Lifetime Data Science: Foundations and Frontiers

Working Version: January 8, 2019

Details for Short Courses

Workshop: Details

Wednesday May 29, 8:30–16:30 Workshop I Building Location

Workshop I

8:30–16:30 **Ørnulf Borgan** (University of Oslo, Norway) and
Sven Ove Samuelsen (University of Oslo, Norway)
Two Phase Studies for Lifetime Data

Wednesday May 29, 8:30–16:30 Workshop II Building Location

Workshop II

8:30–16:30 **Hein Putter** (Leiden University Medical Centre, The Netherlands)
Dynamic Prediction in Survival Analysis

Wednesday May 29, 8:30–16:30 Workshop III Building Location

Workshop III

8:30–16:30 **Jing Qin** (NIH/NIAID)
Biased Sampling, Left Truncation and Survival Analysis

Biased Sampling, Left Truncation and Survival Analysis

Jing Qin (NIH/NIAID)

Overview

Biased sampling occurs when a proper randomization cannot be achieved, the observed sample will not be representative of the population of interest. Biased sampling problems appear in many areas of research, including, Medicine, Epidemiology and Public Health, Social Sciences and Economics. Left truncation and length-biased data are clearly encountered in applications of renewal processes, etiologic studies, genome-wide linkage studies, epidemiologic cohort studies, cancer prevention trials, and studies of labor economy. In observational studies, a prevalent cohort design that draws samples from individuals with a condition or disease at the time of enrollment is generally more efficient and practical. The recruited patients who have already experienced an initiating event are followed prospectively for the failure event (e.g. disease progression or death) or are right censored. Under this sampling design, individuals with longer survival times measured from the onset of the disease are more likely to be included in the cohort, whereas those with shorter survival times are unconsciously excluded. Finding appropriate adjustments for the potential selection bias in analyzing length-biased data or more general biased sampling problems has been a long standing statistical problem.

This workshop discusses various methods to deal with biased sampling problems, exponential tilting models and left truncation and right censored data problems, including prole maximum likelihood method, conditioning likelihood method, composite partial likelihood method as well as general imputation methods.

Aims and Topics Covered

1. Discuss the general methods for handling biased sampling problems, including case and control problems, missing data and casual inference.

2. Derive the Cox partial likelihood from different angles, including rank likelihood method, prole maximum likelihood method, case and control conditional likelihood argument, and optimal estimating equation method.

3. Present the latest results on analyzing length biased survival time data, including Vardi's multiplicative censoring problem, and general imputation principle for missing survival data.

4. Discuss the well known pool adjacent violators algorithm and its combination with the EM algorithm together for estimating shape constrained inference, including estimation of monotonic decreasing density, cumulation hazard or distribution function based on current status data etc.

Learning Outcomes

At the end of the day participants should have some new ideas on handling biased sampling problems and survival data. This course is particularly helpful for those who are interested in learning some "theoretical results" and some "applied problems". The accompanied R programs will be discussed.

Background for the Instructor

Jing Qin is a Mathematical Statistician at the Biostatistics Research Branch in the National Institute of Allergy and Infectious Diseases. Dr. Qin's research interests include empirical likelihood method, case-control study, length bias sampling, econometrics, survival analysis, missing data, causal inference, genetic mixture models, generalized linear models, survey sampling and microarray data analysis. He is the author of "Biased sampling, over-identied parametric problems and beyond" (Springer, 2017). He was elected as a Fellow of the American Statistical Association in 2006.

Dynamic Prediction in Survival Analysis

Hein Putter (Leiden University Medical Center, The Netherlands)

Summary

The medical literature abounds with prediction models. They are statistical models based on patient- and disease characteristics, used to inform treatment decisions, to provide personalized risk estimates for the patient, and also to stratify patients in clinical trials. Important prognostic models include Adjuvant! Online in cancer and the Framingham risk score in cardiovascular disease. The vast majority of these models are focused on prognosis at one well-defined baseline moment, typically at diagnosis, shortly before treatment is initiated. It is at this time that the most important decisions on primary treatment are made. There is little doubt that the available prognostic models are important tools for the treating physician to guide treatment decisions at diagnosis. However, once primary treatment has been initiated, the prognosis of the patient will change over the course of time, as a result of the effect of treatment, possible treatment toxicity, and clinical events such as disease recurrence that may have occurred, and, very simply, because of the fact that the patient is still alive. As a result, these prediction models need to be "updated" to use the knowledge that has become available since baseline. Prediction models that incorporate this dynamic aspect are called dynamic prediction models, and they are the topic of this course.

