Analysis of window-observation recurrence data

by

Jianying Zuo

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
William Q. Meeker, Jr., Co-major Professor
Huaiqing Wu, Co-major Professor
Mark Kaiser
Peng Liu
Steven A. Freeman

Iowa State University

Ames, Iowa

2010

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

ACKNOWLEDGMENTS x

ABSTRACT xi

CHAPTER 1. GENERAL INTRODUCTION 1

1.1 Background 1

1.2 Motivation 2

1.3 Dissertation Organization 3

References 4

CHAPTER 2. ANALYSIS OF WINDOW-OBSERVATION RECURRENCE DATA 5

2.1 Introduction 6

2.1.1 Background and Motivation 6

2.1.2 Previous Work on Analysis of Recurrence Data 7

2.1.3 Overview 7

2.2 Examples 8

2.2.1 Extended Warranty Data 8

2.2.2 AMSAA Vehicle Fleet Data 8

2.3 Nonparametric Estimation Methods for Window-Observation Recurrence Data 9

2.3.1 Notation 9

2.3.2 Nonparametric Estimation of the Mean Cumulative Function 9

2.3.3 Assumptions . . . 11
2.3.4 Variance of $\widehat{\mu}_{\mathrm{NP}}(t)$ and Estimator . . . 12
2.3.5 Confidence Intervals for $\mu(t)$ . . . 14
2.4 Analysis of the Extended Warranty Data . . . 15
2.4.1 Descriptions of the Data . . . 15
2.4.2 Graphical Exploration of the Data . . . 16
2.4.3 Estimation with the Nonparametric Method . . . 16
2.5 Maximum Likelihood Estimation of a Nonhomogeneous Poisson Process (NHPP) Model Using Window-Observation Recurrence Data . . . 18
2.5.1 Assumptions for the Poisson Process . . . 19
2.5.2 NHPP Model . . . 20
2.5.3 NHPP Likelihood for the Window-Observation Data . . . 21
2.5.4 Variance of $\widehat{\mu}_{\mathrm{NHPP}}$ and Estimator . . . 22
2.6 Analysis of the AMSAA Vehicle Fleet Data . . . 23
2.6.1 Descriptions on the Data . . . 23
2.6.2 Graphical Exploration of the Data . . . 24
2.6.3 Nonparametric and Parametric Estimation . . . 27
2.7 The Hybrid Estimator . . . 30
2.7.1 General Approach . . . 30
2.7.2 The Local Hybrid Estimator . . . 31
2.7.3 The NHPP Hybrid Estimator . . . 32
2.7.4 Approximate Variance of the Hybrid Estimator . . . 33
2.7.5 Confidence Intervals for $\mu(t)$ . . . 33
2.7.6 Hybrid Estimation of the MCF for the AMSAA Vehicle Fleet Data . . . 34
2.8 A Simulation Study to Compare Different MCF Estimators . . . 34
2.9 Concluding Remarks and Areas for Further Research . . . 37
Acknowledgments . . . 38
References . . . 39

CHAPTER 3. A SIMULATION STUDY ON CONFIDENCE INTERVAL PROCEDURES OF SOME MEAN CUMULATIVE FUNCTION ESTIMATORS 41

3.1 Introduction and Background 42

3.1.1 Background and Motivation 42

3.1.2 Window-Observation Data 43

3.1.3 Other Previous Related Work 43

3.1.4 Overview 43

3.2 Model and Estimation 44

3.2.1 Notation and Acronyms 44

3.2.2 Model 45

3.2.3 Estimation for the Nonparametric Model 45

3.2.4 Estimation for the NHPP Parametric Model 46

3.2.5 Hybrid MCF Estimators for Window-Observation Recurrence Data . . . 46

3.2.6 Recommendations on Selection of MCF Estimators 46

3.3 Confidence Interval Procedures 47

3.4 Simulation Experimental Design 49

3.4.1 Factors and Factor Levels 49

3.4.2 Simulation Algorithm 50

3.5 The Effect that the Recurrence Rate Shape has on CI Procedures 51

3.5.1 Results from Complete Data 52

3.5.2 Results from Window2 Data 53

3.6 Performances of the CI Procedures Based on the NP Estimator 60

3.6.1 Comparison of Results from Complete Data 60

3.6.2 Comparison of Results from Window2 Data 63

3.6.3 Recommendations for the NP Estimator 65

3.7 Performances of the CI Procedures Based on the NHPP Estimator 66

3.7.1 Comparison of Results from Complete Data and Window2 Data 66

3.7.2 Recommendations for the NHPP Estimator 68
3.8 Performances of the CI Procedures Based on the Hybrid Estimators from Window2 Data 68
3.8.1 Comparison of Results – the Local Hybrid Estimator and the Power Law NHPP Hybrid Estimator 68
3.8.2 Recommendations for the Hybrid Estimators 71
3.9 Concluding Remarks and Areas for Further Research 72
Acknowledgments 73
References 73
CHAPTER 4. ASYMPTOTIC PROPERTIES OF MEAN CUMULATIVE FUNCTION ESTIMATORS FROM WINDOW-OBSERVATION RECURRENCE DATA 75
4.1 Introduction 76
4.1.1 Background and Motivation 76
4.1.2 Other Previous Work 77
4.1.3 Overview 77
4.2 Notation and Assumptions 77
4.2.1 Notation for the Models 77
4.2.2 Notation for the Data 78
4.2.3 Assumptions 78
4.3 Nonparametric MCF Estimator 80
4.3.1 Estimation for the Nonparametric Model 80
4.3.2 Theorems Adapted from Andersen, Borgan, Gill, and Keiding (1993) . . 80
4.3.3 Asymptotic Properties of the Nonparametric Estimator 82
4.4 Nonhomogeneous Poisson Process (NHPP) MCF Estimators 84
4.4.1 Estimation for the NHPP Model 84
4.4.2 Conditions and Theorems Adapted from Andersen, Borgan, Gill, and Keiding (1993) 85

4.4.3 Asymptotic Properties of the NHPP Estimators 88
4.5 Concluding Remarks 94
References 94
CHAPTER 5. CONCLUSIONS 96
APPENDIX A. SUPPLEMENTAL MATERIAL 97
A.1 Calculation of $\widehat{\text{Var}}[\bar{d}^{\dagger}_{\cdot}(t)]$ 97
A.2 A Conservative Estimator of $\text{Var}[\bar{d}(t_k)]$ When the Size of the Risk Set is 1 . . . 98
A.3 Method to Estimate $\widehat{\text{Cov}}\left[\bar{d}^{\dagger}_{\cdot}(t), \bar{d}_{\cdot}(t)\right]$ for the NHPP Hybrid Estimator 99
A.4 Properties of the Complete Data Simulated from the Power Law NHPP Model 102
A.4.1 Distribution of the Number of Recurrences 102
A.4.2 NP MCF Estimator 103
A.4.3 The Power Law NHPP Estimator 104

LIST OF TABLES

Table 2.1 Extended Warranty Data 15

Table 2.2 Simulated Vehicle Data – Comparisons of Risk Set Sizes 27

Table 2.3 AMSAA Vehicle Fleet Data – Comparisons of Parameter Estimates of the Power Law NHPP Model 29

Table 2.4 AMSAA Vehicle Fleet Data – Comparisons of Parameter Estimates of the Loglinear NHPP Model 29

Table 3.1 RSSZ and RSSONE for Window2 Data with $n = 10$, E$(r) = 10$, and Number of Simulations = 5000 60

Table 3.2 RSSZ and RSSONE for Window2 Data with $\beta = 1$ and Number of Simulations = 5000 64

Table A.1 Probability in CI and Error Bound for Complete Data at $n = 10$ . . . 104

LIST OF FIGURES

Figure 2.1 Extended Warranty Data – Event Plot of Selected Units for Labor Code C6050. Note that for lines corresponding to a group of units, $n_1(n_2)$ indicates there are n_1 units in the data and that each of these units has n_2 events. 17

Figure 2.2 Extended Warranty Data – Super-Imposed Event Plot for Labor Code C6050. Note that $n_1(n_2)$ indicates there are n_1 units in the data and that these units had the indicated number of events. 17

Figure 2.3 Extended Warranty Data – Risk Set Plot for Labor Code C6050 . . . 17

Figure 2.4 MCF plot for Labor Code C6050 Extended Warranty Data 18

Figure 2.5 Power Law NHPP Model – Plots of the Recurrence Rate and the MCF with $\eta = 1$ and Different β Values 21

Figure 2.6 Event Plot for AMSAA Vehicle Fleet Complete Data 25

Figure 2.7 Event Plot for AMSAA Vehicle Fleet Random Selection Window Data 25

Figure 2.8 Event Plot for AMSAA Vehicle Fleet Non-random Selection Window Data 25

Figure 2.9 Risk Set Plot for AMSAA Vehicle Fleet Complete Data 26

Figure 2.10 Risk Set Plot for AMSAA Vehicle Fleet Random Selection Window Data 26

Figure 2.11 Risk Set Plot for AMSAA Vehicle Fleet Non-random Selection Window Data 26

Figure 2.12 AMSAA Vehicle Fleet Complete Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, and the True MCF 28

Figure 2.13 AMSAA Vehicle Fleet Random Selection Window Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, and the True MCF 28

Figure 2.14 AMSAA Vehicle Fleet Non-random Selection Window Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, Loglinear NHPP MCF Estimate, and the True MCF 28

Figure 2.15 MCF Plots for AMSAA Vehicle Fleet Non-random Selection Window Data, Comparing the True Model, the NP Estimator, the Power Law NHPP Estimator, the Local Hybrid Estimator, and the NHPP Hybrid Estimator. 35

Figure 2.16 MCF Plots for AMSAA Vehicle Fleet Non-random Selection Window Data, Comparing the True Model, and the bootstrap-t CI of both the Local Hybrid Estimator and the NHPP Hybrid Estimator, based on B=5000 nonparametric re-samplings of the original simulated data. . . 35

Figure 3.1 Comparison of 4 β Values: NP Estimator for the Complete Data . . . 54

Figure 3.2 Comparison of 4 β Values: NHPP Estimator for the Complete Data . 55

Figure 3.3 Comparison of 4 β Values: NP Estimator for the Window2 Data . . . 56

Figure 3.4 Comparison of 4 β Values: NHPP Estimator for the Window2 Data . . 57

Figure 3.5 Comparison of 4 β Values: Local Hybrid Estimator for the Window2 Data 58

Figure 3.6 Comparison of 4 β Values: NHPP Hybrid Estimator for the Window2 Data 59

Figure 3.7 CP Plots for NP Estimator with $\beta = 1$ 61

Figure 3.8 CP Plots for the Power Law NHPP Estimator with $\beta = 1$ 67

Figure 3.9 CP Plots for the Hybrid Estimators for the Window2 Data with $\beta = 1$ 69

ACKNOWLEDGMENTS

I would like to express my gratitudes to those who helped me with various aspects of my Ph.D. study.

First and foremost, thanks to Dr. William Q. Meeker and Dr. Huaiqing Wu for their guidance, patience, and support throughout this research and the writing of this dissertation. Their broad knowledge, keen insights, and encouraging words have often inspired me and will be valuable for my future career and life.

I would also like to thank my committee members, Dr. Mark Kaiser, Dr. Peng Liu, and Dr. Steven A. Freeman, for their efforts and contributions to this work.

Many people have helped me solve different kinds of problems. In particular, thanks to, Mr. Ted Peterson, Ms. Denise Riker, Ms. Sharon Shepard, Ms. Edith Landin, and Ms. Jeanette La Grange. I wish them all the best.

Last but not least, I would like to thank my parents for their continuous love and support. I have been able to pursue my study and research worry-free because I know they are standing by my side.

ABSTRACT

Many systems experience recurrent events. Recurrence data are collected to analyze quantities of interest, such as the mean cumulative number of events. Methods of analysis are available for recurrence data with left and/or right censoring. Due to practical constraints, however, recurrence data are sometimes recorded only in windows, with gaps in between. Nelson (2003, page 75) gives one example, and Chapter 2 describes two other applications that window-observation recurrence data arise. With the need for analytical methods with window-observation recurrence data, our research achieves the following three objectives:

- Extend the existing statistical methods, both nonparametric and parametric, to analyze window-observation recurrence data, and our focus is to estimate the mean cumulative function (MCF).
- Study and compare CI procedures for the MCF estimators with window-observation recurrence data, and make recommendations when the amount of observed information in the data is small.
- Establish the asymptotic (i.e., large-sample) properties for the MCF estimators with window-observation recurrence data.

Our research shows that the existing statistical methods, both nonparametric and parametric, can be extended to estimate the MCF with window-observation recurrence data. Chapter 2 provides the details on four MCF estimators, including the NP estimator, the nonhomogeneous Poisson process (NHPP) estimator, the local hybrid estimator, and the NHPP hybrid estimator. The NP estimator and the NHPP estimator for analysis with window-observation

recurrence data are straight-forward extensions. The NP estimator requires minimum assumptions, but will be inconsistent if the size of the risk set is not positive over the entire period of interest. There is no such difficulty when using a parametric model for the recurrence data, yet the assumption on the recurrence rate form needs careful diagnoses and checking. When risk-set-size-zero (RSSZ) intervals exist, the two hybrid estimators are alternatives to the NP estimator, which generates downwardly biased estimates. Chapter 2 also presents the summary results from a simulation study that can be used as references to select the MCF estimators to use.

The four MCF estimators described in Chapter 2 are relatively easy to calculate. Besides point estimates, however, confidence intervals are useful in many applications. When the amount of observed information is large, for example, the number of units is large, the number of observed recurrences is large, and there is no or very small amount of time with RSSZ, various CI procedures generate similar results for each of the MCF estimators. However, when the number of units is small, and/or the number of observed recurrences is small, and/or there is relatively large amount of time with RSSZ or risk-set-size-one (RSSONE), the choice of which CI procedure to use makes a difference. Our research carries out an extensive simulation study on five CI procedures for each of the four MCF estimators described in Chapter 2, and the details of the simulation studies and the summary results are in Chapter 3. Chapter 3 also makes suggestions on the CI procedures to use based on the number of units, the number of recurrences observed, and the amount of time with RSSZ or RSSONE.

Chapter 4 establishes the asymptotic properties for the NP and NHPP MCF estimators, and outlines the assumptions and conditions that are needed for the MCF estimators to be consistent and asymptotically normal.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Recurrent events in systems are of interest in many applications. Here we use a broad definition of a system. For example, machines or automobiles in a company's fleet break down and are repaired; people become ill and visit a doctor; customers encounter financial need and apply for loans from banks. For such systems, quantities of interest usually include the expected cumulative number of events over a specific time range of system operation. Recurrence data are collected to analyze and estimate these and other quantities of interest.

Recurrence data usually record the type and number of events over time. Often, covariates related to the events, such as cost or operating temperature or pressure, are also recorded. An event could be recorded as having occurred at an exact time or within a particular time interval. Nelson (2003) describes many examples and data analysis methods for such data. Some other references on the methods to analyze the recurrence data, nonparametric and parametric ones, include Cook and Lawless (2007), Meeker and Escobar (1998), Lawless and Nadeau (1995), and Rigdon and Basu (2000).

Modeling of recurrent events can be based on the counts of events, the intensity, and the time between events. Peña, Strawderman, and Hollander (2001) describe nonparametric methods to estimate the interoccurrence times with recurrent event data. Our analysis is based on the count of recurrences, and statistics of interests include the mean cumulative number of recurrences.

The mean cumulative function (MCF) contains very useful information on a system, such as how the likelihood of observing an event changes across the life of the system. Therefore, a good understanding and estimation of the MCF can help better allocate resources across time,

for example, reserve more repair funding and hire more service staff for the time intervals that the MCF has a steeper increase. Comparisons of the MCF estimates from different cohort of systems can help monitor reliability and detect quality improvement or deterioration.

1.2 Motivation

Many analytical methods are available for recurrence data with left and/or right censoring. Due to practical constraints, however, window-observation recurrence data arise when the recurrence histories of some systems are available in disconnected observation windows with gaps in between. Nelson (2003, page 75) gives an example in which window-observation recurrence data arise, and Chapter 2 describes two other applications. Therefore, there arises the need to extend the existing methods to analyze window-observation recurrence data, and our research intends to achieve the following three objectives:

- Extend the existing statistical methods, both nonparametric and parametric, to analyze window-observation recurrence data and estimate the MCF.
- Study and compare CI procedures for the MCF estimators with window-observation recurrence data, and make recommendations when the amount of observed information in the data is small.
- Establish the asymptotic (i.e., large-sample) properties for the MCF estimators with window-observation recurrence data.

For the more familiar recurrence data with left and/or right censoring, there is only one observation window for each system. For the window-observation recurrence data, however, there are usually multiple disconnected observation windows for each system. Note that observation windows can have random length, and the length of the gaps between windows can also be random. Furthermore, there is no requirement to have the same beginning or ending time points of windows for different observational units. Note that window-observation data are different from the "interval-grouped recurrent-event data" (Lawless and Zhan, 1998), which

record the number of recurrent events in time intervals, with no gaps between the intervals (i.e., the number of events is known but their exact times are not known).

Some of the existing MCF estimators, with modifications to accommodate multiple observation windows, can be extended to use with window-observation recurrence data. Difficulties arise, however, when there are time intervals with risk-set-size-zero (RSSZ) and risk-set-size-one (RSSONE). The existence of RSSZ intervals makes the nonparametric MCF estimators to generate biased results, while the variance-covariance components of the variance estimator in the RSSZ and RSSONE intervals are not estimable. For the former problem, we propose hybrid estimators to help reduce the bias, while for the latter one, we suggest a conservative variance estimator for the increase in the MCF when the size of the risk set is 1. For the variance estimator of the nonparametric MCF estimator, we make modifications so that it becomes efficient to handle large recurrence data sets.

For performance comparisons among the MCF estimators and CI procedures, we use simulations, which generate useful information, especially for the scenarios with small sample size or small number of observed recurrences. We also use various graphical tools in our analysis, for examples, event plots and risk set plots to explore and understand the data, and side-by-side plots to compare different CI procedures.

Andersen, Borgan, Gill, and Keiding (1993) provide comprehensive descriptions of the statistical models and methods that can be used to analyze event history observed in continuous time. They derive the asymptotic properties for both the nonparametric and parametric estimators described in their book. Building on their conditions and theorems, we derive the assumptions and conditions needed to establish the asymptotic properties, more specifically the consistency and asymptotic normality, for an MCF estimator with the window-observation recurrence data.

1.3 Dissertation Organization

This dissertation consists of three main chapters, which are closely related yet each is by itself an independent piece that has been submitted for publication or is ready for submission.

Chapter 2 extends the nonparametric and NHPP methods to estimate the MCF with window-observation recurrence data, and proposes two hybrid MCF estimators when there exist RSSZs that cause the NP estimator to be biased. Chapter 3 carries out an extensive simulation study on five CI procedures for each of the four MCF estimators described in Chapter 2, and makes suggestions on the CI procedures to use given factor levels on the number of units, the number of recurrences observed, and the amount of time with RSSZ or RSSONE. Chapter 4 establishes the asymptotic properties for the NP and NHPP MCF estimators. Conclusions are outlined in Chapter 5, and Appendix A provides some supplemental technical details used in Chapters 2 and 3.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Cook, R. J., and Lawless, J. F. (2007), *The Statistical Analysis of Recurrent Events*, New York: Springer-Verlag.

Lawless, J. F., and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.

Lawless, J. F., and Zhan, M., (1998), "Analysis of Interval-grouped Recurrent-event Data Using Piecewise Constant Rate Functions," *The Canadian Journal of Statistics*, 26, 549-565.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Nelson, W. B. (2003), *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Philadelphia: ASA-SIAM.

Peña, E. A., Strawderman, R. L., and Hollander, M. (2001), "Nonparametric Estimation with Recurrent Event Data," *Journal of the American Statistical Association*, 96, 1299-1315.

Rigdon, S. E., and Basu, A. P. (2000), *Statistical Methods for the Reliability of Repairable Systems*, New York: Wiley.

CHAPTER 2. ANALYSIS OF WINDOW-OBSERVATION RECURRENCE DATA

Jianying Zuo

William Q. Meeker

Huaiqing Wu

Statistics Department

Iowa State University

Ames, IA 50011

Abstract

Many systems experience recurrent events. Recurrence data are collected to analyze quantities of interest, such as the mean cumulative number of events. Methods of analysis are available for recurrence data with left and/or right censoring. Due to practical constraints, however, recurrence data are sometimes recorded only in windows. Between the windows, there are gaps over which the process cannot be observed. This paper extends existing statistical methods, both nonparametric and parametric, to window-observation recurrence data. The nonparametric estimator requires minimum assumptions, but will be inconsistent if the size of the risk set is not positive over the entire period of interest. There is no such difficulty when using a parametric model for the recurrence data. For cases in which the size of the risk set is zero for some periods of time, we propose and compare two alternative hybrid estimators. The methods are illustrated with two example applications.

KEY WORDS: Forecast; Mean cumulative function; Nonhomogeneous Poisson process; Nonparametric estimation; Repairable system data.

2.1 Introduction

2.1.1 Background and Motivation

Recurrent events in systems are of interest in many applications. Here we use a broad definition of a system. For example, machines or automobiles in a company's fleet break down and are repaired; people become ill and visit a doctor; customers encounter financial need and apply for loans from banks. For such systems, quantities of interest usually include the expected cumulative number of events over a specific time range of system operation. Recurrence data are collected to analyze and estimate these and other quantities of interest.

Recurrence data usually record the type and number of events over time. Often, covariates related to the events, such as cost or operating temperature or pressure, are also recorded. An event could be recorded as having occurred at an exact time or within a particular time interval. Nelson (2003) describes many examples and data analysis methods for such data.

In some applications, recurrence data are recorded only in observation windows with gaps between the windows, even though the underlying process is continuous in time. Nelson (2003, page 75) describes an example in which window-observation recurrence data arise when "patients may enter and leave a medical study of a disease any number of times." Window-observation data might also arise as a result of a poor data recording procedure (e.g., key data for some periods of time for some or all systems are missing or not recorded correctly). Section 2.2 describes two applications that we have encountered.

Note that observation windows can have random length, and the length of the gaps between windows can also be random. Furthermore, there is no requirement to have the same beginning or ending time points of windows for different observational units. Note that window-observation data are different from the "interval-grouped recurrent-event data" (Lawless and Zhan, 1998), which record the number of recurrent events in time intervals, with no gaps between the intervals (i.e., the number of events is known but their exact times are not known).

2.1.2 Previous Work on Analysis of Recurrence Data

Much work has been done on the development of methods for the analysis of recurrence data. Nelson (1988) presents a nonparametric estimator of the mean cumulative function (MCF), and shows how to use the estimator to make predictions. Nelson (1995) provides an unbiased variance estimator for the MCF estimator, which could (generally with small probability) be negative, and the confidence limits for the MCF. Lawless and Nadeau (1995) provide an alternative variance estimator for the nonparametric MCF estimator. Although their variance estimator is not unbiased, it is always nonnegative. The same paper also presents a flexible semi-parametric regression model that allows for covariates in the analysis of recurrence data.

Nelson (2003) gives a comprehensive treatment of the most important nonparametric methods for analyzing recurrence data. This book also presents many examples, over a wide range of application areas. These methods do not require strong model assumptions, and are easy to apply in practice. Rigdon and Basu (2000) is a good source for parametric statistical models and methods for repairable systems, from which recurrence data are often recorded. The purpose of this paper is to extend the use of these nonparametric and parametric methods to allow analysis of window-observation recurrence data.

2.1.3 Overview

The remainder of this paper is organized as follows. Section 2.2 briefly describes two applications that have recurrence data with observation windows: Extended Warranty Data and AMSAA Vehicle Fleet Data. Section 2.3 reviews the nonparametric estimation method for the MCF and shows how it can be extended to handle window-observation recurrence data. Section 2.4 provides more details on the Extended Warranty Data, and illustrates the nonparametric method with the data. Section 2.5 reviews a parametric estimation method for the MCF with window-observation recurrence data. Section 2.6 shows how to apply the methods to the AMSAA Vehicle Fleet Data in which the gaps between observation windows generate some intervals where the risk set has size zero, resulting in downward bias for the nonparametric

estimator. Section 2.7 proposes two hybrid estimators that correct the downward bias. Section 2.8 summarizes a simulation study that compares different MCF estimators. Section 2.9 contains concluding remarks. The appendices A.1 to A.3 present some technical details.

2.2 Examples

This section briefly describes two examples that we encountered with window-observation recurrence data. More details will be provided when we analyze the data with the MCF estimators in Section 2.4 and Section 2.6. The major difference between the two datasets is that there are no risk-set-size-zero (RSSZ) intervals in the Extended Warranty Data but there are RSSZ intervals in the AMSAA Vehicle Fleet Data.

2.2.1 Extended Warranty Data

Extended warranties are often available to businesses or individuals for high-cost items, such as automobiles, large appliances, and computers. Only recurrence records within the period under the initial warranty and extended warranty are available, and the effective warranty periods form the observation windows. Warranty data for a particular unit can have disconnected warranty periods. For example, a customer might purchase an extended warranty at a point in time after the expiration of the initial warranty. This results in window-observation data.

2.2.2 AMSAA Vehicle Fleet Data

U.S. Army Materiel Systems Analysis Activity (AMSAA) oversees the Army's Field Exercise Data Collection (FEDC) Program. This program maintains a database of part replacement rates for mission-essential weapon systems (e.g., various types of vehicles) that are used during intensive field training exercises. FEDC data are collected from a number of sites around the world and are used to answer such questions as whether a fleet is aging or not, how fast a fleet is aging, and when units should be overhauled or replaced. During an FEDC field exercise, a group of vehicles is used and careful records are taken for all maintenance and repair actions. An exercise generally lasts for approximately 500 miles. Fleet vehicles appear in a number of

such exercises but with considerable gaps in between. Not all vehicles in a fleet are used in all FEDC exercises. Vehicles accrue mileage during the gaps between FEDC exercise uses. That is, gaps may contain non-exercise use and other unobserved field exercises. Recurrences occur but are not observed in these gaps.

