Practical For Session 10: Survival Analysis

Mark Lunt

13/12/2022

1 Practical 10: Survival Analysis

Datasets

The datasets that you will use in this practical can be accessed via http from within stata. However, the directory in which they are residing has a very long name, so you can save yourself some typing if you create a global macro for this directory. You can do this by entering

```
global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
global datadir $basedir/stats/10_survival/data
```

(In theory, the global variable `datadir` could have been set with a single command, but fitting the necessary command on the page would have been tricky. Far easier to use two separate commands as shown above). If you wish to run the practical on a computer without internet access, you would need to:

1. Obtain copies of the necessary datasets
2. Place them in a directory on your computer
3. Define the global macro `$datadir` to point to this directory.

1.1 Life tables and Survival Curves

This section uses the dataset `"leukaemia"`.

1.1 First, set up the data for survival analysis. The time variable is `weeks`, the number of weeks to relapse. The outcome variable is `relapse`, which is 1 if the subject had a relapse at that time and 0 if they did not. Hence the command to set the data up for survival analysis is `stset weeks, fail(relapse)`

1.2 Obtain a life table for the subjects on Drug A with the command `sts list if treatment1 == 1`. What is the median survival in this group (at what time does the survivor function reach 0.5) ?

1.3 How many subjects were lost to followup in this treatment arm ?

1.4 Obtain a life table for the subjects on standard treatment with the command `sts list if treatment1 == 0`. What is the median survival in this group ?

1.5 How many subjects were lost to followup in this treatment arm ?

1.6 Do the answers to your previous questions suggest that Drug A is better, worse, or the same as standard treatment ?

1.7 Produce a Kaplan-Meier curve for each of the treatments with the command `sts graph, by(treatment1)`. Does this confirm your answer to the previous question ?

1.8 Add a horizontal line to the graph by adding the option `yline(0.5)` to the previous command. This line represents half of the group surviving and half having a relapse: the point where it crosses the two survival curves should give you the median survival times you calculated in earlier questions.

1.9 Add the option `lost` to the previous command. This will show how many subjects were censored at each time point. How many subjects were lost to followup in the two treatment arms ? Does this agree with the results you got from `sts list` ?

1.10 Add the option `gwood` to the previous command to obtain confidence bands for the survival curve ? (The odd name for this option is because the formulae used to calculate the confidence bands were developed by a Major Greenwood). Why do the confidence bands get wider over time ?

1.11 Perform a logrank test to compare the survival on Drug A to that on standard treatment, with the command `sts test treatment1`. Is the difference between Drug A and standard treatment statistically significant ?

1.12 Would have had the same answer to the previous question if you had used a Wilcoxon test in place of a logrank test ? (You can do this by adding the option `wilcoxon` to the previous command.)

1.2 Cox Regression

2.1 Have a look at the survival curves by white blood cell count using `sts graph, by(wbc3cat)`. Does the white blood cell count affect survival ?

2.2 Do a cross-tabulation of `treatment1` against `wbc3cat` with `tab wbc3cat treatment1, co` Are the proportions of subjects in each of the white blood cell counts categories the same in the two treatment arms ?

2.3 Given that proportion of subjects in the "High" cell count group is greater in the standard treatment arm than in the Drug A arm, would you expect this to have increased or decreased survival in this arm of the trial ?

2.4 White blood cell count is a potential confounder, so we need to adjust for it. First, we will perform an unadjusted Cox regression to obtain the hazard ratio before adjusting. This is done with the command `stcox treatment1`. What is the hazard ratio for Drug A, and its 95% confidence interval ?

2.5 Now obtain the adjusted hazard ratio with the command `stcox treatment1 i.wbc3cat`. What is the adjusted hazard ratio and its 95% confidence interval ?

2.6 How did the confounding by white blood cell count affect the apparent effect of Drug A ? Is this what you expected from the earlier questions ?

2.7 Now we need to test the proportional hazards assumption. First for treatment: produce a plot of the observed and predicted Kaplan Meier plots with `stcoxkm, by(treatment1)`. Are the observed and predicted curves close to each other ?

2.8 Now we can test the same assumption for the effect of white blood cell count, with `stcoxkm, by(wbc3cat)`. Are the observed and predicted curves close to each other ?

2.9 To obtain a formal test, we need to store the scaled and unscaled Schoenfeld residuals by running the command `stcox treatment1 i.wbc3cat, sca(sca*) sch(sch*)` Now enter the command `stphtest` to get an overall test of proportionality. Is the regression model valid ?

2.10 Use the command `stphtest, detail` to obtain tests of proportionality for each individual variable. Is there any evidence of non-proportional hazards ?

1.3 Non-Proportional Hazards

3.1 There is a second drug used in this trial, stored in `treatment2`. Compare the survival curves for Drug B and standard treatment with the command `sts graph, by(treatment2)` How does the survival on Drug B compare to that on standard treatment during the first 10 weeks ?

3.2 How does the survival on Drug B compare to that on standard treatment after the first 10 weeks ?

3.3 Superimpose the predicted survival curves from the Cox regression model with `stcoxkm, by(treatment2)`. How do the predicted and observed curves differ ?

3.4 Perform a Cox regression of `treatment2` and `wbc3cat` with `stcox treatment2 i.wbc3cat` Does Drug B have a significant effect on survival ?

3.5 To test the proportional hazards assumption, we need to store the Schoenfeld residuals again. First drop the residual from the previous model with `drop sca* sch*` Then rerun the `stcox` command with the options `sca(sca*) sch(sch*)`. Perform the overall test with `stphtest`: is the model appropriate ?

3.6 Test the proportional hazards assumption for each variable separately with `stphtest, detail`. Which variable does not satisfy the assumption ?

3.7 The Kaplan-Meier curves suggest that Drug B has a negative effect on survival initially, then becomes positive. So we will test for different effects before and after 10 weeks. First produce a life-table with `sts list`

3.8 To be able to split the data, you need to have an id for each subject. We can do this with `generate id = _n`. Now each observation has its record number as an identifier.

3.9 Now, for each subject followed for more than 10 weeks, we will split the data into 2 observations, one for the time up to 10 weeks and one for the time after. First, we must include the id information in the stset command with `stset weeks, fail(relapse) id(id)` Then we can split the data with `stsplit split_time, at(10)`

3.10 Check that the life-table remains unchanged by entering the command `sts list` Is it the same as before ?

3.11 Examine the data with `list id weeks relapse split_time _t0 _t`. You should see that for subjects who were followed up for less than ten weeks, there is still a single record. However, for those followed up for more than 10 weeks, there are two records, one with `split_time` == 0, the other with `split_time` == 10. The start of the interval is given by `_t0`, the end by `_t`

3.12 Now we can generate separate treatment variables for the treatment effect before and after 10 weeks. The commands to use are `gen t1 = treatment2*(split_time == 0)` and `gen t2 = treatment2*(split_time == 10)`.

3.13 Now fit the Cox regression model with both `t1` and `t2` as predictors with the command `stcox t1 t2 i.wbc3cat` What is the hazard ratio for t1, with its 95% confidence interval ?

3.14 What is the hazard ratio for `t2` ?

3.15 Do these hazard ratios confirm what you were expecting ?

3.16 Drop the residuals from the previous model with `drop sca* sch*` then create new residuals with `stcox t1 t2 i.wbc3cat, sca(sca*) sch(sch*)` Now test the proportional hazards assumption with `stphtest`. Is the model now appropriate ?

3.17 Test the proportional hazards assumptions for each of the variables separately with `stphtest, detail`. Do any of the predictors show non-proportionality ?