SC968 PANEL DATA METHODS FOR SOCIOLOGISTS

WORKSHEET FOR LECTURE 5

5.0 OBJECTIVES

* To be able to prepare data for EHA
* To be able to summarise time to event data
* To become familiar with Cox regression models in Stata
* To perform appropriate commands in Stata to check the proportional hazards assumption
* To examine the results critically

5.1 GETTING STARTED

 Our teaching data is a subset of the British Household Panel Survey (BHPS).

 Online documentation for BHPS is at
http://www.iser.essex.ac.uk/survey/bhps/about-bhps/latest-release-of-bhps-data

 Set up a working directory. This could be your m: drive, a subset of your m: drive, or a memory stick. In these instructions, m:/ denotes your working directory – so if you are not going to use m:/ as your working directory, substitute the appropriate drive and path.

 If your m: drive is on the ISER server then you need to

 START → RUN → CMD

 and then enter

 Net use M: [\\iserhome\iserhome\users\your](file:///%5C%5Ciserhome%5Ciserhome%5Cusers%5Cyour) user name

 If you have not already done so, download the data from the course website: <http://www.iser.essex.ac.uk/iser/teaching/module-sc968/course-materials> and save it in your working directory as “survival.dta”.

 In order to download the data, you will need the username and password, which you were given in the first practical.

 Try to keep a “clean” copy of the data in your working directory, and save any other datafiles which you may produce under different names.

 Open STATA, and set your working directory as your home directory

 cd m: //substitute the drive and path name where you have saved the data

 Open the teaching dataset

 use survival, clear

The data set is a subset of records from our training.dta data set containing a 20 percent random sample of the British Household Panel Survey. Data from wave 1 (1991) to wave 15 (2005) are included. The file is organised in “long” (person-wave) format so there is one record for each wave that a respondent participated in the survey. For survival.dta, only those who were single, never married at wave 1 are selected and only the variables needed for the practical have been kept. For simplicity, missing values on the independent variables have been filled in using the last recorded value.

The task today is to analyse time to first cohabiting relationship or marriage in the BHPS. We will calculate survival time till first lived with a partner from entry to the study in 1991 until the marital status variable mastat changed to code 1 (married) or 2 (cohabiting).

The first task will be to get to know the data by looking at some descriptive statistics. Some variables are derived from those you are familiar with in preparation for the practical.

describe

summarize

Then comes the most important part before any survival modelling: preparing the data.

5.2 Summarising time to event data

First the data must be declared as survival time data using the stset command.

When setting the data you must consider:

1. What is your population of interest?
	1. In this case, it is single individuals who are at risk of finding a partner.
2. What is your failure event?
	1. In this case it is getting a partner, either via marriage either via cohabitation
3. When should individuals exit your analysis?
	1. In this case, it is either when they experience the event or they are no longer followed up (i.e. censored)
4. When do individuals enter into your analysis?
	1. In this case, it is at the beginning of the study, i.e. in wave 1.
5. When do individuals become at risk of experiencing the event?
	1. In this case, it is the youngest age at which people can find a partner. Clearly, very young people are not at risk of finding a partner. Let’s assume that individuals 16 and older (but not younger) are at risk of finding a partner. Thus, an individual’s risk period begins once (s)he turns 16.

stset wave, id(pid) failure(mastat==1 2) exit(mastat==1 2 .)

Stata creates four new variables when stset is run:

 \_st shows whether the record is to be used in the analysis

 \_d is 1 if the event occurred, 0 if not

 \_t0 is the start of the time span for the current record

 \_t is the end of the time span for the current record.

Have a look at them for the first 100 records and check that they are set correctly.

list pid wave mastat \_st \_d \_t \_t0 in 1/100,sepby(pid) noobs

In the previous setting, Stata considers the start of the risk period to be wave 1. How can you tell that?

Examine the ages of individuals in your dataset

sum age if wave==1, detail

The ages of individuals span from15 to 90 in the first wave of the study. Clearly, it would be wrong to assume that the onset of risk would happen at age 15 for some individuals and at age 90 for others. Instead, you should consider the onset of risk occurs at the same age for all individuals. Let’s assume this age to be 16. Some individuals are observed in the data starting with age 16. For them the onset of risk occurs while they are in the study. For older individuals however, the onset of risk occurs before they enter the study. As you know their age, the time of the onset of risk can be calculated. However, you have to take into account t the period they did not experience the event which occurred before entering the study.

(NB!: this approach assumes that single individuals who are observed older than 16 at the beginning of the study have not had a partner before they entered the study, i.e. they did not have a partner the entire period between turning 16 and wave 1 of the study. If this assumption is violated, the onset of risk for the current spell of being single is unknown. By definition, a partnered person is not at risk of finding a partner and so if a partnership existed before the start of the study but the person is observed to be single at wave 1, the current spell of being single would have to be calculated from the date the previous partnership broke down. This date would be unknown. For the purposes of the following exercises, we will ignore this complication).

stset age, id(pid) failure(mastat=1/2) origin(time 16) entry(wave=1) exit(mastat=1/2 .)

Check again the first 100 observations:

list pid wave age mastat \_st \_d \_t \_t0 in 1/100,sepby(pid) noobs

Then look at the descriptive statistics for the whole file

stdes

Are there any gaps?

