survival_analysis

Vanessa Windausen

3/27/2020

Stages in biomedical research

early stage

Data mining or machine learning techniques can oftentimes be utilized at early stages of biomedical research to analyze large datasets, for example, to aid the identification of candidate genes or predictive disease biomarkers in high-throughput sequencing datasets

surival analysis

Here, we look mainly at a time to a specific event, such as death or disease recurrence. Time to event analysis is an alias of survival analysis. * Time from surgery to death * Time from start of treatment to progression * Time from response to recurrence * Time to machine malfunction

references

- survival analysis tutorial ** https://www.datacamp.com/community/tutorials/survival-analysis-R ** https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html
- creating survival plots: https://rpkgs.datanovia.com/survminer/survminer_cheatsheet.pdf

Typically asked questions

- do patients benefit from therapy regimen A as opposed to regimen B?
- Do patients age and fitness significantly influence the outcome?
- Is residual disease a prognostic biomarker in terms of survival?

important terms

- The term "censoring" refers to incomplete data. Censoring occurs if a subject has not experienced the event of interest by the end of data collection. Censored subjects still provide information and must be appropriately included in the analysis. A subject may be (right) censored due to ** loss of follow-up ** withdrawal fromm study ** no event by end of the fixed study period
- event: death or disease recurrence

important measures

Kaplan-Meier estimator

The statistic estimates the probability that a certain result does not occur. For example, it gives the probability that an individual patient will survive past a particular time t. At t = 0, the Kaplan-Meier estimator is 1 (100% probability to survive) and with t going to infinity, the estimator goes to 0 (0% probability to survive). In theory, with an infinitely large dataset and t measured to the second, the corresponding function of t versus survival probability is smooth. It is based on the assumption that the probability of surviving past a certain time point t is equal to the product of the observed survival rates until time point t.

Log rank test

The logrank test, or log-rank test, is a hypothesis test to compare the survival distributions of two samples or groups of treatments. It is a statistical hypothesis test that tests the null hypothesis that survival curves of two populations do not differ

Hazard function

Another useful function in the context of survival analyses is the hazard function h(t). It describes the probability of an event or its hazard h (i.e. survival). The most common use of the function is to model a patients chance of death as a function of their age.

survival analysis in R

```
## Warning: package 'survminer' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

The probability that a subject will survive beyond any given specified time is

$$S(t) = Pr(T > t) = 1 - F(t) \tag{1}$$

where $S(t)$ is the survival function and $F(t)$ the cumulative distribution function. In theory the function is smooth. The survival probability at a certain time t, is a conditional probability survivng beyond that time, given hat an individual has survived just prior that time. It can be estimated as the number of patients who are alive without loss to follow-up at time t, divided by the number of patients who were alive prior time t. At time t=t, the survival probability is $S(t_0) = 1$. The Kapkan-Meier estimate of survival probability is the product of these conditional probabilities up until that time t.

test data set: dates and time spans

In the test data *data_ex* it is shown how to convert dates and estimate survival time.

```
#create exemplary data set with dates only
date_ex <-
  tibble(
```

```
    sx_date = c("2007-06-22", "2004-02-13", "2010-10-27"),
    last_fup_date = c("2017-04-15", "2018-07-04", "2016-10-31")
    )
#formate dates
date_ex %>%
  mutate(
    sx_date = as.Date(sx_date, format = "%Y-%m-%d"),
    last_fup_date = as.Date(last_fup_date, format = "%Y-%m-%d")
    )
```

```
## # A tibble: 3 x 2
##   sx_date    last_fup_date
##   <date>     <date>
## 1 2007-06-22 2017-04-15
## 2 2004-02-13 2018-07-04
## 3 2010-10-27 2016-10-31
```

```
#calculate survival time
date_ex %>%
  mutate(os_yrs = as.numeric(difftime(last_fup_date,sx_date,units = "days")) / 365.25)
```