2.3 Nonparametric Estimation Methods for Window-Observation Recurrence Data

2.3.1 Notation

We use the following notation. Let $N(t)$ denote the cumulative number of events for a single unit under observation for the period $[0, t]$. The population mean cumulative function is denoted by $\mu(t) = \mathrm{E}[N(t)]$. If $\mu(t)$ is differentiable, then

$$\nu(t) = \frac{d\mu(t)}{dt} \tag{2.1}$$

is the recurrence rate and $\nu(t) \times \Delta t$ can be interpreted as the approximate expected number of events to occur during the next short time interval $(t,\ t + \Delta t)$.

2.3.2 Nonparametric Estimation of the Mean Cumulative Function

First we review the nonparametric method for estimating the population MCF, with some changes in presentation that extend the method to allow for window-observation recurrence data. Nonparametric MCF estimation methods are described, for example, in Nelson (1988), Lawless and Nadeau (1995), Chapter 16 of Meeker and Escobar (1998), and Chapters 3 to 5 of Nelson (2003).

Let n denote the number of observed units and let m denote the number of unique event times. Also, let $t_1, ..., t_m$ be the unique event times. Then the nonparametric estimator of the population MCF is

$$\widehat{\mu}_{\mathrm{NP}}(t_j) = \sum_{k=1}^{j} \left[\frac{\sum_{i=1}^{n} \delta_i(t_k) \times d_i(t_k)}{\sum_{i=1}^{n} \delta_i(t_k)} \right] = \sum_{k=1}^{j} \frac{d_{\cdot}(t_k)}{\delta_{\cdot}(t_k)} = \sum_{k=1}^{j} \bar{d}(t_k), \quad j = 1, ..., m, \tag{2.2}$$

where $d_i(t_k)$ is the number of events recorded at time t_k for unit i, and

$$\delta_i(t_k) = \begin{cases} 1 & \text{if unit } i \text{ is under observation in a time window at time } t_k, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $d.(t_k) = \sum_{i=1}^{n} \delta_i(t_k) \times d_i(t_k)$ is the total number of events reported at time t_k, $\delta.(t_k) = \sum_{i=1}^{n} \delta_i(t_k)$ is the size of the risk set (the number of systems at risk to have an event) at time t_k, taking account of gaps between observation windows and censoring, and $\bar{d}(t_k)$ is the sample mean number of events at time t_k. Thus $\widehat{\mu}_{\mathrm{NP}}(t_j)$ can be viewed as an estimate of the cumulative mean number of events at time t_j, where the mean at the time of each event is computed with respect to the risk set at the time of the event.

It is possible to have ties in the reported recurrence times and observation window endpoint times. If there are ties only in the recurrence times, then either a unit has multiple recurrences at the same time or more than one unit experiences recurrences at the same time. In either case, (2.2) gives the same estimates, because it accumulates the estimates of the increments in the MCF only at the distinct times with recurrences. The same applies for the case when recurrences are reported at either the beginning or the end of a window for a particular unit. When, however, there are ties between a) recurrence times for one unit and b) either the beginning or the end of an observation window for a different unit, additional conventions are needed in the calculation of the size of the risk set. We use the following conventions.

1. For a unit having the beginning of an observation window at the same time as one or more recurrences in other units, the size of the risk set is increased by one, just before the time.

2. For a unit having the end of an observation window at the same time as one or more recurrences in other units, the size of the risk set is reduced by one, just after the time.

If these conventions do not describe the actual situation, it is, of course, possible to override them by breaking any ties in the data before applying our algorithms.

Because the estimate $\widehat{\mu}_{\mathrm{NP}}(t_j)$ is not continuous (there are jumps in the estimate at each event time), we cannot use (2.1) to estimate the recurrence rate. We can, however, obtain a

nonparametric estimate of the recurrence rate over any interval $[t_a, t_b]$ as

$$\hat{\nu}(t_a, t_b) = \frac{\widehat{\mu}_{\text{NP}}(t_b) - \widehat{\mu}_{\text{NP}}(t_a)}{t_b - t_a}.$$

2.3.3 Assumptions

The nonparametric methods described here require no assumptions on the form of the recurrence process that produces the recurrent events. Therefore there is no risk of choosing an incorrect functional form for the MCF. Also, the assumption of independent increments is not needed. Nelson (2003, pages 51-54) lists the assumptions that are required for the nonparametric MCF estimator. Here we modify and adapt those assumptions to the nonparametric MCF estimator with window-observation recurrence data.

1. The units in the sample data are a simple random sample from a well-defined target population. This is a basic assumption needed for extending the information in the sample to the properties of the population. When other sampling methods are used, justification or modification of the statistical procedures is needed for inference about the target population.

2. The population MCF is zero at time zero and exists (i.e., is finite) at any age t of interest up to the greatest censoring age.

3. The stochastic process that generates the observation windows is independent from the stochastic process that generates the recurrences. With this assumption, the position of the windows in the data would provide no information about the recurrence process. This is an extension of the commonly-used non-informative censoring assumption. For the extended warranty example, this assumption would be violated if customers tended to purchase an extended warranty because they have higher recurrence rates than those who do not. For the FEDC example, the assumption would be violated if more intensive use during exercises changes the failure intensity function with respect to miles of service.

4. The size of the risk set must be positive from time zero up to the maximum time that population MCF would be estimated. This assumption is not satisfied by the AMSAA

Vehicle Fleet Data. A naive implementation of the nonparametric estimation method to the data would lead to biased results, as described in Section 2.6.

Under these assumptions, $\widehat{\mu}_{\rm NP}(t_j)$ has the desirable property of being unbiased for window-observation recurrence data. Note that assumptions 1 to 3 are needed for all the MCF estimators that we describe in this paper. Assumption 4 is not needed in parametric estimation.

2.3.4 Variance of $\widehat{\mu}_{\rm NP}(t)$ and Estimator

In addition to the point estimate of $\mu(t_j)$ in (2.2), there is usually need to provide a standard error for $\widehat{\mu}_{\rm NP}(t_j)$. The variance of $\widehat{\mu}_{\rm NP}(t_j)$ is

$$\mathrm{Var}[\widehat{\mu}_{\rm NP}(t_j)] = \sum_{k=1}^{j} \mathrm{Var}[\bar{d}(t_k)] + 2\sum_{k=1}^{j-1}\sum_{v=k+1}^{j} \mathrm{Cov}[\bar{d}(t_k), \bar{d}(t_v)].$$

To estimate the variance of $\widehat{\mu}_{\rm NP}(t_j)$ with window-observation data, we use a modification of the variance estimator given by Lawless and Nadeau (1995). The method and formula described below make it possible to handle large recurrence data sets.

The variance estimate $\widehat{\mathrm{Var}}[\widehat{\mu}_{\rm NP}(t_j)]$ can be computed recursively as follows.

$$\widehat{\mathrm{Var}}[\widehat{\mu}_{\rm NP}(t_1)] = \widehat{\mathrm{Var}}(\bar{d}_1), \tag{2.3}$$

$$\widehat{\mathrm{Var}}[\widehat{\mu}_{\rm NP}(t_j)] = \widehat{\mathrm{Var}}[\widehat{\mu}_{\rm NP}(t_{j-1})] + \widehat{\mathrm{Var}}(\bar{d}_j) + 2\sum_{k=1}^{j-1} \widehat{\mathrm{Cov}}(\bar{d}_k, \bar{d}_j), \text{ for } j = 2, \ldots, m, \tag{2.4}$$

where $\bar{d}_j = \bar{d}(t_j)$ is the average number of events per unit at $t_j, j = 1, \ldots, m$.

To present the details of the calculations, we use the following notation, defined for $j, k = 1, \ldots, m$.

- A_j is the total number of unique unit-event times, summed over all units and accumulated over times $(0, t_j]$ (e.g., if exactly two units have one or more events at a particular point in time, A_j would increase by two at that time).
- $\delta_j = \delta_{\cdot}(t_j)$ is the size of the risk set at time t_j- (i.e., just before t_j).
- $\delta_{j,k} = \sum_{i=1}^{n} \delta_i(t_j)\delta_i(t_k)$ is the number of units that were under observation at both t_j- and t_k-.

- Also,

$$\bar{d}_j^{\,k} = \bar{d}^k(t_j) = \frac{\sum_{i=1}^{n} \delta_i(t_k)\delta_i(t_j)d_i(t_j)}{\delta_{j,k}} \tag{2.5}$$

  is the average increase in number of events at time t_j for all of those units that were under observation at both t_j- and t_k-.

Then, we have

$$\begin{aligned}
\widehat{\mathrm{Var}}(\bar{d}_j) &= \frac{1}{\delta_j^2}\sum_{u=1}^{n}\delta_u(t_j)[d_u(t_j)-\bar{d}_j]^2 \\
&= \frac{1}{\delta_j^2}\left[\sum_{i=A_{j-1}+1}^{A_j}(d_i(t_j)-\bar{d}_j)^2 + (\delta_j - A_j + A_{j-1})\bar{d}_j^{\,2}\right]. \qquad (2.6)
\end{aligned}$$

In (2.6), the term $(\delta_j - A_j + A_{j-1})\bar{d}_j^{\,2}$ accounts for units that were under observation at time t_j- but had no events at that time. Also,

$$\begin{aligned}
\widehat{\mathrm{Cov}}(\bar{d}_k, \bar{d}_j) &= \frac{1}{\delta_k\delta_j}\sum_{u=1}^{n}\delta_u(t_k)d_u(t_k)\delta_u(t_j)[d_u(t_j)-\bar{d}_j^{\,k}] \\
&= \frac{1}{\delta_k\delta_j}\sum_{l=A_{k-1}+1}^{A_k}\left\{d_l(t_k)I(t_j\in O_l)\left[\sum_{i=A_{j-1}+1}^{A_j} I(K_i=K_l)d_i(t_j) - \bar{d}_j^{\,k}\right]\right\} \qquad (2.7)
\end{aligned}$$

where for event l, K_l is the unit ID (identification number) for the system having the event, d_l is the corresponding number of the event, and O_l is the set of observation intervals for the system. In addition, $I(t_j \in O_l)$ is the indicator function for whether the time t_j is in the set of observation intervals O_l while $I(K_i = K_l)$ is the indicator function for whether event i and event l are for the same system. By (2.5), the sample mean number of recurrent events at time t_j for units that are under observation at time t_k is

$$\bar{d}_j^{\,k} = \frac{1}{\delta_{j,k}}\sum_{i=A_{j-1}+1}^{A_j} I(t_k \in O_i)d_i.$$

When there are no gaps in the observation period for any units, our variance estimator is equivalent to that of Lawless and Nadeau (1995). The main difference required for the window-observation data is in the computation of the sample mean in (2.5) and the covariances in (2.7), which must account for which units with an event at time t_j were under observation at event times $t_k < t_j$.

2.3.5 Confidence Intervals for $\mu(t)$

Using $\hat{\mu}(t)$ and $\widehat{\text{Var}}[\hat{\mu}(t)]$ (we omit the indication of a particular estimator because the methods described here apply more generally to any MCF estimator), pointwise normal-approximation confidence intervals are easy to compute. Out of the many possible normal-approximation methods, the following equations (Meeker and Escobar, 1998, Chapter 16, page 400) are two that are relatively easy to program and are commonly used.

- Based on $Z_{\hat{\mu}(t)} = [\hat{\mu}(t) - \mu(t)]/\widehat{\text{se}}_{\hat{\mu}(t)} \dot{\sim} \text{NOR}(0,1)$, an approximate $100(1-\alpha)\%$ confidence interval for $\mu(t)$ is

$$\hat{\mu}(t) \pm z_{(1-\alpha/2)}\widehat{\text{se}}_{\hat{\mu}(t)}, \tag{2.8}$$

  where $\widehat{\text{se}}_{\hat{\mu}(t)} = \sqrt{\widehat{\text{Var}}[\hat{\mu}(t)]}$, and $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

- Based on $Z_{\log[\hat{\mu}(t)]} = \{\log[\hat{\mu}(t)] - \log[\mu(t)]\}/\widehat{\text{se}}_{\log[\hat{\mu}(t)]} \dot{\sim} \text{NOR}(0,1)$, an approximate $100(1-\alpha)\%$ confidence interval for $\mu(t)$ is

$$[\hat{\mu}(t)/w, \quad \hat{\mu}(t) \times w], \tag{2.9}$$

  where $w = \exp[z_{(1-\alpha/2)}\widehat{\text{se}}_{\hat{\mu}(t)}/\hat{\mu}(t)]$.

Equation (2.9) might provide a better approximate confidence interval procedure in applications where $\mu(t)$ is strictly positive (and the interval endpoints will always be positive). Meeker and Escobar (1998, Chapter 16, page 402) point out that when the size of the risk set is small (say, less than 30), using $t_{(p;\nu)}$ (the p quantile of a t distribution with ν degrees of freedom) instead of $z_{(p)}$ in (2.8) and (2.9) can provide a confidence interval procedure with a coverage probability that is closer to the nominal value.

Table 2.1 Extended Warranty Data

VIN Group:	VG1	VG2	VG3	VG4	Total
Observation Window(s):	[0, 36]	[0, 24]	[0, 12]	[0, 12] & [24, 36]	
Number of Automobiles:	48,300	32,395	64,307	16,044	161,046
Percent in Group(%):	30	20	40	10	100
For Labor Code C6050, the number of automobiles					
With Events	82	30	37	25	174
Without Events	48,218	32,365	64,270	16,019	160,872

2.4 Analysis of the Extended Warranty Data

2.4.1 Descriptions of the Data

Actual extended warranty data that we have seen are not available for publication. To help illustrate the extension of existing estimation methods to window-observation recurrence data, we used automobile warranty data to which we do have access to create an extended warranty data set with simulated windows.

The original warranty data are for 161,046 automobiles from model year 1995 that were sold between August 1994 and November 1995. For each automobile, the following information is available: VIN number, build date, sale date, and the complete warranty-report history up to three years of service or until the end of November 1998 (the cutoff date of the data), whichever came first. For these automobiles, there had been 586,750 repair events with 1,745 distinct labor codes. Some of these automobiles never had a warranty report. Many had more than one.

We generated the extended warranty data for these 161,046 automobiles on events with a particular labor code C6050. The automobiles were randomly assigned to one of the four warranty plans shown in Table 2.1, according to the percentages shown there. That is, all automobiles are covered in their first year of service; 30% elect for two extra years, 20% for one additional year, 40% do not extend beyond the first year, while 10% skip year 2, but return to warranty coverage in year 3. The time scale is months of service (months since an automobile had been sold).

2.4.2 Graphical Exploration of the Data

Because of the large number of automobiles in the dataset (161,046 in total), it is not possible to draw a standard event plot for all of the units. To present a more complete picture, we provide two alternative event plots. Figure 2.1 shows an event plot for only 16 units, three units that had one or two events and one that had no events, for each of the four window schemes. Note that the vast majority of units had no events, some units had one event, and very rarely did units have two events. The numbers shown to the right of the life lines reflect the number of units at risk and (in parentheses) the number of events, respectively.

Figure 2.2 shows the super-imposed event plot of the window-observation warranty data, with two lines for each of the four window schemes: one for units with events, and the other for units that had no events. The numbers to the right of the horizontal lines provide information similar to those in Figure 2.1. The change in density of events along time suggests that there was a burn-in period at the beginning of life, after which the recurrence rate stabilized, or even decreased toward the end of observation.

Because of the observation windows imposed on the data, the number of automobiles under observation changed during the period of 0 to 36 months. This is reflected in Figure 2.3, which shows the changing size of the risk set over time. The size of the risk set of the original warranty data remains constant at the value of 161,046 for the same period, and thus the risk set plot is a horizontal line (not shown).

2.4.3 Estimation with the Nonparametric Method

Figure 2.4 shows, for the Extended Warranty Data, the nonparametric MCF estimate of labor code C6050, and the corresponding approximate 95% pointwise confidence intervals. The discontinuities in the estimates reflect the jumps at the time of events. The confidence intervals were computed using (2.8).

Regarding the four model assumptions for the nonparametric method in Section 2.3.3, no severe violations were detected for the Extended Warranty Data. However, one might have questions about assumption 2, because in real life, customers tend to choose to buy an

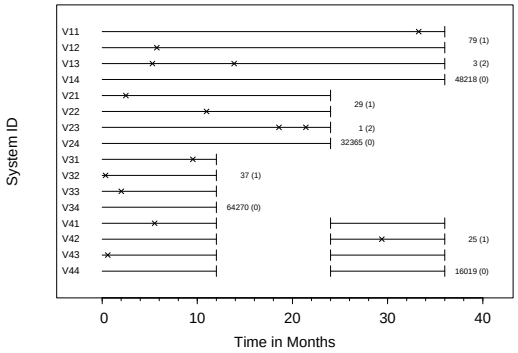


Figure 2.1 Extended Warranty Data – Event Plot of Selected Units for Labor Code C6050. Note that for lines corresponding to a group of units, $n_1(n_2)$ indicates there are n_1 units in the data and that each of these units has n_2 events.

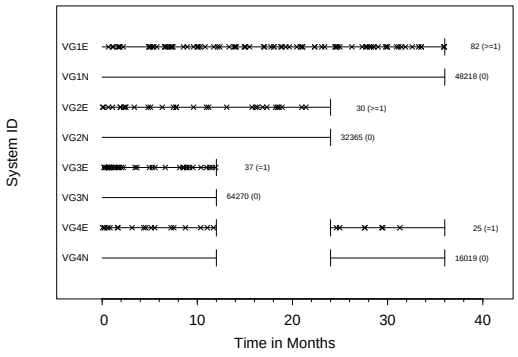


Figure 2.2 Extended Warranty Data – Super-Imposed Event Plot for Labor Code C6050. Note that $n_1(n_2)$ indicates there are n_1 units in the data and that these units had the indicated number of events.

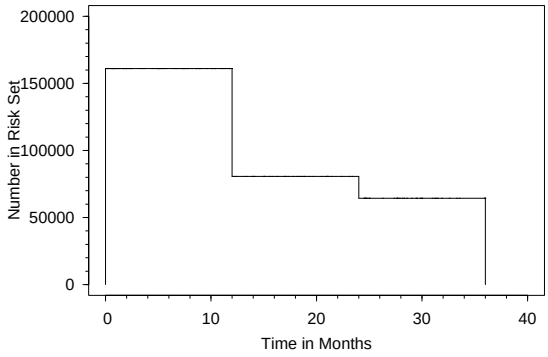


Figure 2.3 Extended Warranty Data – Risk Set Plot for Labor Code C6050

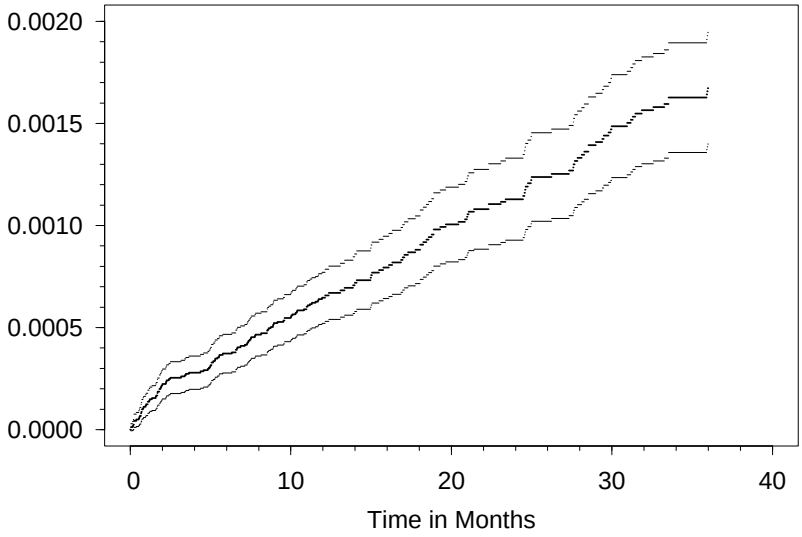


Figure 2.4 MCF plot for Labor Code C6050 Extended Warranty Data

extended warranty because they expect their units to have a higher than usual recurrence rate (e.g., because of harsher than usual environmental conditions). Because there is only a small number of window schemes (four in our example), it is possible to check whether the windowing process is informative by comparing the appropriate parts of the sample MCF's among units having the four different window schemes, with methods described in Doganaksoy and Nelson (1998) and Cook, Lawless, and Nadeau (1996). If important differences are detected, it might be better to segment population by window scheme and do the analyses separately.

2.5 Maximum Likelihood Estimation of a Nonhomogeneous Poisson Process (NHPP) Model Using Window-Observation Recurrence Data

This section describes how to fit a parametric model to the recurrence data. This is, in effect, fitting a curve through the nonparametric MCF estimate. Compared to nonparametric methods, fitting a parametric model has advantages such as a more concise model with just a few parameters and the ability to extrapolate outside the range of the data. It is

straightforward to show that the maximum likelihood (ML) estimator is generally consistent for window-observation recurrence data, even if the size of the risk set is sometimes equal to zero. This can, for example, be done by using methods described in Yuan and Jennrich (1998).

2.5.1 Assumptions for the Poisson Process

The Poisson process is a widely used model for point processes, and is covered in many books, such as Cox and Lewis (1966) and Rigdon and Basu (2000). In addition to assumptions 1-3 described for the nonparametric MCF estimator, two additional assumptions for the NHPP MCF estimator are:

1. The specified recurrence rate function determines the form of the MCF function.

2. The recurrence history for a unit has independent increments (i.e., the numbers of recurrences in non-overlapping intervals are independent). Thompson (1988, page 21) has more technical details on the concept of independent increments.

Many goodness-of-fit tests are available to check the first assumption (e.g., pages 127 and 141 of Rigdon and Basu 2000 for the NHPP power model). For the second assumption, Cox and Lewis (1966, pages 164-167) describe tests based on serial correlation coefficients, and these tests could be adapted for the window-observation recurrence data to detect whether there is evidence of strong correlation among serial recurrences. Another type of non-independence arises from a mixture of populations. Nelson (2003, pages 74-75) describes such an example. For this case, it is possible to check homogeneity of the population by comparing sample MCFs for different groups of units, as described in Doganaksoy and Nelson (1998) and Cook, Lawless, and Nadeau (1996).

The impact of violations to these model assumptions depends on the application. Violation of the independent increments assumption will not cause serious bias in the point estimates of the MCF, but the naive variance estimators would be biased and thus confidence intervals based on such variance estimates would not be accurate. On the other hand, if the assumed

recurrence rate function does not adequately describe the process, point estimates would be biased, and if the departure is large, the bias could be of practical importance.

2.5.2 NHPP Model

A particular NHPP model is specified by its recurrence rate function $\nu(t)$. Let $N(a, b]$ be the number of events in the time range $(a, b]$ from an NHPP with recurrence rate $\nu(t)$. The expectation of $N(a, b]$ over this interval is $\mu(a, b) = \int_a^b \nu(t)dt$. The most commonly used NHPP recurrence rate functions are:

- Constant recurrence rate, also known as homogeneous Poisson process (HPP):

$$\nu(t) = c.$$

- Power recurrence rate, also known as power law process or Weibull process:

$$\nu(t; \beta, \eta) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \quad \beta > 0, \eta > 0.$$

  Note that, for the power law NHPP model, the time to first recurrence follows a Weibull distribution with scale parameter η and shape parameter β. Meeker and Escobar (1998, page 127) point out that η can be interpreted as roughly the 0.632 quantile of a Weibull random variable. Therefore, for the power law NHPP model, η can be interpreted as 0.632 quantile of the time to the first recurrence.

  The shape parameter β determines whether the recurrence rate is increasing ($\beta > 1$), constant ($\beta = 1$) or decreasing ($\beta < 1$). Figure 2.5 shows the plots of the recurrence rate and the MCF for the power law NHPP model with $\beta = 0.8$, 1, 2, and 3, and $\eta = 1$. When $\beta = 1$, the power law NHPP model simplifies to the HPP model with constant recurrence rate $1/\eta$.

- Loglinear recurrence rate:

$$\nu(t; \gamma_0, \gamma_1) = \exp(\gamma_0 \ + \ \gamma_1 \times t).$$

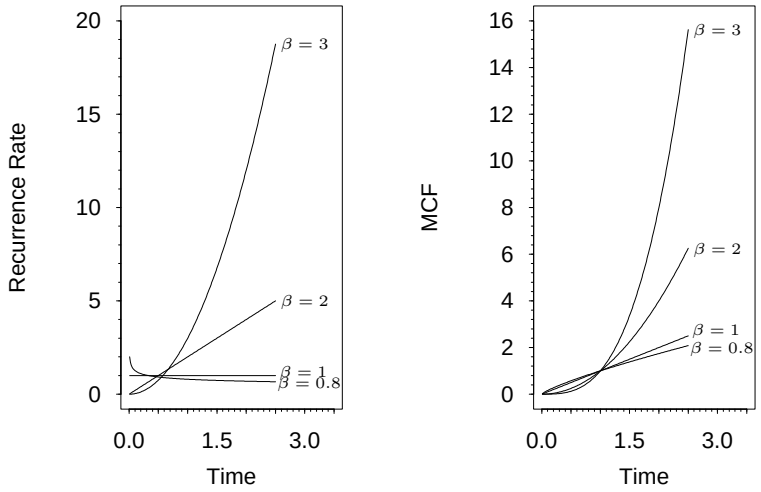


Figure 2.5 Power Law NHPP Model – Plots of the Recurrence Rate and the MCF with $\eta = 1$ and Different β Values

The loglinear NHPP model assumes that the log recurrence function is a linear function of time, where γ_0 is the intercept and γ_1 is the slope. A positive value of γ_1 indicates an increasing recurrence rate, while a negative value of γ_1 indicates a decreasing recurrence rate. When $\gamma_1 = 0$, the NHPP loglinear model simplifies to the HPP model with recurrence rate $\exp(\gamma_0)$.

