ICSA Book Series in Statistics
Series Editor: Ding-Geng (Din) Chen

Jianguo Sun
Ding-Geng Chen *Editors*

Emerging Topics in Modeling Interval-Censored Survival Data

Springer

ICSA Book Series in Statistics

Series Editor

Ding-Geng (Din) Chen, College of Health Solutions, Arizona State University, Chapel Hill, NC, USA

The ICSA Book Series in Statistics showcases research from the International Chinese Statistical Association that has an international reach. It publishes books in statistical theory, applications, and statistical education. All books are associated with the ICSA or are authored by invited contributors. Books may be monographs, edited volumes, textbooks and proceedings.

Jianguo Sun • Ding-Geng Chen
Editors

Emerging Topics in Modeling Interval-Censored Survival Data

Springer

Editors
Jianguo Sun
Department of Statistics
University of Missouri
Columbia, MO, USA

Ding-Geng Chen
College of Health Solutions
Arizona State University
Goodyear, AZ, USA

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-3-031-12365-8 ISBN 978-3-031-12366-5 (eBook)
https://doi.org/10.1007/978-3-031-12366-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is intended primarily to discuss the emerging topics in statistical methods for interval-censored survival data. This book is prepared for booster research, education, and training to advance statistical modeling in interval-censored data, which are commonly collected from public health and biomedical research. However, this type of data can be easily mistaken for typical right-censored data that would result in erroneous statistical inference due to the complexity of this type of data. This book is then constructed to invite a group of nationally and internationally leading researchers to systematically discuss and explore the historical development of the associated methods, their computational implementations, and some newly emerging topics related to interval-censored data. We aim to cover a variety of topics, including univariate interval-censored, multivariate interval-censored, clustered interval-censored, and competing risk interval-censored data, data with interval-censored covariates, interval-censored data in electric medical records, and misclassified interval-censored data. We invited a group of leading experts at the forefront of modeling interval-censored survival data to prepare book chapters, and received many excellent papers on this topic. Fifteen high-quality chapters are included in this wonderful book. Each chapter has been peer reviewed by two editors and revised several times before final acceptance. Therefore, this volume reflects new advances in statistical methods for interval-censored survival data analysis across biostatistics and interdisciplinary areas. This book has the potential to have a significant impact on survival data analysis as both an authoritative source and reference, as it will identify new directions of interval-censored survival data modeling using modern statistical methods. This book will appeal to statisticians, biostatisticians, health-related researchers, graduate students, etc. This book will aid researchers, students, and practitioners on the leading edge of research methods enabling them to tackle problems in research, education, training, and consultation.

This book is organized into three parts. Part I includes three chapters, which present an overview of historical development as well as recent topics in interval-censored modeling. Part II consists of six chapters on emerging topics in methodological development, and Part III is composed of six chapters that present emerging topics in real-life applications of interval-censored data and analysis. All the

chapters are organized as self-contained units with references at the end of each chapter. To aid the reader in recreating these techniques, all the statistical procedures in practice, computer programs, and datasets are included or referenced in this book. The readers may request guidance from the chapter authors to facilitate statistical approaches.

Part I: Introduction and Review (Chapters 1–3)

In first chapter, "Overview of historic development in modeling interval-censored survival data," Dr. Finkelstein presents an overview of the historical development of the methods for the analysis of interval-censored survival data. It begins with a description of how the interval-censored data arise from studies where the subjects are followed periodically, and the time to the event of interest cannot be observed exactly. From a historical perspective, such data became more common with the emergence of new clinical and epidemiological study designs. However, the well-developed methods that are used for right-censored survival data analysis could not be applied. This chapter follows the methodological development of this area, starting in 1970 from a broad historical perspective.

In second chapter, "Overview of recent advances on the analysis of interval-censored failure time data," Dr. Du provides a review of recent advances on several topics related to regression analysis of interval-censored data, mainly from the last five to seven years. These topics include the analysis of univariate interval-censored data with time-varying covariates in the presence of a cured subgroup and in the presence of informative interval censoring, respectively. This chapter discusses some recent advances in the analysis of interval-censored data arising from case-cohort studies and the variable selection based on interval-censored data. Furthermore, some recent work on regression analysis of multivariate interval-censored data is described as well as regression analysis of doubly censored data.

In third chapter, "Predictive accuracy of prediction models for interval-censored data," Dr. Kim proposes a prognostic tool based on survival model to assist in predicting the occurrence of a clinical event, defining better prescription, and assessing cost-effectiveness. In this chapter, she comprehensively reviews several recently proposed time-varying prognostic tools for interval-censored data. A classification index including time-dependent receiver operating characteristic (ROC), time-dependent concordance index, and calibrations such as Brier score and integrated BRIR score have been adopted in the context of interval-censored data. A risk score defined as either a single biomarker or a risk probability combined with potential predictors can have time-varying values. She has also included longitudinal risk scores to illustrate methods using a set of dementia datasets.

Part II: Emerging Topics in Methodology (Chapters 4–9)

In fourth chapter, "A practical guide to exact confidence intervals for a distribution of current status data using the binomial approach," Drs. Kim, Fay, and Proschan consider the construction of pointwise confidence intervals for the distribution of the failure time of interest based on current status data. In particular, they discuss two methods recently developed by the authors using the binomial approach and compare them to other methods developed with the use of the asymptotic approach. One advantage of the methods based on the binomial approach is that they apply to both continuous and discrete assessment distributions. In addition, the related R package **csci** and R codes used are discussed and provided.

