Study Guide

Survival Analysis [SVA]

(Macquarie University unit code STAT827)

Semester 1, 2017

Prepared by: A/Prof Gillian Heller
Department of Statistics, Macquarie University

Edited by: Dr Ken Beath and Dr Petra Graham,
Department of Statistics, Macquarie University

MACQUARIE
University

Contents

SVA is a unit offered by the Department of Statistics, Faculty of Science and Engineering at Macquarie University. This unit is a component of Macquarie University's postgraduate Applied Statistics program, with a value of 4 credit points, and the program of the Biostatistics Collaboration of Australia (BCA). All aspects of the unit, including assessment, will be identical for BCA and MU students. However we will be using separate sites for online learning; elearning for BCA students and ilearn for Macquarie.

1 Contacts

The **unit coordinator** is
Dr Ken Beath
Phone: (02) 9850 8516
E-mail: ken.beath@mq.edu.au

General questions about the unit as well as questions about the assignments should be addressed to Ken. Ken will also be overseeing the elearning discussions. He is also available for consultation on the Macquarie University campus during his consultation hours, or by e-mail, or by appointment.

The **postgraduate administrator** is
Lesley Mooney
Phone: (02) 9850 8550
e-mail: lesley.mooney@mq.edu.au

Enquiries about the receipt of course materials and assessments should be addressed to Lesley. Printed copies of course materials will only be sent upon demand to Lesley within the first two weeks of semester. Her working days are Tuesday, Wednesday and Thursday.

2 Overview

Survival analysis is concerned with the analysis of responses which are times to an event of interest. Such events can be, for example, recurrence of disease (such as a particular type of cancer); death; cessation of smoking; or weaning of breastfed infants. We seek to relate the time-to-event to covariates. In principle, we utilize the same tools that are applicable to other statistical analyses, such as multiple regression, but we have to take account of two issues that are common to time-to-event responses, viz. non-normality and censoring. The models that we study accommodate these features, and enable us to assess the effect of covariates in the usual statistical framework of estimation and hypothesis testing. We will be using the statistical software Stata throughout the unit.

3 Learning outcomes

At the end of this unit you should be able to:

a. Understand the nature of survival data, especially censoring and truncation;

b. Summarise and display survival data using nonparametric methods;

c. Analyse survival data using the Cox proportional hazards model, including:

   i. hypothesis testing;
   ii. diagnostic testing;
   iii. the use of stratification and time-dependent variables, where appropriate;
   iv. interpret results;

d. Analyse survival data using parametric (accelerated failure-time) models;

e. Formulate, implement and interpret multiple-event models, and select appropriate model;

f. Determine sample size for a simple survival analysis;

g. Produce appropriate displays for publication.

The unit is divided into seven modules, each taking two weeks, except for the last, which is of one week duration:

Module	Weeks	Content
1	1,2	The nature of survival data, including censoring; the survival (or survivorship) function: definition and estimation via the Kaplan-Meier curve; the stset command in Stata;Kaplan-Meier estimate of the survival (or survivorship) function: confidence intervals and hypothesis testing.
2	3,4	the density, survival, hazard and cumulative hazard functions; ; the Nelson-Aalen estimate of the cumulative hazard function; Definition of the proportional hazards model; construction of the partial likelihood for the Cox model; the treatment of tied failure times; hypothesis testing on the coefficients, using Wald and partial likelihood ratio tests.
3	5,6	For the Cox PH model: hypothesis testing on the coefficients, contd; estimation of the baseline functions $S_0(t)$ and $H_0(t)$, and their adjustment for covariate values; the effect of a change in scale and origin of units of measurement of covariates.
4	7,8	Model diagnostics for the Cox PH model; the stratified Cox model
5	9,10	Time-dependent covariates in the Cox model; parametric survival time models, in particular the accelerated failure time model, with an exponential and Weibull distrubution; discrete-time logistic model
6	11,12	Correlated survival data; clustered survival data; recurrent events models;competing-risks models
7	13	Sample size determination for comparing two response rates and two survival distributions; good practice for the display of survival analysis results in scientific publications

Module objectives are given in Section A.

4 Assessment

Assessment for the unit consists of three assignments, consist of data analysis and a small number of theoretical questions. The weighting is as follows:

Assignment 1	30%
Assignment 2	40%
Assignment 3	30%

Satisfactory performance in the subject will be based on overall performance: there is no requirement that individual components must be completed satisfactorily. A mark will be supplied to the BCA and hence to your home institution, and their policies applied.

