College of Natural Science
MICHIGAN STATE UNIVERSITY

Department of Statistics and Probability
Analysis of Survival Data

Fall 2020 Syllabus

Course Number STT847
Credit Hours 3
Course meeting days and time: Online
Course location: Online
Course Website (lecture notes, videos, dataset, reference, homework, projects, etc): https://d2l.msu.edu/d2l/loginh/
Course Modality: This course will be taught in an online format, asynchronous mode.

Instructor Information

Name: Hyokyoung G. Hong
Office: Room #C435 Wells Hall
Office hours: By appointment. In FS20, all office hours must be via computer or telephone. We will use Zoom or Microsoft Teams. Please email me with an opening in your schedule to arrange the meeting.
E-mail: hhong@msu.edu
Instructor Introduction: You can find more information about me here https://www.stt.msu.edu/users/hhong/

Grader Information
Alex Pijyan: pijyanal@msu.edu

Coverage Topics

- **L1 Introduction to R Markdown (Rmd)**

R Markdown has now become a core component of many statistics projects. It helps maintain code, ensures reproducibility and consistency. Moreover, it is versatile and people use it to create reports, notes, presentations, and blog posts. You will use R Markdown to write up all R related assignments. In most lectures, R Markdown codes will be provided.

- R Survival package (https://cran.r-project.org/web/packages/survival/survival.pdf)
- R cheat sheet (https://rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf)
- R markdown cheat sheet (https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)
- Advanced R (https://adv-r.hadley.nz/index.html)

- **L2 Chapter 1 Typical censoring and truncation mechanisms**

  Survival Analysis refers to the analysis of time to event data**.** The time to an event is called the failure time and is treated as a random variable. Such data arise in many fields including medicine, engineering, sociology, education, forestry, economics, and many others. The observation of survival time is often incomplete. The statistical term used to describe the process producing incomplete observation is called "censoring" and the observation is referred to as being "censored." In general, incomplete observation of time to an event can occur in several ways and we overview of them.

- **L3 Chapter 2 Key functions in survival analysis; survival function S(t), distribution function, F(t), density function, f(t), hazard function, h(t), cumulative hazard function, H(t), relationship between key functions**

  There are five main functions in the survival analysis; survival function, distribution function, density function, hazard function, and cumulative hazard function. We define each of them and study the relationship between key functions.

- **L4-L6 Chapter 2, Appendix 1, Kaplan-Meier estimator; K-M estimation, K-M analysis using R**

  One of the main goals in the survival analysis is to estimate a population survival curve from a sample. If every patient is followed until death, the curve may be estimated simply by computing the fraction surviving at each time. However, in most studies patients tend to drop out, become lost to follow-up, move away, etc. Kaplan-Meier analysis allows estimation of survival over time, even when patients drop out or are studied for different lengths of time. The Kaplan-Meier estimator of the survival function has been, and continues to be, the most frequently used estimator, largely due to the fact that it is routinely calculated by most software packages.

- **L7-L8 Chapter 2 Nelson-Aalen estimator, N-A analysis using R**

  Estimators of the survival function could be based on an estimator of H(t). Aalen (1975, 1978), Nelson (1969, 1972) and Altshuler (1970) have proposed an easily computed estimator of H(t), which we refer to as the Nelson-Aalen estimator. The work by Aalen is considered to be one of the landmark contributions to the field, as virtually all recent statistical developments for the analysis of survival time data have been based on the counting process approach he used to derive his version of the estimator of H(t).

- **L9-L10 Chapter 2 Comparison of survival functions (formally and visually); log-rank test, Wilcoxon test, plotting curves for more than two groups**

  We are often interested in assessing whether there are differences in survival among different groups of participants. For example, in a clinical trial with a survival outcome, we might be interested in comparing survival between participants receiving a new drug as compared to a placebo (or standard therapy). In an observational study, we might be interested in comparing survival between men and women, or between participants with and without a particular risk factor (e.g., smoking). We investigate several tests to compare survival among independent groups.

Project #1: Survival data analysis I (L11-L12)

- **L13-L14 Likelihood methods for uncensored and censored data**

  We develop the likelihood functions for censored data while assuming the failure times, T, follows certain parametric distributions such as exponential and Weibull distributions. More specifically, the following examples will be demonstrated: compute the MLE of the hazard, estimate the standard error of the estimated hazard, conduct Wald and likelihood ratio tests if the T follows the exponential distribution or Weibull distributions.

- **L15-L16 Chapter 8 Parametric regression models (accelerated failure time model); the exponential regression model, the Weibull regression model**

  So far, we have assumed that covariates do not play any roles in influencing survival time data. However, it is well known that some covariates such as age, gender, and race are important risk factors for cancer survival (regression component). Moreover, there may be settings in which the distribution of survival time, through previous research, has a known parametric form that justifies use of a fully parametric model to address the goals of the analysis better (parametric component). A fully parametric regression model has some advantages: (1) full maximum likelihood may be used to estimate the parameters, (2) the estimated coefficients or transformations of them can provide clinically meaningful estimates of effect, (3) fitted values from the model can provide estimates of survival time, and (4) residuals can be computed as differences between observed and predicted values of time. An analysis of censored time-to-event data using a fully parametric model can almost have the look and feel of a normal-errors linear regression analysis.