The next task is to produce a summary of survival times and rates for the complete sample.

stsum

What is the total person years of follow-up?

What is the cohabiting relationship rate per year of follow up?

Now look at the summaries for men and women.

stsum, by(sex)

You can plot the hazard function for men and women using the Kaplan-Meier graph

sts graph, by (sex)

What is the median time to cohabitation for men and women?

Now test for a difference in the time to cohabitation between men and women using the log rank tests

sts test sex

Is there a significant difference in time to cohabitation between men and women?

5.3 Cox regression models

Now let’s fit some Cox proportional hazard models to the data. The command for this is stcox. For example, to predict time to partnership from gender:

xi:stcox i.sex

First look at gender and age group. Run the Cox proportional hazard model for this and fill in the values in table 1 below:

Table 1. Hazard ratios by gender for time to first cohabiting partnership

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hazard ratio | 95% C.I. | P value |
| Unadjusted |  |  |  |
| Adjusted for age group |  |  |  |

What happens to the hazard ratio for gender when you adjust for age group?

What does the hazard ratio for gender represent when you adjust for age?

What’s the hazard ratio of women finding a partner relative to men when they are aged 35 and over?

Now you will look at different measures of socio-economic position (SEP) at baseline as potential predictors of cohabitation.

The file contains several SEP measures:

 nssec\_w1 Occupational class at baseline

 nssec Occupational class at wave(time+1)

 hqual\_w1 Education at baseline

 hqual Education at wave(time+1)

 income\_w1 Income at baseline

 income Income at wave(time+1)

Type stvary to check this out.

stvary nssec\* hqual\* income\*

Try some models with gender, age group and the three SEP measures at baseline in turn. Then fill in the values in table 2 below:

Table 2 Hazard ratios for time to first cohabiting partnership

|  |  |  |  |
| --- | --- | --- | --- |
| Category  | ClassHR (95% CI) | EducationHR (95% CI) | IncomeHR (95% CI) |
| Unadjusted |  |  |  |
| Adjusted for age group and gender |  |  |  |
| Adjusted for age group and gender and other SEP measures |  |  |  |

Which measure of SEP has the highest hazard ratios in the unadjusted models? And which the lowest?

Are they all significant predictors of time to cohabitation?

How would you interpret the differences between the unadjusted hazard ratios and the hazard ratios adjusted for age and sex?

Does each SEP measure still predict time to cohabitation when you control for the other SEP measures?

What do you conclude from this?

5.4 The proportional hazards assumption

We ought to check the proportional hazards assumption.

Plot a Kaplan Meier graph for men and women.

stcoxkm, by(sex)

What do you notice? Are the observed lines close to the predicted lines?

Now get a survival probability plot for each gender

sts graph, by(sex) cumhaz

Do the cumulative survival curves cross?

Another way to assess proportional hazards is to use stphplot. For example:

stphplot, strata(sex) adjust(nssec\_w1 hqual\_w1 income\_w1 agegroup)

If the proportional hazards assumption is met, then the lines should be approximately parallel. Are they?

You can also carry out a formal test of the proportional hazards assumption for gender. One way to do this is to introduce an interaction term between gender and time. This can be done with the options tvc and texp in the stcox command.

xi: stcox i.sex i.agegroup i.nssec\_w1 i.hqual\_w1 i.income\_w1, /// tvc(i.sex)texp(log(\_t))

The tvc option specifies that gender should interact with a function of time and the texp specifies that this function is log(time). This allows the effect of gender on survival to vary by time since entry to the study.

You can see that the output gives you two rows for sex. The second relates to the interaction of sex with log(time). If the interaction is significant, there is evidence that the effect of gender varies by time. Does it?

The second method to formally test the proportional hazards assumption is on the basis of Schoenfeld residuals.

xi:stcox i.sex i.agegroup i.hqual\_w1 i.nssec\_w1 i.income\_w1, ///

schoenfeld(sch\*) scaledsch(sca\*)

estat phtest, rank detail

Is the test statistic significant for sex?

Is there any evidence that hazards are non proportional for one of the other covariates?

Run the graphical and formal checks on the proportional hazards function for the variable agegroup.

If there is evidence that the proportional hazards assumption is not met, then re-run the analyses treating the agegroups as separate strata. This estimates separate baseline hazard functions for each group thus allowing them to be non-proportional. To do this, use the strata option.

xi:stcox i.sex i.nssec\_w1 i.hqual\_w1 i.income\_w1, strata(agegroup)

If the proportional hazards function for the SEP measures had been violated, an option would have been to re-run the analysis using the time-varying versions of the SEP measures.

xi:stcox i.sex i.agegroup i.hqual i.income i.nssec

Compare these estimates with the ones in table 2. Why do you think they are different?

5.5 Trying a new survival analysis on your own!

Some respondents could have withdrawn from the survey before they have found a partner. We are now going to investigate predictors of drop-out.

The variable wdrawn takes a value 1 if the respondent dropped out of the survey before wave 15 and 0 otherwise.

Use stset to declare time to withdrawal.

Use list, stdes, stsum to check that you have set the survival variables up correctly.

Is gender, age or SEP related to withdrawal from the survey?