2.5.3 NHPP Likelihood for the Window-Observation Data

For unit i with exact event times for the observation period $(0, t_{a_i}]$ and no gaps in observation, the NHPP likelihood is

$$L_i(\boldsymbol{\theta}) \quad = \quad \left\{ \prod_{j=1}^{r_i} \nu(t_{ij}; \boldsymbol{\theta}) \right\} \{\exp\left[-\mu(0, t_{a_i}; \boldsymbol{\theta})\right]\}, \tag{2.10}$$

where $\boldsymbol{\theta}$ is the unknown parameter vector (e.g., $\boldsymbol{\theta} = (\beta, \eta)'$ for the power law NHPP model), r_i is the total number of events being observed for unit i, and $t_{i1}, ..., t_{ir_i}$ are the corresponding unique event times.

For window-observation recurrence data, denote the non-overlapping windows of observation for unit i as $(t_{i1L}, t_{i1U}]$, $(t_{i2L}, t_{i2U}]$, ..., $(t_{ip_iL}, t_{ip_iU}]$ (with $t_{i1L} \geq 0$, $t_{i,k-1,U} < t_{ikL}$, and $t_{ip_iU} < t_{a_i}$). The NHPP likelihood for unit i with events in those observation windows reported at exact times is

$$L_i(\boldsymbol{\theta}) = \left\{\prod_{j=1}^{r_i} \nu(t_{ij};\boldsymbol{\theta})\right\}\left\{\prod_{k=1}^{p_i} \exp\left[-\mu(t_{ikL}, t_{ikU};\boldsymbol{\theta})\right]\right\}. \quad (2.11)$$

It is easy to establish (2.10) and (2.11) by writing the probability of the data for interval censored data (event times are known to be in specific intervals) and then allowing the width of the intervals to approach 0. The observation windows are assumed to be prefixed, and thus the random variables in the likelihood functions are r_i and $t_{i1}, ..., t_{ir_i}$, the number of recurrences for unit i and the corresponding recurrence times.

For a sample of n independent NHPP systems with the same intensity function, the overall likelihood is simply the product of the likelihoods for the individual units,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}), \quad (2.12)$$

and $\hat{\boldsymbol{\theta}}$, the ML estimator of $\boldsymbol{\theta}$, is obtained by maximizing (2.12) or its logarithm. Generally, this must be done numerically. Given $\hat{\boldsymbol{\theta}}$, the ML estimator of the NHPP MCF is

$$\widehat{\mu}_{\text{NHPP}}(t) = \int_0^t \nu(x;\hat{\boldsymbol{\theta}})dx.$$

2.5.4 Variance of $\widehat{\mu}_{\text{NHPP}}$ and Estimator

Using the delta method, an approximate variance of $\widehat{\mu}_{\text{NHPP}}$ is

$$\text{Var}[\widehat{\mu}_{\text{NHPP}}(t)] \doteq \left[\frac{\partial \mu(t)}{\partial \boldsymbol{\theta}}\right]' \Sigma_{\hat{\boldsymbol{\theta}}} \left[\frac{\partial \mu(t)}{\partial \boldsymbol{\theta}}\right], \quad (2.13)$$

where $\Sigma_{\hat{\boldsymbol{\theta}}}$ is the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$. An estimator of $\Sigma_{\hat{\boldsymbol{\theta}}}$ can be obtained by evaluating the inverse of the negative Hessian matrix at $\hat{\boldsymbol{\theta}}$.

$$\widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} = -\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

An estimator for the variance in (2.13) is

$$\widehat{\mathrm{Var}}[\widehat{\mu}_{\mathrm{NHPP}}(t)] = \left[\frac{\partial \mu(t)}{\partial \boldsymbol{\theta}}\right]'_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} \left[\frac{\partial \mu(t)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

For example, using the power law NHPP model,

$$\mu(t) = \int_0^t \nu(x;\beta,\eta)dx = \left(\frac{t}{\eta}\right)^{\beta}$$

and

$$\frac{\partial \mu(t)}{\partial \boldsymbol{\theta}} = \left[\left(\frac{t}{\eta}\right)^{\beta} \ln\left(\frac{t}{\eta}\right),\ -\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta}\right]', \quad \text{where } \boldsymbol{\theta} = [\beta,\ \eta]'. \tag{2.14}$$

Thus an estimator of the variance of $\widehat{\mu}_{\mathrm{NHPP}}(t)$ for the power law NHPP model is

$$\begin{aligned}
\widehat{\mathrm{Var}}[\widehat{\mu}_{\mathrm{NHPP}}(t)] &= \left[\left(\frac{t}{\hat{\eta}}\right)^{\hat{\beta}} \ln\left(\frac{t}{\hat{\eta}}\right)\right]^2 \widehat{\mathrm{Var}}(\hat{\beta}) \\
&+ 2\left[\left(\frac{t}{\hat{\eta}}\right)^{\hat{\beta}} \ln\left(\frac{t}{\hat{\eta}}\right)\right]\left[-\frac{\hat{\beta}}{\hat{\eta}}\left(\frac{t}{\hat{\eta}}\right)^{\hat{\beta}}\right] \widehat{\mathrm{Cov}}(\hat{\beta},\ \hat{\eta}) \\
&+ \left[-\frac{\hat{\beta}}{\hat{\eta}}\left(\frac{t}{\hat{\eta}}\right)^{\hat{\beta}}\right]^2 \widehat{\mathrm{Var}}(\hat{\eta}).
\end{aligned}$$

2.6 Analysis of the AMSAA Vehicle Fleet Data

2.6.1 Descriptions on the Data

Because actual FEDC data are sensitive, we were asked to analyze data that had been simulated by an analyst at AMSAA, according to a process that is similar to the process that generates actual FEDC data.

Three datasets were provided to us, one with complete history of each vehicle up to the end of the observation time, while the other two with different levels of gaps (i.e., intervals with recurrence history missing). With all three datasets and the true model of the recurrence process, it is possible for us to compare the performances of different estimators. In particular,

- The recurrences in the AMSAA Vehicle Fleet *Complete Data* were simulated from a power law NHPP model with $\beta = 2.76$ and $\eta = 5447$ miles, using the method described

in Section 16.7 of Meeker and Escobar (1998). For each of the ten vehicles in the fleet, the simulation was run until the end of the observation period had been reached. For each vehicle, the end of the observation period was simulated from a uniform distribution between 20 and 30 thousand miles.

- The AMSAA Vehicle Fleet *Random Selection Window Data* were generated from the Complete AMSAA Vehicle Fleet Data, by screening the recurrence history of each vehicle through the simulated observation windows (FEDC exercises) and the gaps (time between these exercises). The length of each exercise (observation window) was simulated from a uniform distribution between 400 to 600 miles. The gaps between exercises were simulated from a uniform distribution between 600 and 1400 miles.

- The AMSAA Vehicle Fleet *Non-random Selection Window Data* used a model in which vehicles with more miles of service are less likely to be chosen for an exercise (in effect increasing the length of the gaps for vehicles as the number of miles of service gets larger). The data also started with the Complete AMSAA Vehicle Fleet Data. The length of each exercise was also simulated from a uniform distribution between 400 to 600 miles. However, the length for gap i was simulated from a uniform distribution between $200 \times 2^{(i-1)}$ and $400 \times 2^{(i-1)}$ miles. For example, the length of time to the first exercise, gap 1, was simulated from a uniform distribution between 200 and 400 miles. Such data arise, for example, when commanders are permitted to choose lower-mileage vehicles that they might perceive as being less likely to fail during the exercise.

2.6.2 Graphical Exploration of the Data

Figures 2.6 to 2.8 are the event plots for the three datasets, showing the recurrent events and the observation windows. Because of the method that generated the datasets, the events shown on the plots for the same units are the same. The difference is in the observation windows for the Random Selection Window Data and the Non-random Selection Window Data.

Figures 2.9 to 2.11 are the risk set plots of the three datasets. The size of the risk set for the Complete Data is positive until the maximum end-of-observation time. There are some

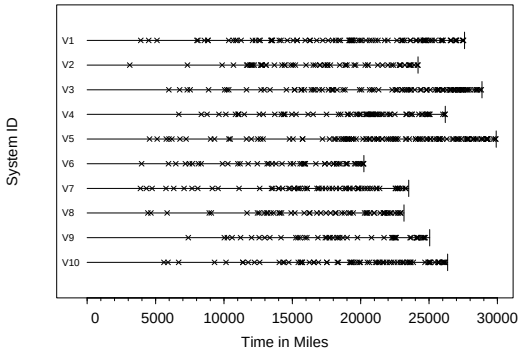


Figure 2.6 Event Plot for AMSAA Vehicle Fleet Complete Data

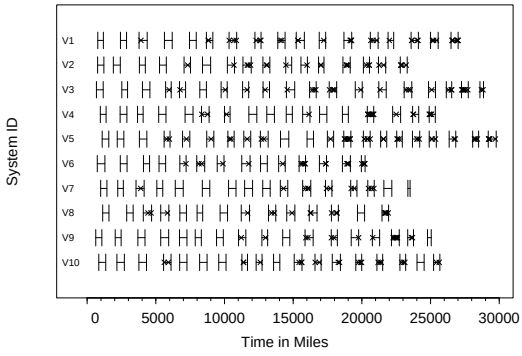


Figure 2.7 Event Plot for AMSAA Vehicle Fleet Random Selection Window Data

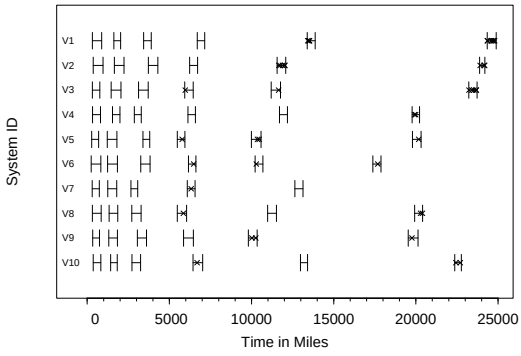


Figure 2.8 Event Plot for AMSAA Vehicle Fleet Non-random Selection Window Data

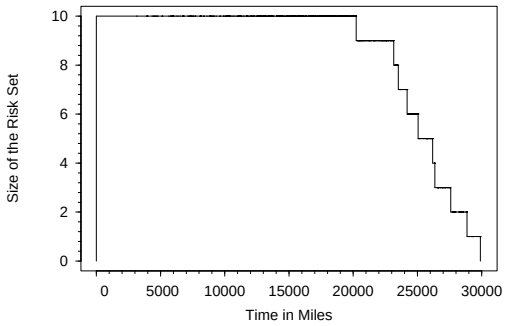


Figure 2.9 Risk Set Plot for AMSAA Vehicle Fleet Complete Data

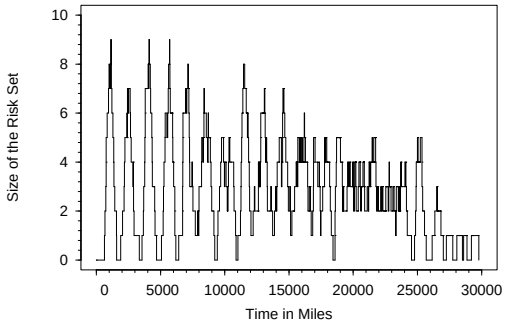


Figure 2.10 Risk Set Plot for AMSAA Vehicle Fleet Random Selection Window Data

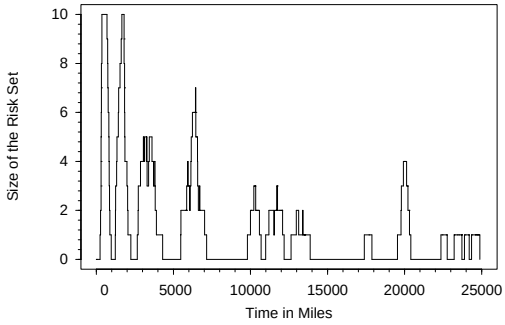


Figure 2.11 Risk Set Plot for AMSAA Vehicle Fleet Non-random Selection Window Data

Table 2.2 Simulated Vehicle Data – Comparisons of Risk Set Sizes

Risk Set	*Complete Data*		*Random Selection Window Data*		*Non-random Selection Window Data*	
size	time	pct (%)	time	pct (%)	time	pct (%)
0	0	0.00	3949	13.26	13371	53.73
1	1042	3.48	5349	17.96	5235	21.04
2	1271	4.25	5444	18.28	2493	10.02
> 2	27593	92.27	15037	50.49	3786	15.21
Total	29906	100	29779	100	24885	100

intervals in time over which the size of the risk set is zero (RSSZ intervals) for the Random Selection Window Data and there are much wider RSSZ intervals for the Non-random Selection Window Data.

Table 2.2 shows the amount of time that the risk set size is at the values of 0, 1, 2 and greater than 2, for the Complete Data, the Random Selection Window Data, and the Non-random Selection Window Data. Note that for the Non-random Selection Window Data, the size of the risk set is zero for 53.73% of the total observation time, while for the Random Selection Window Data, the figure is 13.26%.

2.6.3 Nonparametric and Parametric Estimation

In this section, we apply the nonparametric (NP) MCF estimation method, the power law NHPP model, and the loglinear NHPP model to the three simulated AMSAA Vehicle Fleet data sets. Figures 2.12 to 2.14 compare the resulting MCF estimates with the "true" model that generated the simulated data. Because the MCF estimates of the power law NHPP model and those of the loglinear NHPP model are very close to each other for the Complete Data and the Random Selection Window Data, we did not show the loglinear MCF estimates for these two datasets on the plots. Tables 2.3 and 2.4 summarize the estimated model parameters, as well as the value of the maximum log likelihood, of the power law NHPP model and loglinear NHPP model for the three data sets. We note the following:

- All of the NHPP estimates have good agreement with the true MCF.
- For all of the data sets, the values of the maximum log likelihood are somewhat larger

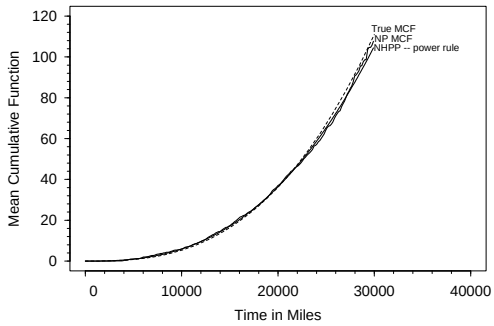


Figure 2.12 AMSAA Vehicle Fleet Complete Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, and the True MCF

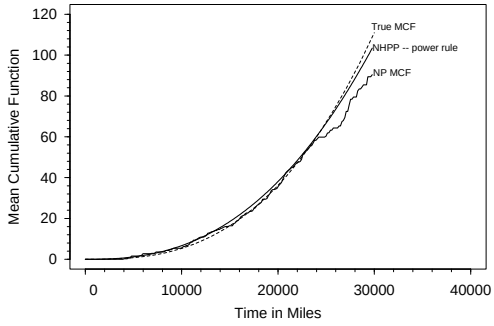


Figure 2.13 AMSAA Vehicle Fleet Random Selection Window Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, and the True MCF

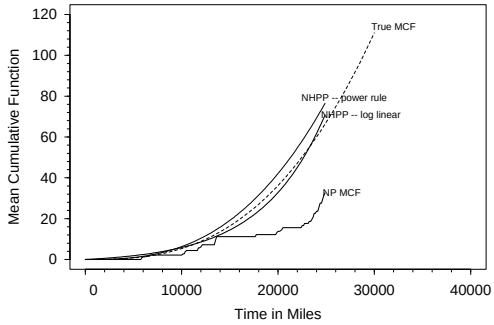


Figure 2.14 AMSAA Vehicle Fleet Non-random Selection Window Data: Nonparametric MCF Estimate, Power Law NHPP MCF Estimate, Loglinear NHPP MCF Estimate, and the True MCF

Table 2.3 AMSAA Vehicle Fleet Data – Comparisons of Parameter Estimates of the Power Law NHPP Model

Parameter		*Complete Data*	*Random Selection Window Data*	*Non-random Selection Window Data*
	MLE	5063.070	4686.747	5098.635
η	Std.Err.	310.798	515.508	801.295
(true value: 5447)	95% Lower Bound	4453.920	3676.370	3528.126
	95% Upper Bound	5672.223	5697.123	6669.144
	MLE	2.617	2.509	2.736
β	Std.Err.	0.095	0.156	0.276
(true value: 2.76)	95% Lower Bound	2.430	2.202	2.195
	95% Upper Bound	2.804	2.815	3.276
Maximum Log Likelihood		-4606	-1564	-298

Table 2.4 AMSAA Vehicle Fleet Data – Comparisons of Parameter Estimates of the Loglinear NHPP Model

Parameter		*Complete Data*	*Random Selection Window Data*	*Non-random Selection Window Data*
	MLE	-7.728	-7.558	-8.296
γ_0	Std.Err.	0.114	0.190	0.362
	95% Lower Bound	-7.952	-7.931	-9.006
	95% Upper Bound	-7.504	-7.185	-7.586
	MLE	0.0001140	0.0001060	0.0001523
γ_1	Std.Err.	0.0000057	0.0000095	0.0000196
	95% Lower Bound	0.0001030	0.0000870	0.0001139
	95% Upper Bound	0.0001250	0.0001250	0.0001906
Maximum Log Likelihood		-4624	-1570	-303

for the power law NHPP model (from which the data were simulated), when compared with the loglinear NHPP model.

- For the Complete Data, the nonparametric estimate is very close to the true MCF.
- For the Random Selection Window Data, the nonparametric estimate is somewhat smaller than the true MCF, especially after 24,000 miles.
- For the Non-random Selection Window Data, the nonparametric estimate is importantly below the true MCF.

The reason for the downward bias in the NP MCF estimator for the Random Selection Window Data and the Non-random Selection Window Data is the existence of the RSSZ intervals for these two datasets, a violation of Assumption 4 in Section 2.3.3. To help solve the problem, Section 2.7 suggests two hybrid estimators.

When estimating the variance of the NP MCF estimator, there can be estimation problems when the size of the risk set is one or two. In particular, when the size of the risk set is one at event time t_k, then $\text{Var}[\bar{d}(t_k)]$ is not estimable. Also, unless at least two units are being observed at both event times t_k and t_v, $\text{Cov}[\bar{d}(t_k), \bar{d}(t_v)]$ is not estimable.

For the NHPP model assumptions, close agreement between the power law model and loglinear model estimates and their correspondence with the true model indicates that both recurrence forms describe the data well. Also, there is no evidence of violation of the independent-increments assumption (as expected because of the way that the data were simulated). Unlike the Extended Warranty Data, where different operating environments and driving habits are likely to cause inhomogeneity of the automobile population, the vehicles in the AMSAA data are expected to be more homogeneous.

2.7 The Hybrid Estimator

2.7.1 General Approach

For window-observation recurrence data, which has both RSSZ intervals and risk-set-size-positive (RSSP) intervals, we suggest using a hybrid estimator obtained from the following three steps.

1. Apply the NP method to estimate the increase in the MCF over the RSSP intervals.

2. Use information from the RSSP intervals to estimate increase in the MCF over the RSSZ intervals.

3. Calculate the hybrid estimate for the MCF by summing over time the estimated increase in the MCF over the RSSP intervals and RSSZ intervals, as obtained in the above two steps.

Additional assumptions are required for Step 2, and we propose two alternatives, the local hybrid estimator (suggested to us by a referee) and the NHPP hybrid estimator. Details are provided in the following subsections.

The hybrid estimators require additional notation. Let t be the time of interest at which the MCF will be estimated, let t_{max} be the largest end-of-observation time among all units, and let q be the number of RSSZ intervals. Define the beginning and ending time points of the q RSSZ intervals as $(t_{1L}, t_{1U}]$, $(t_{2L}, t_{2U}]$, ..., $(t_{qL}, t_{qU}]$ (with $t_{1L} \geq 0$, $t_{i-1,U} < t_{iL}$, and $t_{qU} < t_{max}$). Then the corresponding beginning and ending time points of the q RSSP intervals following these RSSZ intervals are $(t_{1U}, t_{2L}]$, $(t_{2U}, t_{3L}]$, ..., $(t_{qU}, t_{max}]$.

Within the RSSP intervals, the estimate of the MCF changes only at times that have events. At each unique event time t_k, the estimated increase in the MCF is

$$\bar{d}(t_k) = \frac{\sum_{i=1}^{n} \delta_i(t_k) \times d_i(t_k)}{\sum_{i=1}^{n} \delta_i(t_k)}.$$

Therefore, for the RSSP interval $(t_{iU}, t_{i+1,L}]$, the estimated increase in the MCF is

$$\bar{d}.(t_{iU}, t_{i+1,L}) = \sum_{k: t_{iU} < t_k \leq t_{i+1,L}} \bar{d}(t_k),$$

and the sample mean recurrence rate over this RSSP interval is

$$\bar{d}.(t_{iU}, t_{i+1,L})/(t_{i+1,L} - t_{iU}).$$

2.7.2 The Local Hybrid Estimator

The local hybrid estimator is purely nonparametric, and uses a weighted mean of the recurrence rates of the two neighboring RSSP intervals, weighted by the length of those RSSP intervals, to estimate the recurrence rate over an RSSZ interval. If the data begin with an RSSZ interval, the estimate of the recurrence rate in the first RSSZ interval is the same as that of the first RSSP interval. Thus to calculate recurrence rate for each of the q RSSZ intervals, we use

$$\nu_i = \begin{cases} \frac{\bar{d}.(t_{1U},t_{2L})}{(t_{2L}-t_{1U})} & \text{if } i=1 \text{ and data begin with an RSSZ interval} \\ \frac{\bar{d}.(0,t_{1L})+\bar{d}.(t_{1U},t_{2L})}{(t_{1L}-0)+(t_{2L}-t_{1U})} & \text{if } i=1 \text{ and data begin with an RSSP interval} \\ \frac{\bar{d}.(t_{i-1,U},t_{iL})+\bar{d}.(t_{iU},t_{i+1,L})}{(t_{iL}-t_{i-1,U})+(t_{i+1,L}-t_{iU})} & \text{for } i=2,3,...,q-1 \\ \frac{\bar{d}.(t_{q-1,U},t_{qL})+\bar{d}.(t_{qU},t_{max})}{(t_{qL}-t_{q-1,U})+(t_{max}-t_{qU})} & \text{for } i=q. \end{cases}$$

Then the estimated increase in the MCF for RSSZ interval i is

$$d_i^{\ddagger} = (t_{iU} - t_{iL}) \times \nu_i,$$

and the local hybrid estimator of the MCF at time t is

$$\widehat{\mu}_{\text{Localhybrid}}(t) = \bar{d}.(t) + d_{\cdot}^{\ddagger}(t) \tag{2.15}$$

where $\bar{d}.(t) = \sum_{k:t_k \le t} \bar{d}(t_k)$ and

$$d_{\cdot}^{\ddagger}(t) = \begin{cases} \sum_{i:t_{iU} \le t} d_i^{\ddagger} & \text{if } t \text{ is not in an RSSZ interval} \\ \sum_{i:t_{iU} \le t} d_i^{\ddagger} + (t - t_{jL}) * \nu_j & \text{if } t \text{ is in the } j\text{th RSSZ interval } (t_{jL} < t \le t_{jU}). \end{cases}$$

2.7.3 The NHPP Hybrid Estimator

The NHPP hybrid estimator uses an NHPP model to estimate the increase in the MCF over the RSSZ intervals. Let $\hat{\nu}(t)$ be the estimated recurrence rate of the NHPP model. Then the estimated increase in the MCF for the RSSZ interval i is

$$d_i^{\dagger} = \int_{t_{iL}}^{t_{iU}} \hat{\nu}(t)dt,$$

and the NHPP hybrid estimator of the MCF at time t is

$$\widehat{\mu}_{\text{NHPPhybrid}}(t) = \bar{d}.(t) + d_{\cdot}^{\dagger}(t) \tag{2.16}$$

where $\bar{d}.(t) = \sum_{k:t_k \le t} \bar{d}(t_k)$ and

$$d_{\cdot}^{\dagger}(t) = \begin{cases} \sum_{i:t_{iU} \le t} d_i^{\dagger} & \text{if } t \text{ is not in an RSSZ interval} \\ \sum_{i:t_{iU} \le t} d_i^{\dagger} + \int_{t_{jL}}^{t} \hat{\nu}(x)dx & \text{if } t \text{ is in the } j\text{th RSSZ interval } (t_{jL} < t \le t_{jU}). \end{cases}$$

The resulting NHPP hybrid estimator $\widehat{\mu}_{\text{NHPPhybrid}}(t)$ consists of two parts: the contribution from the nonparametric model $\bar{d}_{\cdot}(t)$ and the contribution from the parametric adjustment $d_{\cdot}^{\dagger}(t)$. If the proportion of time with RSSZ intervals is relatively small, the hybrid estimator will be dominated by the nonparametric estimator; otherwise, it will be more strongly affected by the parametric estimator.

2.7.4 Approximate Variance of the Hybrid Estimator

Direct computation from (2.15) and (2.16) provides the following equations for the variances of $\widehat{\mu}_{\text{Localhybrid}}(t)$ and $\widehat{\mu}_{\text{NHPPhybrid}}(t)$, respectively:

$$\text{Var}[\widehat{\mu}_{\text{Localhybrid}}(t)] = \text{Var}[\bar{d}_{\cdot}(t)] + \text{Var}[d_{\cdot}^{\ddagger}(t)] + 2\text{Cov}[d_{\cdot}^{\ddagger}(t),\ \bar{d}_{\cdot}(t)],$$

$$\text{Var}[\widehat{\mu}_{\text{NHPPhybrid}}(t)] = \text{Var}[\bar{d}_{\cdot}(t)] + \text{Var}[d_{\cdot}^{\dagger}(t)] + 2\text{Cov}[d_{\cdot}^{\dagger}(t),\ \bar{d}_{\cdot}(t)].$$

Note that $\text{Var}[\widehat{\mu}_{\text{Localhybrid}}(t)]$ and $\text{Var}[\bar{d}_{\cdot}(t)]$ are both obtained as the sum of a set of $\text{Var}[\bar{d}(t_k)]$ and $\text{Cov}(\bar{d}(t_k), \bar{d}(t_v))$ values, and can be estimated by the method described in Section 2.3.4. However, $\text{Var}[\bar{d}(t_k)]$ is not estimable if the size of the risk set is less than two at time t_k. For such event times, Appendix A.2 suggests a conservative approach for estimating $\text{Var}[\bar{d}(t_k)]$.

