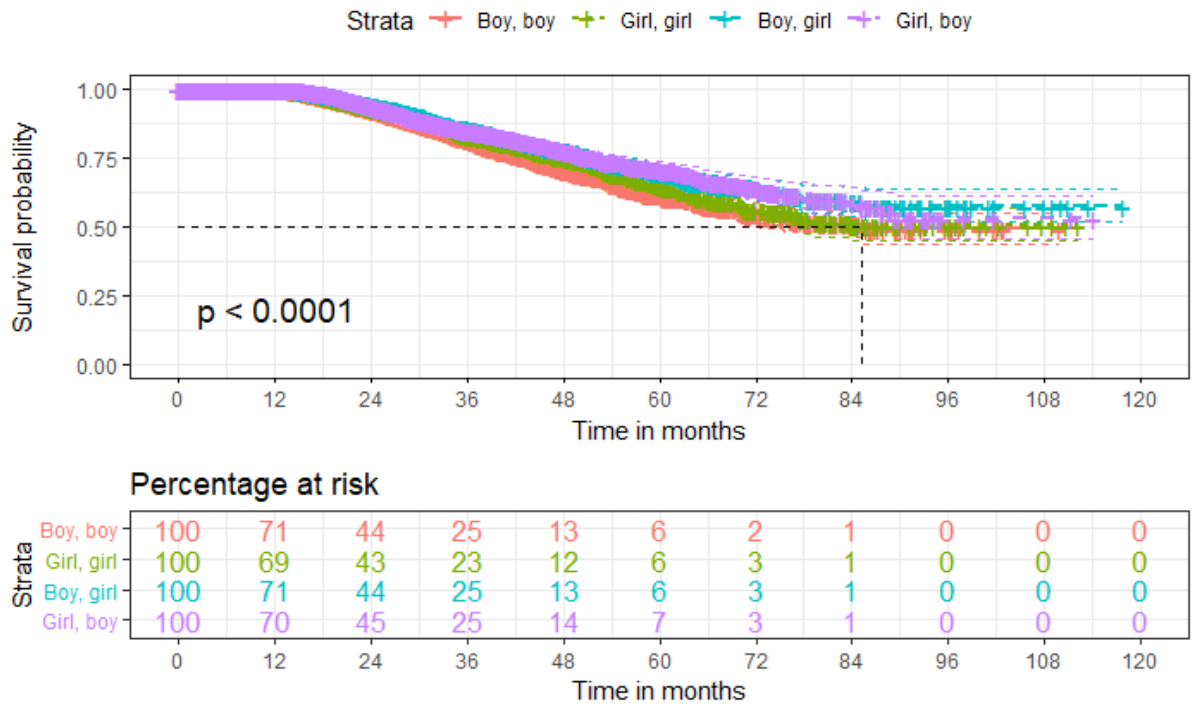

Figure 2.Survival probability of the third birth for the four groups of women by their previous children genders

More precisely, according to the results of the log-rank test the difference in survival probability curves between the four groups of women by gender was significant at the .0008% level of significance (χ^2 (3) = 26.4, p = 8e-06). Therefore, we can reject the null hypothesis that there is no difference in survival probability over time between the groups of women. As shown in Table 2, the largest weighted number of events (both observed and expected) was in the group of women who had two boys.

Table 2. The weighted number of childbirths in each group of women by the first two children's genders

		Weighted number of events (childbirth)	
Genders	***n***	**Observed**	**Expected**
Boy, boy	4 334	559	475
Girl, girl	3 759	411	397
Boy, girl	4 067	394	447
Girl, boy	3 956	397	442

N = 16 116.

Multivariate Cox regression model

Model estimation

According to results of the multivariate Cox regression analysis (Table 2), the hazard of the third birth was not affected by the mother's age at the first birth (p = .73), while time from the first to the second birth and gender of the first two children were significant predictors.

One-day longer time from first to the second birth, holding the other variables constant, reduces the hazard of the third birth by a factor of HR = .9997, or for 0.3%. Therefore, with longer time from first to the second birth, probability of the third birth should also be somewhat lower. This difference is extremely small, but highly significant (p = .0002), due to a very small standard error of the regression coefficient.

