lsdwrangling

Charles Ingulli

June 21, 2020

Introduction

The `lsdwrangling` package seeks to provide methods for transforming data to a different format. For example, statistical packages such as the `survival` package or `msm` package require their input data to be configured in different ways. In order to use the same data set in both packages, it is necessary to reconfigure the same information in two forms. The `lsdwrangling` package offers several functions that offer a seemless transition between one or more data formats.

Data Formats

The relevant data formats that are used in this package are described below. They include counting process data, right censored data, longitudinal data (panel data)

Counting Process Data

One common data setup for survival analysis is the counting process format. This data is set up with two time variables and an event indicator.

Let us consider a survival study of 3 individuals. Individuals are followed from the time of enrollment in the study until the event of interest occurs with monthly follow up visits. The initial visit will be referred to as the baseline where all individuals are assumed to be event free. At the monthly follow ups, individuals are tested to assess whether or not they have received the event. Time is then represented by intervals between visits with an associated event indicator variable. A visual depiction of the data is shown below in Table 1 where time is measured in days.

Table 1: Table 1: Counting Process Data

sid	start	stop	status
1	0	31	0
1	31	64	0
1	64	126	1
2	0	33	0
2	33	59	0
3	0	28	1

The variable representation of each column is :

- *sid* - the individual's identification number.
- *start* - the left end point of the time interval.
- *stop* - the right end point of the time interval.

- *status* - the event indicator (0 denotes no event and 1 denotes having the event).

Each row of the data corresponds to a particular time interval for each individual. We can see in the first row that individual 1 started the study at time 0 and had a follow up visit 31 days later. At day 31, individual 1 is tested for the event and is determined to have not experienced the event. Time intervals are considered open on the left and closed on the right. That is, for times, t_1 and t_2 where $t_1 < t_2$, the time interval is $(t_1, t_2]$. Looking at the third row, individual 1 starts a new time interval at day 64 and is seen for another follow up at day 126. At time 126, individual 1 is tested and is marked as experiencing the event.

Some data sets may also incorporate one or more explanatory variables. Often is the case in survival studies where other variables are measured that relate to the event of interest. These variables may be constant over time or may vary over time. Additionally, they may be qualitative variables or quantitative variables. Consider our previous example of 3 individuals. Let's say that in addition to measuring for the event of interest, individuals were also measured for assigned sex and age. The former can be considered a constant qualitative variable while the latter can be considered a time varying quantitative variable. Both constant and time varying variables have a big effect on how the data set is organized as each row is used for a single time interval.

Table 2 shows how additional explanatory variables would be incorporated to counting process format.

```
#>   sid start stop status sex  age
#> 1   1     0   31      0   0   46
#> 2   1    31   64      0   0   79
#> 3   1    64   96      1   0    0
#> 4   2     0   33      0   1 8766
#> 5   2    33   59      0   1 8799
#> 6   3     0   28      1   0    0
```

The coding for new variables is:

- *age* - the numerical age of the individual.
- *sex* - the assigned sex of the individual at birth (0 represents female and 1 represents male).

The constant variables have an easier interpretation since at every time point the variable will be the same. Individual one was assigned female at birth which stays true at baseline and each follow up visit.

Right Censored Data

A simplified form of the data is then shown below.

```
#>   sid start stop status
#> 1   1     0  126      1
#> 2   2     0   33      0
#> 3   3     0   28      1
```

This form conveys the similar information as the previous. Within the time interval $(0, 126]$, individual 1 experiences the event. The data set shown in Table 2 is considered right censored data. The topic of censoring is outside the scope of this vignette but see (some sources) for more information. Additional simplifications are also possible. Since each individual starts at the same time point, in this example it is time of enrollment, the start variable can be ommitted. When each row corresponds to only one individual, some data sets may ommit the identification variable.

Longitudinal Data

The longitudinal data format is format used in repeated observation studies. It is also known as panel data and similarly panel studies. The respective terminologies are equivalent and refer to the same topics.

Application

The data formats described thus far are used for different statistical models. The counting process format is widely used in the `survival` package for implementing different proportional hazards models such as the Cox-proportional hazards model. The `flexsurv` package also uses this format and extends the `survival` package to add more flexible parametric proportional hazards models. The `msm` package uses the logitudinal data format to fit multi-state models.

Implementation

```
#library(mypackage)
```

References

Therune survival vignette Ther book Collett