$\text{Var}[d_{\cdot}^{\dagger}(t)]$ can be estimated using the delta method, as described in Section 2.5.4. Appendix A.1 gives more detail on the computation of $\widehat{\text{Var}}[d_{\cdot}^{\dagger}(t)]$.

Both $d_{\cdot}^{\dagger}(t)$ and $\bar{d}_{\cdot}(t)$ are functions of the number of recurrences r_i, the recurrence times t_{ij}, and the size of the risk set at the time of the recurrences $\delta_{\cdot}(t_{ij})$. When there is a closed form for the estimators of the model parameters of the NHPP model, an expression for $\text{Cov}\left[d_{\cdot}^{\dagger}(t), \bar{d}_{\cdot}(t)\right]$ can be obtained directly. When a closed form for the estimators of the model parameters is not available (e.g., for the case of ML estimator of $(\beta, \eta)'$ for the power law NHPP model for window-observation recurrence data), we use numerical methods based on the delta method, as shown in Appendix A.3.

2.7.5 Confidence Intervals for $\mu(t)$

To compute confidence intervals (CIs) for $\mu(t)$ with the hybrid estimator, $\widehat{\mu}_{\text{Localhybrid}}(t)$ or $\widehat{\mu}_{\text{NHPPhybrid}}(t)$, one can use the normal approximation method outlined in Section 2.3.5 with

the above estimates of the variance. Obtaining CIs using Bootstrap methods is another option. For example, Efron and Tibshirani (1993) describe how to calculate Bootstrap-t CI on pages 160-161, as well as other bootstrap methods.

2.7.6 Hybrid Estimation of the MCF for the AMSAA Vehicle Fleet Data

Figure 2.15 shows, for the Non-random Selection Window Data, the MCF point estimates of the nonparametric estimator, the power law NHPP estimator, the local hybrid estimator, and the NHPP hybrid estimator (power law model), as well as the true MCF. The nonparametric estimator substantially under-estimates the true MCF, while the other three estimators agree well with the true MCF.

Figure 2.16 shows, also for the Non-random Selection Window Data, the true MCF and the 95% Bootstrap-t CIs of both the local hybrid estimator and the NHPP hybrid estimator (power law model). We used Bootstrap-t CI because it is second-order accurate. Both CIs capture the true MCF, and, as expected, the CI of the NHPP hybrid estimator has a shorter CI length and is smoother. The estimates for the upper bound and the lower bound of the CI might not be monotone. To ensure monotonicity, we made adjustments to the raw estimates of the CI bands with the method similar to the one described in Section 3.4 of Weston and Meeker (1991). Making such an adjustment has no effect on the actual coverage probability of the procedure.

2.8 A Simulation Study to Compare Different MCF Estimators

We conducted an extensive simulation study to compare the performance of the four MCF estimators described in this paper: the NP estimator, the power law NHPP estimator, the local hybrid estimator, and the NHPP hybrid estimator (power law model). The estimators were compared, primarily, in terms of mean square error (MSE), although we evaluated and studied bias and variance separately to better understand the results. There were three experimental factors in the simulation study: the true MCF model, the process to generate observation windows, and the number of units in the data. Each of these factors had several levels. Details

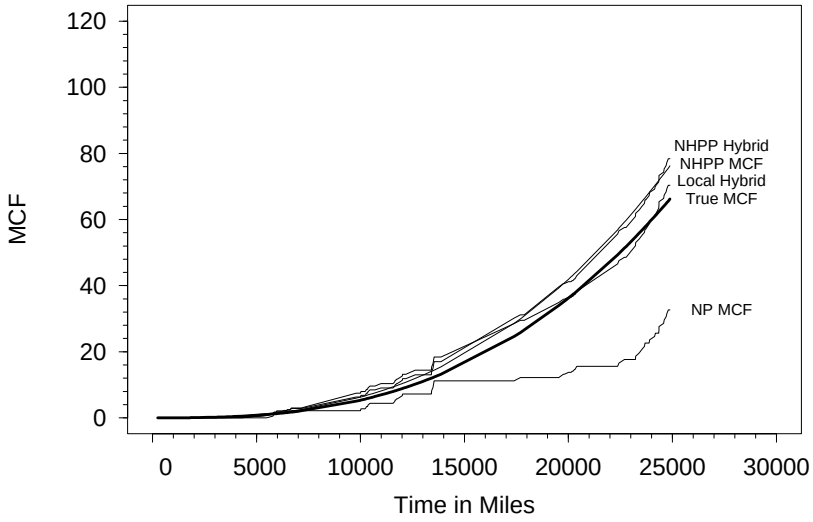


Figure 2.15 MCF Plots for AMSAA Vehicle Fleet Non-random Selection Window Data, Comparing the True Model, the NP Estimator, the Power Law NHPP Estimator, the Local Hybrid Estimator, and the NHPP Hybrid Estimator.

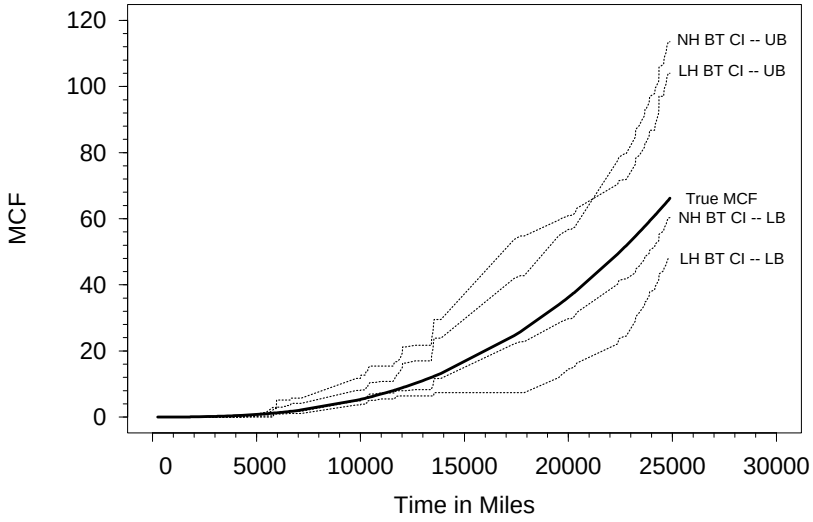


Figure 2.16 MCF Plots for AMSAA Vehicle Fleet Non-random Selection Window Data, Comparing the True Model, and the bootstrap-t CI of both the Local Hybrid Estimator and the NHPP Hybrid Estimator, based on B=5000 nonparametric re-samplings of the original simulated data.

of the simulation study will be described elsewhere, but we provide here a summary of the results.

- When there are no RSSZ intervals, the two hybrid estimators are the same as the NP estimator. Compared to the power law NHPP estimator, the NP estimator generally has a slightly higher variance, but (as we know from theory) is always unbiased. The power law NHPP estimator is approximately unbiased only when the true model is close to the power law NHPP model. When the true model is not close to the power law NHPP model, the NP estimator has a smaller MSE than the power law NHPP estimator because of the bias of the power law NHPP estimator that results from using the wrong model.

- When the proportion of time with RSSZ intervals is relatively small, the NP estimator, the local hybrid estimator, and the NHPP hybrid estimator generate very similar results. The power law NHPP estimator performs better than the other three estimators in terms of variance and MSE when the true model is the power law NHPP model or when the model deviation is modest. However, when the model deviation is more substantial, bias causes the power law NHPP estimator to have a larger MSE.

- When the proportion of time with RSSZ intervals gets larger, the local hybrid estimator remains approximately unbiased, but the NP estimator can have large downward bias, and thus a larger MSE. The bias of the power law NHPP estimator and the NHPP hybrid estimator depends on deviation of the true model from the assumed power law NHPP model. When the model deviation is small, the power law NHPP estimator has the best MSE because of the small variance, but when the model deviation is large, the NHPP hybrid estimator or local hybrid estimator will have a smaller MSE because of the relatively smaller bias.

- The differences between the local hybrid estimator and NHPP hybrid estimator can be large when the proportion of time with RSSZ intervals is large and the assumed NHPP model is seriously inadequate. Under the power law NHPP model, the NHPP hybrid estimator has a smaller MSE because it has a smaller variance and it is approximately

unbiased. When model deviation between the true model and the power law NHPP model is not large, the NHPP hybrid estimator still has a smaller MSE because its gain from having smaller variance compensates for the loss from being somewhat biased. However, when the model deviation is more substantial, the bias is larger, and the NHPP hybrid estimator could have a larger MSE than the local hybrid estimator. When the assumed NHPP model substantially over-estimates the increases in the MCF over the RSSZ intervals, the NHPP hybrid estimator will generate MCF estimates with positive bias and larger variance, and therefore, much larger MSE than local hybrid estimator.

- For each estimator, changing the number of units has very little impact on bias, but increasing the number of units can decrease the variance and thus decrease the MSE. For the factor levels that we used, 10 units and 100 units, the decreases in variance estimates among the different estimators were not influential enough to change the ordering of the MSE performance. However, as the number of units gets much larger, those estimators that assume an inappropriate model would have a relatively poor MSE, because the bias resulting from using the wrong model would not improve as the number of units increases.

2.9 Concluding Remarks and Areas for Further Research

We have shown how to extend existing nonparametric and parametric methods for recurrence data to analyze window-observation recurrence data. We see the following as important areas for further research:

- The methods described in the paper apply to the number of events, events where each event causes jump of size one. These methods can, however, be extended to apply to a more general jump of arbitrary size (e.g., when there is a cost or other event size). For the nonparametric method, Nelson (2003) describes this extension. For the parametric method, the compound Poisson processes could be used. Descriptions of compound Poisson processes are available in many books, such as Cox and Isham (1980) and Parzen (1962).

- In some applications, there will be covariate information on units (e.g., concerning the operating environment), and it will be important to have a model that allows for such information. Extending models and methods for window-observation recurrence data with covariates is an important area for future research.

- Population inhomogeneity (e.g., a mixture of units with high intensity and units with low intensity) or other extra sources of variability lead to the violation of independent-increment assumptions. Although the nonparametric estimator does not depend on the independent increments assumption, the NHPP model does. Naive application of the NHPP model will result in confidence intervals for the MCF that are too narrow. A random-effects model that describes such extra sources of variability might be able to make the NHPP type model applicable to a much larger range of real-life datasets.

- The nonparametric estimator and the local hybrid estimator apply to both continuous and discrete time recurrence processes (i.e., when the probability of a tie is non-negligible). Estimators involving NHPP models strictly apply only to continuous time recurrence processes. Another area for further research would be to extend methods for fitting the NHPP and NHPP hybrid models to handle discrete-time data.

ACKNOWLEDGMENTS

We would like to thank Dr. Michael J. Cushing from AMSAA for suggesting that we work on the development of statistical methods for window-observation data, for providing the background information on FEDC exercises, and for providing the simulated AMSAA Vehicle Fleet data. We would also like to thank Professor Luis A. Escobar for helpful comments on an earlier version of this paper. This material is based upon work supported, in part, by the National Science Foundation under Grant No. 0421916. In addition, an Associate Editor and two referees provided valuable comments and suggestions, which led to significant improvements in the paper.

References

Cook, R. J., Lawless, J. F., and Nadeau, C. (1996), "Robust Tests for Treatment Comparisons Based on Recurrent Event Responses," *Biometrics*, 52, 557-571.

Cox, D. R., and Isham, V. (1980), *Point Processes*, New York: Chapman and Hall.

Cox, D. R., and Lewis, P. A. W. (1966), *The Statistical Analysis of Series of Events*, New York: Wiley.

Doganaksoy N., and Nelson W. (1998), "A Method to Compare Two Samples of Recurrence Data," *Lifetime Data Analysis*, 4, 51-63.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Lawless, J. F., and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.

Lawless, J. F., and Zhan, M., (1998), "Analysis of Interval-grouped Recurrent-event Data Using Piecewise Constant Rate Functions," *The Canadian Journal of Statistics*, 26, 549-565.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Nelson, W. (1988), "Graphical Analysis of System Repair Data," *Journal of Quality Technology*, 20, 24-35.

Nelson, W. (1995), "Confidence Limits for Recurrence Data: Applied to Cost or Number of Product Repairs," *Technometrics*, 37, 147-157.

Nelson, W. (2003), *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Philadelphia: ASA-SIAM.

Parzen, E. (1962), *Stochastic Processes*, San Francisco: Holden-Day.

Rigdon, S. E., and Basu, A. P. (2000), *Statistical Methods for the Reliability of Repairable Systems*, New York: Wiley.

Thompson, W. A. (1988), *Point Process Models with Applications to Safety and Reliability*, New York: Chapman & Hall.

Weston, S. A. and Meeker, W. Q. (1991), Coverage Probabilities of Nonparametric Simultaneous Confidence Bands for a Survival Function, *Journal of Statistical Computation and Simulation*, 38, 83-97.

Yuan, K. H. and Jennrich, R. I. (1998), Asymptotics of Estimating Equations under Natural Conditions, *Journal of Multivariate Analysis*, 65, 245-260.

CHAPTER 3. A SIMULATION STUDY ON CONFIDENCE INTERVAL PROCEDURES OF SOME MEAN CUMULATIVE FUNCTION ESTIMATORS

Jianying Zuo

William Q. Meeker

Huaiqing Wu

Statistics Department

Iowa State University

Ames, IA 50011

Abstract

Recurrence data are collected to study the recurrent events on biological, physical, and other systems. Quantities of interest include the mean cumulative number of events and the mean cumulative cost of events. The mean cumulative function (MCF) can be estimated with nonparametric methods or by fitting parametric models, and many procedures have been suggested to construct the confidence interval (CI) for the MCF. This paper summarizes the results of a large simulation study that was designed to compare five CI procedures for both the nonparametric and parametric estimation. When doing parametric estimation, we assume the power law non-homogeneous Poisson process (NHPP) model. Our results include evaluation of these procedures when they are used for window-observation recurrence data where recurrence histories of some systems are available only in observation windows with gaps in between.

KEY WORDS: Mean cumulative function; MCF; Nonhomogeneous Poisson process; Nonparametric estimation; Recurrence data.

3.1 Introduction and Background

3.1.1 Background and Motivation

Recurrence data arise in many applications, including business processes, medical statistics, and repairable system reliability. Both nonparametric and parametric methods are available to estimate the mean cumulative function (MCF). Although many important questions can be answered by using nonparametric methods, in some applications, parametric models are needed. Nelson (2003) presents graphical and nonparametric statistical methods for recurrence data and describes many applications. Cook and Lawless (2007) provide a comprehensive account of statistical methods to analyze recurrent event data. They describe nonparametric, parametric, and semipamametric methods, give mathematical and statistical background, and also present many applications.

There are several procedures that can be used to construct approximate confidence intervals (CIs) for the MCF. Normal approximation CIs are relatively easy to implement and compute. With modern powerful computers, bootstrap or simulation-based procedures can also be used.

The adequacy of approximate CI procedures depends on the underlying model as well as the number of units and the length of time that each unit is observed. With constraints on resources to observe the process and collect the data, however, it may be difficult to have data with both a large number of observational units and a long time of observation. To help explore the impact of various factors on the performances of CI procedures, we carried out an extensive simulation study to compare five CI procedures. Two of these procedures are normal approximation procedures, and the other three are bootstrap-based procedures. Section 3.3 gives technical details on how to compute the intervals.

3.1.2 Window-Observation Data

In some applications, window-observation recurrence data arise because recurrence history of some units are observed in disconnected windows with gaps between the windows. Nelson (2003, page 75) describes an example. Zuo, Meeker and Wu (2008) provide more examples, and extend some nonparametric (NP) and parametric estimation methods to window-observation recurrence data. If there are intervals of time over which the size of the risk set is zero, the NP method can be seriously biased. For such scenarios, Zuo, Meeker and Wu (2008) describe two hybrid MCF estimators that help correct the bias. In this paper, we evaluate CI procedures for both complete and window-observation data.

3.1.3 Other Previous Related Work

Cook and Lawless (2007) describe different approaches to the modeling of recurrent events such as models based on counts of events, intensity, and time between events. Many publications provide descriptions of analysis and modeling of counting processes, such as Cox and Lewis (1966) and Andersen, Borgan, Gill and Keiding (1993).

There are a number of useful text books that describe bootstrap procedures for computing CIs. For example, Efron and Tibshirani (1993) present basic background of bootstrap methods and many applications to statistical procedures. More specifically, Chapters 12 to 14, and 22 in Efron and Tibshirani (1993) present bootstrap procedures to construct CIs. Hall (1992) interweaves the topics of bootstrap and Edgeworth expansion, and applies Edgeworth expansion methods to characterize the performance of some bootstrap methods.

3.1.4 Overview

The remainder of this paper is organized as follows. Section 3.2 describes the model and estimation of the MCF. Section 3.3 explains the five CI procedures in the simulation study, and Section 3.4 outlines the details of the simulation study. Section 3.5 shows the impact of the recurrence rate function on the performances of the CI procedures. Sections 3.6 to 3.8 summarize the performances of the CI procedures for the NP estimator, the power law NHPP

estimator, and the hybrid estimators, respectively. Concluding remarks and areas for future research are outlined in Section 3.9. The appendix A.4 presents some necessary technical details for the NP and NHPP estimators.

3.2 Model and Estimation

In this section, we describe briefly the recurrence data model and provide formulas for point estimators.

3.2.1 Notation and Acronyms

We will use the following notation:

- n: the number of units under observation
- t_{endobs}: the pre-specified end-of-observation time for the observational units in the simulation. For the simulation study, recurrences and observation windows up to this time point are recorded in the data
- E(r): expected total number of observed recurrences for all n units over the time range (0, t_{endobs})
- RSSZ: risk-set-size-zero
- RSSONE: risk-set-size-one
- RSSP: risk-set-size-positive
- NP: nonparametric
- NHPP: non-homogeneous Poisson process
- CP: coverage probability

3.2.2 Model

The models in our study are based on counts of events (i.e., the number of events in some time range of interest, say $[0, t]$). Let $N(t)$ denote the number of events in the time range $[0, t]$. Then one statistic of interest is $\mu(t)$, the expectation of $N(t)$, which is also known as the mean cumulative function (MCF). Our goal is to estimate the MCF.

If the MCF is differentiable, then $\nu(t) = d\mu(t)/dt$ is the recurrence rate and $\nu(t) \times \Delta t$ can be interpreted as the approximate expected number of events to occur during the next short time interval $(t,\ t + \Delta t)$.

The nonparametric approaches do not make assumptions about the form of the recurrence process, while the parametric ones do. Model assumptions for the nonparametric estimation methods are stated in Nelson (2003) and Zuo, Meeker and Wu (2008). Model assumptions for the NHPP parametric estimation methods are given in Rigdon and Basu (2000).

3.2.3 Estimation for the Nonparametric Model

Detailed descriptions of nonparametric MCF estimation methods are available in Nelson (1988), Lawless and Nadeau (1995), Chapter 16 of Meeker and Escobar (1998), and Chapters 3 to 5 of Nelson (2003). Zuo, Meeker and Wu (2008) extend the nonparametric method to window-observation recurrence data.

Let m denote the number of unique event times. Also, let $t_1, ..., t_m$ be the unique event times. Then the nonparametric estimator of the population MCF is

$$\widehat{MCF}_{NP}(t_j) = \sum_{k=1}^{j} \left[\frac{\sum_{i=1}^{n} \delta_i(t_k) \times d_i(t_k)}{\sum_{i=1}^{n} \delta_i(t_k)} \right] = \sum_{k=1}^{j} \frac{d_{\cdot}(t_k)}{\delta_{\cdot}(t_k)} = \sum_{k=1}^{j} \bar{d}(t_k), \quad j = 1, ..., m, \qquad (3.1)$$

where $d_i(t_k)$ is the number of events recorded at time t_k for unit i, and

$$\delta_i(t_k) = \begin{cases} 1 & \text{if unit } i \text{ is under observation in a time window at time } t_k, \\ 0 & \text{otherwise.} \end{cases}$$

Details on estimators of $\text{Var}[\widehat{MCF}_{NP}(t_j)]$ are available in Nelson (1995), Lawless and Nadeau (1995), and Zuo, Meeker and Wu (2008).

3.2.4 Estimation for the NHPP Parametric Model

NHPP models and estimation methods for recurrence data are described, for example, in Rigdon and Basu (2000, Chapter 2) and Meeker and Escobar (1998, Chapter 16). Zuo, Meeker and Wu (2008) extend the NHPP estimation methods to window-observation recurrence data. Given the maximum likelihood (ML) estimates of the model parameters $\hat{\boldsymbol{\theta}}$, the ML estimator of the NHPP MCF is

$$\widehat{MCF}_{NHPP}(t) = \int_0^t \nu(x;\hat{\boldsymbol{\theta}})dx. \tag{3.2}$$

With the estimate of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, the delta method can be used to estimate $\text{Var}[\widehat{MCF}_{NHPP}(t)]$. Zuo, Meeker and Wu (2008) give more details, using the power law NHPP model as an example.

3.2.5 Hybrid MCF Estimators for Window-Observation Recurrence Data

Zuo, Meeker and Wu (2008) introduce two hybrid MCF estimators for window-observation recurrence data – the local hybrid estimator and the NHPP hybrid estimator. Such estimators are needed because the existence of RSSZ intervals can cause $\widehat{MCF}_{NP}(t)$ to be seriously biased.

The local hybrid estimator is $\widehat{MCF}_{LH}(t) = \bar{d}_{\cdot}(t) + d_{\cdot}^{\ddagger}(t)$, where $\bar{d}_{\cdot}(t)$ is the nonparametric estimator of the increase in the MCF from RSSP intervals, while $d_{\cdot}^{\ddagger}(t)$ is the estimator of the increase in the MCF from RSSZ intervals, assuming the recurrence rate of the RSSZ interval is the weighted average of the recurrence rates of the two neighboring RSSP intervals.

The NHPP hybrid estimator is $\widehat{MCF}_{NHPPH}(t) = \bar{d}_{\cdot}(t) + d_{\cdot}^{\dagger}(t)$, where $d_{\cdot}^{\dagger}(t)$ is the estimator of the increase in the MCF from RSSZ intervals, assuming the counts of recurrences in the RSSZ intervals follow an NHPP model. The NHPP model is estimated with the data in the RSSP intervals.

3.2.6 Recommendations on Selection of MCF Estimators

Zuo, Meeker and Wu (2008) present a brief summary from a simulation study that compared the NP estimator, the power law NHPP estimator, the local hybrid estimator, and the NHPP

hybrid estimator. The nonparametric approaches, such as the NP estimator and the local hybrid estimator, generally have little or no bias, but might have a large variance. On the other hand, more parametric based approaches generally have small variance, but could be seriously biased if the assumed model form is very different from the true model. Therefore, selection among different MCF estimators becomes a bias-variance tradeoff, and model assumption diagnosis is important in the model selection and estimation process. This paper presents the results of a much more extensive simulation to compare the properties of CI procedures based on the same estimators.

3.3 Confidence Interval Procedures

In this section, we outline two normal approximation CI procedures and three bootstrap CI procedures. These five CI procedures are evaluated in our simulation experiments, and applied to the four MCF estimators described in Section 3.2. We use the following general notation.

- $\widehat{MCF}$: estimate of the MCF from the original data.
- $\widehat{SE}_{\widehat{MCF}}$: standard error of $\widehat{MCF}$ (i.e., estimate of the standard deviation of $\widehat{MCF}$) from the original data.
- $\widehat{MCF}^*$: estimate of the MCF from the bootstrap re-sampled data.
- $\widehat{SE}^*_{\widehat{MCF}}$: standard error of $\widehat{MCF}$ (i.e., estimate of the standard deviation of $\widehat{MCF}$) from the bootstrap re-sampled data.
- t^*: t-like ratio from the bootstrap re-sampled data, computed as $t^* = \left(\widehat{MCF}^* - \widehat{MCF}\right) / \widehat{SE}^*_{\widehat{MCF}}$.