This course will focus on methodology for dynamic prediction. The dynamic aspect of dynamic prediction involves using information on events and/or measurements up to the present, in order to "update" the prediction. It will be shown in this course how dynamic predictions may be obtained using the concept of landmarking and using multi-state models. Analyses will be illustrated using R, in particular the `mstate` and `dynpred` packages. Implementation of the methods in other statistical software packages like SAS, Stata and SPSS will be discussed.

Aims

1. Discuss situations where dynamic prediction is relevant;
2. Illustrate how the Cox model can be used to obtain dynamic predictions with time-fixed covariates;
3. Introduce multi-state models as an extension of survival analysis and competing risks;
4. Show how multi-state models can be used to obtain dynamic predictions;
5. Introduce landmarking as a way of dealing with time-dependent covariates;
6. Show how landmarking can be used to include time-dependent information in the dynamic predictions;
7. Discuss robustness properties;
8. Illustrate how to carry out the analyses discussed during the course using R.

Learning Outcomes

At the end of the course participants should:

1. Understand the connection between hazards and dynamic prediction probabilities;

2. Know how to obtain dynamic prediction probabilities from time-fixed Cox models;
3. Understand the difficulties of predicting with time-dependent covariates;
4. Be acquainted with concepts in multi-state models like transition intensities, transition probabilities, state occupation probabilities, the Markov assumption;
5. Understand the relation between transition intensities and transition probabilities, and be acquainted with the Aalen-Johansen estimator;
6. Understand how landmarking can be used for dynamic prediction.

Topics Covered

The course material will be presented in a lecture format, changing between theory and illustrations. Ample attention will be devoted to the practical implementation of the methods covered in the course, using R.

Topics covered include:

- *Dynamic use of familiar survival analysis techniques*
  A short overview of survival analysis will be given, including the Cox model. The emphasis in this overview will be on how these familiar techniques can be used to obtain dynamic predictions. We will introduce conditional survival (the effect of being alive) and the fixed width failure function, and their relation to the familiar hazard function. Extensions to competing risks will briefly be mentioned.
- *Time-dependent covariates and landmarking*
  We will then introduce time-dependent covariates and discuss techniques to handle them such as time-dependent Cox regression and landmarking. The differences between these approaches and the relative merits will be discussed.
- *Multi-state models*
  A brief overview of multi-state models will be given, including how they can be used to obtain dynamic predictions. The overview includes discussion of concepts like transition intensities and transition probabilities, and ways of estimating transition intensities. The Aalen-Johansen estimator of the transition probabilities will be presented, and the assumptions needed for validity of the Aalen-Johansen estimator, in particular the Markov assumption will be discussed.
- *Landmarking and dynamic prediction*
  Then we will show how landmarking can be used to include time-dependent information in the dynamic predictions. We will briefly discuss more traditional methods that can also be used for dynamic prediction, such as multi-state models. Advantages and disadvantages of different approaches will be discussed.
- *Practical implementation*
  Methods discussed during the lectures will be illustrated using R, and in particular the `mstate` and `dynpred` packages. Data used is available from the presenter upon request.

Learning Strategy

The material will be presented using slides and through class discussion. Attendees will be given a booklet containing the slides , which will contain clear descriptions of the methodology, of applications, and of how to implement analyses in R.

Pre-requisites

This course is directed at statisticians or epidemiologists in academia, government or industry interested in dynamic prediction in survival analysis. Participants are expected to have a fair knowledge of the techniques from classical survival analysis.

About the Instructor

Hein Putter is Professor at the Leiden University Medical Center (Department of Biomedical Data Sciences). His research interests include competing risks and multi-state models, frailty models and dynamic prediction. He is co-author of the book “Dynamic Prediction in Clinical Survival Analysis”, with Hans van Houwelingen.

