Biostatistics & Epidemiological Data Analysis using R

13

Survival analysis

Stefan Konigorski

Health Intervention Analytics Group, HPI

February 10, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2021.10.28
	2	First steps in data analysis using R	2021.11.04
	3	Second steps in data analysis using R	2021.11.11
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2021.11.18
	5	Estimation	2021.11.25
	6	Hypothesis testing & study planning	2021.12.02
	7	Missing data	2021.12.09
Data analysis w/ regression models	8	Linear regression I	2021.12.16
	9	Linear regression II	2022.01.13
	10	Regression models for binary and count data	2022.01.20
	11	Analysis of variance & Linear mixed models I	2022.01.27
	12	Linear mixed models II & Meta analysis	2022.02.03
	13	Survival analysis	2022.02.10
	14	Causal inference & Data analysis challenge	2022.02.17

(see full schedule online)

Review of last week

- Linear mixed models for hierarchical data.
- Application of mixed models to meta-analysis.

1. Survival analysis
   - Introduction and basic concepts
   - Estimation and inference
2. Data analysis

Survival analysis

Survival analysis

Alternative terminology

- Failure-time analysis
- Time-to-event analysis

Survival analysis

Alternative terminology

- Failure-time analysis
- Time-to-event analysis

Aim

- Describe and model a binary outcome (for example: death vs. alive) over time.

Survival analysis

Questions

- What is the probability of surviving 10 years after being diagnosed with stage 2 lung cancer?
- What is the mean time to relapse after doing a 6-week treatment in a rehab clinic for alcohol addiction?
- Which variables are associated with the risk of developing diabetes?

Survival analysis

Questions

- What is the probability of surviving 10 years after being diagnosed with stage 2 lung cancer?
- What is the mean time to relapse after doing a 6-week treatment in a rehab clinic for alcohol addiction?
- Which variables are associated with the risk of developing diabetes?

Statistical models

- Which statistical models that we have learned so far can we use to investigate such questions?
- What are their limitations?

Review: Epidemiological cohort study

- Longitudinal observational study (i.e. study over time)
- Define a group of people that are free of the disease, assess all exposures of interest at baseline, then observe over time who develops the disease.
- I.e. assess exposures first, observe disease later.
- Can calculate incidence rates.
- Important: response rate, drop-out rate.

Review: Epidemiological cohort study

- Longitudinal observational study (i.e. study over time)
- Define a group of people that are free of the disease, assess all exposures of interest at baseline, then observe over time who develops the disease.
- I.e. assess exposures first, observe disease later.
- Can calculate incidence rates.
- Important: response rate, drop-out rate.

$\longrightarrow$ data usually contains censoring

Censoring

- In cohort studies, in general, the event (e.g. death) is only observed in some people.
- In all other people, we only know that the time until the event occured was longer than the time they were observed
  $\longrightarrow$ (right) censoring.
- Reason for censoring: people are only observed for a restricted time (potential other reason: drop-out etc.).

Censoring

- In cohort studies, in general, the event (e.g. death) is only observed in some people.
- In all other people, we only know that the time until the event occured was longer than the time they were observed
  $\longrightarrow$ (right) censoring.
- Reason for censoring: people are only observed for a restricted time (potential other reason: drop-out etc.).
- Censoring more generally: some lifetimes are known to have occurred only within certain intervals.

Censoring

- In cohort studies, in general, the event (e.g. death) is only observed in some people.
- In all other people, we only know that the time until the event occured was longer than the time they were observed
  $\longrightarrow$ (right) censoring.
- Reason for censoring: people are only observed for a restricted time (potential other reason: drop-out etc.).
- Censoring more generally: some lifetimes are known to have occurred only within certain intervals.
- Variable that denotes whether a person was censored or not is often denoted by C (or δ).

Censoring

- In cohort studies, in general, the event (e.g. death) is only observed in some people.
- In all other people, we only know that the time until the event occured was longer than the time they were observed
  $\longrightarrow$ (right) censoring.
- Reason for censoring: people are only observed for a restricted time (potential other reason: drop-out etc.).
- Censoring more generally: some lifetimes are known to have occurred only within certain intervals.
- Variable that denotes whether a person was censored or not is often denoted by C (or δ).

$\longrightarrow$ How can you integrate that information into the statistical model?