Only the $\widehat{MCF}$ and the $\widehat{SE}_{\widehat{MCF}}$ are needed to construct normal approximation CIs. There are many possible normal approximation CI procedures, depending on the transformation used. The two normal approximation procedures a) and b) below are from Meeker and Escobar (1998, Chapter 16, page 400). In the formulas, $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

Efron and Tibshirani (1993) describe various bootstrap procedures. The common first step in any bootstrap procedure is to generate bootstrap samples from the original data, and these bootstrap samples are used to calculate $\widehat{MCF}^*$ and $\widehat{SE}^*_{\widehat{MCF}}$ and, for the bootstrap-t procedures, t^*. To construct a bootstrap interval, we repeat the sampling and estimation process a large number of times (say B times), and sort $\widehat{MCF}^*$ or t^* values (depending on the procedure). Let k be the largest integer less than or equal to $(B+1)\alpha/2$, where α is the complement of the desired nominal CP. The three bootstrap CIs can be obtained by c), d), and e) below, where $y_{(k)}$ indicates the k^{th} ordered value in a sequence where y has been ordered from smallest to largest. Specifically, the five CI procedures that we evaluate in our simulation are

a) *Normal Approximation (NORMA)*: $\widehat{MCF} \pm z_{(1-\alpha/2)}\widehat{SE}_{\widehat{MCF}}$;

b) *Lognormal Approximation (LNORMA)*: $\left[\widehat{MCF}/w, \quad \widehat{MCF} \times w\right]$,
where $w = \exp\left[z_{(1-\alpha/2)}\widehat{SE}_{\widehat{MCF}}/\widehat{MCF}\right]$;

c) *Bootstrap Percentile (BootP)*: $\left[\widehat{MCF}^*_{(k)}, \quad \widehat{MCF}^*_{(B+1-k)}\right]$;

d) *Bootstrap-t (Boott)*: $\left[\widehat{MCF} - t^*_{(B+1-k)}\widehat{SE}_{\widehat{MCF}}, \quad \widehat{MCF} - t^*_{(k)}\widehat{SE}_{\widehat{MCF}}\right]$,
where $t^* = \frac{\widehat{MCF}^* - \widehat{MCF}}{\widehat{SE}^*_{\widehat{MCF}}}$;

e) *Bootstrap-t Based on Log Transformation (LBoott)*:
$$\left[\frac{\widehat{MCF}}{\exp\left(t^*_{(B+1-k)}\widehat{SE}_{\widehat{MCF}}/\widehat{MCF}\right)}, \quad \frac{\widehat{MCF}}{\exp\left(t^*_{(k)}\widehat{SE}_{\widehat{MCF}}/\widehat{MCF}\right)}\right],$$
where $t^* = \left(\log(\widehat{MCF}^*) - \log(\widehat{MCF})\right)/\widehat{SE}^*_{\log(\widehat{MCF})}$ and $\widehat{SE}^*_{\log(\widehat{MCF})} = \widehat{SE}^*_{\widehat{MCF}}/\widehat{MCF}^*$.

For simplification, we use the acronyms in the parentheses to represent the five CI procedures.

3.4 Simulation Experimental Design

3.4.1 Factors and Factor Levels

In previous simulation experiments to study confidence interval properties for censored lifetime data (e.g., Jeng and Meeker 2000), it was shown that the adequacy of asymptotic approximations tend to depend on the number of failures, rather than the sample size. Thus, an important experimental factor in our simulation experiment is the expected number of events, E(r), with four factor levels at 10, 20, 50, and 100. E(r) is, however, affected by the following factors:

- **The pattern of the observation windows for the units in the data set (Window Schemes)**. Three window schemes, corresponding to data sets analyzed in Zuo, Meeker and Wu (2008), are used in our study, and they are

  1. *Complete* data: All units in the data are observed continuously in the same single window $[0, t_{endobs}]$.
  2. *Window1* data: There are some gaps between the observation windows for each unit. Length of the observation window follows a uniform distribution between 0.08 and 0.12, while length of the gap follows a uniform distribution between 0.12 to 0.28. Whether a unit begins with a window or a gap follows a Bernoulli (0.5) distribution.
  3. *Window2* data: Similar to *Window1* data, except that the length for gap i is simulated from a uniform distribution between $0.04 \times 2^{(i-1)}$ and $0.08 \times 2^{(i-1)}$. Therefore, for units in a *Window2* data set, as time gets larger, the probability of observing a given recurrence gets smaller.

  We did simulations using all three window schemes. However, the results of Window1 data are similar to those of Complete data, mainly because the percentage of times with RSSZ and RSSONE is zero or very small. Therefore, we focus on the results from the Complete data and Window2 data.

- **Number of Units**: n, with five levels at the values 10, 20, 50, 100, and 200 units being observed.

- **Form of the MCF of the recurrence process**. We use the power law NHPP model with the recurrence function as

$$\nu(t;\beta,\eta) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \quad \beta > 0, \eta > 0. \tag{3.3}$$

  Without loss of generality, we use $\eta = 1$. For the shape parameter β, there are four levels at values 0.8 (decreasing recurrence rate), 1 (constant recurrence rate), 2 (moderately increasing recurrence rate), and 3 (rapidly increasing recurrence rate).

Note that when **Window Schemes**, **Number of Units**, and **Form of the MCF** are given, t_{endobs} depends only on the value of E(r).

3.4.2 Simulation Algorithm

Given a set of factor levels from our simulation design, the following procedure was carried out to estimate the CPs of the five CI procedures.

1. Generate simulated data based on the inputs for the window scheme, n, β, and t_{endobs}.

2. If the window scheme is Window1 or Window2, use the simulated data to compute the $\widehat{MCF}$ at t_{endobs} for the four different estimators – the NP estimator, the power law NHPP estimator, the local hybrid estimator, and the power law NHPP hybrid estimator. For the Complete data case, we compute only the NP estimator and the power law NHPP estimator. We also compute the corresponding standard errors.

3. Compute the NORMA and the LNORMA CIs described in Section 3.3.

4. Generate bootstrap samples, and construct the BootP, the Boott, and the LBoott CIs described in Section 3.3. In our simulation study, we used simple random sampling with replacement (SRSWR) to draw the whole history of a unit and thus created bootstrap re-sampled data that had the same number of units in the original data. We used $B = 2000$ bootstrap samples.

5. Check whether the five CIs obtained in Steps 3 and 4 capture the true MCF at t_{endobs}, assign to coverage indicator a value of 1 if the true MCF is in the CI and a value of 0 otherwise.

6. Repeat Steps 1 to 5 a large number of times (5000 in our simulation study) and then calculate the average of coverage indicators as the estimate of the CP for each of the five CI procedures.

When the number of observed recurrences is small (e.g., 4 or fewer), MCF estimates are poor and there can be estimation problems (e.g., $\boldsymbol{\theta}$ is not estimable if the dimension of $\boldsymbol{\theta}$ is larger than the number of observed recurrences for $\widehat{MCF}_{NHPP}$, or some components of the variance estimator of $\widehat{MCF}_{NP}$ are not estimable if fewer than two units are being observed for the time with observed recurrences). Thus we used only simulated data sets with 5 or more distinct recurrence times, as well as bootstrap re-sampled data with 3 or more distinct recurrence times, to estimate CP of the MCF estimators. That is, our simulation results are conditional on $\sum_{i=1}^{n} X_i \geq 5$, where $\sum_{i=1}^{n} X_i$ is the total number of observed recurrences among all n units. As described in Appendix A.4.1, the values of $\Pr\left(\sum_{i=1}^{n} X_i \leq 4\right)$ are 0.0293, 1.69×10^{-5}, 5.45×10^{-17}, and 1.61×10^{-37} respectively for $\mathrm{E}(r) = 10$, 20, 50, and 100. Therefore, when $\mathrm{E}(r) = 20$, 50, and 100, the conditional probabilities are close to 1.

3.5 The Effect that the Recurrence Rate Shape has on CI Procedures

In our simulation study, we assume that the true model is the power law NHPP model. For a power law NHPP model, the value of β determines whether the recurrence rate is increasing (for $\beta > 1$), constant (for $\beta = 1$), or decreasing (for $\beta < 1$). One question of interest is how the shape of the recurrence rate across time affects the performances of the five CI procedures. In particular, we will study whether the estimated CP values are approximately the same or have some pattern across the different values of β in this section, and we will discuss the comparisons among the five CI procedures in Sections 3.6 to 3.8.

In order to graphically show the impact of n and $\mathrm{E}(r)$, there are six plots for each of the

estimators in Figures 3.1 to 3.6, arranged in three rows and two columns. In rows 1 to 3, n is 10, 20, and 50, respectively. E(r) is 10 on the left and 20 on the right.

3.5.1 Results from Complete Data

First we describe the simulation results based on Complete data, where each unit in the data is observed from time zero to t_{endobs}. Because there are no observation gaps for the Complete data, we only need to compare CI procedures for the NP method and the power law NHPP method. Figures 3.1 and 3.2 are for the NP estimator and the power law NHPP estimator respectively, and some noticeable patterns for both estimators are:

- CP values are about parallel across the four values of β for each of the five CI procedures, and a parallel pattern exists for all n and E(r) values.
- At each n level, when E(r) increases from 10 to 20, differences among the five CI procedures get smaller and the CP values are more closely clustered around the nominal value 0.95. This trend continues with higher E(r) values, and thus plots with E(r) $\geq$ 50 are not shown.
- For E(r) = 20, differences among the five CI procedures get smaller when n increases. However, when E(r) = 10, the maximum distance among the CP values of the five CI procedures does not get smaller when n increases. This indicates that E(r) must be in the order of 20 for the large sample approximations to be adequate, even if n is large.

The explanation for the observed parallel patterns in Figures 3.1 and 3.2 is that

$$\widehat{MCF}_{NP}(t_{endobs}) = \widehat{MCF}_{NHPP}(t_{endobs}) = \sum_{i=1}^{n} X_i/n,$$

where X_i is the number of observed recurrences for unit i and $\sum_{i=1}^{n} X_i$ follows a Poisson distribution with λ = E(r), as shown in Appendix A.4. Therefore, $\widehat{MCF}_{NP}(t_{endobs})$ and $\widehat{MCF}_{NHPP}(t_{endobs})$ are proportional to a Poisson random variable with a mean λ = E(r), not depending on the value of β. Appendix A.4 also shows that the distributions of

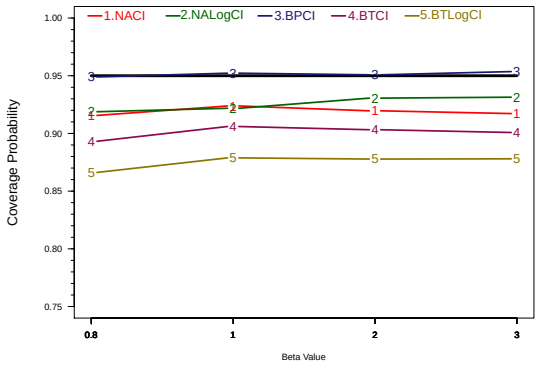
$\widehat{\text{Var}}[\widehat{MCF}_{NP}(t_{endobs})]$ and $\widehat{\text{Var}}[\widehat{MCF}_{NHPP}(t_{endobs})]$ depend only on E(r) and n as well. Therefore, how the recurrences are distributed across time (more clustered at the beginning of life for $\beta < 1$, equally likely across time for $\beta = 1$, and with higher density as time increases for $\beta > 1$) does not have a strong effect on the performances of the CI procedures and MCF estimators, as long as the expected number of recurrences for the time of estimation is the same.

3.5.2 Results from Window2 Data

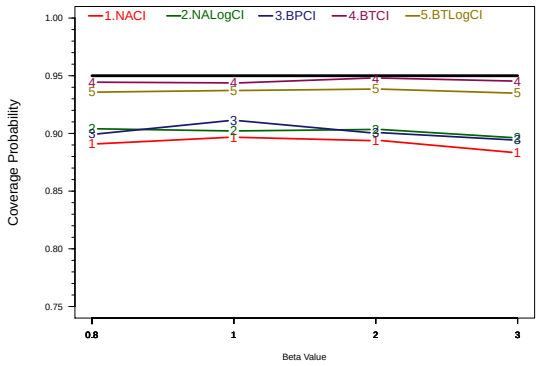
Figures 3.3 to 3.6 show the plots of simulation results using Window2 data, for the four estimators, organized as in Figure 3.1. The main observations from the four sets of plots are:

- For the NHPP estimator in Figure 3.4, there is again a parallel pattern across the values of β for all six plots.
- For the NP estimator and the two hybrid estimators, in Figures 3.3, 3.5 and 3.6 respectively, CP values are close to parallel across the β values when $n = 20$ and E(r) = 10 or when $n = 50$. When $n = 10$, or $n = 20$ and E(r) = 20, there is noticeable increasing or decreasing trend of the CP values as β changes, indicating that the asymptotic approximations are far from adequate.

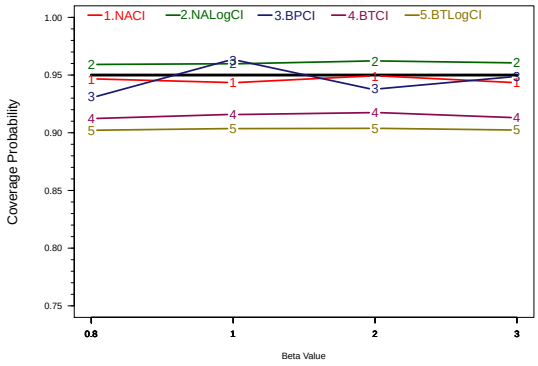
One important difference between the Complete data and the Window2 data is that it is possible to have a high percentage of time with RSSZ and RSSONE for the Window2 data. For the Complete data, there is no RSSZ or RSSONE. As observed in Figures 3.1 and 3.2, the performances of the CI procedures do not depend on the β values when there is no RSSZ or RSSONE, for the Complete data case. For the Window2 data, the percentage of time with RSSZ and RSSONE is zero or negligible when $n = 50$, and is somewhat close to 5% for all four β values when $n = 20$ and E(r) = 10. This percentage is, however, relatively large and different among the four β values when $n = 10$, as well as $n = 20$ and E(r) = 20. Therefore, the existence of RSSZ and RSSONE is the main reason that CP depends strongly on β in Figures 3.3, 3.5 and 3.6. For example, Table 3.1 shows, for E(r) = 10 and $n = 10$, the percentage of time with RSSZ and RSSONE increases as β decreases.



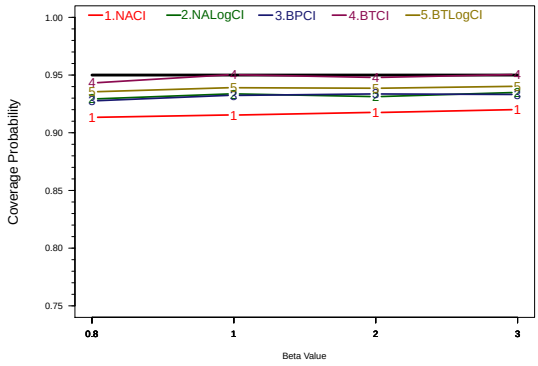
(a) $n = 10$ and $\mathrm{E}(r) = 10$



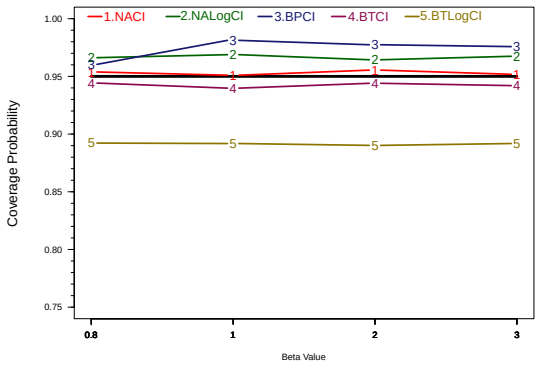
(b) $n = 10$ and $\mathrm{E}(r) = 20$



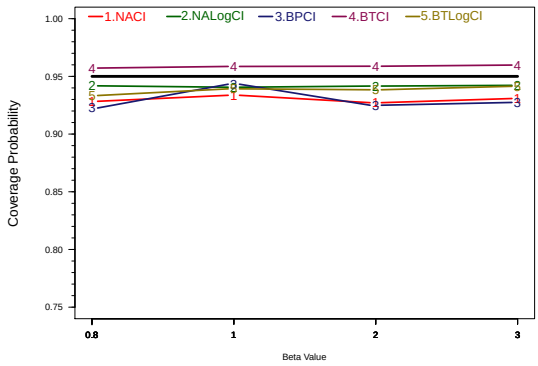
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$

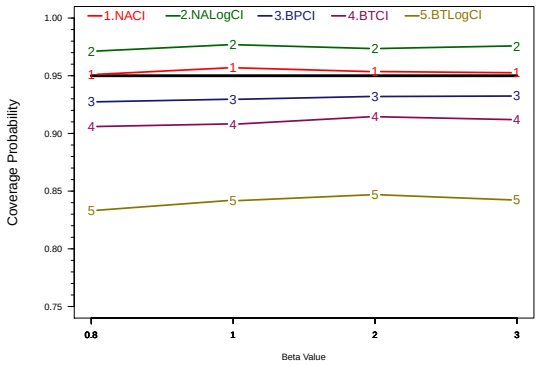


(e) $n = 50$ and $\mathrm{E}(r) = 10$

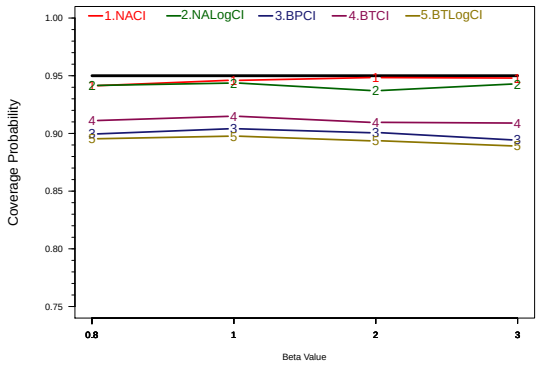


(f) $n = 50$ and $\mathrm{E}(r) = 20$

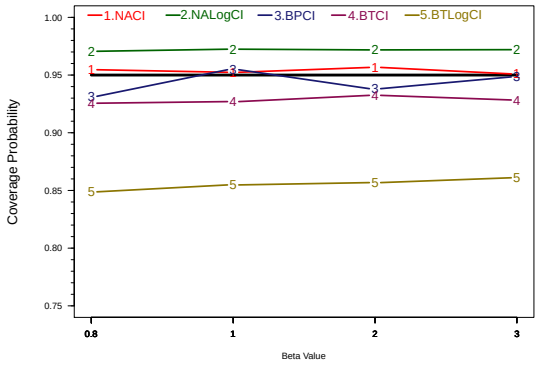
Figure 3.1 Comparison of 4 β Values: NP Estimator for the Complete Data



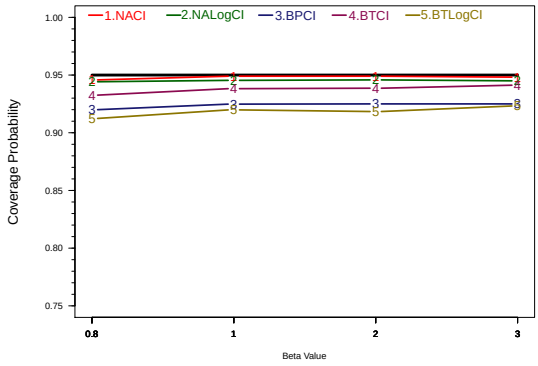
(a) $n = 10$ and $\mathrm{E}(r) = 10$



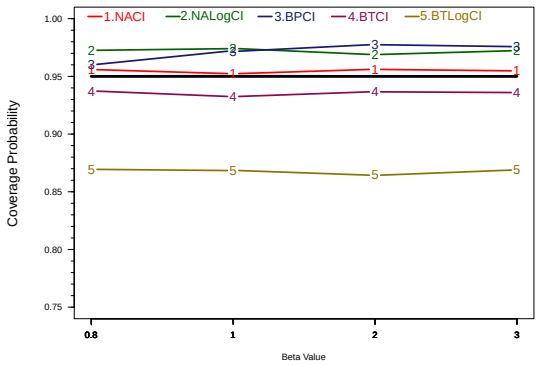
(b) $n = 10$ and $\mathrm{E}(r) = 20$



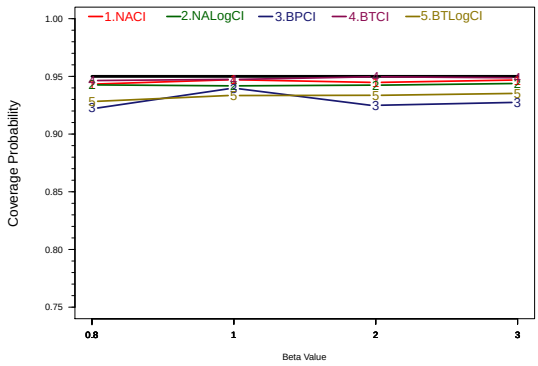
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$

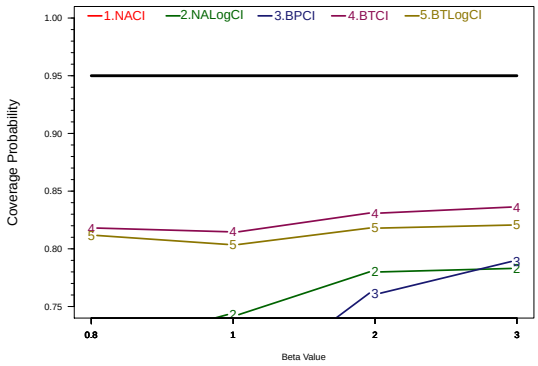


(e) $n = 50$ and $\mathrm{E}(r) = 10$

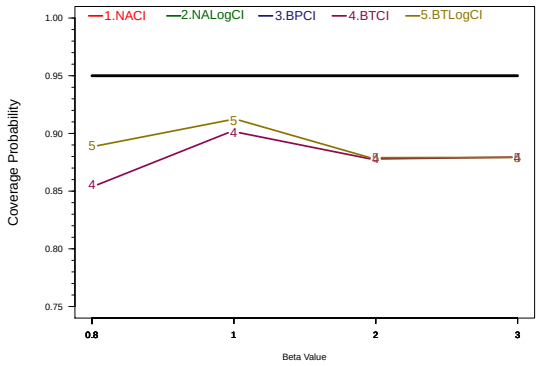


(f) $n = 50$ and $\mathrm{E}(r) = 20$

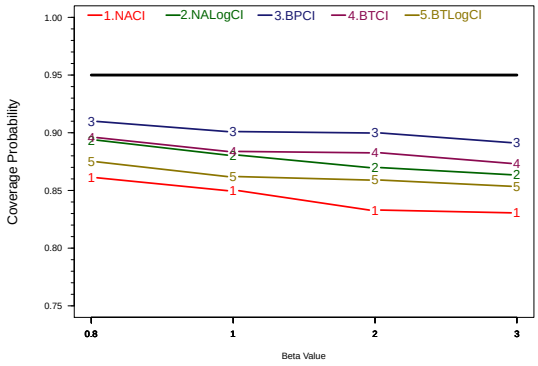
Figure 3.2 Comparison of 4 β Values: NHPP Estimator for the Complete Data



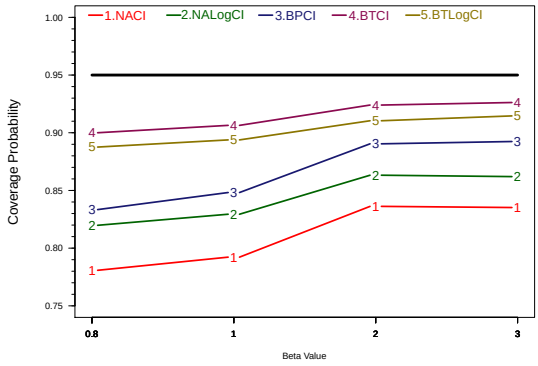
(a) $n = 10$ and $\mathrm{E}(r) = 10$



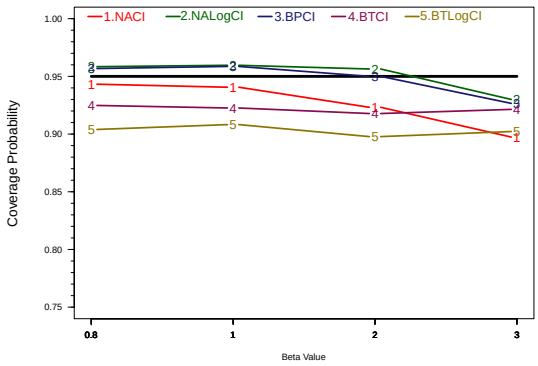
(b) $n = 10$ and $\mathrm{E}(r) = 20$



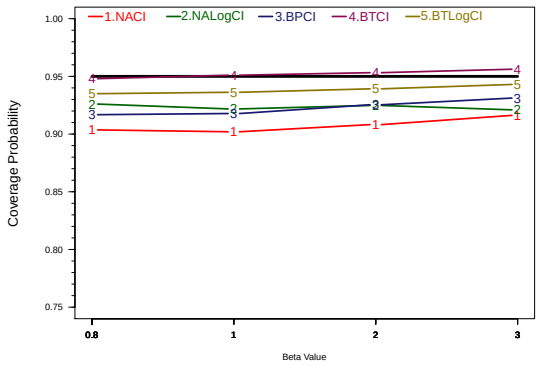
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$

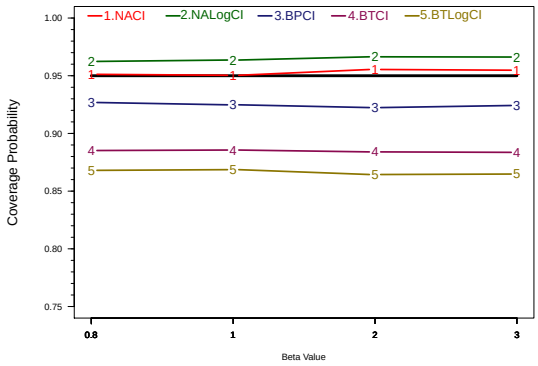


(e) $n = 50$ and $\mathrm{E}(r) = 10$

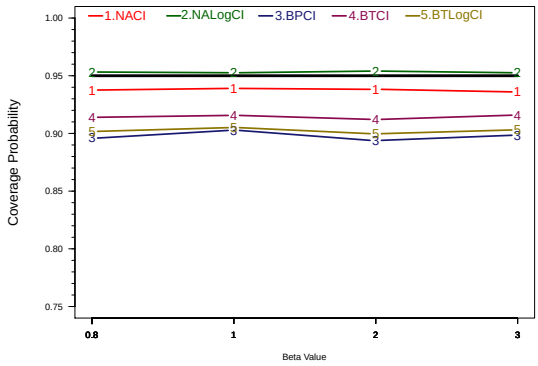


(f) $n = 50$ and $\mathrm{E}(r) = 20$

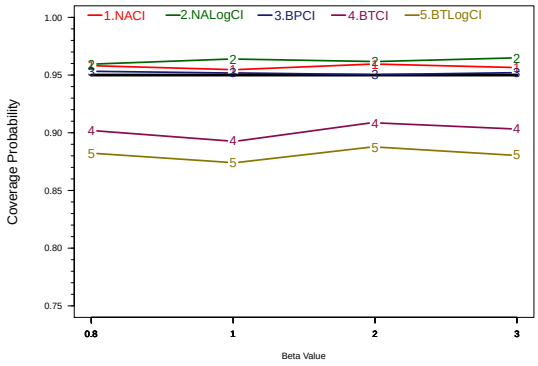
Figure 3.3 Comparison of 4 β Values: NP Estimator for the Window2 Data



