Survival Analysis

P. Schwarz

21/09/2021

Data

The surgery data was split between two files and contains the following columns

Variable name	Description
id	Patient ID
surgery_date	Date of surgery
event_date	Date of event
event	Yes=Death, No=Censored due to end of study
T	1=Surgery procedure A, 0= other surgery procedure
inflammation	Inflammation score; higher score is more inflammation
bmi	Body Mass Index (BMI)
age	Age, in years, at date of surgery
hospital	Hospital ID
volume	Hospital volume; mean number of patients undergoing surgery at the hospital annually
surgery_year	Year of surgery
prior_treatment	1=No prior treatment for the disease, 2=Prior treatment A, 3=Prior treatment A+B
srh	Self reported health status; higher score is better health
surgery_type	1= Surgery type A, 2=Surgery type B, 3=Surgery type C
severity	Severity of disease as classified by a physician; higher score is more severe
sex	Sex
technique	Surgery technique; Open or Keyhole
dead90	Yes=Dead within 90 days from date of surgery, No=Not dead within 90 days from date of surgery

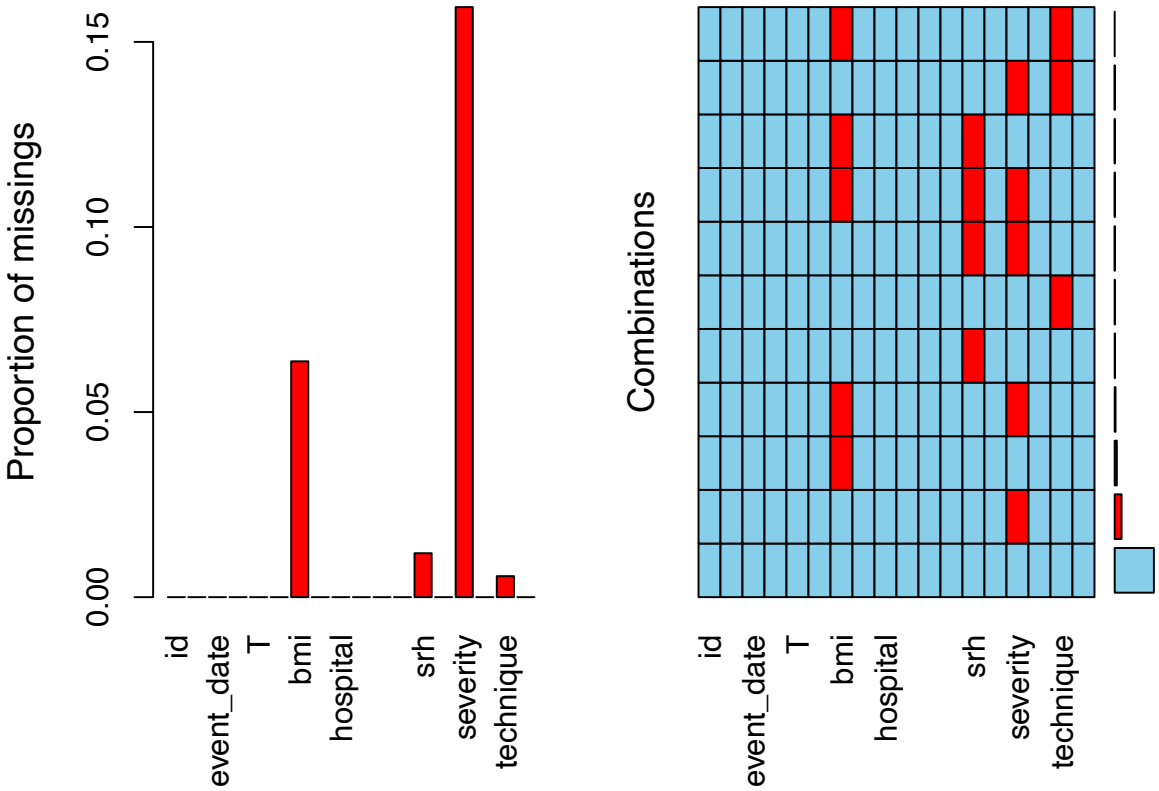
The goal is to conduct a survival analysis and a Cox proportional hazards regression based on the data sets.