```
## # A tibble: 3 x 3
##   sx_date    last_fup_date os_yrs
##   <chr>      <chr>          <dbl>
## 1 2007-06-22 2017-04-15      9.82
## 2 2004-02-13 2018-07-04     14.4
## 3 2010-10-27 2016-10-31      6.01
```

R package data sets

The *ovarian* dataset comprises a cohort of ovarian cancer patients and respective clinical information, including the time patients were tracked until they either died or were lost to follow-up (futime), whether patients were censored or not (fustat, 0=right censored, 1= event observed at time t), patient age, treatment group assignment (rx), presence of residual disease (1=no, 2=yes) and performance status (1=good, 2=bad).

The *lung* dataset contains data of survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities. Status 1 indicates censored data and 2 that the patient died. The patients were subdivided in two groups (1=male, 2=female).

fit the Kaplan-Meier curves

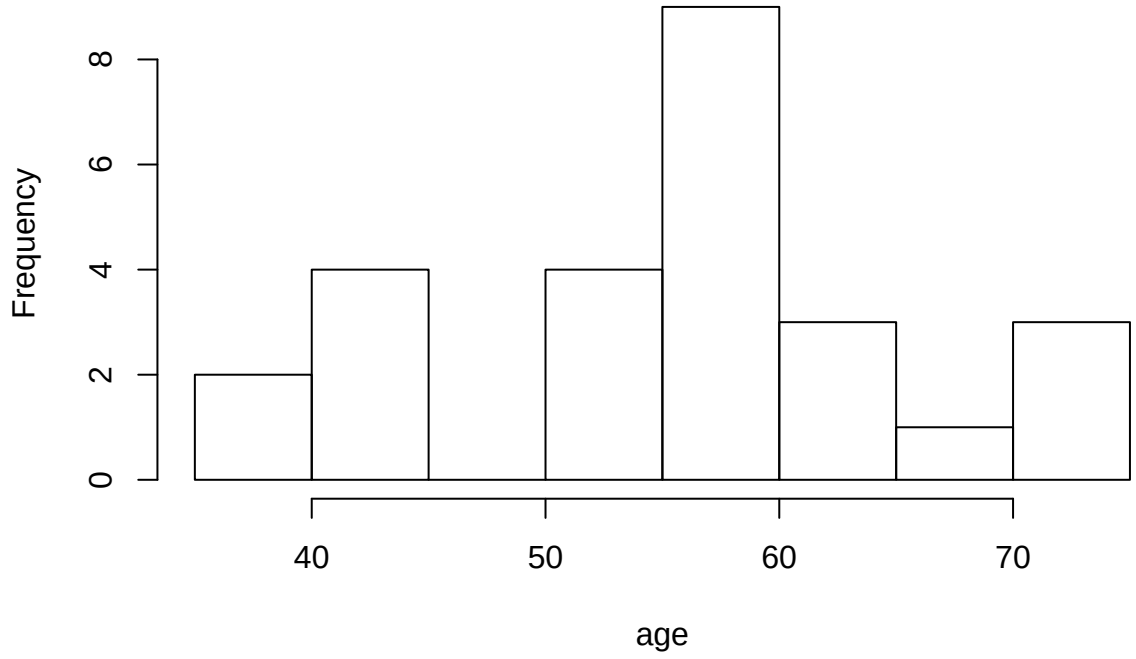
ovarian data set

```
data(ovarian)      #A + behind survival times indicates censored data points
head(ovarian)
```

```
##   futime fustat     age resid.ds rx ecog.ps
## 1     59      1 72.3315        2  1       1
## 2    115      1 74.4932        2  1       1
## 3    156      1 66.4658        2  1       2
## 4    421      0 53.3644        2  2       1
## 5    431      1 50.3397        2  1       1
## 6    448      0 56.4301        1  1       2
```

```r
# Dichotomize age and change data labels
ovarian$rx <- factor(ovarian$rx,
                     levels = c("1", "2"),
                     labels = c("A", "B"))
ovarian$resid.ds <- factor(ovarian$resid.ds,
                           levels = c("1", "2"),
                           labels = c("no", "yes"))
ovarian$ecog.ps <- factor(ovarian$ecog.ps,
                          levels = c("1", "2"),
                          labels = c("good", "bad"))