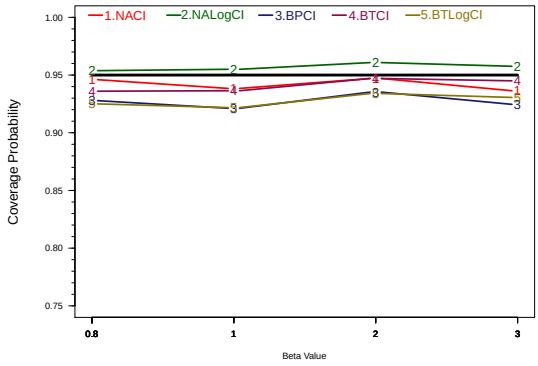
(a) $n = 10$ and $\mathrm{E}(r) = 10$



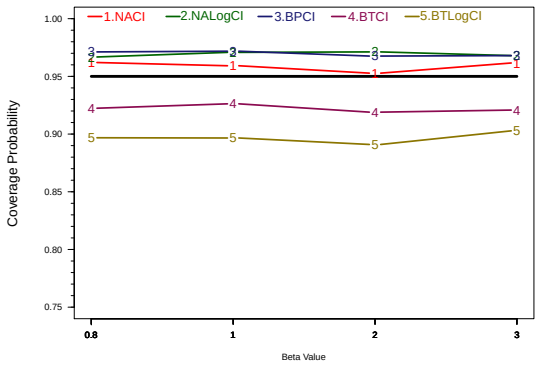
(b) $n = 10$ and $\mathrm{E}(r) = 20$



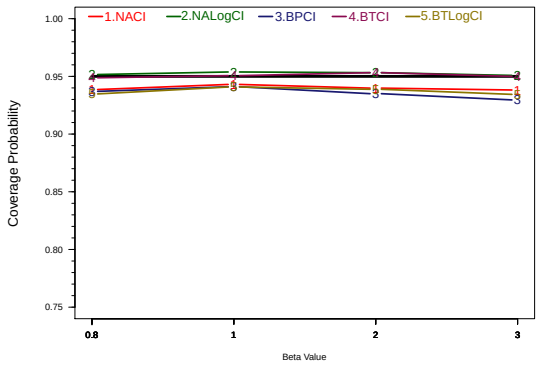
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$

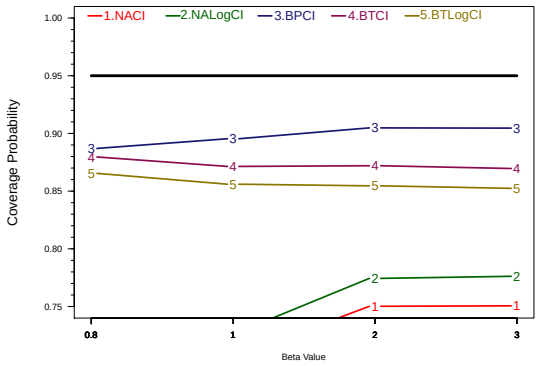


(e) $n = 50$ and $\mathrm{E}(r) = 10$

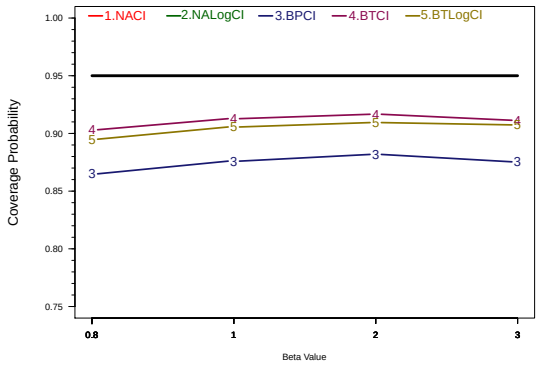


(f) $n = 50$ and $\mathrm{E}(r) = 20$

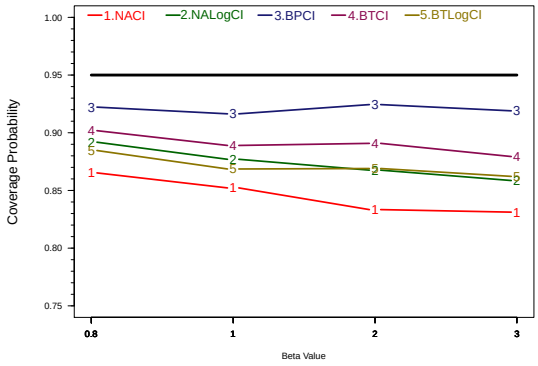
Figure 3.4 Comparison of 4 β Values: NHPP Estimator for the Window2 Data



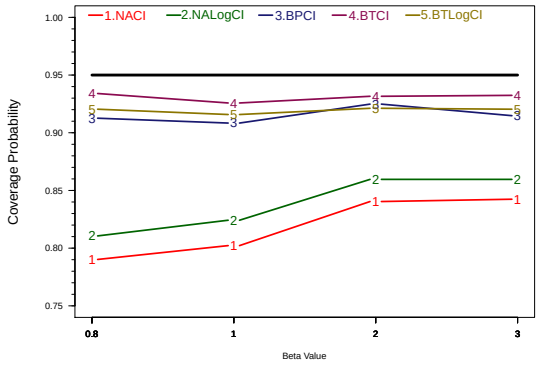
(a) $n = 10$ and $\mathrm{E}(r) = 10$



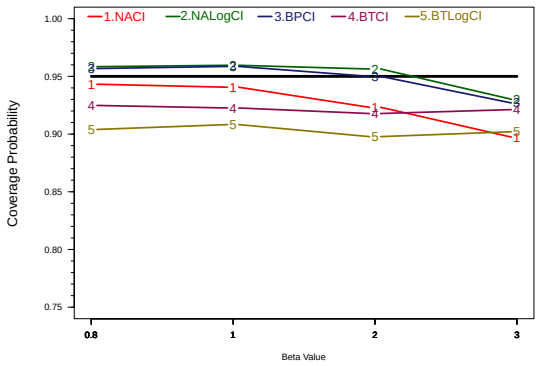
(b) $n = 10$ and $\mathrm{E}(r) = 20$



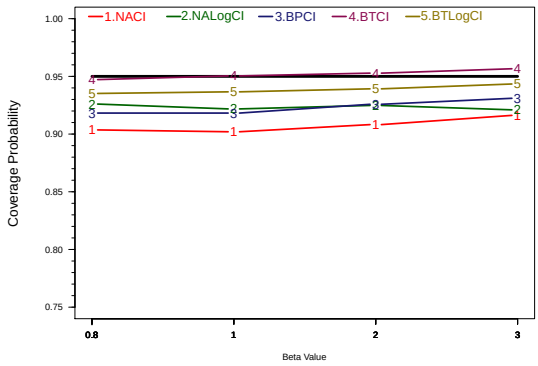
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$

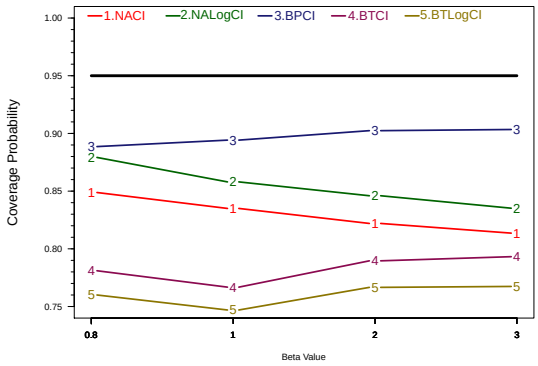


(e) $n = 50$ and $\mathrm{E}(r) = 10$

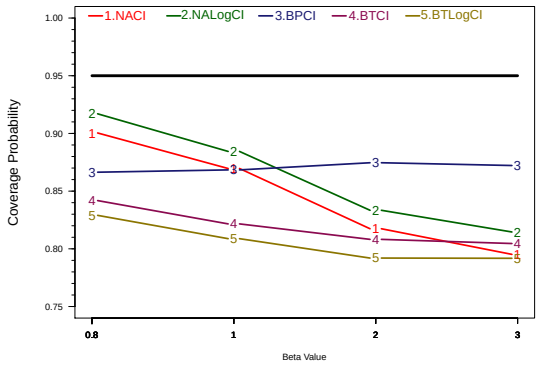


(f) $n = 50$ and $\mathrm{E}(r) = 20$

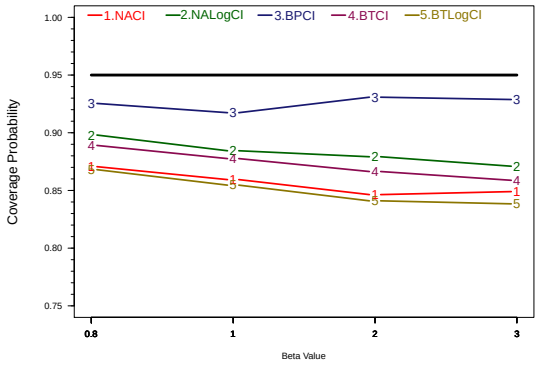
Figure 3.5 Comparison of 4 β Values: Local Hybrid Estimator for the Window2 Data



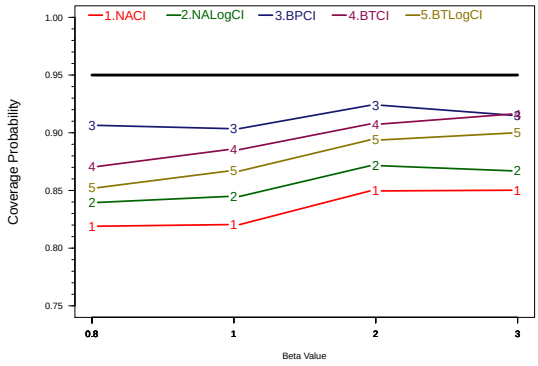
(a) $n = 10$ and $\mathrm{E}(r) = 10$



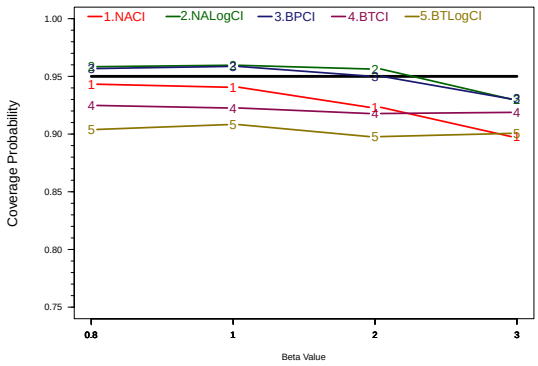
(b) $n = 10$ and $\mathrm{E}(r) = 20$



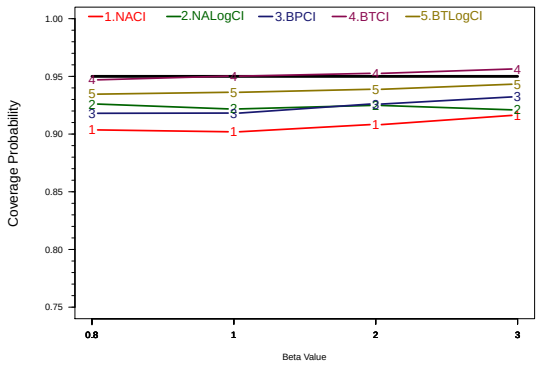
(c) $n = 20$ and $\mathrm{E}(r) = 10$



(d) $n = 20$ and $\mathrm{E}(r) = 20$



(e) $n = 50$ and $\mathrm{E}(r) = 10$



(f) $n = 50$ and $\mathrm{E}(r) = 20$

Figure 3.6 Comparison of 4 β Values: NHPP Hybrid Estimator for the Window2 Data

Table 3.1 RSSZ and RSSONE for Window2 Data with $n = 10$, E$(r) = 10$, and Number of Simulations = 5000

		Average		Percentage of time relative to t_{endobs}		
β	t_{endobs}	RSSZ	RSSONE	RSSZ	RSSONE	RSSZ+RSSONE
0.8	6.83	1.29	2.01	18.9%	29.5%	48.3%
1	4.30	0.60	1.07	14.0%	24.9%	38.8%
2	1.95	0.15	0.30	7.5%	15.4%	22.9%
3	1.53	0.09	0.20	5.6%	12.9%	18.5%

Among the four estimators, the NHPP estimator is robust to having a high percentage of time with RSSZ, as observed by the parallel patterns in Figure 3.4. This nice property is, however, based on the condition that the assumed NHPP model adequately describes the true process.

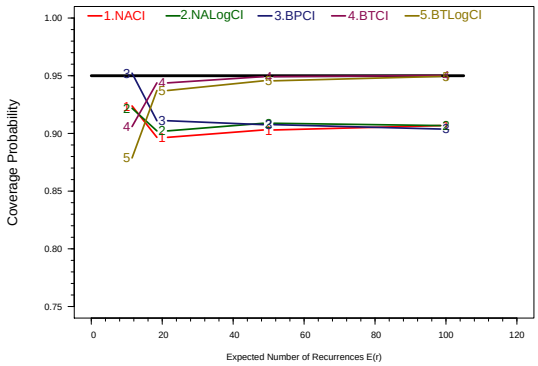
Based on the results of the Complete data and Window2 data in this section, the performances of the CI procedures for each of the four MCF estimators depend more on the existence of RSSZ and RSSONE rather than on the value of β. On the other hand, we carried out simulations on all four β values, but did not find special patterns that depend solely on the value of β. Therefore, in the subsequent sections on the CI performances for each of the MCF estimators, we will focus on the simulation results for $\beta = 1$, and discuss the impact of n, $E(r)$, and the existence of RSSZ and RSSONE in more detail.

3.6 Performances of the CI Procedures Based on the NP Estimator

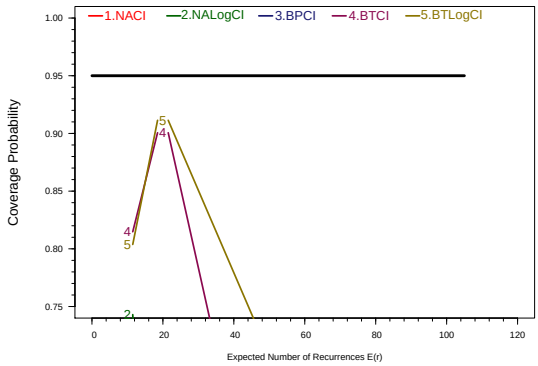
This section summarizes the simulation results for the NP estimator, and makes some recommendations. Figure 3.7 shows six plots for the NP estimator, with factor $E(r)$ as the x-axis in each plot. From top down, n increases from 10 to 20 and then to 50, and plots on the left are for Complete data while those on the right are for Window2 data.

3.6.1 Comparison of Results from Complete Data

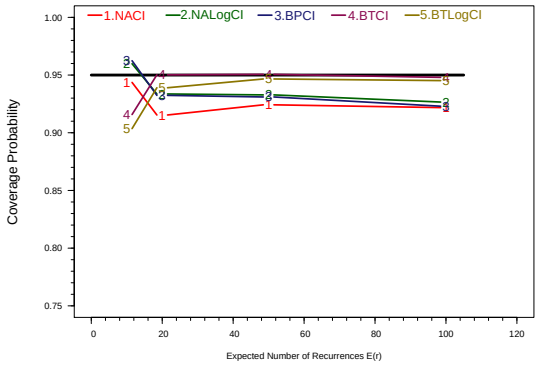
The main observations from the three plots for the Complete data results, shown in Figures 3.7 (a), (c) and (e), are:



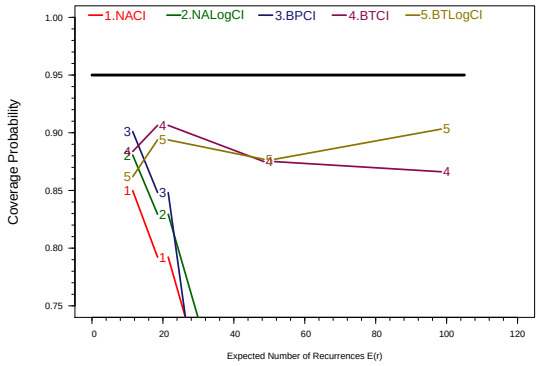
(a) Complete Data with $n = 10$



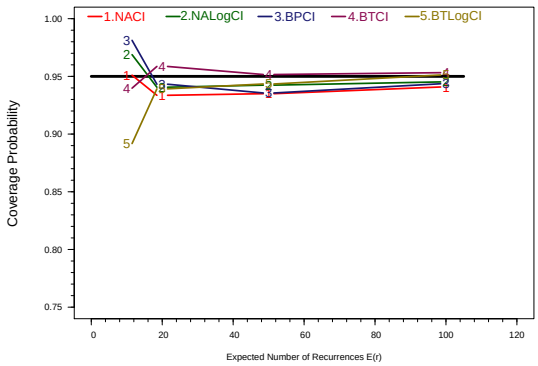
(b) Window2 Data with $n = 10$



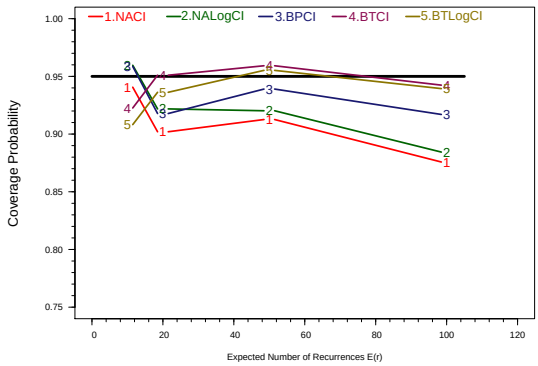
(c) Complete Data with $n = 20$



(d) Window2 Data with $n = 20$



(e) Complete Data with $n = 50$



(f) Window2 Data with $n = 50$

Figure 3.7 CP Plots for NP Estimator with $\beta = 1$

- For each n value, CP values stabilize when $E(r) \geq 20$, and over the parallel part, the Boott and LBoott procedures consistently have CP values that are very close to the nominal value.

- When $E(r) \geq 20$, with the increase of n, the performances of the NORMA procedure, the LNORMA procedure, and the BootP procedure improve, with CP values getting closer to the nominal value.

- When $E(r) = 10$, increasing the value of n does little to help decrease the differences among the five CI procedures, and which CI procedures to use depends on the value of n. When $n = 10$, the BootP procedure has the best CP, while the other four CI procedures have CP values that are less than the nominal 0.95 value. When $n \geq 20$, the NORMA procedure generates the best CP, followed by the LNORMA procedure (with a CP that is larger than the nominal 0.95). For all the n values, the LBoott procedure severely underestimates the CP, an indication that when $E(r)$ is small and thus there is smaller variation in the bootstrap re-samples, the log-transformation is a less appropriate choice.

For the two normal approximation procedures, Appendix A.4.2 shows that, when $n = 10$, the CP values are close to 0.9 for E(r) = 10, 20, 50, and 100, even though the nominal CP is 0.95. Therefore, it is somewhat surprising to observe CP at about 0.92 when E(r) = 10 in Figure 3.7 (a). The reason for this behavior is that the CP values in our simulation study are conditional on $\sum_{i=1}^{n} X_i \geq 5$. The impact of this condition is very small for E(r) $\geq$ 20; therefore the conditional CP values are close to 0.9 as well when E(r) $\geq$ 20. However, when E(r) = 10, Pr $(\sum_{i=1}^{n} X_i \leq 4) = 0.0293$ is not negligible. There are 12 distinct scenarios that have $\sum_{i=1}^{n} X_i \leq 4$ when $n = 10$, such as none of the 10 units have a recurrence, as well as 1 unit has 4 recurrences and 9 units have no recurrences. Among these 12 scenarios, there is only one case for which the CI captures the true MCF – one unit has 4 recurrences and 9 units have no recurrences, and the corresponding probability for this case is 1.89×10^{-5}. As a result, the conditional CP given $\sum_{i=1}^{n} X_i \leq 4$ is $1.89 \times 10^{-5}/0.0293 = 6.45 \times 10^{-4}$. With Pr(MCF $\in$ CI) = 0.89415 from Table A.1 in Appendix A.4.2, we have

$$\Pr(\text{MCF} \in \text{CI} | \sum_{i=1}^{n} X_i \geq 5) = \frac{\Pr(\text{MCF} \in \text{CI}) - \Pr(\text{MCF} \in \text{CI} | \sum_{i=1}^{n} X_i \leq 4)}{1 - \Pr(\sum_{i=1}^{n} X_i \leq 4)} = 0.92.$$

This explains why in Figure 3.7 (a), the CP for the NORMA and LNORMA procedures are about 0.92 when $\text{E}(r) = 10$, while about 0.90 when $\text{E}(r) \geq 20$.

3.6.2 Comparison of Results from Window2 Data

Compared to the Complete data, outcomes for the Window2 data are more complicated, because the amount of time with RSSZ and RSSONE depends on the simulation experiment factor levels. Table 3.2 shows the average length in time of RSSZ and RSSONE, as well as the corresponding ratios to t_{endobs}, for $n = 10$, 20, and 50. Even though the NP estimator is biased when there are times with RSSZ, we still summarize below all the results that we have observed, because some CI procedures are relatively robust to the existence of RSSZ and RSSONE, and sometimes the NP estimator might be the only applicable option to use. Recommendations, including the scenario that the NP estimator should not be used, are outlined in Section 3.6.3.

- Consider Figure 3.7 (b), when $n = 10$. All five CI procedures perform poorly when the amount of time with RSSZ and RSSONE is large, especially the NORMA, the LNORMA, and the BootP procedures. These three CI procedures generate CP values that are below 0.75 even when $\text{E}(r) = 10$, and the CP values deteriorate fast to 0 when $\text{E}(r) \geq 50$, and thus no lines are shown for these three procedures in the plot. By comparison, the Bootт and the LBootт CI procedures have better performance, and their CP values, even though still well below the nominal value at 0.95, are much higher, especially when $\text{E}(r) \leq 20$. Table 3.2 shows that, when $n = 10$, the percentage of time with RSSZ ranges from 14% to 74.3%, and as $\text{E}(r)$ increases, the time with RSSZ and RSSONE becomes more dominant.

- Consider Figure 3.7 (d), when $n = 20$. The Bootт and the LBootт CI procedures show strong robustness to the existence of RSSZ and RSSONE, with all CP values above 0.85. The NORMA, the LNORMA, and the BootP procedures still perform poorly when $\text{E}(r) \geq 20$, but the BootP procedure has CP value that is closest to the nominal value

Table 3.2 RSSZ and RSSONE for Window2 Data with $\beta = 1$ and Number of Simulations = 5000

			Average		Percentage of time relative to t_{endobs}		
n	E(r)	t_{endobs}	RSSZ	RSSONE	RSSZ	RSSONE	RSSZ+RSSONE
10	10	4.30	0.60	1.07	14.0%	24.9%	38.8%
10	20	14.58	4.31	4.98	29.6%	34.2%	63.7%
10	50	81.42	46.65	24.42	57.3%	30.0%	87.3%
10	100	312.91	232.63	65.19	74.3%	20.8%	95.2%
20	10	1.40	0.02	0.05	1.2%	3.8%	5.0%
20	20	4.30	0.14	0.42	3.1%	9.6%	12.8%
20	50	21.97	3.37	5.92	15.3%	26.9%	42.3%
20	100	81.42	28.80	27.65	35.4%	34.0%	69.3%
50	10	0.38	0.00	0.00	0.0%	0.0%	0.0%
50	20	1.00	0.00	0.00	0.0%	0.1%	0.1%
50	50	4.30	0.00	0.02	0.1%	0.4%	0.5%
50	100	14.58	0.09	0.44	0.6%	3.0%	3.6%

when E(r) = 10. Table 3.2 shows that, when $n = 20$, the percentage of time with RSSZ ranges from 1.2% to 35.4%.

- Consider Figure 3.7 (f), when $n = 50$. All CI procedures have performances that are much closer to those of the Complete data. This agrees with what is shown in Table 3.2 that the amount of time with RSSZ is zero or negligible when $n = 50$. The CP values, however, show some drop from E(r) = 50 to 100, and one contributing factor is the existences of RSSONE and intervals with relatively small size of the risk set. As shown in Table 3.2, the percentage of time with RSSONE is 3% when E(r) = 100. As for the comparisons among the five CI procedures, when E(r) = 10, the NORMA, the LNORMA, and the BootP procedures have CP that are closer to the nominal value, while the Boott and the LBoott CI procedures have CP values that are farther away and lower than the nominal value. When E(r) $\geq$ 20, however, the Boott and the LBoott CI procedures have CP values that are very close to the nominal value, while the other three procedures are not as good.

3.6.3 Recommendations for the NP Estimator

Based on the simulation results, we recommend the following for the NP estimator. For Complete data, or window data with very small percentage of time as RSSZ and RSSONE, we recommend:

1. When E(r) = 10 and $n \geq 20$ ($n \geq 50$ for Window2 data), one should use the NORMA procedure, because of simplicity in calculation and because it has a CP that is close to the nominal value.

2. When both E(r) and n are small, the BootP procedure provides a CP that is closest to the nominal value.

3. When E(r) ≥ 20, one can use either the Boott or the LBoott procedure to have a CP that is close to the nominal value.

For window data with non-negligible amounts of time with RSSZ and RSSONE, we recommend:

1. When the amount of time with RSSZ and RSSONE is more than 70% of t_{endobs}, the NP estimator should not be used, because it is seriously biased, and none of the CI procedures behave well.