Survival function

- Let's denote the time-to-event random variable by T.
- Example: T = time until death.
- T_i = time from study start until the i-th person in the study died.

Survival function

- Let's denote the time-to-event random variable by T.
- Example: T = time until death.
- T_i = time from study start until the i-th person in the study died.

- Cumulative distribution function F of t: probability that T is smaller or equal to the value t: $F(t) = P(T \leq t)$.

Survival function

- Let's denote the time-to-event random variable by T.
- Example: T = time until death.
- T_i = time from study start until the i-th person in the study died.

- Cumulative distribution function F of t: probability that T is smaller or equal to the value t: $F(t) = P(T \leq t)$.
- Interpretation: $F(t)$ = probability of dying earlier or at time t.

Survival function

- Let's denote the time-to-event random variable by T.
- Example: T = time until death.
- T_i = time from study start until the i-th person in the study died.

- Cumulative distribution function F of t: probability that T is smaller or equal to the value t: $F(t) = P(T \leq t)$.
- Interpretation: $F(t)$ = probability of dying earlier or at time t.

- (Cumulative) survival function $S(t) = 1 - F(t)$.
- $S(t)$ = probability of surviving longer than time t.

Estimation and inference in survival analysis

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?
- Just count how many had the event, compute the relative frequency!

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?
- Just count how many had the event, compute the relative frequency!

- How can we estimate the mean/median survival time?

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?
- Just count how many had the event, compute the relative frequency!

- How can we estimate the mean/median survival time?
- Just compute the mean/median survival time!

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?
- Just count how many had the event, compute the relative frequency!

- How can we estimate the mean/median survival time?
- Just compute the mean/median survival time!

- How can we estimate the distribution/survival function?

Estimation of survival times

Let's look at time-to-death and assume that all study participants have been observed completely (i.e. all have died).

- How can we estimate the probability of ever having an event?
- Just count how many had the event, compute the relative frequency!

- How can we estimate the mean/median survival time?
- Just compute the mean/median survival time!

- How can we estimate the distribution/survival function?
- At each time point, compute the probability of having/not having an event and sum them up!

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

- How can we estimate the mean/median survival time?

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

- How can we estimate the mean/median survival time?
- Trickier.

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

- How can we estimate the mean/median survival time?
- Trickier.

- How can we estimate the distribution/survival function?

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

- How can we estimate the mean/median survival time?
- Trickier.

- How can we estimate the distribution/survival function?
- Same.

Estimation of survival times

Let's look at time-to-relapse and assume that all study participants have been observed completely (i.e. all that will have the event, had the event).

- How can we estimate the probability of ever having an event?
- Same.

- How can we estimate the mean/median survival time?
- Trickier.

- How can we estimate the distribution/survival function?
- Same.

Estimation of the survival function with censoring? See next slide.

Estimation of the survival function

The Kaplan Meier estimate of the survival function takes the following approach:

- The estimate of the survival function at time t is
- the product of the probabilities of surviving each time point up to and including t,
- which are computed based on the number of events and number of people at risk:

Estimation of the survival function

The Kaplan Meier estimate of the survival function takes the following approach:

- The estimate of the survival function at time t is
- the product of the probabilities of surviving each time point up to and including t,
- which are computed based on the number of events and number of people at risk:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{y_i}\right) & \text{if } t \geq t_1 \end{cases}$$

- where t_1 is the first time point in the study when someone had the event,
- d_i is the number of events at time t_i and
- y_i are the number of people at risk at time t_i.

Construction of the Kaplan Meier curve

Example how the Kaplan Meier estimate of the survival function can be constructed for a dataset. For that:

- first count how many people had an event and when,
- make a table with one row for each time someone had an event
- (and with one row for each time someone was censored)
- and fill it out as follows to get $\hat{S}(t)$:

Construction of the Kaplan Meier curve

day t	at risk	died at day t	Prob to survive t	$\hat{S}(t)$
10	10	1	$9/10$	$9/10$
12	9	2	$7/9$	$9/10 \cdot 7/9 = 7/10$
50	7	1	$6/7$	$7/10 \cdot 6/7 = 6/10$
70	6	0	1	$6/10$
90	5	0	1	$6/10$
100	3	1	$2/3$	$6/10 \cdot 2/3 = 4/10$
300	2	1	$1/2$	$4/10 \cdot 1/2 = 2/10$

where

- 'at risk' is the number of people still in the study at the beginning of day t,
- 'died at day t' is the number of people that have died at day t
- the probability of surviving day t is a conditional probability since it is conditional on having survived up to day t.