Already having two boys, compared to having two girls, does not significantly affect the hazard of the third birth (p = .07). With having first a boy and then a girl, compared to having two

girls, the hazard of the third birth is reduced by a factor of HR = .8461, or for 15.4%. With having first a girl, and then a boy, compared to having two girls, the hazard of the third birth is reduced by HR = .8627, or for 13.7%. Therefore, with already having two children of different genders, compared to having two girls, likelihood of having the third birth becomes lower with time.

Table 3. Multivariate Cox regression coefficients of predictors of the hazard of the third birth

	Regression coefficients					**95.0% C.I. for Exp(B)**	
Predictor	**B**	**Exp(B)**	**Exp(-B)**	**S.E.(B)**	**Sig.**	**L.B.**	**U.B.**
Age at first birth	.0049	1.0050	1.9951	.0145	.7341	.9768	1.0339
Time from 1st to 2nd birth	-.0003	.9997	1.0003	.0001	.0002	.9996	.9999
Boy, boy	.1174	1.1250	.8892	.06504	.0711	.9900	1.2774
Boy, girl	-.1671	.8461	1.1818	.0705	.0178	.7369	.9716
Girl, boy	-.1477	.8627	1.1592	.0704	.0359	.7515	.9903

N = 16 116, *n* of events = 1 761.
Abbreviations: B, regression coefficient; S.E., standard error; Sig., significance level; C.I., confidence interval; L.B., lower bound; U.B., upper bound.

The concordance index of this model was .54 (standard error = .008), which was acceptable. The model's predictors explained only 0.3% of the variance of the hazard of the third birth (R^2 = .003). However, we can reject the null hypothesis that survival is just the function of time, based on the highly significant results of likelihood ratio, Wald, and score (logrank) test ($p \leq$ 4e-08). This means that the model's predictors do significantly predict the survival rate of the third birth, i.e. the model is acceptable.

Model evaluation

According to the test of the Schoenfeld residuals (Table 4), due to the absent of significant values ($p > .05$), the hypothesis of dependence of residuals on time, both for covariates and globally, can be rejected. The proportional hazards can hence be assumed for this multivariate Cox regression model.

Table 4. Shoenfield residuals

Predictor	**φ**	**χ^2**	**Sig.**
Age at first birth	.05	3.70	.055
Time from 1st to 2nd birth	-.01	.19	.661
Boy, boy	-.02	.67	.413
Boy, girl	-.01	.31	.578
Girl, boy	-.02	.83	.363
Global	NA	5.70	.336

Symbols: φ, rho coefficient; χ^2, chi-squared coefficient.

Figure 3 presents the deviance residuals for each of the *N* = 16 116 women, used for testing influential observations among the data. Deviance residuals are normalized transformations of the Martingale residuals. There were more negative than positive values, indicating that there were more women who did not have the third child, or had it later, compared to the women who had the third child earlier. There were not too many very large or small values, i.e. outliers, regarding the sample size. The assumption of the absence of outliers was not seriously violated.