In fifth chapter, "Accelerated hazards model and its extension for interval-censored data," Dr. Xiang discusses the analysis of interval-censored data under the accelerated hazards model and their generalizations. In particular, a generalized accelerated hazard mixture cure model is presented for situations where there exists a subgroup of cured subjects. For example, she investigates the use of sieve maximum likelihood estimation approach based on spline functions. She provides extensive simulation results and two real data applications in this chapter.

In sixth chapter, "Maximum likelihood estimation of semiparametric regression models with interval-censored data," Drs. Lin and Zeng consider regression analysis of interval-censored data with time-dependent covariates under the semiparametric Cox proportional-hazards model. For example, the nonparametric maximum likelihood estimation approach was developed that treats the unknown cumulative hazard function to be a step function and a simple and stable EM algorithm based on Poisson latent variables was provided. Furthermore, the methodology was generalized to competing risks interval-censored data as well as multivariate or clustered interval-censored data.

In seventh chapter, "Use of the INLA approach for the analysis of interval-censored data," Drs. van Niekerk and Rue present the integrated nested Laplace approximation (INLA) methodology for interval-censored data. Most survival models, including those with interval censoring, can be shown to be a latent-Gaussian model and as such INLA can be used for near real-time Bayesian inference. They provide a brief summary of the INLA methodology and illustrate the approach on real data examples with interval censoring, including a joint model. The analysis is done using the R package INLA and all code is available for reproducibility.

In eighth chapter, "Copula models and diagnostics for multivariate interval-censored data," Drs. Ding and Sun discuss the use of the copula model-based approach for regression analysis of multivariate interval-censored data and the goodness-of-fit test for the assumed copula model with a focus on bivariate interval-censored data. On the regression analysis, a class of flexible semiparametric transformation models was employed to describe covariate effects and a sieve maximum likelihood estimation approach was developed for inference. To test the assumed copula model, they introduce a general goodness-of-fit test procedure based on the information ratio this method applies to any copula family with a

parametric form. Finally, the authors discuss the R package **CopulaCenR** for the implementation of the presented methods and illustrate it through two sets of real multivariate interval-censored data.

In ninth chapter, "Efficient estimation of the additive risks model for interval-censored data," Drs. Wang, Bandyopadhyay, and Sinha discuss the fitting of the semiparametric additive risks model to interval-censored data. Under the case-II interval censoring scenario, in contrast to the commonly used EM algorithm, the authors presented a minorize-maximize (MM) algorithm for nonparametric maximum likelihood estimators of both nonparametric and finite-dimensional components of the model. The method applies to both time-independent and time-varying covariates and has the advantage of allowing separate maximization over the nonparametric and finite components, thus yielding a stable and fast computation process. The operating characteristics of the proposed MM approach are assessed via simulation studies and a corresponding R package, **MMIntAdd,** is provided and illustrated through a set of real data.

Part III: Emerging Topics in Applications (Chapters 10–15)

In tenth chapter, "Modeling and analysis of chronic disease processes under intermittent observation," Drs. Cook and Lawless describe independence conditions needed for valid likelihood-based inference about multistate disease processes under intermittent observation schemes. They further describe how joint models for disease and observation processes can be used to address disease-related clinic visits and how joint models can be used to deal with internal time-dependent markers when marker values are observed only at clinic visits. They also investigate the limiting values of regression coefficients of marker effects when the common approach of carrying forward the most recently recorded value is used.

In eleventh chapter, "Case-cohort studies with time-dependent covariates and interval-censored outcome," Drs. Gao, Hudgens, and Zou provide an inverse probability weighting likelihood approach for fitting a parametric model to interval-censored data with both fixed and time-dependent covariates arising from case-cohort studies. The method is a generalization of that given in Sparling et al. (2006) for usual interval-censored data with time-dependent covariates. Simulation results demonstrated that the proposed estimator is approximately unbiased and the standard errors are well estimated from the sandwich estimators. The method was applied to an observational study that examined the association between hormonal contraceptive use and the risk of HIV acquisition.

In twelfth chapter, "The **BivarIntCensored**: An R package for nonparametric inference of bivariate interval-censored data," Drs. Zhou, Wu, and Zhang consider nonparametric estimation of a bivariate cumulative distribution function based on bivariate interval-censored data. After reviewing two existing sieve nonparametric maximum likelihood estimation approaches, they present and discuss the use of an R package, **BivarIntCensored,** which implements the two estimation procedures.

In the methods, B- or I-spline functions were used. In addition, an association test is provided and discussed.

In thirteenth chapter, "Joint modeling for longitudinal and interval-censored survival data: application to IMPI multi-center HIV/AIDS clinical trial," Drs. Chen and Singini discuss the joint models for longitudinal and interval-censored survival data using a cardiology multi-center clinical trial with the illustration of R statistical software.

In fourteenth chapter, "Regression with interval-censored covariates: application to liquid chromatography," motivated by the data from the metabolomic analysis area, Drs. Melis, Marhuenda-Muñoz, and Langohr discuss the analysis of generalized linear models when there exists a covariate that suffers interval censoring. They use an extension of the method from the linear regression model given in Gómez, Espinal, and Lagakos (2003) to accommodate non-normal responses belonging to an exponential family. In addition, they discuss two goodness-of-fit measures that accommodate interval-censored covariates and apply the methods to determine the association between glucose, a completely observed response variable, and the sum of carotenoids, an interval-censored explanatory variable. The implementation of the discussed methods in R is also discussed.

