

Exercise, Tuesday 13, Survival and event history analysis

Ram/Ayushi

Data

The analysis is based on the Nepal study data presented in the lecture. We look at the time to the next occurrence of Pneumonia. The subjects were children recruited to the study immediately after they had recovered from their first case of Pneumonia. The children were under observation over a certain period, some for about 3 months, others for about 6 months. The time variable `time_14` measures the time from inclusion to the next occurrence of pneumonia, or until the end of follow-up, whatever comes first. If the end of the follow-up comes first, the time observation is said to be censored. If pneumonia comes first the child is said to have had an event. We assume that the probability of being censored due to loss to follow-up or death from causes other than the one being studied is unrelated to the probability of pneumonia, i.e. independent censoring. The children were randomized to two treatments, either placebo or zinc.

Load the data file **Pneumonia_time.dta**.

There is a number of variables in the data set. The variables we will use are:

age: Age of the child (in months)

sex: Sex of the child (0 = male, 1 = female)

treat_1: Treatment by zinc (0 = placebo, 1 = zinc)

time_14: Time since last visit (or time to next pneumonia)

event: Event (0 = censored, 1 = pneumonia)

Describe data

Questions a) and b) below give rough descriptive statistics of the data file. They do not appropriately account for censoring.

a) Get an overview of the variables in the data file:

- **su age, detail**
- **hist age, freq**
- **tab sex**
- **tab event treat_1**
- **su time_14, detail**
- **hist time_14, freq**

b) Study the distribution of time to pneumonia (`time_14`) for each sex and treatment:

- **bysort sex: su time_14, detail**
- **hist time_14, freq by(sex)**

- **bysort treat_1: su time_14, detail**
- **hist time_14, freq by(treat_1)**

Prepare survival data

We have to declare data to be survival-time data by using the command `stset`. The `stset` command creates 4 new variables. These variables contain all the necessary information for STATA to do a survival analysis (except covariates, which must be given explicitly in the later commands). The variables are

- `_t` analysis time when the record ends
- `_d` 1 if failure, 0 if censored
- `_t0` analysis time when record begins
- `_st` 1 if the record is included, 0 if excluded

All the survival analysis (`st`) commands use these variables, as all the information regarding survival times is contained within these four variables.

c) Declare data to be survival-time data:

- **`stset time_14, failure(event = 1)`**

Describe what you get in the variable window and explain the screen output.

d) Describe and give short summaries for the important variables in the survival- time data:

- **`stdes`**
- **`sts list`**
- **`stsum, by(treat_1)`**
- **`stsum, by(sex)`**

Kaplan-Meier analyses

e) Create a single overall Kaplan-Meier survival curve:

- **`sts graph`**

Estimate (roughly) the median survival time based on the Kaplan-Meier plot. The median survival time is the time t at which $S(t) = 0.5$.

f) Create and compare Kaplan-Meier survival curves in the two groups with and without treatment of zinc:

- **`sts graph, by(treat_1)`**
- ***Add confidence interval**
- **`sts graph ,by (treat_1)`**
- ***Add risk table**
- **`sts graph ,by (treat_1) ci risktable legend(pos (6))`**

g) Carry out h) once more, but this time for sex:

- **sts graph, by(sex)**

h) Test for a significant difference between two or several survival curves:

- **sts test treat_1, logrank**
- **sts test treat_1, wilcoxon**

Is there any statistically significant difference between the two treatments? What test (log-rank or Wilcoxon test) would you prefer here and why?

Cox regression

The survival times depend on other factors such as age, sex, etc. We can take these factors into account when estimating or testing for a difference between two survival curves by using Cox regression (Stata command `stcox`).

i) Estimate the effect of treatment (`treat_1`) on the risk of pneumonia:

- **stcox treat_1**

Compare these results with the results in h).

We can also create two or more survival curves adjusted for these factors by the command `stcurve` after the command `stcox` is performed with the option `basesurv`.

j) Create the survival curves based on the Cox regression (i.e. after running the command `stcox`):

- **stcurve, survival at (treat_1 = 1) at (treat_1 = 0) lcol(blue red)**

Compare these curves with the survival curves in f).

k) Taking into account age and sex in the Cox regression model, estimate the effect of treatment on the risk of pneumonia:

- **stcox treat_1 age sex**

Note: In this file, age measures age at inclusion. We might prefer to adjust for age as a time-dependent covariate, but we ignore that here.

l) To assure that the Cox regression model fits the survival data, we may check the assumptions of this method. An important assumption to check is the proportional hazard. Use the following command:

- **stphplot, strata(treat_1) adjust(sex)**

What would be your opinion about the proportional hazards assumption, based on this plot?