2. When the amount of time with RSSZ and RSSONE is non-negligible but less dominating:

   a When E(r) is 20 or more, the Boott and the LBoott procedures should be used because of their robustness to the existence of RSSZ and RSSONE and because their CP values are close to the nominal value.

   b When E(r) is small, and the amount of time with RSSZ and RSSONE is about 5%, as in the case of $n = 20$, the BootP procedure is recommended, because its CP values are relatively close to the nominal values.

   c When E(r) is small, yet the amount of time with RSSZ and RSSONE is relatively large, as about 40% for the case of $n = 10$, the Boott and the LBoott procedures are recommended because their CP values are relatively close to the nominal values.

3.7 Performances of the CI Procedures Based on the NHPP Estimator

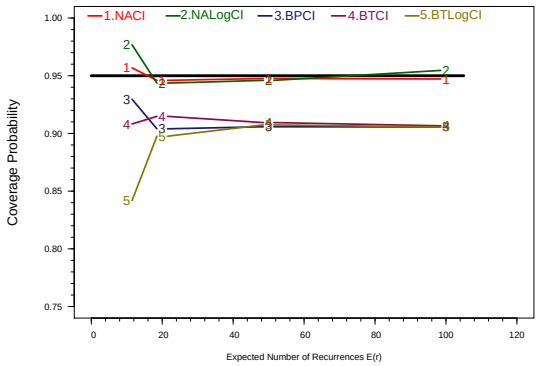
This section summarizes the simulation results for the power law NHPP estimator, and makes some recommendations. Figure 3.8 shows six plots for the NHPP estimator, similar to those in Figure 3.7 for the NP estimator.

3.7.1 Comparison of Results from Complete Data and Window2 Data

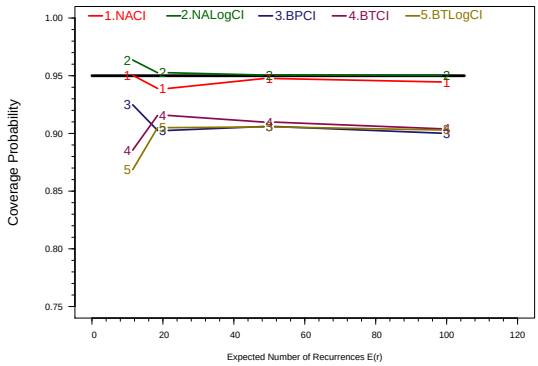
From the plots in Figure 3.8, we have the following primary observations for the NHPP estimator:

- The two plots on the same row, left for the Complete data and right for the Window2 data, are very similar. This indicates that the performances of the CI procedures for the NHPP estimator is not strongly affected by the existence of RSSZ intervals.
- For each n value, CP values level off when $E(r) \geq 20$.
- When $E(r) \geq 20$, the two normal-approximation-based procedures have CP values that are close to the nominal value. The three bootstrap methods, on the other side, are not as good when n is small because of the discreteness in the re-sampling procedure. However, when n increases, the performances of the three bootstrap methods improve, and they are close to the two normal-approximation-based procedures when $n = 50$.
- When $E(r) = 10$, the NORMA procedure has a CP that is consistently close to the nominal value. The LNORMA procedure has a CP that is consistently larger than the nominal 0.95, while the Boott procedure has a CP that is smaller than the nominal value. The LBoott procedure has a CP that is much smaller than the nominal value. When n increases, however, there is only a small improvement in CI performances for the LNORMA and LBoott procedures, and this differs from the observation when $E(r) \geq 20$ that the CP values become closer to the nominal value of 0.95 as n increases.

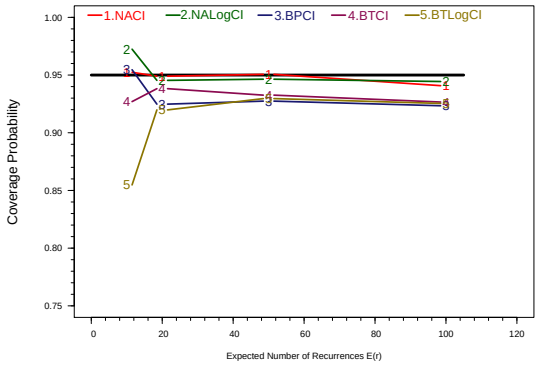
The simulation results suggest that the adequacy of the CI procedures for the NHPP estimator depends mainly on E(r), and is relatively independent of n and tends to be robust to the



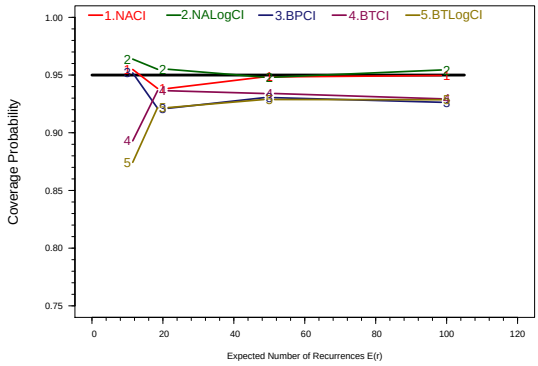
(a) Complete Data with $n = 10$



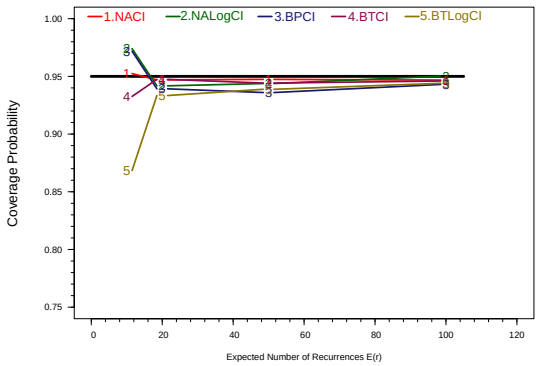
(b) Window2 Data with $n = 10$



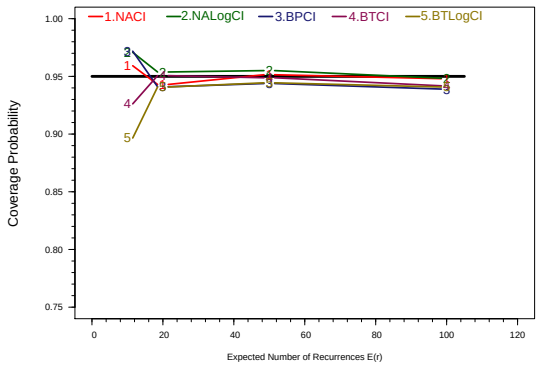
(c) Complete Data with $n = 20$



(d) Window2 Data with $n = 20$



(e) Complete Data with $n = 50$



(f) Window2 Data with $n = 50$

Figure 3.8 CP Plots for the Power Law NHPP Estimator with $\beta = 1$

existence of RSSZ and RSSONE. When n is small, however, say 20 or smaller, the bootstrap procedures can have CP values that are importantly less than the nominal value.

3.7.2 Recommendations for the NHPP Estimator

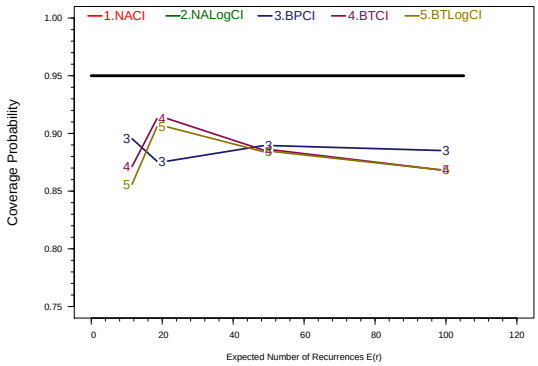
When the assumed NHPP model provides an adequate description of the underlying stochastic process, the NORMA procedure would be a good choice, because of simplicity in calculation and because the CP is close to the nominal value.

3.8 Performances of the CI Procedures Based on the Hybrid Estimators from Window2 Data

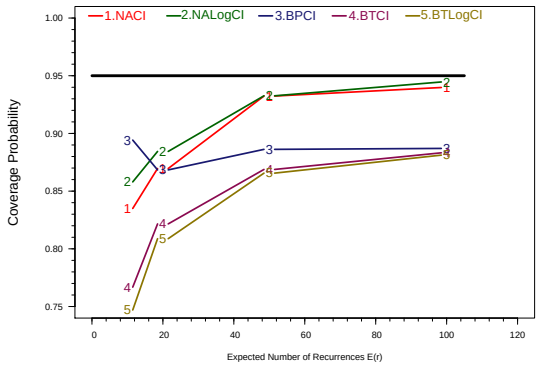
This section summarizes the simulation results for the local hybrid estimator and the power law NHPP hybrid estimator for the Window2 data, and makes some recommendations. Figure 3.9 shows six plots for these two hybrid estimators, with factor $E(r)$ as the x-axis in each plot. From the top down, n increases from 10 to 20 and then to 50, and the plots on the left are for the local hybrid estimator while those on the right are for the power law NHPP hybrid estimator.

3.8.1 Comparison of Results – the Local Hybrid Estimator and the Power Law NHPP Hybrid Estimator

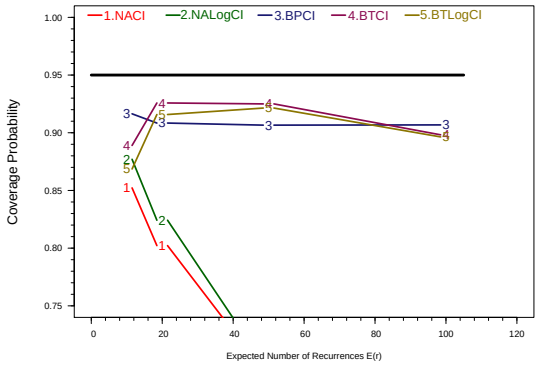
As shown in Section 3.2.5, the local hybrid estimator and the power law NHPP hybrid estimator have the same $\bar{d}_\cdot(t)$, the nonparametric estimator of the increase in the MCF from the RSSP intervals. Thus, when there is no RSSZ (or the amount of time with RSSZ is small), the two hybrid estimators are identical (or close) to the NP estimator. For example, there is hardly any differences observed in Figure 3.9 (e) for the local hybrid estimator and Figure 3.9 (f) for the NHPP hybrid estimator, because when $n = 50$, Table 3.2 shows that the percentage of time with RSSZ is less than 1%. Therefore, meaningful comparisons come from the simulation settings where the time with RSSZ as a percentage of t_{endobs} is not negligible.



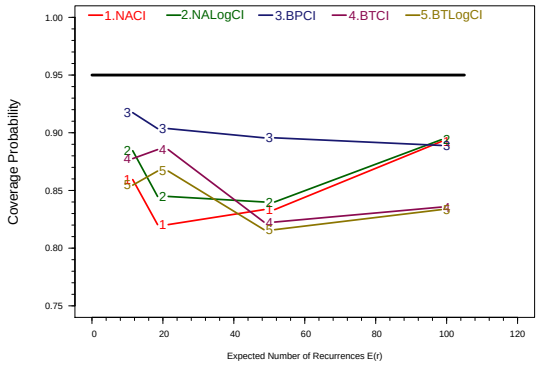
(a) Local Hybrid Estimator with $n = 10$



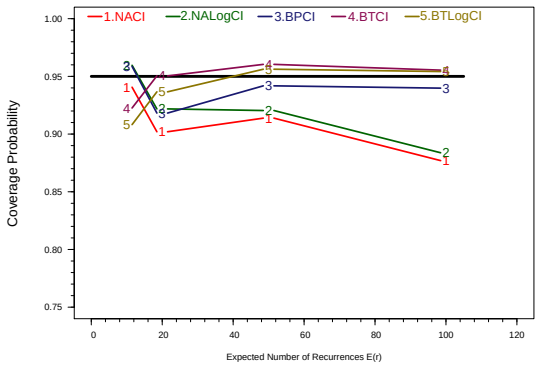
(b) NHPP Hybrid Estimator with $n = 10$



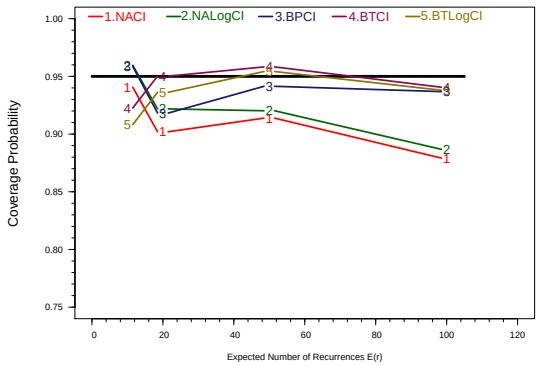
(c) Local Hybrid Estimator with $n = 20$



(d) NHPP Hybrid Estimator with $n = 20$



(e) Local Hybrid Estimator with $n = 50$



(f) NHPP Hybrid Estimator with $n = 50$

Figure 3.9 CP Plots for the Hybrid Estimators for the Window2 Data with $\beta = 1$

When $n = 10$, Table 3.2 shows that the average amount of time of RSSZ plus RSSONE as a percentage of t_{endobs} increases from 38.8% to 95.2% when E(r) increases from 10 to 100. Correspondingly, the differences among the two hybrid estimators also increase, as shown in Figure 3.9 (a) and (b). For the local hybrid estimator, all five CI procedures, especially the three bootstrap-based procedures, have better performances than those of the NP estimator, which are partially observable by comparing Figure 3.9 (a) and Figure 3.7 (b). The CP values for the NORMA and LNORMA procedures are all below 0.75 and fall to 0.49 when E(r) = 100, and thus are not shown in Figure 3.9 (a). One major difficulty for the normal-approximation-based procedures is that the normal distribution assumption is hard to satisfy when the size of the risk set is small. Compared to the local hybrid estimator, the five CI procedures for the power law NHPP hybrid estimator have much better performances: the CP values increase and get closer to the nominal value as E(r) increases. The increasing dominance of RSSZ and RSSONE does not impair the performances of the CI procedures. The main reason is that the increases in the MCF from the RSSZ intervals are estimated from the assumed power law NHPP model, which is the correct model form that is used to generate the simulation data, and which, as observed in Section 3.7, is robust to high percentage of time with RSSZ and RSSONE.

When holding E(r) at the same values and increasing n from 10 to 20, the proportion of time with RSSZ, as well as time with RSSZ and RSSONE, gets smaller. Thus we see better performances of the CI procedures with this increase in the values of n. For the local hybrid estimator, despite the different levels of improvement, the two normal-approximation-based procedures are still far from being desirable, as shown in Figure 3.9 (c), and the CP values fall to 0.57 at E(r) = 100. For the power law NHPP hybrid estimator, however, the BootP procedure is the only procedure that has CP values closer to the nominal value for all four E(r) values, while the other four CI procedures show different amounts of deterioration at E(r) = 50 and E(r) = 100, as observed by comparing Figure 3.9 (b) and (d). One contributing factor for this behavior is the relatively large proportion of time with RSSONE. Even though the percentage of time with RSSZ and RSSONE gets smaller when n increases from 10 to

20, Table 3.2 shows that the percentage of time with RSSONE only drops slightly from 30% to 26.9% at E(r) = 50, and even inceases from 20.8% to 34% at E(r) = 100. Covariances among the recurrence times over periods of time with RSSONE are not estimable, and also covariances between RSSONE intervals and other time intervals are not estimable. Therefore, the existence of a relatively large percentage of time with RSSONE has a negative effect on the estimation of the variances of the MCF estimators. The BootP procedure only depends on the point estimates of the MCF, and larger values of n ensure more randomized bootstrap re-samples, and thus its performance improved.

3.8.2 Recommendations for the Hybrid Estimators

Among our simulation experiment factor levels, when $n \geq 50$, the percentage of time with RSSZ is zero or close to zero. In these cases, the hybrid estimators are the same as or close to the NP estimator. Therefore, our observations and the recommendations below for the hybrid MCF estimators mainly apply to data with relatively small n (e.g., ≤ 20) where there is a non-negligible amount of time with RSSZ.

1. When the percentage of time with RSSZ and RSSONE is large, the differences in CI performances between the local hybrid estimator and those of the NHPP hybrid estimator are large. When there is no strong indication of model deviation from the NHPP model, the NHPP hybrid estimator is preferred, because the variance estimate tends to be smaller in the bias-variance tradeoff; otherwise, the local hybrid estimator is preferred.

2. For the NHPP hybrid estimator, assuming that the NHPP model provides a relatively good description of the underlying point process,
   - When the percentage of time with RSSZ is very large (e.g., 57.3% when $E(r) = 50$ and $n = 10$ or 74.3% when $E(r) = 100$ and $n = 10$), the two normal-approximation-based procedures are preferred, because their CP values are closer to the nominal value, and they are simple to construct. For these cases, the NHPP hybrid estimator is close to the NHPP estimator, and as shown in Section 3.7.2, the NORMA procedure is recommended for the NHPP estimator.

- When the percentage of time with RSSZ is moderate (e.g., $\leq 15.3\%$ for the cases of $E(r) <= 50$ and $n = 20$, as well as $E(r) = 10$ and $n = 10$), the BootP procedure is preferred, because its CP is the closest to the nominal value. The relative robustness to the existence of RSSONE is one contributing factor for the comparatively better performance of the BootP procedure.

3. For the local hybrid estimator,
   - When E(r) is small, for example around 10, the BootP procedure is preferred, because its CP is the closest to the nominal value.
   - When E(r) ≥ 20, the three bootstrap-based procedures generate similar CP results and these results are substantially better than those of the two normal-approximation-based procedures. Among the three bootstrap-based procedures, the BootP procedure might be preferred because it is, by comparison, simpler to construct.

3.9 Concluding Remarks and Areas for Further Research

In our simulation study, we have shown the performances of five of the commonly used CI procedures. All five CI procedures perform well when both the number of units n and the expected number of observed recurrences $E(r)$ are reasonably large, and there is no time with RSSZ and RSSONE or the percentage of time with RSSZ and RSSONE is small. However, when n is small, or $E(r)$ is small, or the percentage of time with RSSZ and RSSONE is not negligible, choices among the MCF estimators and the CI procedures can lead to very different results. We have made some recommendations based on the simulation results. There are, however, some important areas for further research.

- The five CI procedures in our simulation study are relatively simple and easy to construct. With improvements in the computing power, more complicated bootstrap CI procedures could be included in the comparison study. The BC_a method and the ABC method described in Efron and Tibshirani (1993) are both second-order accurate and transformation respecting, and their finite-sample properties would be of interest.

- In addition to normal-approximation and bootstrap-based CI procedures, likelihood-based methods could be used to construct CIs for the NHPP model. Similarly, empirical likelihood methods could be used for the NP estimators to construct CIs. These likelihood and empirical likelihood methods could also be extended to the hybrid estimators. In particular, methods described in Owen (2001) could be extended for these cases.

- It is sometimes possible to obtain better estimation of the MCF by including the explanatory variables in the model. Some such models are described in Lawless and Nadeau (1995), and Cook and Lawless (2007). Confidence interval procedures for such models could also be studied.

ACKNOWLEDGMENTS

This material is based upon work supported, in part, by the National Science Foundation under Grant No. 0421916.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Cook, R. J., and Lawless, J. F. (2007), *The Statistical Analysis of Recurrent Events*, New York: Springer-Verlag.

Cox, D. R., and Lewis, P. A. W. (1966), *The Statistical Analysis of Series of Events*, New York: Wiley.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Jeng, S. L., and Meeker, W. Q. (2000), "Comparisons of Approximate Confidence Interval Procedures for Type I Censored Data," *Technometrics*, 42, 135-148.

Lawless, J. F., and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Nelson, W. B. (1988), "Graphical Analysis of System Repair Data," *Journal of Quality Technology*, 20, 24-35.

Nelson, W. B. (1995), "Confidence Limits for Recurrence Data: Applied to Cost or Number of Product Repairs," *Technometrics*, 37, 147-157.

Nelson, W. B. (2003), *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Philadelphia: ASA-SIAM.

Owen, A. B. (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC.

Rigdon, S. E., and Basu, A. P. (2000), *Statistical Methods for the Reliability of Repairable Systems*, New York: Wiley.

Zuo, J., Meeker, W. Q., and Wu, H. (2008), "Analysis of Window-Observation Recurrence Data," *Technometrics*, 50, 128-143.

CHAPTER 4. ASYMPTOTIC PROPERTIES OF MEAN CUMULATIVE FUNCTION ESTIMATORS FROM WINDOW-OBSERVATION RECURRENCE DATA

Jianying Zuo

Huaiqing Wu

William Q. Meeker

Statistics Department

Iowa State University

Ames, IA 50011

Abstract

A variety of nonparametric and parametric methods have been used to estimate the mean cumulative function (MCF) for the recurrence data collected from the counting process. When the recurrence histories of some systems are available in disconnected observation windows with gaps in between, Zuo, Meeker, and Wu (2008) show that both the nonparametric and parametric methods can be extended to estimate the MCF. In this article, we establish some asymptotic properties of the MCF estimators for the window-observation recurrence data.

KEY WORDS: Asymptotic normality; Consistency; Nonhomogeneous Poisson process; Nonparametric estimation.

4.1 Introduction

4.1.1 Background and Motivation

Window-observation recurrence data arise when the recurrence histories of some systems are available in disconnected observation windows with gaps in between. Nelson (2003, page 75) gives an example in which window-observation recurrence data arise, and Zuo, Meeker, and Wu (2008) describe two other applications. Nonparametric and parametric methods to analyze the recurrence data are available in many publications. Examples include Nelson (2003), Cook and Lawless (2007), Meeker and Escobar (1998), Lawless and Nadeau (1995), and Rigdon and Basu (2000). Among the quantities to be analyzed with the recurrence data are the locations and counts of recurrences, and statistics of interest include the mean cumulative number of recurrences. Zuo, Meeker, and Wu (2008) show that both the nonparametric and parametric methods can be extended to estimate the mean cumulative function (MCF) for window-observation recurrence data. When there are time intervals with risk-set-size-zero (RSSZ) (i.e., time intervals in which no system is under observation), they propose the local hybrid estimator and the nonhomogeneous Poisson process (NHPP) hybrid estimator. As a continuation of Zuo, Meeker, and Wu (2008), this article establishes some asymptotic properties of the MCF estimators. Note that the hybrid estimators are finite-sample alternatives to the nonparametric (NP) estimator because the latter is downward-biased with the existence of RSSZ. Because there is no information available for estimating the MCF in an RSSZ interval, RSSZs need to go away asymptotically for the MCF estimators to be consistent. Therefore, in presenting the asymptotic properties of the MCF estimators, we only include the NP estimator and the NHPP estimators.

Andersen, Borgan, Gill, and Keiding (1993) provide comprehensive descriptions of the statistical models and methods that can be used to analyze event history observed in continuous time. They derive the asymptotic (i.e., large-sample) properties for both the nonparametric

and parametric estimators described in their book. Building on their conditions and theorems, we show that the NP and the NHPP MCF estimators of Zuo, Meeker, and Wu (2008) have the desirable asymptotic properties under some mild conditions that are generally satisfied in practical analysis.

4.1.2 Other Previous Work

Peña, Strawderman, and Hollander (2001) describe nonparametric methods to estimate the inter-occurrence times with recurrent event data. Ghosh and Lin (2000) also focus on the nonparametric analysis of recurrent event data, possibly with a terminal event, and they estimate the mean frequency function, which is defined the same as the MCF. Both articles establish the asymptotic properties of their respective nonparametric estimators, and use simulation studies for the finite-sample properties.

4.1.3 Overview

The remainder of this article is organized as follows. Section 4.2 outlines some common notation and assumptions to be used. Sections 4.3 and 4.4 establish the asymptotic properties for the NP and NHPP estimators, respectively. Section 4.5 gives some concluding remarks.

4.2 Notation and Assumptions

4.2.1 Notation for the Models

Let $N(t)$ denote the number of events in the time interval $(0, t]$. Then $\mu(t)$, the expectation of $N(t)$, is the MCF. If the MCF is differentiable, then $\nu(t) = d\mu(t)/dt$ is the recurrence rate, and $\nu(t) \times \Delta t$ can be interpreted as the approximate expected number of events to occur during the next short time interval $(t,\ t + \Delta t]$.

Let $K^{(n)}$ denote the counting process with the multiplicative intensity model $\lambda(t) = \nu(t)\delta_{\cdot}(t)$, where n is the total number of units, and $\delta_{\cdot}(t)$ is the size of the risk set at time t (i.e., the number of units that are in observation windows at time t). Note that $\delta_{\cdot}(t) = \sum_{i=1}^{n} \delta_i(t)$, where

$$\delta_i(t) = \begin{cases} 1 & \text{if unit } i \text{ is under observation in a time window at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Then $M^{(n)}(t) = K^{(n)}(t) - \int_0^t \lambda(s)ds$ is a local square integrable martingale (see section IV.1.1, Andersen et al. 1993).

4.2.2 Notation for the Data

For window-observation recurrence data, let w_i be the number of observation windows for unit i, and r_i be the number of recurrences recorded in these observation windows for unit i. Let t_{i1}, t_{i2}, ..., t_{ir_i} be the corresponding recurrence times, and $(t^{(n)}_{i1L}, t^{(n)}_{i1U}]$, ..., $(t^{(n)}_{iw_iL}, t^{(n)}_{iw_iU}]$ be the corresponding observation windows for unit i.

When considering the superposition of all n units, let t_{max} be the largest end-of-observation time among all units. Then $\delta.(t)$ is piece-wise constant over $(0, t_{max}]$. Let z be the number of intervals with constant $\delta.(t)$, and denote the z intervals as $(t^{(n)}_{1L}, t^{(n)}_{1U}]$, $(t^{(n)}_{2L}, t^{(n)}_{2U}]$, ..., $(t^{(n)}_{zL}, t^{(n)}_{zU}]$, where $t^{(n)}_{1L} = 0$, $t^{(n)}_{zU} = t_{max}$, and $t^{(n)}_{iL} = t^{(n)}_{(i-1)U}$ for $i = 2, ..., z$. Note that z, t_{max}, and these intervals all depend on the data.