In fifteenth chapter, "Misclassification simulation extrapolation procedure for interval-censored log-logistic accelerated failure time model," Drs. Sevilimedu, Yu, Chen, and Lio discuss the misclassification of binary covariates since it often occurs in survival data. Any survival data analysis ignoring such misclassification will result in estimation bias. To handle such misclassification, the misclassification simulation extrapolation (MC-SIMEX) procedure is a flexible method proposed in survival data analysis, which has been investigated extensively for right-censored survival data. However, the performance of the MC-SIMEX method has not been explored much for interval-censored survival data. This chapter is then aimed at investigating the performance of the MC-SIMEX procedure with interval-censored survival data through Monte-Carlo simulations and real data analysis. They focus this investigation on the log-logistic accelerated failure time (AFT) model since the log-logistic distribution plays an important role in evaluating non-monotonic hazards for survival data.

We sincerely thank all of the people who have given us strong support for the publication of this book on time. Our acknowledgments go to all the chapter authors (in the "List of Contributors") for submitting their excellent works to this book. We also thank Ms. Anne Rubio at the College of Health Solutions, Arizona State University, and Ms. Jenny K. Chen at Morgan-Stanley Wealth Management, for their professional editing of this book, which has substantially improved the quality of the chapters and the entire book. Furthermore, we are so grateful to Dr. Eva Hiripi and Ms. Faith Su (Statistics Editors, Springer Nature) from Springer and Kirthika Selvaraju (Project Coordinator of Books, Springer Nature) for their full support during the long publication process. In addition, this book was made possible through funding provided by DST-NRF-SAMRC-SARChI Research Chair in Biostatistics, Grant number: 114613.

We look forward to receiving the comments about the book from our readers. If the readers have any suggestions about further improvements to the book, please contact us: Drs. Sun and Chen by email.

Columbia, MO, USA Jianguo Sun
Phoenix, AZ, USA Ding-Geng Chen
Pretoria, South Africa

Contents

Editors and Contributors

About the Editors

Jianguo Sun is Curator's Distinguished Professor in the Department of Statistics at the University of Missouri, USA. He is a world-leading researcher in survival data analysis and has been working on the analysis of interval-censored data, the topic of this book, for over 30 years. He has published over 200 papers and three books and has been invited to write review articles on the analysis of interval-censored failure time several times. Professor Sun is a fellow of the American Statistical Association and Institute of Mathematical Statistics and an elected member of the International Statistical Institute.

Ding-Geng Chen received his Ph.D. in Statistics from the University of Guelph (Canada) in 1995 and is now the executive director and professor of biostatistics at the College of Health Solutions, Arizona State University. He served as a professor in biostatistics at the University of North Carolina-Chapel Hill, a biostatistics professor at the University of Rochester Medical Center, and the Karl E. Peace endowed eminent scholar chair in biostatistics from the Jiann-Ping Hsu College of Public Health at Georgia Southern University. Dr. Chen is an elected fellow of the American Statistical Association and a senior expert consultant for biopharmaceutical and government agencies with

extensive expertise in clinical trial biostatistics. Dr. Chen has more than 200 scientific publications and has co-authored/co-edited 33 books on clinical trials, survival data, meta-analysis, Monte-Carlo simulation-based statistical modeling, and statistical modeling for public health applications. His research has been funded as PI/Co-PI from NIH R01s and other governmental agencies.

Contributors

Dipankar Bandyopadhyay Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Richard J. Cook Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Ying Ding Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

Mingyue Du School of Mathematics, Jilin University, Changchun, China

Michael P. Fay National Institute of Allergy and Infectious Diseases, Rockville, MD, USA

Dianne Finkelstein Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

Xiaoming Gao Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Michael G. Hudgens Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Sungwook Kim University of the Sciences in Philadelphia, Philadelphia, PA, USA

Yang-Jin Kim Department of Statistics, Sookmyung Women's University, Seoul, Korea

Klaus Langohr Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Jerald F. Lawless Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Danyu Lin Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Yuhlong Lio Department of Mathematical Science, University of South Dakota, Vermillion, SD, USA

Maria Marhuenda Department of Nutrition, Food Sciences and Gastronomy, University of Barcelona, Barcelona, Spain

Guadalupe Gómez Melis Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Michael A. Proschan National Institute of Allergy and Infectious Diseases, Rockville, MD, USA

Havard Rue King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

Varadan Sevilimedu Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Issac L. Singini Department of Statistics, University of Pretoria, Pretoria, South Africa

Samiran Sinha Department of Statistics, Texas A&M University, College Station, TX, USA

Tao Sun Center for Applied Statistics and School of Statistics, Renmin University, Beijing, China

Janet van Niekerk King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

Tong Wang School of Statistics and Data Science, Nankai University, Tianjin, ChinaDepartment of Statistics, Texas A&M University, College Station, TX, USA

Yuan Wu Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

Liming Xiang School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore

Lili Yu Department of Biostatistics, Georgia Southern University, Statesboro, GA, USA

Donglin Zeng Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Ying Zhang Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE, USA

Junyi Zhou Amgen Inc., Thousand Oaks, CA, USA

Fei Zou Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Part I
Introduction and Review

Overview of Historical Developments in Modeling Interval-Censored Survival Data

Dianne Finkelstein

Abstract This chapter serves as an overview of the historical development of the methods for the analysis of interval-censored survival data. It begins with a description of how the interval-censored data arise from the studies where the subjects are followed periodically, and the time to the event of interest cannot be observed exactly. From a historical perspective, such data become more common with the emergence of new clinical and epidemiologic study designs. However, the well-developed methods that are used for right-censored survival data analysis could not be applied. This chapter follows on the methodological development of this area, starting in 1970 from a broad historical perspective.

