Business Intelligence I
Supplement
Time-to-Event Analysis

W. Grossmann

Content

- Problem formulation
- Terminology
- Analysis Template
- Estimation of the survival function
- Cox Regression

Problem Formulation

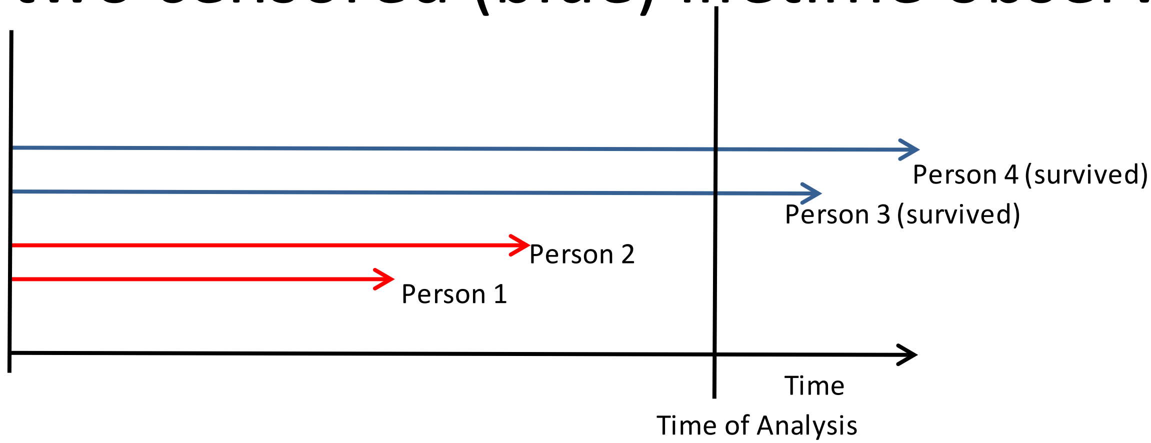
- In Time-to-Event Analysis we are interested in modeling and predicting the time up to a certain event
- Examples:
  - Prediction of the duration until a customer will quit her/his relationship with a company
  - Prediction of the duration of the lifetime of a certain device

Problem Formulation and Terminology

- Other notions for such problems:
  - Event History Analysis
  - Survival Analysis
- The time up to the event is called life time
- Main characteristic of the available data:
  - The data about the lifetime are ***censored***, i.e. for some customers the event is observed, for others the event will occur in the future
- This type of censoring is called ***right censored***

Problem Formulation

- Graphical representation for two complete (red) and two censored (blue) lifetime observations



- Besides the censored lifetime usually other information about the customers is known, e.g. age, occupation, type of machine, ….

Terminology

- The time up to the event is denoted by T and is a random variable
- The probability that the event occurs before time t is denoted by

$$F(t) = P(T \le t)$$

- The survival function is the probability that the event occurs after time t

$$S(t) = 1 - F(t)$$

Terminology

- The mean of the survival function is called the expected survival time
- The hazard function gives the likelihood that the event occurs at time t, given that the event has not occurred up to time t
- Formula

$$h(t) = F'(t) / (1 - F(t)) = \frac{f(t)}{1 - F(t)}$$

Analysis Template

Template: Time to Event Analysis

- **Relevant Business and Data:** Customer behavior represented by cross-sectional data and time sequences containing censored information about a terminal event.
- **Analytical Goals:** Predict from the uncensored data the duration up to the event for the censored time sequences
- **Modeling Tasks:**
  - Definition of a a survival table
  - Definition of a Cox regression model for the time to event
- **Analysis Tasks:**
  - Estimate the time to event using the Kaplan Meier estimate
  - Estimation of the coefficients in the Cox regression model
- **Evaluation and Reporting Task:** Evaluate the results using method for evaluation of regression

Modeling the survival function

- A frequently used class of model in time-to-event analysis are Weibull distributions defined