# Data seems to be bimodal
hist(ovarian$age,xlab="age",main="")
```



The obviously bi-modal distribution in the ovarian data set suggests a cutoff of 50 years. Use the mutate function to add an additional age_group column to the data frame. Convert the future covariates into factors.

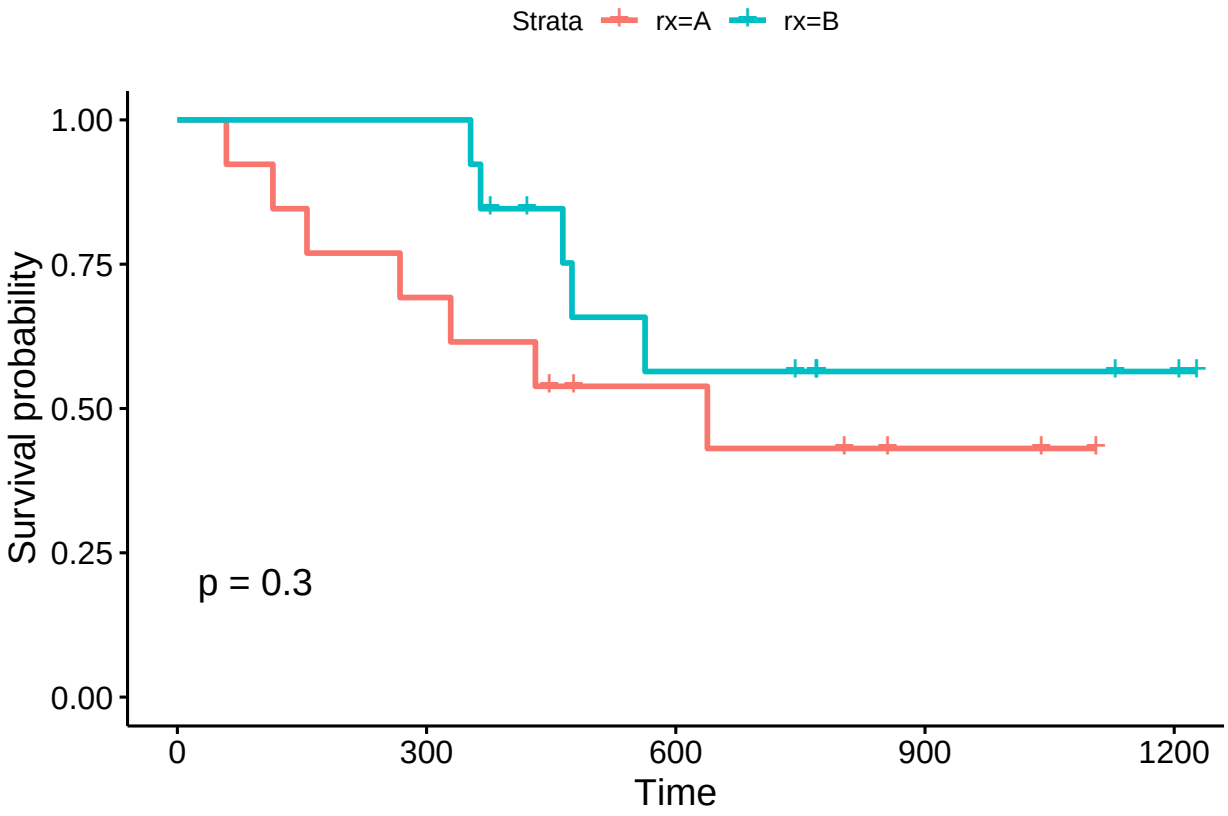
```r
ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)
#summary(ovarian)