Keywords Clinical trials · Interval-censored data · Non-parametric estimation · Periodical follow-up · Semiparametric analysis

1 Emerging Interval-Censored Data

In survival analysis, we are interested in the time to an event, such as death or the onset of disease. Interval-censored data arise when the event is not observed exactly and instead is only known to have occurred within a window of time. Such data can be encountered in studies where subjects are followed up periodically, such as in a clinical trial or longitudinal observational study. For example, suppose that a survey is performed at regular intervals (say annually) and the observations are the age of the participants and whether they have experienced a specific outcome (such as the onset of puberty). Some subjects may miss observations and return with a change in status. Thus, the data from the study will consist of overlapping intervals of age at

D. Finkelstein (✉)
Harvard Medical School, Boston, MA, USA

Biostatistics Center, Massachusetts General Hospital, Harvard University, Boston, MA, USA
e-mail: dfinkelstein@mgh.harvard.edu

J. Sun, D.-G. Chen (eds.), *Emerging Topics in Modeling Interval-Censored Survival Data*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-031-12366-5_1

the time of completed surveys and whether or not the event has occurred within the interval.

The methods that have been developed for right-censored and exact data may not be directly applied for the analysis of interval-censored data. It is possible to broadly group the data and apply methods for discrete data, but some loss of information may occur since only summary information is used in this way (Sun, 2006). This is especially the case if wide grouping intervals are used.

2 Emerging Methods in Analyzing Interval-Censored Data

The methods and theory for the analysis of interval-censored data emerged in the statistical literature at least five decades ago, but were not widely applied until the 1990s, when these data became more common with the arrival of new diseases such as HIV/AIDS. Also, new technology for the diagnosis and monitoring of diseases, such as imaging studies and laboratory-detected markers, resulted in the discovery of a clinical event only at the time the clinical test was done.

An early work by Sir Richard Peto (Peto, 1973) described estimation of the survival curve from interval-censored data in the context of the analysis of the onset of puberty from a study using annual surveys in New Guinea (https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2346307). In 1976, Bruce Turnbull (Turnbull, 1976) published a paper on the empirical distribution function for arbitrarily grouped and truncated data. He noted that interval-censored data can arise in medical or correctional follow-up or industrial life-testing settings, where subjects enter the data set at different ages. He also noted that the bioassay problem of Ayer et al. (1955) resulting in right- and left-censored data could be considered a special case of interval-censored data. Finkelstein and Wolfe (1985) published a paper proposing a test for interval-censored data, which included a full data set from a breast cancer study with a treatment difference in time to cosmetic deterioration following radiation therapy. The data were censored into an interval between clinic visits. These data were applied in many subsequent studies, contributing to the 322 citations in Google Scholar. In 1986, Finkelstein (1986) proposed a proportional hazards model for the analysis of interval-censored data, which was cited 774 times according to Google Scholar. These papers are really the seminal works in the analysis of interval-censored data with Turnbull, for example, having over 2200 citations noted in Google Scholar.

While only two papers had the keywords “interval-censored data” between 1900 and 1970, and 12 additional in the 1970s, followed by 25 in 1980–85, the work in this field grew rapidly as shown in Fig. 1, with over 2700 references in just the last five years alone. The growth of this field was likely partially in response to the new data that arose in the medical field, and the first book on the topic that was published by Sun in 2006, which provides a relatively comprehensive review of the literature up to 2006. The AIDS epidemic, in particular, produced a wealth of new data that were interval-censored. For example, the incubation period of the disease

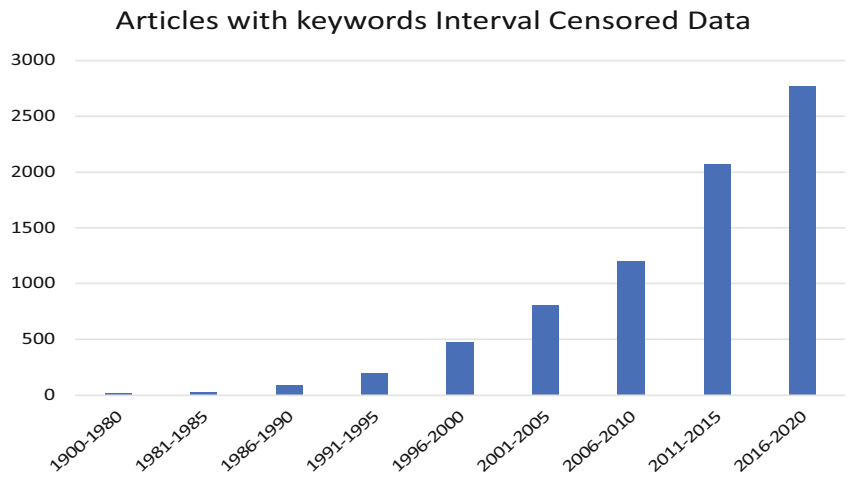

Fig. 1 The number of published articles with key words "Interval-Censored Data"

was the time between the last time a subject tested HIV negative and the first time he or she tested HIV positive (De Gruttola & Lagakos, 1989). Similarly, in cancer, where preventing disease spread was the primary evidence of treatment efficacy, the patient was monitored by newer imaging and laboratory technology, and the time of progression was thus censored into the interval between clinic visits in which the event occurred, often before there were symptoms reported by the patient.