- **L17-L20 Chapters 3,4 Proportional hazards regression model (or Cox model); Fitting, Interpretation, Model diagnostic**

  The Cox proportional-hazards model (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables. Cox's semi-parametric proportional hazards model is more widely used than fully parametric models such as AFT model.

Project #2: Survival data analysis II (L21-L22, Choose either Project #2 or Project #3)

- **L23-L24 Generation of survival times for simulation studies**

  We discuss techniques to generate survival times for simulation studies regarding Cox proportional hazards models. Generation of survival times is not currently offered by most statistical software. Simulation study is an invaluable tool for statistical research, particularly for the evaluation of new methods and for the comparison of alternative methods. In the Cox model, which is formulated based on the hazard function, the effect of the covariates has to be translated from the hazards to the survival times.

- **L25-L26 High-dimensional variable selection with survival outcome**

Project #3: High-dimensional survival data analysis (L27-L28, Choose either Project #2 or Project #3)

Modern biomedical technology has led to a wealth of data in which the number of variables (e.g. features) exceeds the number of observations (e.g. patients). In this case, one might be interested in (i) identifying variables that are associated with survival, and (ii) developing a regression model for predicting survival when a new observation is available. Due to the high dimensionality, most classical statistical methods (covered in L15-L20) for survival analysis cannot be applied directly. We review a number of methods from the literature that address these problems.

Homework/Project due dates

Submission should be made to

D2L: Assessments>Assignment>…

If you complete an assignment on paper, you will need to scan the documents to a single PDF file. It is important to make sure the scan of your work is clear and legible. Improperly formatted or illegible scans may not be accepted. If you use a phone, use a scanning app such as Microsoft Lens, Scannable (iOS), or Genius Scan and not just the camera.

For any R related assignments, it is required to submit homework as both RMD and PDF (or MS Word) file by the due date of the assignment. All electronic submissions should follow the following naming convention: last name, first name, assignment number, and proper extension. So, for example, if Matthew Stacy is turning in Homework 1, he would name the file Matthew_ Stacy _HW1.pdf. The associated code would be Matthew _ Stacy _HW1.Rmd. If you wish to break up your code into separate files, you may submit them as Matthew_ Stacy _HW1a.Rmd, Matthew_ Stacy _HW1b.Rmd, and so on. There will be a 20% penalty per day that your homework is late.

- HW1: 9/18 (L2-L3)
- HW2: 10/2 (L4-L8)
- HW3: 10/9 (L9-L10)
- **Project #1: 10/16 (L11-L12)**

- HW4: 10/30 (L13-L16)
- HW5: 11/13 (L17-L20)
- **Project #2: 11/16-11/20 (L21-L22) (Choose either Project #2 or Project #3, 15 minutes zoom presentation)**

- HW6: 12/4 (L23-L26)
- **Project #3 12/7-12/11 (L27-L28) (Choose either Project #2 or Project #3, 15 minutes zoom presentation)**

Prerequisite

Calculus through multivariate calculus and basic knowledge of regression methods, statistical methods for estimation and inferences.

Text: Applied survival analysis: Regression modeling of time-to-event data, second edition by David W. Hosmer, Stanley Lemeshow, and Susanne May. ISBN: 978-0-471-75499-2

Optional Useful Text:

- David G. Kleinbaum and Mitchel Klein. *Survival Analysis*
- Lawless, J.F. (2002). *Statistical Models and Methods for Lifetime Data*, 2nd Edition, New York: Wiley.
- J.P. Klein, and Moeschbeger, M.L. (2003). Survival Analysis: Techniques for Censored and Truncated Data, 2nd Edition, New York: Springer.

Grades:

- Homework (60%)
  - There will be homework approximately every two weeks
  - Students are encouraged to work in groups, but are required to write up their own individual solutions
- Project 1 (20%)
  - More information will be provided the two weeks before the Project 1 due date.
- Project 2 or Project 3 (20%) 15 minutes Zoom presentation will be scheduled for each student.
  - More information will be provided the two weeks before the Project 2,3 due date.

Score (5)	<50	50-59	60-69	70-74	75-79	80-84	85-89	≥90
Grade	0	1	1.5	2	2.5	3	3.5	4

Academic Honesty: The Departments of Statistics & Probability adheres to the policies of academic honesty as specified in the General Student Regulations 1.0, Protection of Scholarships and Grades, and in the All-University of Integrity of scholarship and Grades which are included in Spartan Life: Student Handbook and Resource Guide. Students who plagiarize will receive a grade 0.0 on the homework or project.

Withdrawal: Deadlines for withdrawal from courses are published in each semester's course catalog. A faculty member cannot drop or withdraw a student. It is the student's responsibility to handle withdrawal procedures from any class to avoid receiving a grade of "F".

Incomplete grades

As per university policy, incomplete grades are granted only in the case of work unavoidably missed (and excused) and not already covered by the professor's policy on missed work or activities, and only if at least 70% of the course work has been completed. An incomplete grade must be resolved within eight weeks from the first day of the subsequent long semester. If the required work to complete the course and to remove the incomplete grade is not submitted by the specified deadline, the incomplete grade becomes changed automatically to F.

ADA: To arrange for accommodation a student should contact the Resource Center for People with Disabilities at http://www.rcpd.msu.edu/ or (517) 353-9642

Disclaimer: Changes on the syllabus/important dates will be announced in class and on the course web site. It is students' responsibility to keep up with any changed policies and assignments.