#creating a survival object
surv_object_ovarian <- Surv(time = ovarian$futime, event = ovarian$fustat)
surv_object_ovarian #time until event ~ censored+ or not
```

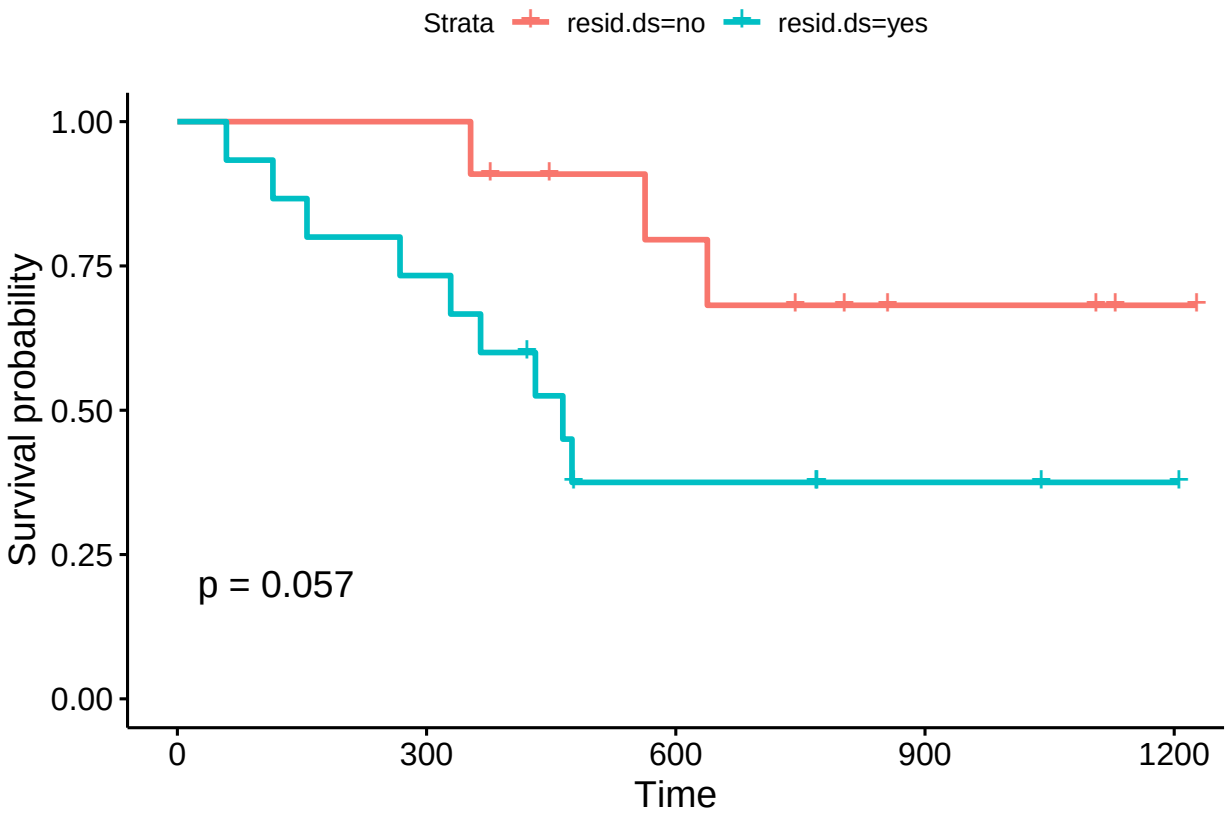
```
##  [1]   59   115   156   421+  431   448+  464   475   477+  563   638
```

```
## [12]  744+  769+  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+  268
## [23]  329   353   365   377+
```

```
#estimate survival curves with the Kaplan-Meier method
fit_ovarian <- survfit(surv_object_ovarian ~ rx, data = ovarian) #time to event ~ treatment
#summary(fit_ovarian)
#plot(fit_ovarian, mark.time=TRUE,xlab="Time", ylab = "Survival probability")
ggsurvplot(fit_ovarian, data = ovarian, pval = TRUE)
```



The log-rank p-value of 0.3 indicates a non-significant result. In this study, none of the treatments examined were significantly superior, although patients receiving treatment B are doing better in the first month of follow-up.



A follow-up study with an increased sample size could validate these results, that is, that patients with positive residual disease status have a significantly worse prognosis compared to patients without residual disease.

lung data set

```
#creating a survival object
surv_object_lung <- Surv(lung$time, lung$status)

#estimate survival curves with the Kaplan-Meier method
#fit_lung <- survfit(surv_object_lung ~ 1, data=lung)  #analysis across groups
fit_lung <- survfit(surv_object_lung ~ sex, data=lung) #analysis by group
fit_lung
```