3 More on Emerging Methods in Analyzing Interval-Censored Data

The focus of the methods for interval-censored data has been on non-parametric or semi-parametric methods. Before this field became widely known, it was common practice in medical studies to simplify the interval censoring structure of the data into a more standard right-censoring situation by, for instance, imputing the midpoint of the censoring interval. The availability of software for the analysis of right-censored and exact data could have contributed to this practice. The software for the analysis of interval-censored data lagged behind the methodology. Prior to 2000, it was challenging to find professional software that provided these methods to handle interval-censored data. Consequently, often times the data that were interval-censored were masked as grouped or right-censored in the presentation, as the methodology was not widely available or understood.

Each observation of interval-censored data is represented by the two endpoints of the time interval in which the subject's event occurred. We note that standard right-censored and exact data can be seen as a subset of interval-censored data in which

the right endpoint is infinity (if the observation time is right-censored) and the two endpoints are equal if the event is observed exactly. Similarly, cross-sectional data can be viewed as left- and right-censored depending on whether the event has or has not already occurred, and is also a subset of interval-censored data, where one of the endpoints is negative or positive infinity. The book by Sun (2006) provides several complete interval-censored data sets.

Failure time data can also be doubly interval-censored, which occurs when the failure time is the time between two events, both of which are interval-censored (Sun, 2006). For example, the latency time for HIV/AIDS is the time from infection (interval-censored as noted above) until HIV diagnosis (De Gruttola & Lagakos, 1989). The onset of HIV/AIDS could be asymptomatic and only diagnosed at a clinic visit, for example by a blood test. It is possible that interest could focus on multiple interval-censored events. For example, an AIDS opportunistic infection could be diagnosed by a laboratory test, and thus the time to onset of the infection could be interval-censored. The analysis of the relationship between the various infections (such as CMV and MAC) would require multivariate interval-censored methods, or possibly methods that allow for an interval-censored time-varying covariate.

The challenge in developing non-parametric and semi-parametric methods for the analysis of interval-censored data is that the paradigm used in exact/right-censored survival analysis cannot be applied because these rely on identification at each follow-up time and for each subject at risk whether they are free of the event of interest. However, for interval-censored data, during the interval in which an event is known to occur, we only know they were free of the event at the left endpoint, but we do not know when during the interval to assign the occurrence of the event. This impacts the estimation of the event-time (survival) curve as well as regression methods used to predict events given demographic, treatment and clinical variables (covariates) measured on each patient. Sometimes these covariates are longitudinal and time-varying, and also subject to interval censoring. The availability of the programs that can be used to analyze interval-censored data is also important. The book by Chen, Sun and Peace (2013) provides some of these programs. In this book, we will focus on the methods that have been developed to directly handle these issues in the context of interval-censored data.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics, 26*(4), 641–647.

Chen, D. G., Sun, J., & Peace, K. E. (2013). *Interval-censored time-to-event data*. Chapman & Hall/CRC Press.

De Gruttola, V., & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics, 45*, 1–12.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics, 42*(4), 845–854.

Finkelstein, D. M., & Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics, 41*(4), 933–945.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 22*(1), 86–91. https://doi.org/10.2307/2346307

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological), 38*(3), 290–295. https://doi.org/10.1111/j.2517-6161.1976.tb01597.x

Overview of Recent Advances on the Analysis of Interval-Censored Failure Time Data

Mingyue Du

Abstract As discussed by Dr. Finkelstein in Chap. 1, interval-censored failure time data are a general type of failure time or time-to-event data that often occur in many areas, including demographical studies, epidemiological studies, medical or public health research and social science. In contrast to the historic review of Chap. 1, this chapter will provide a brief review of some recent advances on several topics concerning the analysis of interval-censored data. These include the analysis of interval-censored data with time-dependent covariates, the presence of informative censoring, or the presence of a cured subgroup, respectively. Also it will cover the analysis of interval-censored data arising from case-cohort studies and the variable selection based on interval-censored data as well as the analysis of doubly interval-censored data.

Keywords Correlated failure times · Cured subgroup · Informative censoring · Time-dependent covariates · Variable selection

1 Introduction

As discussed by Dr. Finkelstein in Chap. 1, interval-censored failure time data are a general type of failure time or time-to-event data that often occur in many areas, including demographical studies, epidemiological studies, medical or public health research and social science. Although a large literature, including four books, Bogaerts et al. (2018), Chen et al. (2012), Sun (2006) and Van den hout (2017), and several review papers (Du & Sun, 2021; Sun et al., 2018), has been established for the analysis of interval-censored data, there still exist many open questions or more research is needed for many existing or new issues. In contrast to the historic review of Chap. 1, this chapter will provide a brief review of some recent advances on

M. Du (✉)
School of Mathematics, Jilin University, Changchun, China

J. Sun, D.-G. Chen (eds.), *Emerging Topics in Modeling Interval-Censored Survival Data*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-3-031-12366-5_2

several important topics with the focus on regression analysis of interval-censored data but not to provide a comprehensive review of the recent literature.

An example of interval-censored failure time data is given by Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal follow-up study that started in 2004 and was designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of the Alzheimer's disease (AD) (Li et al., 2017a, 2020a; Wu et al., 2020). In the study, the participants were recruited across North America and followed and reassessed periodically to track the pathology of the disease as it progresses. Among others, one variable of interest is the time from the baseline visit date to the AD conversion. Since the participants were only examined intermittently, the AD conversion thus cannot be observed exactly and is known only to lie between the last examination time when the AD had not occurred and the first examination time when the AD had already occurred. In other words, only interval-censored data on the AD conversion are available.