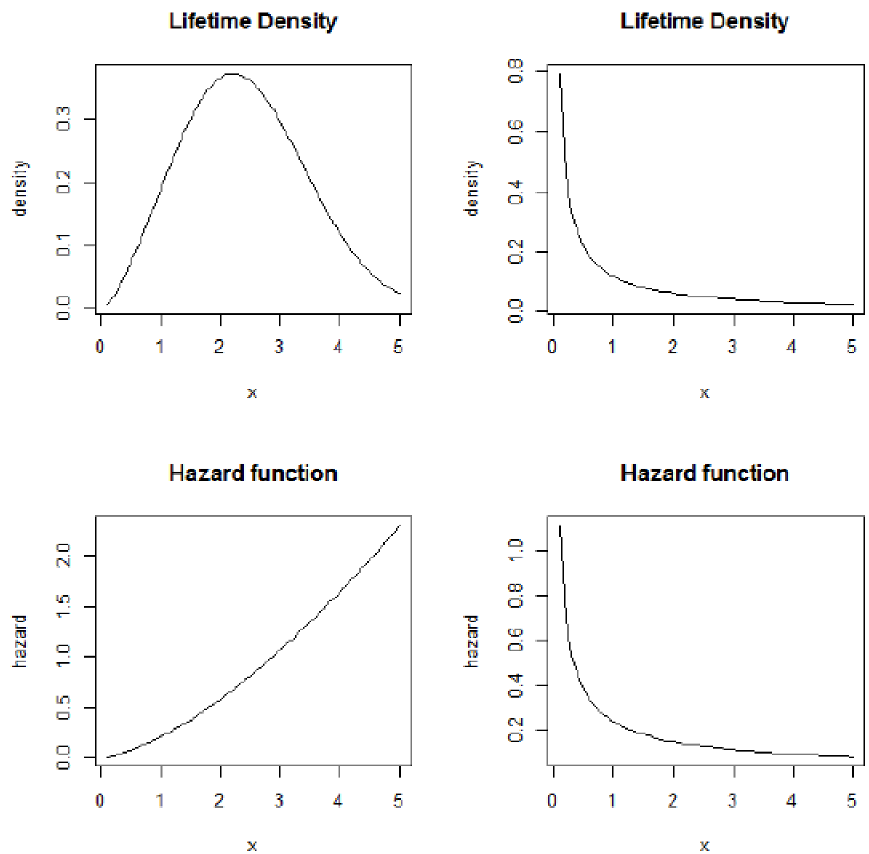
$$F(t) = 1 - \exp\left[-(\alpha t)^{\beta}\right]$$

$$f(t) = \beta \cdot (\alpha t)^{\beta-1} \alpha \cdot \exp\left[-(\alpha t)^{\beta}\right]$$

which allows adaptation to different scenarios like increasing hazard or decreasing hazard by choosing appropriate parameters

Modeling the survival function

- Examples of survival functions



Estimation of the survival function

- The basic information about the survival function is given by the Kaplan Meier estimate, which is summarized in the survival table with the following columns:
  - Time interval
  - Number of persons entering the interval (*n.risk*)
  - Number of events occurred in the interval (*n.event*)
  - Value of the survival function at the end of the time interval (*survival*)

Estimation of the survival function

- The standard error of the estimate for the survival function
- Confidence interval for the survival function

- Example of a survival table:

  305 patients with different types of melanoma observed from 2006 - 2010

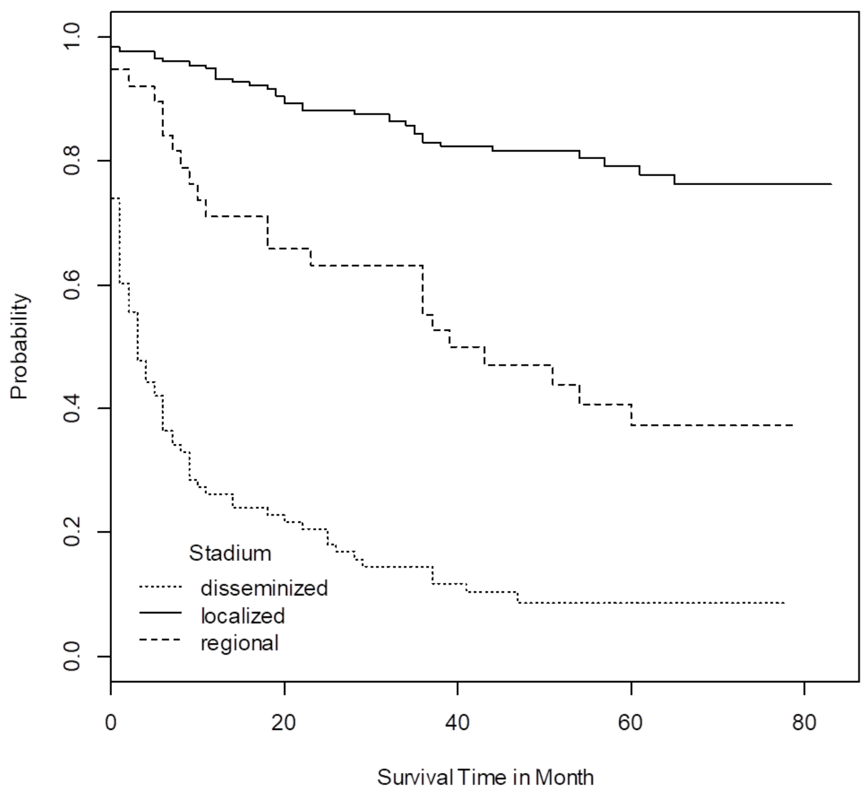
Estimation of the survival function

Survival table

Year	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0	305	69	0.774	0.0240	0.728	0.822
1	236	23	0.698	0.0263	0.649	0.752
2	213	19	0.636	0.0275	0.584	0.692
3	174	16	0.578	0.0286	0.524	0.637
4	136	6	0.552	0.0292	0.498	0.612
5	86	4	0.526	0.0305	0.470	0.590

Estimation of the survival function

- Plot of survival function for the three groups



Cox Regression

- If there are additional explanatory variables for the occurrence of the event one can estimate the hazard rate with ***Cox regression***, also known as ***proportional hazard model***
- The model defines a time dependent baseline hazard for all observations which is modified according to the explanatory variables

Estimation of the survival function

- Formula

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

- Interpretation of the parameters:
  - For a quantitative explanatory variable x the relative risk changes by $\exp(\beta)$ if x is increased by one unit

Cox Regression

- For a dummy variable representing a factor level the relative risk changes by $\exp(\beta)$ compared to a reference level

- Example
  - For the 305 patients the influence of the explanatory variables age at diagnosis and stadium of the tumor is of interest
  - The results are shown on the next slide

Cox Regression

```
coxph(formula = Surv(Time, Event) ~ Age_Diagnosis + Stadium,
    data = vie1)

  n= 305, number of events= 137

                     coef exp(coef) se(coef)       z Pr(>|z|)
Age_Diagnosis     0.02991   1.03036  0.00653   4.580 4.64e-06 ***
Stadiumlocalized -2.64494   0.07101  0.21324 -12.404  < 2e-16 ***
Stadiumregional  -1.41158   0.24376  0.24521  -5.756 8.59e-09 ***
Stadiumunknown         NA        NA  0.00000      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
Age_Diagnosis      1.03036     0.9705   1.01726    1.0436
Stadiumlocalized   0.07101    14.0826   0.04675    0.1079
Stadiumunknown          NA         NA        NA        NA

Concordance= 0.835  (se = 0.027 )
Rsquare= 0.456   (max possible= 0.992 )
Likelihood ratio test= 185.5  on 3 df,   p=0
Wald test            = 166.9  on 3 df,   p=0
Score (logrank) test = 236.6  on 3 df,   p=0
```

References

G. Broström: Event History Analysis with R. CRC Press Taylor & Francis Group 2012

R package: `survival`