```
## Call: survfit(formula = surv_object_lung ~ sex, data = lung)
##
##         n events median 0.95LCL 0.95UCL
## sex=1 138    112    270     212     310
## sex=2  90     53    426     348     550
```

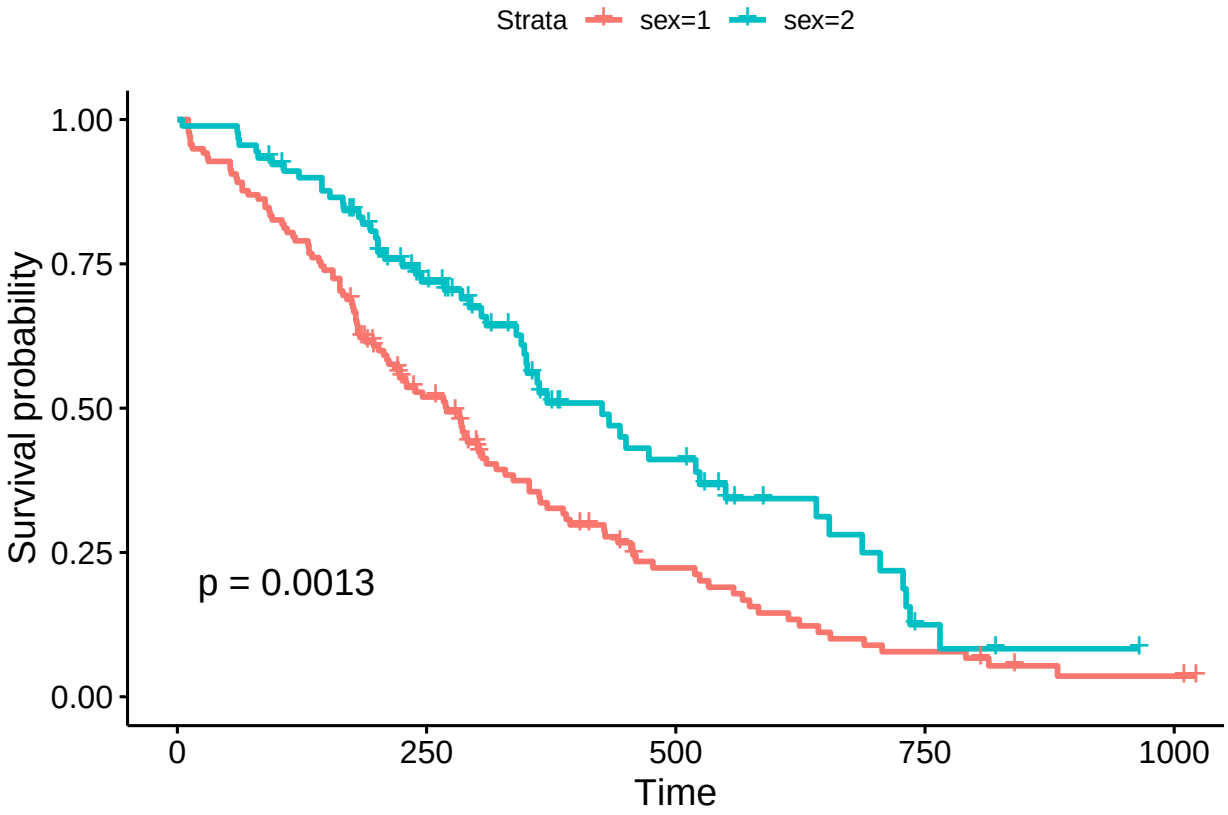
```
#p-value
sd_lung <- survdiff(Surv(time, status) ~ sex, data = lung)
1 - pchisq(sd_lung$chisq, length(sd_lung$n) - 1)
```

```
## [1] 0.001311165
```

```
#names(fit_lung)
#key components: time (start and endpoint of each interval; surv (survival probability at time t))
```

Survival times are not expected to be normally distributed so the mean is not an appropriate summary. Median survival is the time corresponding to a survival probability of 0.5. The median survival time across groups is 310 days. The median survival time off male patients is 270 days, that of female patients 426 days. The survival of female patients is significantly better than that of male patients (p-value=0.0013).

```
#plot(fit_lung, xlab = "Days", ylab = "Overall survival probability", mark.time=TRUE)
ggsurvplot(fit_lung, pval = TRUE)
```




```
  #solid line: step function
   ##horizontal line: survival duration for the interval
   ##an interval is terminated by an event
   ##the height of vertical lines show the change in cumulative probability
   ##Censored observations, indicated by tick marks, reduce the cumulative survival between intervals.
  #dotted line: confidence interval
```

The log-rank test equally weights observations over the entire follow-up time and is the most common way to compare survival times between groups. Here, there is a significant difference in survival time between male (2) and female (1) patients.