Two-Phase Studies for Lifetime Data

Ørnulf Borgan (University of Oslo Norway)
Sven Ove Samuelsen (University of Oslo, Norway)

Overview

In cohort studies, regression methods are commonly applied to assess the influence of risk factors and other covariates on mortality or morbidity; in particular Cox-regression is much used. Estimation in Coxs model is based on a partial likelihood that at each observed death or disease occurrence ("failure") compares the covariate values of the failing individual to those of all individuals at risk. Thus Cox regression requires collection of covariate information for all individuals in the cohort, even when only a small fraction of them actually get diseased or die. This may be very expensive, or even logistically impossible. Further, when covariate measurements are based on biological material stored in biobanks, it will imply a waste of valuable material that one may want to save for future studies. Cohort sampling designs, where covariate information is collected for all failing individuals ("cases"), but only for a sample of the individuals who do not fail ("controls"), then offer useful alternatives that may save biological material and drastically reduce the workload of data collection and error checking. Such cohort sampling designs may be considered as two-phase designs, where the cohort is the phase I sample (selected from a superpopulation) and the case-control sample is the phase II sample selected from the cohort.

There are two main types of two-phase designs for life time data: nested case-control and case-cohort designs, and the two types of designs differ in the way controls are selected. The course presents the two types of designs both in their original form and later extensions and describes how the statistical analysis of such two-phase studies may be performed. The focus is on estimation of relative risks using partial likelihoods and pseudo-likelihoods (or weighted likelihoods) that resemble the full cohort partial likelihood. Other topics like estimation of absolute risk and model checking will also be discussed, and methods that use all available data in the full cohort will be mentioned. There will be practical exercises in analyzing two-phase life time data, and the participants should bring their own laptop with R installed. Information on R packages that are needed will be given closer to the course.

Aims

1. Introduce the most common two-phase designs for life time data: nested case-control and case-cohort.
2. Discuss classical statistical methods for estimating relative risks for two-phase life time data, and give an outline of their theoretical properties.
3. Discuss methods for absolute risk estimation and model assessment.
4. Describe two-phase methods that use all available data from the full cohort.
5. Illustrate how to carry out statistical analyses of two-phase life time data using R.

Learning Outcomes

At the end of the day participants should:

1. Know the characteristics of the two common types of two-phase designs for life time data and understand the pros and cons of the designs.

2. Know how to estimate relative and absolute risks from nested case-control and case-cohort data.
3. Have some knowledge of methods that make use of data that are available for the full cohort.
4. Have some experience in analyzing nested case-control and case-cohort data using R.

Topics Covered

The material will be presented in a lecture format, where the theory and methods will be motivated and illustrated by examples from health research. In addition, the participants will get hands-on experience with the methods from practical exercises using R. Topics covered include:

1. Summary of methods for analyzing cohort life time data.
2. Nested case-control designs, including counter-matched sampling of the controls.
3. Case-cohort designs, including stratified sampling of the subcohort.
4. Classical methods for estimating relative and absolute risk from nested case-control and case-cohort data.
5. Analysis of general models for nested case-control and case-cohort data using inverse probability weighting.
6. Calibration of inverse probability weights for case-cohort data.
7. Methods for two-phase data that use all available cohort information (multiple imputation and maximum likelihood).
8. Practical examples and exercises using R.

Learning Strategy

The material will be presented using slides, class discussion, and practical exercises using R. Attendees will be given a booklet containing the slides, which will contain clear descriptions of the methodology, of applications, and of how to implement analyses in R.

Pre-requisites

The short course will be directed at statisticians in academia, government or industry interested in learning about two phase designs for life time data. It will be assumed that the participants are familiar with the basic concepts and methods in survival analysis and that they have some experience in using the R software.

Recommended reading

- Chapters 7 and 8 of Keogh & Cox: *Case-Control Studies*, Cambridge University Press, 2014.
- Part IV of *Handbook of Statistical Methods for Case-Control Studies*, eds Borgan, Breslow, Chatterjee, Gail, Scott & Wild, CRC Press, 2018.

About the Instructors

Ørnulf Borgan is professor of Statistics at the University of Oslo. His main research interest has been statistical methods for survival and event history data, including nested case-control and case-cohort designs. He is co-author of two books on the use of counting processes and martingales in survival and event history analysis, and he is one of the editors of the recent *Handbook of Statistical Methods for Case-Control Studies* (CRC Press, 2018). Borgan has been editor of the Scandinavian Journal of Statistics, and he is a Fellow of the American Statistical Association and member of the Norwegian Academy of Science and Letters.

Sven Ove Samuelsen is professor of Statistics at the University of Oslo. His main research interest has been statistical methods for survival and event history data, in particular case-cohort and nested case-control designs. He has been involved in planning and analyzing many case-control and other epidemiological studies. Samuelsen is on the editorial board of Lifetime Data Analysis.