Interval-censored failure time data occur in or can have different forms, or several formulations are commonly used in the literature (Sun, 2006). Among them, an important type is case I interval-censored data, also often referred to as current status data, meaning that each subject is observed only once for the occurrence of the failure event of interest. In consequence, the failure time T is either left- or right-censored and the observation on a study subject has the form $\{C, I(T \leq C)\}$, where C represents the observation time. One type of studies that usually produce current status data is cross-sectional studies, which are commonly used in, for example, demographical studies among others.

Corresponding to case I interval-censored data, another formulation for interval-censored data that is often seen in the literature is case II interval-censored data, which assume that there exist two observation times for each study subject. For the situation, the observation has the form $\{U, V, \delta_1 = I(T < U), \delta_2 = I(U \leq T < V)\}$ with $U < V$, where U and V denote the two observation times. A more general formulation or type of interval-censored data is case K interval-censored data, meaning that there exists a sequence of observation times for each subject. For the case, the data have the form $\{K, U_0 < U_1 < \cdots < U_K, \delta_k = I(U_{j-1} < T \leq U_j);\ j = 1, \ldots, K\}$, where K denotes the number of observation times with the U_j's being the observation times. In practice, both K and the U_j's can be subject-dependent, and it is easy to see that many observation schemes such as these commonly used in medical follow-up or longitudinal studies can be naturally represented by this formulation.

The formulation that is used most to describe interval-censored data in practice is perhaps $I = (L, R]$ with $T \in I$, which will be referred to as general interval-censored data below. Under this formulation, it is easy to see that case I interval-censored data correspond to the situation where either $L = 0$ or $R = \infty$, while right-censored data mean either $L = R$ or $R = \infty$ for all study subjects. Also it is apparent that both case II and case K interval-censored data can be reduced to this format as often happened in reality.

It is worth noting that the analysis of interval-censored data is quite different from and much more challenging than that of right-censored data. One such difference

is that for the latter, the counting process approach could be easily adopted, which makes the analysis much easier, while this is not true for the former. A more specific difference can be seen from their regression analyses under the Cox model. For the situation with right-censored data, a simple partial likelihood function could be conveniently derived and commonly used for inference about regression parameters, while with interval-censored data, a more complicated full likelihood function has to be used in general.

One fundamental and important feature of failure time data is censoring and different formations of the data correspond to different censoring structures. In reality, one can classify censoring as either independent censoring or dependent or informative censoring, meaning that the failure time of interest and the censoring mechanism are correlated (Sun, 2006; Kalbfleisch & Prentice, 2002; Wang et al., 2018a, 2020a). With the former, the analysis is usually performed conditional on the censoring process no matter the formats of the data. In contrast, with the latter, the analysis can be very different and also difficult as one usually has to make certain assumptions or model the censoring mechanism. In particular, for right-censored data, the modeling is relatively easy partly as only one variable is needed to describe the censoring, while for interval-censored data, as discussed below, two or more variables are usually required to characterize the censoring mechanism. As pointed out in the literature, in the presence of informative censoring, the analysis that ignores it may result in biased results or misleading conclusions.

The remainder of this chapter is organized as follows. In Sect. 2, we will first discuss some recent advances on several topics related to regression analysis of univariate interval-censored failure time data. They are the analysis with time-dependent covariates, in the presence of a cured subgroup, and with the focus on variable selection, respectively. Section 3 will also consider univariate interval-censored data as in Sect. 2 but with dependent or informative censoring. In Sect. 4, the attention will be on regression analysis of clustered and multivariate interval-censored data, and Sect. 5 will briefly discuss several other topics related to regression analysis of interval-censored data. They include the analysis of the data arising from case-cohort studies, the analysis of doubly censored data, and the analysis of the data with missing covariates. Section 6 will give some concluding remarks and point out some topics for which more research is needed.

2 Regression Analysis of Univariate Interval-Censored Failure Time Data

As mentioned above, a great deal of literature has been established for the analysis of univariate interval-censored failure time data and especially, many methods have been developed for their regression analysis (Chen et al., 2012; Sun, 2006). In this section, we will discuss three topics on the regression analysis that have recently attracted a good amount of attentions. They are the regression analysis

when covariates are time-dependent, there exists a cured subgroup, or when variable selection is of main interest, respectively.

2.1 Regression Analysis with Time-Dependent Covariates

Consider a failure time study that yields interval-censored data with the main goal being making inference about the effects of time-dependent covariates. For estimation of such effects, two approaches are commonly used. One is the marginal maximum likelihood approach and the other is the joint modeling approach. Among others, Zeng et al. (2016) considered the former approach to the problem under a class of semiparametric transformation (ST) models. More specifically, they developed the maximum likelihood estimation (MLE) approach and showed that the maximum likelihood (ML) estimators of regression parameters are consistent and asymptotically efficient and normal. In addition, they developed a flexible and computationally efficient EM algorithm following Wang et al. (2016a), who considered the same problem but under the Cox model with time-independent covariates.