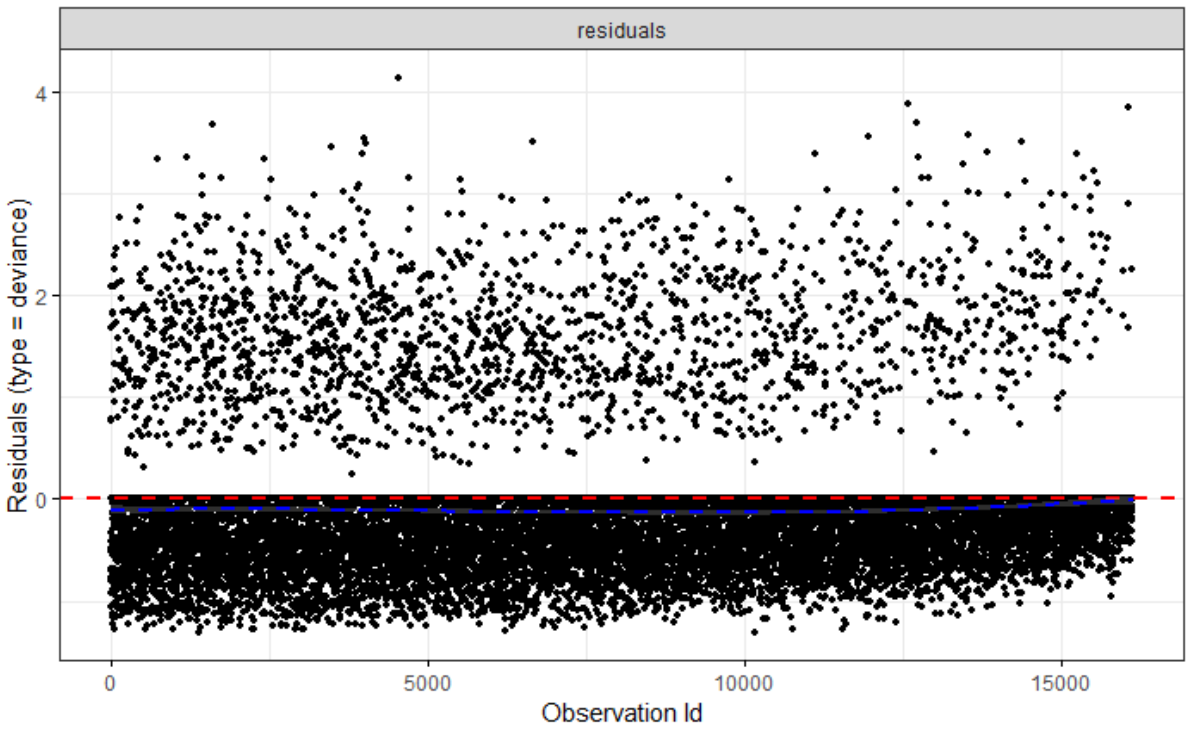

Figure 3. Deviance residuals for each observation

The linearity assumption was tested by plotting Martingale residuals for each covariate. According to the plots displayed in Figure 4, there seemed to be trends neither for *age* and *spacing* (time between the second and the third birth), nor for *sibs* (gender of the first two children). Regarding the continuous variables age and the time between second and third birth, there is possibly a cluster structure, due to some hidden factor. More possible is, however, that martingale residuals have negative values for longer surviving values (i.e. longer time to the third birth, or no third birth at all), and positive values for earlier third births.