```
summary(fit_lung, times = 365.25)$surv
```

```
## [1] 0.3360878 0.5264630
```

Ignoring censoring leads to an overestimate of the overall survival probability. This is mainly, because the censored subjects only contribute information for part of the follow-up time, and then fall out of the risk set, thus pulling down the cumulative probability of survival. In the lung study 121 out of 228 patients died by 1 year so the naive estimate of survival probability after 1 year is (1-121/228)*100=47%. This is an incorrect estimate of the 1-year probability of survival when you ignore the fact that 42 patients were sensored before 1

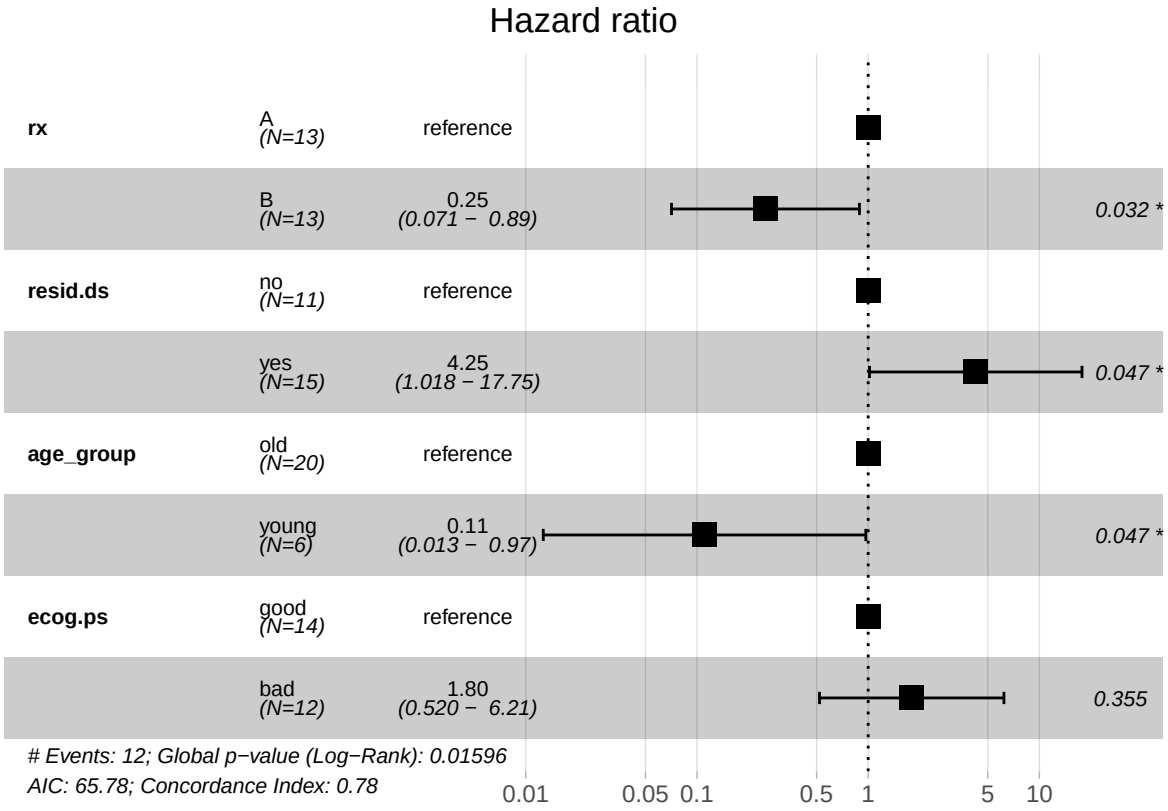
year. The correct estimate of the 11-year probability across groups is 40.9%. The survival probabillity of male patients (1) is 33.61% and that of female patients 52.65%.

Cox proportional hazards models

Cox proportional hazards models allow you to include covariates. Forest plot show so-called hazard ratios (HR) which are derived from the model for all covariates that we included in the formula.An HR > 1 indicates an increased risk of death.

```
fit.coxph <- coxph(surv_object_ovarian ~ rx + resid.ds + age_group + ecog.ps,
                   data = ovarian)
ggforest(fit.coxph, data = ovarian)
```

```
## Warning: Removed 4 rows containing missing values (geom_errorbar).
```



A hazard ratio of 0.25 for treatment groups tells you that patients who received treatment B have a reduced risk of dying compared to patients who received treatment A (which served as a reference to calculate the hazard ratio). As shown by the forest plot, the respective 95% confidence interval is 0.071 - 0.89 and this result is significant.

to do

look at https://rpkgs.datanovia.com/survminer/survminer_cheatsheet.pdf logistic regression repeat course from conference publish on github