4.1 Assignments

Assignments will be made available at least 2 weeks before the due date. They will generally require data analysis, but may also involve calculations and mathematical analysis.

Rules for assignment submission are given in the document BCA Assessment Guide, which is included with these notes and is also on `http://bca.edu.au`. *Please read this document carefully.* Assessed exercises and assignments should be submitted on the eLearning website. Please consult the BCA Assessment Guide (*Please read this document carefully.*) for details about submitting your assignments, and guidelines for written work. For SVA submissions:

- Your assessment submission should be one Word or preferably PDF document. Multiple files/documents will not be accepted. Either embed in the Word document or there are some free tools for combining pdfs, for example PDFill.
- Do not send EXCEL spreadsheets.
- Solutions for both exercises and assignments should include any formulae that you have used to arrive at your conclusion. Marks are awarded for correct formulae and working.
- If using WORD, please use the equation editor for writing mathematical formulae.
- Hand written solutions are acceptable for mathematics. Please scan and include in your Word or PDF document.
- You must include your name both in the filename and in the footer of each page.

Competed assignments must be submitted in the *Turnitin* submission dropbox.

Academic Honesty You will be required to accept a statement on academic honesty before being given access to assignments and exercises.

Late submission Requests for an extension of the due date for an assignment must be made in advance of the due date for that assessment. These requests must be made directly to the unit coordinator by email. The unit coordinator will reply by email with the decision as to whether an extension has been granted and the new due date.

If no extension has been granted, late submissions will incur a penalty. Five percent of the total marks for the assessment will be deducted for each day the assessment is late, up to a maximum of 50%. It is not the intention of this late penalty policy to cause a student to fail the unit when otherwise they would have passed. If deductions for late assignments result in the final unit mark for a student being less than 50, when otherwise it would have been 50 or greater, the student's final mark will be exactly 50.

Note: I have made the submission date a week after the end of the material covered, except for Assignment 3. Therefore there should be no reason for requiring an extension except for very unforeseen circumstances.

5 Timetable

We will be following the University of Sydney semester, with dates as follows:

Week	Starting date	Module	Assessment due
1	6 March	1	
2	13 March	1	
3	20 March	2	
4	27 March	2	
5	3 April	3	
6	10 April	3	Assignment 1
Mid-semester break (1 week)			
7	24 April	4	
8	1 May	4	
9	8 May	5	
10	15 May	5	Assignment 2
11	22 May	6	
12	29 May	6	
13	5 June	7	
	12 June	None	Assignment 3

Assessment is due on the Monday of the indicated weeks, at 11.55pm, except for Assignment 3 which will be due on the Tuesday 13 June due to the public holiday.

6 Method of delivery and communication

The unit is offered in distance mode. Our means of communication will be notes which will be available on elearning, e-mail, and elearning (http://elearning.sydney.edu.au/), the University of Sydney's e-learning site. Our primary communication method is via elearning and we expect you to log in at least weekly to check for announcements and release of assignments and so on.

The unit relies heavily on the prescribed text Hosmer, Lemeshow and May (see below). The study notes provide a guide to readings in this text, as well as sometimes to other readings, which will be provided. They also provide additional explanation where this is needed. In the study notes for each module, tutorial exercises are given, mostly referring to exercises in Hosmer, Lemeshow and May. Electronice versions of this text may be available from your universities' library.

Study notes will be posted on the elearning site. A printed copy may also be requested, but must be requested from Lesley Mooney (lesley.mooney@mq.edu.au) by the end of the second week, that is March 17.

elearning

We will be using elearning for online discussions, posting of course notes, assignments, solutions and data sets, and submission of assignments.

As elearning is the primary medium for communication in this unit, we expect that students will access the site at least every week during semester. Your usual uniKey will be required. There will be a introduction thread on elearning where you can describe your background. Photos are welcome.

7 Textbooks

The prescribed text is

Hosmer DW, Lemeshow S and May S (2008). *Applied Survival Analysis*, John Wiley and Sons, Second Edition.

There are numerous texts on survival analysis which you may wish to consult, but the following may be particularly helpful because of its use of Stata:

Cleves MA, Gould WW, Gutierrez RG and Marchenko Y (2010). *An Introduction to Survival Analysis using Stata*, Third Edition, Stata Press.