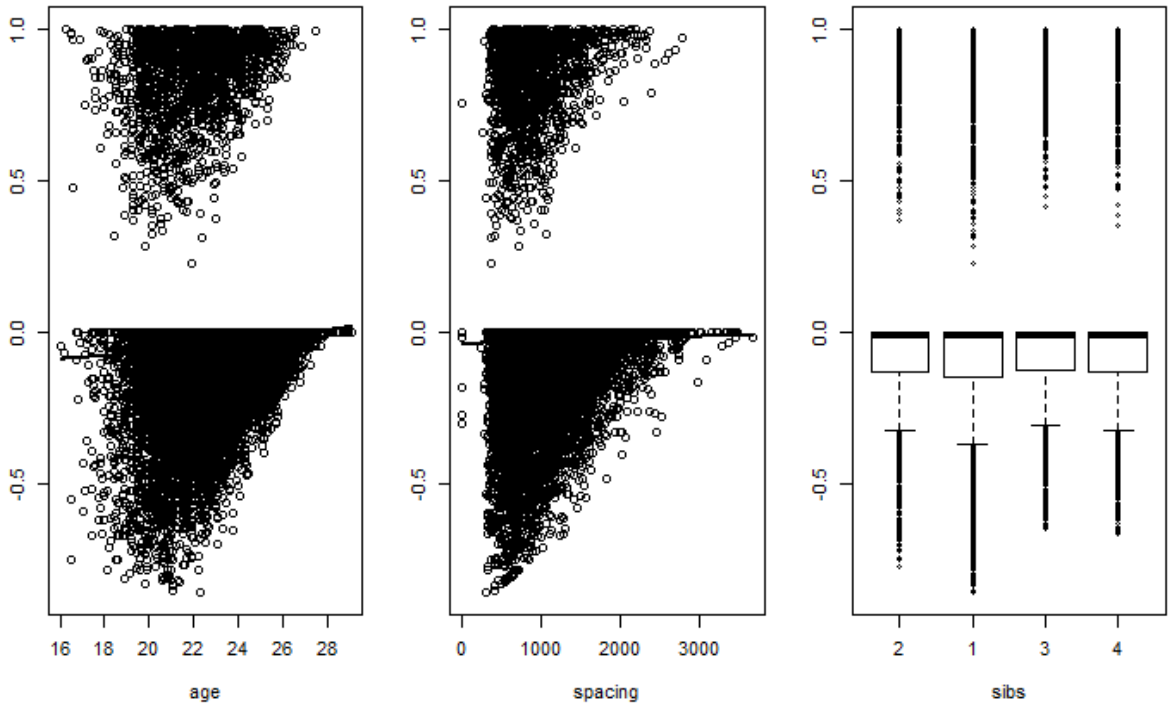

Figure 4. Martingale residuals for each covariate

This model can be evaluated as generally good, due to no serious violation of any of the assumptions for multivariate Cox regression.

Predicting top eighty women with the highest risk rates

Using the multivariate Cox regression model, including age at first birth, time from the first to the second birth, as well as gender of the first two children, the first eighty women with the highest likelihoods to give the third birth earliest (i.e. the highest risk) were identified on the randomly chosen unlabeled sample of 500 from the Medical Birth Registry of Norway data. The results for the first ten of them are presented in Table 3. Among this sample, the highest risk scores that a woman will give the third birth after the second one are, therefore, around 32.0%. The babies' pairs from the top eighty women with the highest risk scores were both boys in 95.0% of the cases (four women around the end of the list had the two-girl pairs).

Table 5. Top ten women with highest risk scores for giving the third birth

Rank	ID	Age at first birth	Time from 1st to 2nd birth (days)	Time from 1st to 2nd birth (months)	Gender	Risk score
1	308	23.3	381	12.49	Boy, boy	.328
2	130	24.0	415	13.61	Boy, boy	.323
3	386	19.8	357	11.70	Boy, boy	.317
4	412	28.0	515	16.89	Boy, boy	.316
5	370	19.2	361	11.84	Boy, boy	.313
6	298	22.4	429	14.07	Boy, boy	.311
7	205	26.6	507	16.62	Boy, boy	.311
8	93	27.1	526	17.25	Boy, boy	.308
9	37	20.7	411	13.48	Boy, boy	.307
10	405	26.4	527	17.28	Boy, boy	.305

N = 500.

Conclusions

Survival probability curves indicated that within seven years after the second birth, more than one in two women are not likely to have the third birth. Eight years after the second birth no woman should have a likelihood of having the third birth. Likelihood of not having the third birth within the measured time after the second birth is slightly lower in those who already have two boys, compared to having two girls, or children of both genders, which could be also confirmed by the predictions made on the smaller unlabeled sample of women, as almost all of the eighty of them had only boys.

Multivariate Cox regression model indicated that probability that a woman has the third birth at a specific time, within the measured time, is not affected by her age of having the first baby. The longer is the time from first to the second birth, probability of the third birth is somewhat lower at any specific time point. Additionally, those who have two children of different genders, compared to those who have two girls, are less likely to have the third birth is substantially lower at any time point. It is worth mentioning that although the model was evaluated as acceptable, particularly with no assumption seriously violated, the time between the first two births, and genders of the first two children explained very tiny amount of the probability of the third birth at a particular time. Therefore, the model should include more factors potentially influential to the likelihood of giving the third birth, which would be possible with a data set that contains more information about mothers.

Appendix

Code in R markdown

```
```{r}
install.packages(c('survival', 'survminer', 'dplyr', 'summarytools'))
```

```{r}
library('survival')
library('survminer')
library('dplyr')
library('summarytools')
```

```{r}
data <- read.delim('ThirdBirth.txt', sep = '')
data <- tibble::rowid_to_column(data, "ID")
data$spacing_months = data$spacing / 30.5
data$time_months = data$time / 30.5
data$sibs = factor(data$sibs)
data$sibs = relevel(data$sibs, ref = 2)
col_order <- c('ID', 'age', 'spacing', 'spacing_months', 'sibs', 'time',
'time_months', 'status')
data <- data[, col_order]
head(data)
```

```{r}
str(data)
```

```{r}
summary(data[c('age', 'spacing', 'spacing_months', 'time', 'time_months')])
```

Descriptive statistics.

