Project 5

Your data set (data_project_5.csv) comes from a study aiming to test the validity of the "27 club" (Wolkewitz et al, 2011). This "club" is a set of famous musicians who died at age 27 (Jimi Hendrix, Jim Morrison, Janis Joplin, etc). The respective data dictionary is the following:

- Person.num, Person ID number

- Band.num, Band ID number

- Date.of.birth, date of birth

- Date.of.number.one, Date of UK number one album (see: Wikipedia pages, e.g., http://en.wikipedia.org/wiki/List_of_number-one_albums_from_the_1970s_(UK))

- Date.of.death.censored, Date of death for those who died and censored data (August 2011) for those still alive

- Status, 1=dead, 0=alive

Before conducting any formal analysis, present and describe the data with appropriate statistics.

Create a new categorical variable in your dataset where you divided the musicians by the decade of birth (e.g., 1950-1959, 1960-1969, 1970-1979, etc). In the first analysis, your event of interest is time to death. Calculate the respective survival curve for musicians born in each decade using an appropriate method (be aware of censored data). Compare these survival curves using an appropriate non-parametric statistical test. Justify the chosen test. Use an appropriate parametric regression model for comparing the same curves.

In the second analysis, consider the event of interest as the time to reach number 1 in the UK charts (i.e. age of the musician when the album reached the top position in the chart). Calculate the respective survival curve for musicians born in each decade using an appropriate method (be aware of censored data if any). Compare these survival curves using an appropriate non-parametric statistical test. Justify the chosen test. Test different parametric regression models for comparing the same curves. Which is the best parametric model? What interpretation can you extract from it?

In both analysis, which results, parametric or non-parametric, are for you more trustworthy from a statistical standpoint?

Important:

Prepare a 15-minute presentation with your <u>main findings</u>. There will be a penalty of 0.5 points in your project grade if you exceed the time of your presentation. Upload your R script/R Markdown for code verification. Also upload your presentation as a pdf file. Failure to upload these files before classroom evaluation leads to a penalty of 0.5 points in your project grade.

Reference:

Wolkewitz M, Allignol A, Graves N, Barnett A G. Is 27 really a dangerous age for famous musicians? Retrospective cohort study BMJ 2011; 343 :d7799. doi:10.1136/bmj.d7799