Data Cleaning

Initially, the data had to be merged and cleaned. The two files provided were merged by their `id` column. The files both came with headers, but different character delimiters.

An analysis of missing data was conducted, as is shown in the plot above, in which `severity` is shown as the column with more than 15% missing data points. After investigating the variable `technique` the missing values were re-coded to show as `NA` and therefore missing as well.

After this procedure the missingness within the data set was assessed again.



The plot above shows now also the missing values in the `technique` column.

Further data cleaning was conducted by transforming `inflammation`, `prior_treatment`, `srh` and `severity` into ordered factors. `T`, `hospital` and `surgery_type` were transformed into regular factors. All of these were initially classified as a different type, which could cause problems in the analysis of categorical variables.

Additionally, the variables `surgery_date` and `event_date` were transformed to dates.

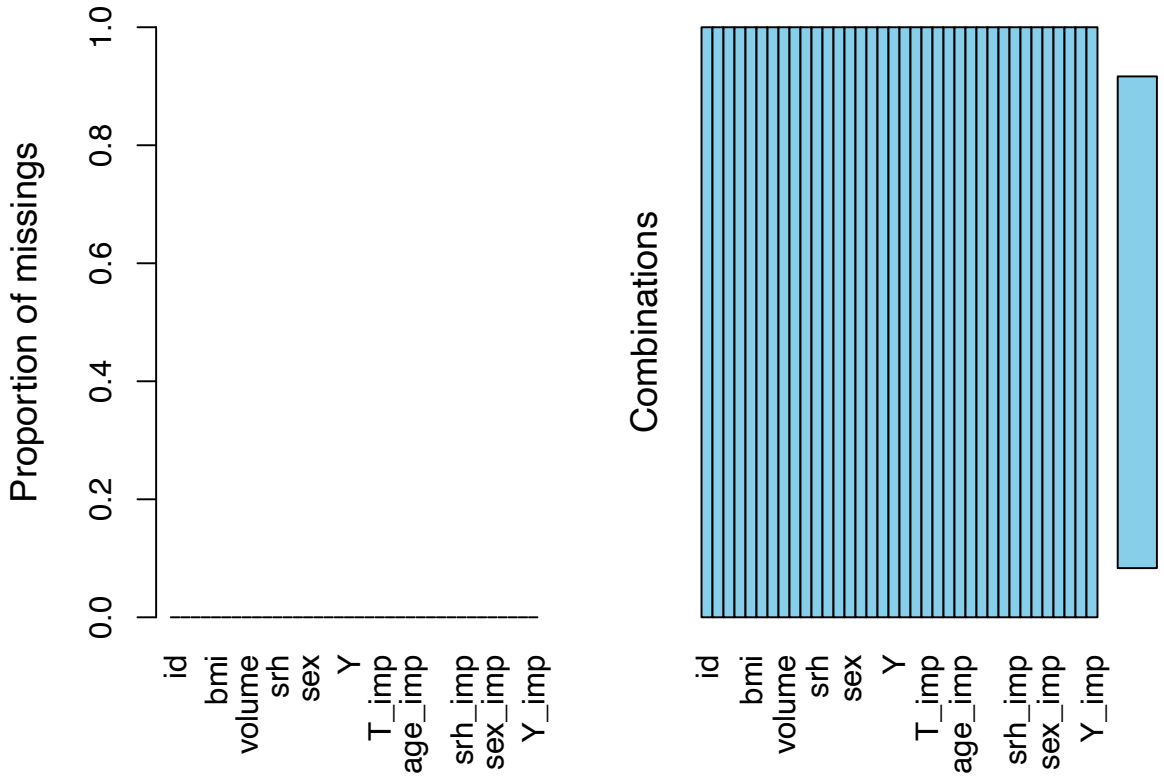
A new variable called `severity_ind` was created based on the value in the column `severity`. When `severity` is larger than `2` `severity_ind` takes value `1` and `0` if `severity` is lower.