In contrast to the marginal approach, the joint modeling approach treats the time-dependent covariates as longitudinal processes and is usually preferred when there also may exist measurement errors on covariates. For this, one commonly used method is to model the failure time of interest and the longitudinal covariate process jointly by using, for example, the latent variable approach. Among others, Yi et al. (2020) proposed a Cox frailty model and developed a MLE procedure under this framework along with a MCEM algorithm. Note that many methods have been developed in the literature for joint analysis of longitudinal data and failure time data with either the failure time or the longitudinal variable as the variable of interest. However, most of them only focused on right-censored data on the failure time except Chen et al. (2018). One major difference between the methods given in Chen et al. (2018) and Yi et al. (2020) is that the former treats the failure time as the dropout or stopping variable and assumed that there is no more observation after the dropout. In other words, it cannot give efficient or valid estimation if there exist more observations after the failure time. In contrast, the latter takes into account all observations and also the algorithm given in Yi et al. (2020) is faster and more stable than that given in Chen et al. (2018). More discussion on regression analysis of interval-censored data with time-dependent covariates can be found in Chaps. 6, 9 and 11 of this book.

2.2 Regression Analysis in the Presence of a Cured Subgroup

By the existence of a cured subgroup, we usually mean that there exists a portion of study subjects who never experience or are non-susceptible to the failure event

of interest for various reasons. These individuals are usually considered to be cured or immune from the failure event and referred to as long-term survivors or cured subjects. To deal with this, two types of models or methods are commonly used and they are two-component mixture cure model approach and non-mixture cure model approach (Hu & Xiang, 2016; Li et al., 2019a). The former models the effects of covariates on the cure rate of the population and the survival function of non-cured subjects through two separate regression models, and a drawback of this is that it does not have the usual survival model property for the whole population. In contrast, the latter assumes that cured subjects have infinity survival time and uses a single model to describe the survival function of the entire population (Hu & Xiang, 2016). Sometimes the latter model is also referred to as the promotion time cure model.

Specifically, under the two-component mixture cure model, the failure time of interest T is usually written as $T = Y T^* + (1 - Y)\infty$, where T^* denotes the failure time of a susceptible subject and Y indicates, by value 1 or 0, whether the study subject is susceptible or not. To describe the effects of covariates, one could employ a regular failure time regression model such as the Cox model for the effect on the failure time and the logistic model for the possible effect on the cure rate. Among others, Hu and Xiang (2016) discussed this approach when one observes interval-censored data and proposed a sieve ML method under a class of ST models and the logistic model.

As mentioned above, the non-mixture cure model uses a single model to describe the survival function of the entire population, and for the situation, one could easily extend a regular regression model such as the Cox model to the Cox cure model. An attractive feature of the Cox cure model is that it inherits the Cox model structure for the whole population and thus regression parameters have relatively appealing, easy interpretations. Among others, Li et al. (2019a) recently considered this approach under a class of ST cure models and developed the MLE procedure for fitting the model to interval-censored data. Other authors who recently investigated the analysis of interval-censored data with a cured subgroup include Liu et al. (2020) and Zhou et al. (2018a). The former discussed the situation when there exist mismeasured covariates, and the latter proposed a generalized odds rate mixture cure model.

2.3 Variable Section for Interval-Censored Data

Variable selection has recently attracted a great deal of attention with a huge amount of literature established under various contexts. This is particularly true for the analysis of failure time data and a few penalized variable selection methods have been proposed for interval-censored data under different situations. Among others, Zhao et al. (2020a) discussed the problem under the Cox model and proposed a broken adaptive ridge regression procedure. Furthermore, they proved that the resulting variable selection and estimation procedure has both the oracle property

and the grouping property, and the approach works too with the use of other commonly used penalty functions such as LASSO, ALASSO and SCAD. Following Zhao et al. (2020a), Li et al. (2020a) and Zhao et al. (2020b) generalized the method to the situations where the failure time of interest follows a class of ST models and the interval-censored data arise from case-cohort studies, respectively. As Zhao et al. (2020b), Du et al. (2022) also investigated the variable selection based on case-cohort interval-censored data but their method allows for informative interval censoring.

An assumption behind the methods mentioned above is that although it can diverge with the sample size, the number of covariates cannot be larger than the sample size. To address this, Wu et al. (2020) generalized the method given in Zhao et al. (2020a) to the case where the number of covariates to be larger than the sample size. Furthermore, their generalized procedure allows for the existence of a vector of low-dimensional covariates that may have non-linear effects on the failure time of interest. Other authors who also recently studied variable selection for interval-censored data include Chen and Sun (2022), Du and Sun (2022), Sun et al. (2019), Xu et al. (2021) and Yi et al. (2022). In particular, Chen and Sun (2022) considered the situation where covariate effects may be time-varying, and Sun et al. (2019) and Xu et al. (2021) proposed some variable selection procedures for interval-censored data where there may exist a cured subgroup. Yi et al. (2020) considered the variable selection under the context of joint analysis of longitudinal data and interval-censored data, and Du et al. (2021a) gave a uniform approach for the problem under the Cox model that allows for informative censoring.

Except Wu et al. (2020), all of the methods mentioned above assume linear covariate effects, and corresponding to this, Li and Sun (2020) discussed the same problem but under high-dimensional quadratic Cox model. Note that all of the work discussed above on interval-censored data only investigated either low- or high-dimensional situations and sometimes one may face ultra-high-dimensional covariates. To address the latter situation, Hu et al. (2020) developed a model-free or nonparametric screening and feature selection procedure based on the idea of cumulative residuals for interval-censored data. In particular, they proved that their method has the sure independent screening property and tends to rank the active or significant covariates above the inactive or non-significant ones in terms of their association with the failure time of interest. Following Hu et al. (2020), Zhang et al. (2022) discussed the same problem and gave another model-free screening procedure. More discussion and more references on variable selection based on interval-censored data can be found in a recent review paper (Du & Sun, 2022).