```{r}
mean(data$age)
sd(data$age)
var(data$age)
```

```{r}
mean(data$spacing_months)
sd(data$spacing_months)
var(data$spacing_months)
```

```{r}
mean(data$time_months)
sd(data$time_months)
var(data$time_months)
```
```

```{r}
data_birth <- data[which(data$status == 1),]
str(data_birth)
```

```{r}
summary(data_birth$time_months)
```

```{r}
mean(data_birth$time_months)
sd(data_birth$time_months)
var(data_birth$time_months)
```

```{r}
freq(data$sibs)
```

```{r}
freq(data$status)
```

Kaplan-Meier and Log-rank for comparisons of survival probabilities.

```{r}
fit <- survfit(Surv(time_months, status) ~ 1, data = data)
print(fit)
```

```{r}
plot(fit, lwd = 1, xlab = 'Time in months', ylab = 'Survival probability')
```

```{r}
ggsurvplot(fit, data = data, color = '#2E9FDF',
           conf.int = TRUE,
           conf.int.style = "step",
           risk.table = "percentage",
           xlab = 'Time in months',
           break.time.by = 12,
           ggtheme = theme_minimal())
```

```{r}
fit_sibs <- survfit(Surv(time_months, status) ~ sibs, data = data)
print(fit_sibs)
```

```{r}
plot(fit_sibs, col = c('blue', 'deeppink', 'forestgreen', 'orange'), lwd =
2, xlab = 'Time in months', ylab = 'Survival probability')
legend('topright', lty = 1, col = c('blue', 'deeppink', 'forestgreen',
'orange'), lwd = 2, legend = c('Boy, boy', 'Girl, girl', 'Boy, girl', 'Girl,
boy'))
```

```{r}
ggsurvplot(fit_sibs,
          data = data,
          pval = TRUE,
```

```
          conf.int = TRUE,
          conf.int.style = "step",
          surv.plot.height = 1,
          risk.table = "percentage",
          risk.table.col = "strata",
          tables.height = 0.35,
          legend.labs=c('Boy, boy', 'Girl, girl', 'Boy, girl', 'Girl, boy'),
          linetype = "strata",
          surv.median.line = "hv",
          xlab = "Time in months",
          break.time.by = 12,
          ggtheme = theme_bw())
```

```{r}
surv_diff <- survdiff(Surv(time_months, status) ~ sibs, data = data)
surv_diff
```

Multivariate Cox regression model

```{r}
fit <- coxph(Surv(time, status) ~ age + spacing + sibs, data = data)
summary(fit)
```

Model evaluation.

Testing proportionality of hazards.

```{r}
test.ph <- cox.zph(fit)
test.ph
```

```{r}
ggcoxzph(test.ph)
```

Testing for influential observations.

```{r}
ggcoxdiagnostics(fit, type = 'deviance',
                 linear.predictions = FALSE, ggtheme = theme_bw())
```

```{r}
ggcoxdiagnostics(fit, type = 'dfbetas',
                 linear.predictions = FALSE, ggtheme = theme_bw())
```

Testing for linearity.

```{r}
data$residual <- residuals(fit, type = 'martingale')
```

```{r}
par(mfrow = c(1, 3), mar = c(4.2, 2, 2, 2))
```

```
with(data, {

  plot(age, residual)
  lines(lowess(age, residual), lwd = 2)

  plot(spacing, residual)
  lines(lowess(spacing, residual), lwd = 2)

  plot(residual ~ sibs, caption = 'Gender')

})
```


Data segmentation.

```{r}
new_data <- read.delim('ThirdBirthTest.txt', sep = '')
head(new_data)
```

```{r}
new_data <- tibble::rowid_to_column(new_data, "ID")
new_data <- select(new_data, -time, -status)
new_data$spacing_months = new_data$spacing / 30.5
col_order <- c('ID', 'age', 'spacing', 'spacing_months', 'sibs')
new_data <- new_data[, col_order]
head(new_data)
```

```{r}
new_data <- mutate(new_data, sibs = relevel(factor(sibs), ref = 2))
```

```{r}
new_data_segmented <-
  new_data %>%
  mutate(risk_score = predict(fit, newdata = new_data, type = "lp"))
head(new_data_segmented)
```

```{r}
new_data_segmented %>%
  arrange(desc(risk_score)) %>%
  top_n(80)
```