Another new variable `Y` was added that contains the difference between `event_date` and `surgery_date` and will be used as the basis for the survival analysis later.

Lastly `surgery_date` and `event_date` were removed from the data set because the new variable made them obsolete.

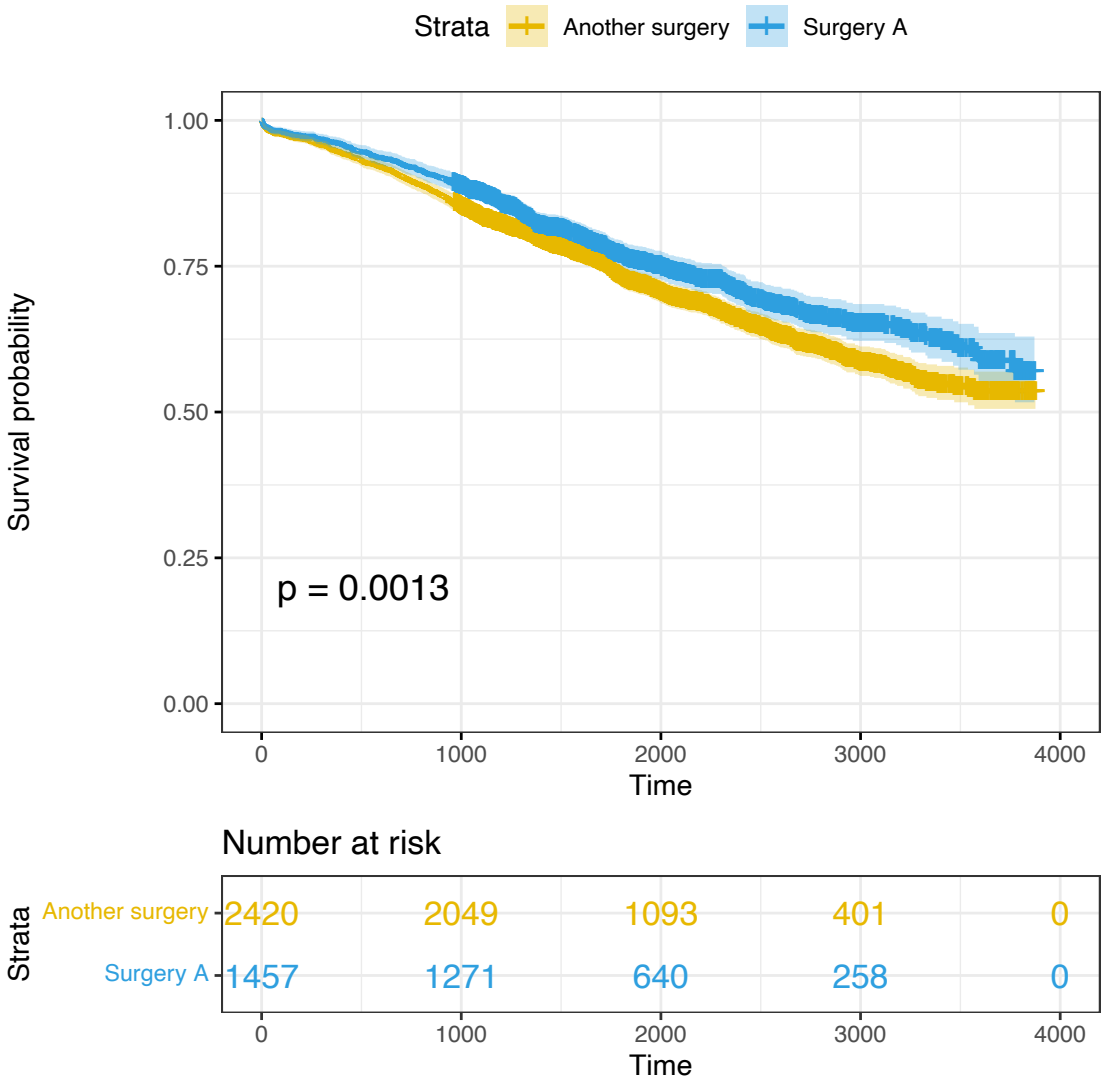
Dealing With Missing Data

To deal with the missing data a function was created that allows the deal with the missing data in 3 ways. The standard setting is to simply drop empty rows from the data set. Alternatively, hotdeck imputation or imputation based on the `mice` package can be conducted.