Other useful texts are:

- Klein JP and Moeschberger ML (2003). *Survival analysis : techniques for censored and truncated data*, Springer.
- Kleinbaum DG (2012). *Survival analysis : a self-learning text*, Springer-Verlag.

- Allison PD (2010). *Survival Analysis Using SAS: A Practical Guide*, SAS Instiutute. (for those using SAS)
- Moore, DF (2010). *Applied Survival Analysis using R*, Springer. (for those using R)
- Therneau, TM and Grambsch, PM (2001). *Modeling Survival Data: Extending the Cox Model*, Springer. (more advanced treatment with code for SAS and R)

8 Software

We will be using Stata (version 12 or later, preferably at least version 13). While Stata has a GUI we will be using the command language. It is still useful to experiment with the GUI, as the corresponding commands are available in the Review pane. It is also fine to use R or SAS provided that you already have experience with either of these. There are brief notes describing the use of each in the Other section of the Notes and also some code.

A Module objectives

A.1 Module 1

To understand:

a. the nature of survival data, including censoring;

b. the survival (or survivorship) function: definition and estimation.

c. the Kaplan-Meier estimate of the survival (or survivorship) function: confidence intervals and hypothesis testing;

d. the Nelson-Aalen estimate of the cumulative hazard function;

e. the relationship between the density, survival, hazard and cumulative hazard functions, and their derivation.

To be able to compute and interpret:

a. the density, survival, hazard and cumulative hazard functions over a grid of parameter values.

b. the stset command in Stata;

c. the Kaplan-Meier curve.

d. the Kaplan-Meier curve; logrank and Wilcoxon tests;

e. the Nelson-Aalen estimate of the cumulative hazard function;

A.2 Module 2

To understand:

a. the definition of the proportional hazards model;

b. construction of the partial likelihood for the Cox model;

c. the treatment of tied failure times;

d. hypothesis testing on the coefficients, using Wald and partial likelihood ratio tests.

To be able to compute and interpret:

a. the Cox regression model with both categorical and continuous predictors;

b. Estimates using the different methods for treating tied times;

c. Hypothesis tests on the βs, individually and jointly, using both the Wald and partial likelihood ratio tests.

A.3 Module 3

To understand, for the Cox proportional hazards model:

a. hypothesis testing on the coefficients, contd;

b. estimation of the baseline functions $S_0(t)$ and $H_0(t)$, and their adjustment for covariate values;

c. the effect of a change in scale and origin of units of measurement of covariates.

To be able to compute and interpret:

a. Hypothesis tests on the βs, individually and jointly, using both the Wald and partial likelihood ratio tests;

b. $\widehat{S}_0(t)$ and $\widehat{H}_0(t)$, adjusted for covariate values, plotted against time;

c. hazard ratios and confidence intervals for scaled covariates.

A.4 Module 4

To understand:

a. model diagnostics for the Cox PH model;

b. the stratified Cox model.

To be able to compute and interpret:

a. martingale residuals for the correct functional form of a covariate;

b. categorical forms of continuous covariates;

c. log-log survival plots;

d. Schoenfeld residuals;

e. Cox-Snell residuals for the overall goodness of fit;

f. dfbeta residuals for identification of influential observations;

g. the stratified Cox model.

A.5 Module 5

To understand:

a. Time-dependent covariates in the Cox model;

b. Difficulties with time-dependent covariates;

c. Parametric survival time models, in particular the accelerated failure time model, with an exponential and Weibull distrubution, and the discrete-time survival model;

To be able to compute and interpret:

a. Cox regression with time-dependent covariates;

b. Coefficients and hazard ratios from exponential and Weibull regressions;

c. Coefficients and hazard ratios from discrete-time survival model;

A.6 Module 6

To understand:

a. Models for clustered survival data;

b. Models for recurrent events, and appropriate model choice;

c. Models for competing risks.

To be able to compute and interpret:

a. Cox regression with robust SE for clustered data;

b. Marginal model for recurrent events;

c. Competing risks models, and the cumulative incidence function.

A.7 Module 7

To understand:

a. Sample size determination for comparing two response rates and two survival distributions;

b. Good practice for the display of survival analysis results in scientific publications.

To be able to compute and interpret:

a. Sample size calculations;

b. Survival plots in a variety of forms.