3 Regression Analysis with Informative Interval Censoring

In the presence of informative censoring, unlike the non-informative case where the analysis is usually performed conditional on the censoring mechanism or observation process, one needs to model the censoring mechanism or observation

process together with the failure time of interest. For this, two types of approaches are commonly used and they are the frailty or latent variable-based approach and the copula model-based approach. The former employs some frailty or latent variables to characterize the relationship between the censoring mechanism and the failure time of interest, while the latter uses copula functions to achieve the purpose.

Among others, Li et al. (2017b, 2019b) and Xu et al. (2022) recently discussed regression analysis of case I informatively interval-censored data and proposed some latent variable-based sieve MLE procedures. More specifically, Li et al. (2017b) considered the situation where both the failure time of interest and the censoring variable follow Cox frailty models and developed a three-stage data augmentation EM algorithm. Li et al. (2019b) and Xu et al. (2022) studied the same problem as Li et al. (2017b) except that the failure time of interest follows an additive frailty model and a class of generalized odds rate frailty models, respectively. On the analysis of case K interval-censored data, some recent work can be found in Wang et al. (2016b, 2018a, 2020a), which generalized the methods given in Li et al. (2017b, 2019b). In addition, Wang et al. (2018b) also discussed the same problem but under a class of ST models. It is worth noting that with case I informative censoring, one only needs to deal with one censoring or observation variable but with case K informative censoring, one usually has to make use of a stochastic process such as Poisson process to describe the censoring or observation process.

As mentioned above, to deal with informative censoring, an alternative to the latent variable-based approach is to employ the copula model-based approach, which connects the failure time of interest and censoring variables through some copula functions. Among others, Cui et al. (2018), Du et al. (2019), Xu et al. (2019a), Xu et al. (2020) and Zhao et al. (2019) recently applied this approach to regression analysis of case I interval-censored data with informative censoring. More specifically, they considered the situation where the failure time of interest marginally follows the Cox model, the generalized probit model, the ST model, the accelerated failure time model, or the additive hazards (AH) model, respectively. Note that as other similar methods, all of the methods mentioned above except that given in Cui et al. (2018) assume that both the copula function and the association parameter are known. Cui et al. (2018) proposed a two-step estimation procedure that allows for the association parameter to be estimated.

Some recent work on the application of the copula model-based approach to regression analysis of general, informatively interval-censored data can be found in Ma et al. (2016) and Xu et al. (2019b). More specifically, they discussed the data given by the formulation $I = (L, R]$ where the dependence between the failure time T of interest and the censoring mechanism can be characterized by the correlation between T and $W = R - L$, the length of the censoring interval. For inference, they developed the MLE procedures for the situation where T follows the Cox model or a class of ST models, respectively, and W follows the Cox model. Instead of the two approaches discussed above, Zhou et al. (2022) discussed a third approach, the marginal approach, for regression analysis of informatively interval-censored data. For inference, they developed some estimating equations by using

the inverse probability weighted technique, and it has the advantage of avoiding to model the censoring process.

4 Regression Analysis of Clustered and Multivariate Interval-Censored Data

In this section, we will first briefly discuss some recent advances on regression analysis of clustered interval-censored failure data and then on regression analysis of multivariate interval-censored data. It is well-known that for the analysis of these data, one key issue is how to describe or model the correlation among the correlated failure times.

For regression analysis of clustered interval-censored data, one commonly used type of procedure is the latent variable-based approach, which employs latent variables to characterize the relationship among the correlated failure times of interest. Among others, Lee et al. (2022) and Zeng et al. (2017) recently discussed the use of this approach for fitting a class of ST model to case $I\,I$ clustered interval-censored data. In particular, the MLE procedure proposed in Zeng et al. (2017) can apply to both time-dependent covariates and the combination of multivariate and clustered interval-censored data. For the two methods described above, the latent variable has been assumed to follow a known distribution with some unknown parameters that can be estimated along with other parameters. Sometimes one may not want to specify the distribution of the latent variable or prefer to leave the correlation among the failure times of interest arbitrary. For this purpose, Yang et al. (2021, 2022) and Zhao et al. (2018) proposed some within-cluster-resampling estimation procedures for general clustered interval-censored data under the Cox model and a class of ST models, respectively. Both of the methods given in Yang et al. (2022) and Zhao et al. (2018) allow for the presence of informative cluster size, while the methods provided by Yang et al. (2021, 2022) apply to the case where there exists a cured subgroup.

On regression analysis of multivariate interval-censored data, two types of approaches, the latent variable-based and copula model-based approaches, are commonly used similarly to the analysis of informatively interval-censored data. Among others, Li et al. (2020b) and Zhou et al. (2017a) recently investigated the problem and gave some latent variable-based methods. The former focused on multivariate current status data under a class of ST frailty models and developed a MLE procedure. Under similar models, in contrast, the latter proposed a sieve MLE method with the use of Bernstein polynomials for general bivariate interval-censored data. Furthermore, Liu and Qin (2018) studied the same problem under a class of probit models, and Gao et al. (2019) discussed the situation with time-dependent covariates. In addition, Li et al. (2022) and Yu et al. (2022) also investigated regression analysis of multivariate current status data and case $I\,I$ interval-censored data, respectively, under the marginal AH frailty model. Unlike