Kaplan Meier curve in R

- Use the function `survfit` in the `survival` package: `survfit(Surv(T, C) ~ 1, data = dat)` where
- `T` is the follow-up time, and `C` is the censoring indicator which is 0 if the person was alive and 1 if the person had an event.
- The curve can be plotted by using the `plot` function.
- From the plot, e.g. the median survival time can be seen (value of t where $\hat{S}(t) = 50\%$).
- See `R_13_survival.Rmd`.

Testing differences in survival times

- Whether the survival is different between two groups can be tested with the log-rank test.
- The log-rank tests is based on the Kaplan-Meier curve and compares the observed number of events at each time point in each group with the expected number of events in each group, assuming that there is no difference.
- Then, a χ^2 test statistic can be computed (which tests the null hypothesis that there are no differences in the hazard functions of the two groups).
- In R, this can be done using the `survdiff` function in the `survival` package: `survdiff(Surv(T, C) ~ sex, data = dat)`

Testing effect of predictors in survival times

- Build regression-type models, by predicting the survival times or a function of them.
- Most often used model: Cox proportional hazards regression model.

Cox proportional hazards model

Predict the hazard function of the survival times by the predictors:

Cox proportional hazards model

Predict the hazard function of the survival times by the predictors:

- $log(h(t)) = log(h_0(t)) + \sum_{j=1}^{k} \beta_j x_j$ where
- $h(t)$ is the hazard function, which describes the "immediate/instantaneous" risk of dying at time point t, given that you are alive at time point t.
- $h_0(t)$ is some baseline risk, that is left unspecified.

Cox proportional hazards model

Predict the hazard function of the survival times by the predictors:

- $log(h(t)) = log(h_0(t)) + \sum_{j=1}^{k} \beta_j x_j$ where
- $h(t)$ is the hazard function, which describes the "immediate/instantaneous" risk of dying at time point t, given that you are alive at time point t.
- $h_0(t)$ is some baseline risk, that is left unspecified.
- i.e. the model is semiparametric, since the form of $h(t)$ is not assumed to follow a specific form, only the effect of the predictors is assumed to be linear.
- Proportional hazards: the model contains the assumption that the hazard rates of two individuals are constant, and independent of time (c.f. odds ratio is constant).
- In R: e.g. use the `coxph` function in the `survival` package: `coxph(Surv(T, C) ~ X, data = dat)`

Further topics in survival analysis

- Parametric survival models: assume that the survival function $S(t)$ i.e. the time T follows a specific distribution, similar to GLMs. This can be done by using the `survreg` function in the `survival` package.

Further topics in survival analysis

- Parametric survival models: assume that the survival function $S(t)$ i.e. the time T follows a specific distribution, similar to GLMs. This can be done by using the `survreg` function in the `survival` package.
- Time-varying (dependent) covariates: The effect of covariates that vary over time on the survival time can also be considered in survival models, e.g. in the `coxph` and `survreg` functions.

Further topics in survival analysis

- Parametric survival models: assume that the survival function $S(t)$ i.e. the time T follows a specific distribution, similar to GLMs. This can be done by using the `survreg` function in the `survival` package.
- Time-varying (dependent) covariates: The effect of covariates that vary over time on the survival time can also be considered in survival models, e.g. in the `coxph` and `survreg` functions.
- Recurrent events/competing risks: some data have the challenge that individuals can have multiple events/events of multiple types, that are "competing". E.g. people can have a heart attack or die, but they cannot have a heart attack if they have died already. This requires extra modeling.

Questions?

References

- Klein & Moeschberger (2003). Survival analysis. Techniques for censored and truncated data. Springer.

Data analysis

- Get together in groups of 3-5 with at least one German speaking person (for reading of the KiGGS data dictionary).
- Investigate variables that are associated with needing a reading help (`e025a`).
- Choose 3-5 variables that you think are relevant.
- Think of a suitable statistical model how to investigate this (you could also include the age when the child got their reading help into the analysis: `e025az1`).
- Format the variables.
- Do the analysis.
- Check the assumptions of the analysis.
- Interpret the results.
- See `R_13_data_analysis.Rmd`.

Homework

Homework

See file `R_13_homework.Rmd`.