The missing data was imputed using hotdeck imputation. The graph above shows how the missing values in the data set are gone. The resulting data was used for a survival analysis based on the time between the surgery and the end of the observed time frame or the death of the patient. The strata are based on the two different treatments surgery A and another type of surgery. To conduct this procedure the time and the event variable were re-coded to be numeric values.

Survival Analysis



The result of the analysis is represented on the plot above.It can be seen that the comparison between the two surgery methods shows that surgery A results in a higher probability of patient survival in the time frame of the study. The difference between the two methods is statistically significant with a p value of 0.0013.

Cox proportional hazards regression

Based on the the survival model a Cox proportional hazards regression was performed using the following model specification:

$$h_i(t) = exp(\beta_1 T + \beta_2 inflammation + \beta_3 srh + \beta_4 surgery\_type + \beta_5 sex + \beta_6 bmi + \beta_7 age) h_0(t)$$

```
## Call:
## coxph(formula = Surv(as.numeric(survival_data$Y), as.numeric(survival_data$event)) ~
##     (survival_data$T + inflammation + srh + surgery_type + sex +
##         bmi + age), data = survival_data)
##
##   n= 3877, number of events= 1192
##
##                        coef exp(coef)  se(coef)      z Pr(>|z|)
## survival_data$T1 -0.052921  0.948455  0.062484 -0.847 0.397023
## inflammation.L    0.087694  1.091654  0.077394  1.133 0.257179
## inflammation.Q   -0.047195  0.953901  0.053220 -0.887 0.375188
## srh.L             0.544773  1.724217  0.071335  7.637 2.23e-14 ***
## srh.Q             0.178696  1.195658  0.050973  3.506 0.000455 ***
## surgery_type2     0.488102  1.629220  0.090895  5.370 7.88e-08 ***
## surgery_type3     0.568305  1.765272  0.086860  6.543 6.04e-11 ***
## sexMale          -0.208008  0.812201  0.060857 -3.418 0.000631 ***
## bmi              -0.027978  0.972409  0.007616 -3.674 0.000239 ***
## age               0.046912  1.048030  0.003459 13.562  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## survival_data$T1    0.9485     1.0543    0.8391    1.0720
## inflammation.L      1.0917     0.9160    0.9380    1.2705
## inflammation.Q      0.9539     1.0483    0.8594    1.0588
## srh.L               1.7242     0.5800    1.4992    1.9830
## srh.Q               1.1957     0.8364    1.0820    1.3213
## surgery_type2       1.6292     0.6138    1.3634    1.9469
## surgery_type3       1.7653     0.5665    1.4889    2.0929
## sexMale             0.8122     1.2312    0.7209    0.9151
## bmi                 0.9724     1.0284    0.9580    0.9870
## age                 1.0480     0.9542    1.0409    1.0552
##
## Concordance= 0.705  (se = 0.008 )
## Likelihood ratio test= 658.8  on 10 df,   p=<2e-16
## Wald test            = 654.6  on 10 df,   p=<2e-16
## Score (logrank) test = 712.5  on 10 df,   p=<2e-16
```

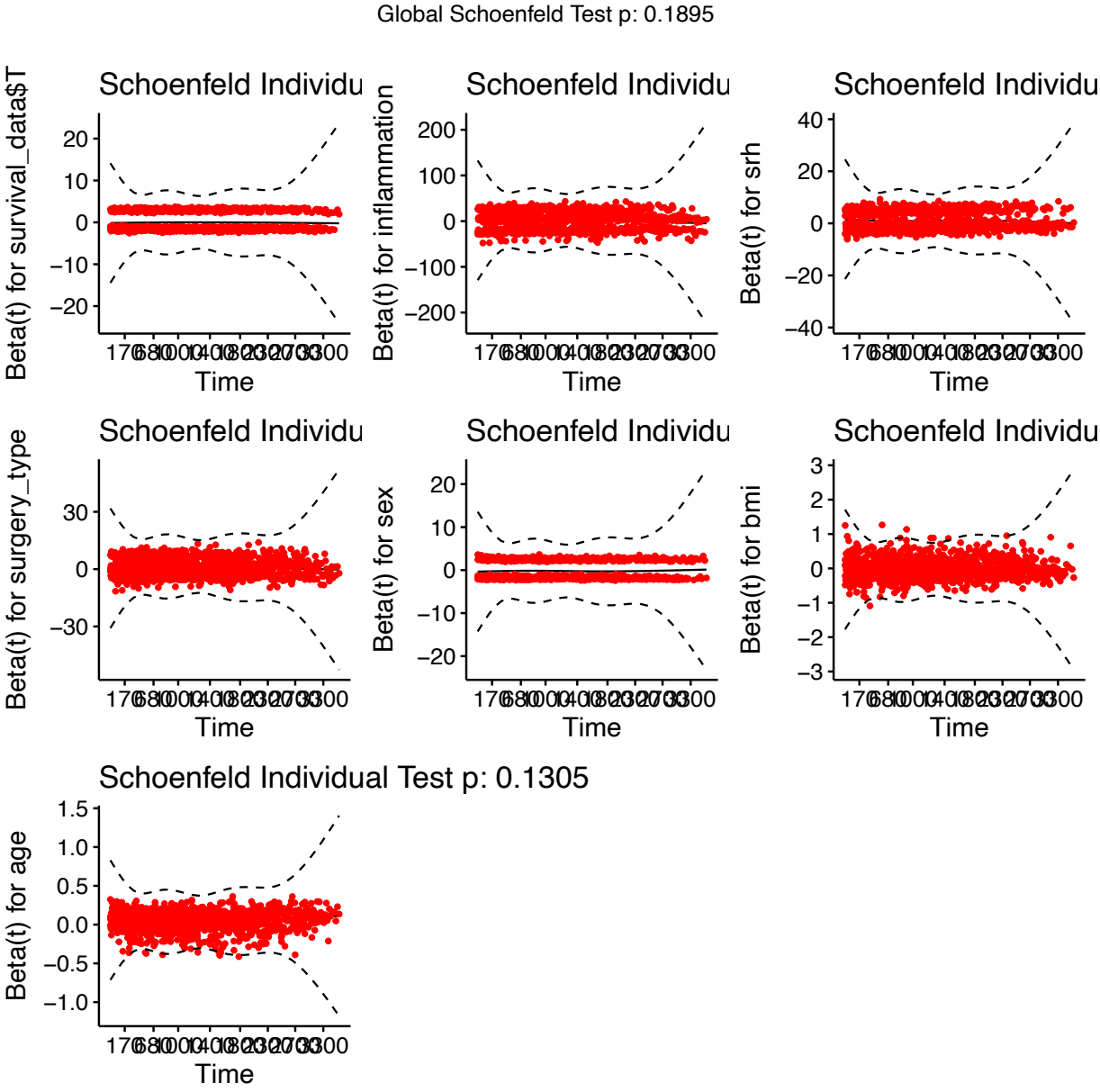
The summary of the survival model above provides an exponentiated coefficient for the treatment variable of 0.9471.

The 95% confidence interval for the treatment variable is [0.8379, 1.0705].

All of the covariates except the inflammation and the treatment are statistically significant with a p-value lower than 0.05.

```
##                   chisq df    p
## survival_data$T   0.052  1 0.82
## inflammation      2.905  2 0.23
## srh               1.310  2 0.52
## surgery_type      3.033  2 0.22
## sex               0.178  1 0.67
## bmi               0.992  1 0.32
## age               2.287  1 0.13
## GLOBAL           13.651 10 0.19
```

As the table above shows, the test for proportional hazards is not statistically significant for any of the covariates, nor the global test. Proportional hazards can be assumed.




```
ggcoxdiagnostics(cox,
 type = "schoenfeld"
```

```
)
```

```
## `geom_smooth()` using formula 'y ~ x'
```