4.2.3 Assumptions

We assume that the set of observation windows for each unit is a simple event from the sample space consisting of a finite number of prespecified sets of non-overlapping observation windows, and that the probability of each simple event is positive. Furthermore, we assume that the overlap of all these sets of windows leaves no gaps that result in RSSZ. Let $\tau < \infty$ be the ending time point for all these windows. We have $t_{max} \leq \tau$ for all n, and $t_{max} \to \tau$ as $n \to \infty$. These assumptions ensure that at any time $t \in (0, \tau]$, there is a positive probability for a unit to be observed.

Let i_0 be the number of simple events and p_1, ..., p_{i_0} be the corresponding probabilities. Then the expectation of the size of the risk set $\delta.(t)$ with n units is

$$\mathrm{E}[\delta.(t)] = np(t),$$

where $p(t) = \sum_{j=1}^{i_0} p_j I(t \text{ is in the } j\text{th set of observation windows})$ is the probability that a unit is observed at time t. Here $I(S)$ is the indicator function that equals 1 if the statement S is true and 0 otherwise. We have

$$\begin{aligned} p(t) &\geq (\min_{1\leq j\leq i_0} p_j) \sum_{j=1}^{i_0} I(t \text{ is in the } j\text{th set of observation windows}) \\ &\geq \min_{1\leq j\leq i_0} p_j \equiv p_0 > 0. \end{aligned}$$

Note that $\delta.(t) = \sum_{i=1}^{n} \delta_i(t)$ and $\delta_1(t)$, ..., $\delta_n(t)$ are independent and identically distributed (i.i.d.) Bernoulli random variables that take the value of 1 with probability $p(t)$. Thus $\delta.(t)$ is a Binomial (n, $p(t)$) random variable.

Consider the superposition of all i_0 sets of observation windows in the sample space. Let $h+1$ be the number of unique endpoints of all these observation-window intervals and denote the ordered unique endpoints by $0 = t_{1L} < t_{1U} < t_{2U} < ... < t_{hU} = \tau$. Let $t_{iL} = t_{(i-1)U}$ for $i = 2, ..., h$. Then, for $t \in (t_{iL}, t_{iU}]$, $i = 1, ..., h$,

$$p(t) = p(t_{iU}) \equiv p_{ci} \quad \text{and} \quad \delta.(t) = \delta.(t_{iU}) \equiv \delta_i.$$

That is, $p(t)$ is constant and $\delta.(t)$ is a random variable not depending on t for all $t \in (t_{iL}, t_{iU}]$, $i = 1, ..., h$. Thus as $n \to \infty$, we have

$$\frac{\delta.(t)}{n} = \frac{\delta.(t_{iU})}{n} = \frac{\delta_i}{n} \to p_{ci} \quad \text{for } t \in (t_{iL}, t_{iU}],\ i = 1, ..., h.$$

We can link this condition to practical applications. For example, for the extended warranty data in Zuo, Meeker, and Wu (2008), we have, as $n \to \infty$

$$\delta.(t)/n \to \begin{cases} 1 & \text{for } t \in (0, 1], \\ 0.5 & \text{for } t \in (1, 2], \\ 0.4 & \text{for } t \in (2, 3]. \end{cases}$$

4.3 Nonparametric MCF Estimator

4.3.1 Estimation for the Nonparametric Model

Let m denote the number of unique event times, and let $t_1, ..., t_m$ be the unique event times. Zuo, Meeker, and Wu (2008) define the NP MCF estimator as

$$\widehat{MCF}_{NP}(t_j) = \sum_{k=1}^{j} \left[\frac{\sum_{i=1}^{n} \delta_i(t_k) \times d_i(t_k)}{\sum_{i=1}^{n} \delta_i(t_k)} \right] = \sum_{k=1}^{j} \frac{d_\cdot(t_k)}{\delta_\cdot(t_k)} = \sum_{k=1}^{j} \bar{d}(t_k), \quad j = 1, ..., m,$$

where $d_i(t_k)$ is the number of events recorded at time t_k for unit i.

4.3.2 Theorems Adapted from Andersen, Borgan, Gill, and Keiding (1993)

For the NP MCF estimator, the following two theorems from Andersen et al. (1993) are simplified and adapted with our notation. Note that "$\xrightarrow{\mathbf{P}}$" is used to denote "converges in probability to" and "$\xrightarrow{\mathscr{D}}$" is used to denote "converges in distribution to." Also $\sup_{s\in[0,t]} f(s)$ and $\inf_{s\in[0,t]} f(s)$ stand for the supremum and infimum of $f(s)$ over $[0, t]$, respectively.

Theorem IV.1.1 of Andersen et al. (1993, page 190). Let $t \in (0, \tau]$ and assume that, as $n \to \infty$,

$$\int_0^t \frac{I[\delta_\cdot(s) > 0]}{\delta_\cdot(s)} \nu(s) ds \xrightarrow{\mathbf{P}} 0 \tag{4.1}$$

and

$$\int_0^t \{1 - I[\delta_\cdot(s) > 0]\} \nu(s) ds \xrightarrow{\mathbf{P}} 0. \tag{4.2}$$

Then, as $n \to \infty$,

$$\sup_{s\in[0,t]} |\hat{\mu}(s) - \mu(s)| \xrightarrow{\mathbf{P}} 0,$$

where $\hat{\mu}(t)$ is the NP estimator of $\mu(t)$ defined as

$$\hat{\mu}(t) = \int_0^t \delta.(s)^{-1} dN(s) = \widehat{MCF}_{NP}(t).$$

This theorem can be used to show that the NP estimator of $\mu(t)$, based on window-observation recurrence data, is uniformly consistent on compact intervals.

Theorem IV.1.2 of Andersen et al. (1993, page 191). Let $t \in (0, \tau]$ and assume that there exists a sequence of positive constants $\{a_n\}$, increasing to infinity as $n \to \infty$, and a positive function y such that ν/y is integrable over $[0, t]$. Assume that

(NP(A)) For each $s \in [0, t]$,

$$a_n^2 \int_0^s \frac{I[\delta.(u) > 0]}{\delta.(u)} \nu(u) du \xrightarrow{\mathbf{P}} \sigma^2(s) \quad \text{as} \quad n \to \infty,$$

where

$$\sigma^2(s) = \int_0^s \frac{\nu(u)}{y(u)} du.$$

(NP(B)) For all $\epsilon > 0$,

$$a_n^2 \int_0^t \frac{I[\delta.(u) > 0]}{\delta.(u)} \nu(u) I\left\{ a_n \frac{I[\delta.(u) > 0]}{\delta.(u)} > \epsilon \right\} du \xrightarrow{\mathbf{P}} 0 \quad \text{as} \quad n \to \infty.$$

(NP(C))

$$a_n \int_0^t (1 - I[\delta.(u) > 0]) \nu(u) du \xrightarrow{\mathbf{P}} 0 \quad \text{as} \quad n \to \infty.$$

Then

$$a_n(\hat{\mu} - \mu) \xrightarrow{\mathscr{D}} U \quad \text{as} \quad n \to \infty$$

on $D[0, t]$, where U is a Gaussian martingale with $U(0) = 0$ and $\text{Cov}(U(s_1), U(s_2)) = \sigma^2(s_1 \wedge s_2)$. Here $D[0, t]$ is the Skorohod space on $[0, t]$, that is, the space of right-continuous functions with left-hand limits on $[0, t]$; $s_1 \wedge s_2$ is the smaller of s_1 and s_2. Theorem IV.1.2 can be used to establish the asymptotic normality of the NP estimator of $\mu(t)$ when it is based on window-observation recurrence data.

4.3.3 Asymptotic Properties of the Nonparametric Estimator

The following lemma will be useful for verifying the conditions in Section 4.3.2.

Lemma. Let $\{X_n\}$ be a sequence of nonnegative random variables. If $\mathrm{E}X_n \to 0$ as $n \to \infty$, then $X_n \xrightarrow{\mathbf{P}} 0$ as $n \to \infty$.

The lemma follows by observing that, for any $\epsilon > 0$,

$$\Pr(X_n \geq \epsilon) \leq \frac{\mathrm{E}X_n}{\epsilon} \to 0 \quad \text{as } n \to \infty.$$

Assume that $\mu(t) < \infty$ for $t \in (0, \tau]$. By the assumptions in Section 4.2.3, for any $t \in (0, \tau]$ and $s \in [0, t]$, we have

$$\begin{aligned}
\mathrm{E}\left[\frac{I[\delta.(s) > 0]}{\delta.(s)}\right] &= \sum_{x=1}^{n} \frac{1}{x} \frac{n!}{x!(n-x)!} p(s)^x [1-p(s)]^{n-x} \\
&= \sum_{x=1}^{n} \frac{(n+1)!}{(x+1)![(n+1)-(x+1)]!} p(s)^{x+1} [1-p(s)]^{(n+1)-(x+1)} \\
&\quad \left(\frac{1}{n+1} \frac{x+1}{x} \frac{1}{p(s)}\right) \\
&\leq \frac{1}{n+1} \frac{2}{p_0} \sum_{u=0}^{n+1} \frac{(n+1)!}{u![(n+1)-u]!} p(s)^u [1-p(s)]^{(n+1)-u} \\
&\quad \left(\text{because } \frac{x+1}{x} \leq 2 \text{ for } x \geq 1 \text{ and } \frac{1}{p(s)} \leq \frac{1}{p_0}\right) \\
&= \frac{2}{p_0(n+1)}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{E}\left[\int_0^t \frac{I[\delta.(s) > 0]}{\delta.(s)} \nu(s) ds\right] &= \int_0^t \mathrm{E}\left[\frac{I[\delta.(s) > 0]}{\delta.(s)}\right] \nu(s) ds \\
&\leq \frac{2}{p_0(n+1)} \int_0^t \nu(s) ds \\
&= \frac{2\mu(t)}{p_0} \frac{1}{n+1} \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Thus by the lemma, (4.1) is satisfied.

Similarly, to verify (4.2), we have

$$\begin{aligned}
\mathrm{E}\left[\int_0^t \{1 - I[\delta.(s) > 0]\}\nu(s)ds\right] &= \int_0^t \{1 - \Pr[\delta.(s) > 0]\}\nu(s)ds \\
&= \int_0^t \Pr[\delta.(s) = 0]\nu(s)ds \\
&= \int_0^t [1 - p(s)]^n \nu(s)ds \\
&\leq \int_0^t [1 - p_0]^n \nu(s)ds \\
&= \mu(t)[1 - p_0]^n \\
&\to 0 \quad \text{as } n \to \infty
\end{aligned}$$

because $0 < p_0 < 1$. Thus by the lemma, (4.2) is satisfied. Therefore, the uniform consistency of the NP estimator of $\mu(t)$ is established.

To establish the asymptotic normality of the NP estimator of $\mu(t)$, we need to show that conditions NP(A) to NP(C) are satisfied.

Let $a_n = \sqrt{n}$ and $y(t) = p(t)$, the probability that a unit is being observed at time t. From the verification of (4.2), we have

$$\mathrm{E}\left[\sqrt{n}\int_0^t \{1 - I[\delta.(u) > 0]\}\nu(u)du \leq \sqrt{n}\mu(t)[1 - p_0]^n\right] \quad \to \quad 0 \quad \text{as } n \to \infty.$$

Thus by the lemma, condition NP(C) is satisfied.

To verify conditions NP(A) and NP(B), first note that as $n \to \infty$,

$$\frac{\delta_i}{n} \xrightarrow{\mathbf{P}} p_{ci} \quad \text{and} \quad I(\delta_i > 0) \xrightarrow{\mathbf{P}} 1,$$

where $\delta_i = \delta.(t_{iU})$, for $i = 1, ..., h$. Thus

$$\begin{aligned}
\left| n\int_0^s \frac{I[\delta.(u)>0]}{\delta.(u)}\nu(u)du - \sigma^2(s)\right| &= \left|\int_0^s \left[\frac{I[\delta.(u)>0]}{\delta.(u)/n} - \frac{1}{p(u)}\right]\nu(u)du\right| \\
&\leq \int_0^s \left|\frac{I[\delta.(u)>0]}{\delta.(u)/n} - \frac{1}{p(u)}\right|\nu(u)du \\
&\leq \int_0^s \nu(u)du \max_{1\leq i\leq h}\left|\frac{I[\delta_i>0]}{\delta_i/n} - \frac{1}{p_{ci}}\right| \\
&\xrightarrow{\mathbf{P}} 0 \quad \text{as } n\to\infty.
\end{aligned}$$

Therefore, condition NP(A) is satisfied.

Furthermore,

$$\begin{aligned}
& n\int_0^t \frac{I[\delta.(u)>0]}{\delta.(u)}\nu(u)I\left\{\sqrt{n}\frac{I[\delta.(u)>0]}{\delta.(u)}>\epsilon\right\}du \\
\leq & \int_0^t \nu(u)du \max_{1\leq i\leq h}\frac{I[\delta_i>0]}{\delta_i/n} I\left\{\max_{1\leq i\leq h}\frac{I[\delta_i>0]}{\delta_i/n}\frac{1}{\sqrt{n}}>\epsilon\right\} \\
\xrightarrow{\mathbf{P}} & 0 \quad \text{as } n\to\infty.
\end{aligned}$$

Thus, condition NP(B) is satisfied. This concludes the proof of the asymptotic properties of the NP estimator of $\mu(t)$.

4.4 Nonhomogeneous Poisson Process (NHPP) MCF Estimators

4.4.1 Estimation for the NHPP Model

For a parametric model, let $\boldsymbol{\theta}$ denote the model parameter, which is a q-dimensional real vector. Also let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$, and $\boldsymbol{\Theta}_0$ be a neighborhood of $\boldsymbol{\theta}_0$.

A particular NHPP model is specified by its recurrence rate function $\nu(t;\boldsymbol{\theta})$, and the expected number of events in the time range $(a,b]$ is $\mu(a,b;\boldsymbol{\theta}) = \int_a^b \nu(t;\boldsymbol{\theta})dt$.

Given the maximum likelihood (ML) estimator, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$, the ML estimator of the NHPP MCF $\mu(t;\boldsymbol{\theta})$ is

$$\widehat{MCF}_{NHPP}(t) = \int_0^t \nu(s;\hat{\boldsymbol{\theta}})ds.$$

The most commonly used NHPP recurrence rate functions are:

- Constant recurrence rate, also known as homogeneous Poisson process (HPP):

$$\nu(t) = c.$$

- Power recurrence rate, also known as the power law process or the Weibull process:

$$\nu(t;\beta,\eta) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \quad \beta > 0, \eta > 0.$$

  Here, $\boldsymbol{\theta} = (\beta,\eta)'$. The power law NHPP model is flexible enough to model a counting process with either an increasing or a decreasing recurrence rate function. When the shape parameter $\beta > 1$, the recurrence function is increasing, and when $\beta < 1$, the recurrence rate function is decreasing. When $\beta = 1$, the power law NHPP model simplifies to the HPP model with constant recurrence rate $1/\eta$.

- Loglinear recurrence rate:

$$\nu(t;\gamma_0,\gamma_1) = \exp(\gamma_0 \ + \ \gamma_1 \times t).$$

  Here, $\boldsymbol{\theta} = (\gamma_0,\gamma_1)'$. The loglinear NHPP model assumes that the log recurrence rate function is a linear function of time, where γ_0 is the intercept and γ_1 is the slope. A positive value of γ_1 indicates an increasing recurrence rate, while a negative value of γ_1 indicates a decreasing recurrence rate. When $\gamma_1 = 0$, the loglinear NHPP model simplifies to the HPP model with recurrence rate $\exp(\gamma_0)$.

4.4.2 Conditions and Theorems Adapted from Andersen, Borgan, Gill, and Keiding (1993)

To establish the asymptotic properties of the NHPP MCF estimators, the following conditions and theorems from Andersen et al. (1993) are simplified and adapted with our notation.

From page 402 of Andersen et al. (1993), the log-partial-likelihood function for the parametric model is

$$\ell_\tau(\boldsymbol{\theta}) \quad = \quad \int_0^\tau \log\lambda(t;\boldsymbol{\theta})dN(t) - \int_0^\tau \lambda(t;\boldsymbol{\theta})dt, \tag{4.3}$$

and, assuming that we may interchange the order of differentiation and integration, the vector $\mathbf{U}_\tau(\boldsymbol{\theta})$ of score statistics $U_\tau^j(\boldsymbol{\theta})$, $j = 1, ..., q$, is given by

$$U_\tau^j(\boldsymbol{\theta}) \quad = \quad \int_0^\tau \frac{\partial}{\partial \theta_j} \log \lambda(t; \boldsymbol{\theta}) dN(t) - \int_0^\tau \frac{\partial}{\partial \theta_j} \lambda(t; \boldsymbol{\theta}) dt.$$

Note that under the assumptions of Section 4.2.3, the partial likelihood is equivalent to the likelihood (up to a constant scalar not depending on $\boldsymbol{\theta}$). Thus the partial likelihood equation $\mathbf{U}_\tau(\boldsymbol{\theta}) = \mathbf{0}$ is also the likelihood equation.

Condition VI.1.1 of Andersen et al. (1993, pages 420 and 421) can be simplified below for our model.

(A) There exists a neighborhood $\boldsymbol{\Theta}_0$ of $\boldsymbol{\theta}_0$ such that for all n and $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and almost all $t \in (0, \tau]$, the partial derivatives of $\lambda(t; \boldsymbol{\theta})$ and $\log \lambda(t; \boldsymbol{\theta})$ of the first, second, and third order with respect to $\boldsymbol{\theta}$ exist and are continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Moreover, the log-likelihood function (4.3) may be differentiated three times with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ by interchanging the order of integration and differentiation.

(B) There exists a sequence $\{a_n\}$ of nonnegative constants increasing to infinity as $n \to \infty$ and finite functions $\sigma_{jl}(\boldsymbol{\theta})$ defined on $\boldsymbol{\Theta}_0$ such that for all j, l

$$a_n^{-2} \int_0^\tau \left\{ \frac{\partial}{\partial \theta_j} \log \lambda(t; \boldsymbol{\theta}_0) \right\} \left\{ \frac{\partial}{\partial \theta_l} \log \lambda(t; \boldsymbol{\theta}_0) \right\} \lambda(t; \boldsymbol{\theta}_0) dt \xrightarrow{\mathbf{P}} \sigma_{jl}(\boldsymbol{\theta}_0) \tag{4.4}$$

as $n \to \infty$.

(C) For all j and all $\epsilon > 0$, we have that

$$a_n^{-2} \int_0^\tau \left\{ \frac{\partial}{\partial \theta_j} \log \lambda(t; \boldsymbol{\theta}_0) \right\}^2 I\left(\left| a_n^{-1} \frac{\partial}{\partial \theta_j} \log \lambda(t; \boldsymbol{\theta}_0) \right| > \epsilon \right) \lambda(t; \boldsymbol{\theta}_0) dt \xrightarrow{\mathbf{P}} 0 \tag{4.5}$$

as $n \to \infty$.

(D) The matrix $\boldsymbol{\Sigma} = \{\sigma_{jl}(\boldsymbol{\theta}_0)\}$ with $\sigma_{jl}(\boldsymbol{\theta}_0)$ defined in Condition B is positive definite.

(E) For any n there exist predictable processes G_n and H_n not depending on $\boldsymbol{\theta}$ such that for all $t \in (0, \tau]$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \left| \frac{\partial^3}{\partial\theta_j \partial\theta_l \partial\theta_m} \lambda(t; \boldsymbol{\theta}) \right| \leq G_n(t),$$

and

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \left| \frac{\partial^3}{\partial\theta_j \partial\theta_l \partial\theta_m} \log \lambda(t; \boldsymbol{\theta}) \right| \leq H_n(t),$$

for all j, l, m. Moreover

$$a_n^{-2} \int_0^\tau G_n(t) dt,$$

$$a_n^{-2} \int_0^\tau H_n(t) \lambda(t; \boldsymbol{\theta}_0) dt$$

as well as (for all j, l)

$$a_n^{-2} \int_0^\tau \left\{ \frac{\partial^2}{\partial\theta_j \partial\theta_l} \log \lambda(t; \boldsymbol{\theta}_0) \right\}^2 \lambda(t; \boldsymbol{\theta}_0) dt \tag{4.6}$$

all converge in probability to finite quantities as $n \to \infty$, and for all $\epsilon > 0$,

$$a_n^{-2} \int_0^\tau H_n(t) I \left\{ a_n^{-1} \left[H_n(t) \right]^{1/2} > \epsilon \right\} \lambda(t; \boldsymbol{\theta}_0) dt \xrightarrow{\mathbf{P}} 0. \tag{4.7}$$

Theorem VI.1.1 of Andersen et al. (1993, page 422). Assume that Condition VI.1.1 holds. Then, with a probability tending to one, the equation $\mathbf{U}_\tau(\boldsymbol{\theta}) = \mathbf{0}$ has a solution $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} \xrightarrow{\mathbf{P}} \boldsymbol{\theta}_0$ as $n \to \infty$.

Theorem VI.1.2 of Andersen et al. (1993, pages 424 and 425). Assume that Condition VI.1.1 holds, and let $\hat{\boldsymbol{\theta}}$ be a consistent solution of the equation $\mathbf{U}_\tau(\boldsymbol{\theta}) = \mathbf{0}$. Then

$$a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathscr{D}} \mathscr{N}(\mathbf{0}, \boldsymbol{\Sigma}^{-1}),$$

where $\boldsymbol{\Sigma} = \{\sigma_{jl}(\boldsymbol{\theta}_0)\}$ defined in Condition (D) may be estimated consistently by $a_n^{-2}\mathscr{J}_\tau(\hat{\boldsymbol{\theta}})$. Here $-\mathscr{J}_\tau(\hat{\boldsymbol{\theta}})$ is the matrix of the second-order partial derivatives of the log-likelihood function (4.3). That is, the (j,l)th element of the matrix $\mathscr{J}_\tau(\boldsymbol{\theta})$ can be written as

$$\mathscr{J}_\tau^{jl}(\boldsymbol{\theta}) \;=\; \int_0^\tau \frac{\partial^2}{\partial\theta_j\partial\theta_l}\lambda(t;\boldsymbol{\theta})dt - \int_0^\tau \frac{\partial^2}{\partial\theta_j\partial\theta_l}\log\lambda(t;\boldsymbol{\theta})dN(t).$$

Theorem VI.1.1 establishes the existence of a consistent MLE of $\boldsymbol{\theta}$, and Theorem VI.1.2 establishes the asymptotic normality of the MLE.

4.4.3 Asymptotic Properties of the NHPP Estimators

First note that the consistency of $\widehat{MCF}_{NHPP}(t)$ in (4.3) follows from that of $\hat{\boldsymbol{\theta}}$, provided that $\mu(t;\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Furthermore, the asymptotic normality of $\widehat{MCF}_{NHPP}(t)$ follows from that of $\hat{\boldsymbol{\theta}}$ by the δ-method, provided that $\mu(t;\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ (Billingsley 1986, page 402). To establish the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ for the NHPP models, we need to show that the Conditions (A) to (E) in Condition VI.1.1 are satisfied.

By the assumptions in Section 4.2.3, $\delta_\cdot(t) = \delta_\cdot(t_{iU}) \equiv \delta_i$ for $t \in (t_{iL}, t_{iU}] \equiv J_i$, for $i = 1, ..., h$, and $\Pr[\delta_\cdot(t) > 0] \to 1$ as $n \to \infty$. Note that for the multiplicative intensity function, we have $\lambda(t;\boldsymbol{\theta}) = \nu(t;\boldsymbol{\theta})\delta_\cdot(t)$. Thus Condition (A) can be restated as

(A′) There exists a neighborhood $\boldsymbol{\Theta}_0$ of $\boldsymbol{\theta}_0$ such that for all n and $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and almost all $t \in (0, \tau]$, the partial derivatives of $\nu(t;\boldsymbol{\theta})$ and $\log\nu(t;\boldsymbol{\theta})$ of the first, second, and third order with respect to $\boldsymbol{\theta}$ exist and are continuous in $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Moreover, the log-likelihood function (4.3) may be differentiated three times with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ by interchanging the order of integration and differentiation.

For Conditions (B) to (E), let $a_n = \sqrt{n}$. Then the left side of (4.4) can be written as

$$
\begin{aligned}
& n^{-1}\int_0^\tau \left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}\left\{\frac{\partial}{\partial\theta_l}\log\nu(t;\boldsymbol{\theta}_0)\right\}\nu(t;\boldsymbol{\theta}_0)\delta.(t)dt \\
= & \sum_{i=1}^{h}\frac{\delta_i}{n}\int_{J_i}\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}\left\{\frac{\partial}{\partial\theta_l}\log\nu(t;\boldsymbol{\theta}_0)\right\}\nu(t;\boldsymbol{\theta}_0)dt \\
\xrightarrow{\mathbf{P}} & \sum_{i=1}^{h}p_{ci}\sigma_{ijl}\equiv\sigma_{jl}(\boldsymbol{\theta}_0) \quad \text{as } n\to\infty,
\end{aligned}
$$

where

$$
\sigma_{ijl}=\int_{J_i}\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}\left\{\frac{\partial}{\partial\theta_l}\log\nu(t;\boldsymbol{\theta}_0)\right\}\nu(t;\boldsymbol{\theta}_0)dt.
$$

That is, Condition (B) is satisfied if the following Condition (B′) holds.

(B′)

$$
\int_0^\tau\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}^2\nu(t;\boldsymbol{\theta}_0)dt<\infty \quad \text{for all } j.
$$

Similarly, the left side of (4.5) can be written as

$$
\sum_{i=1}^{h}\frac{\delta_i}{n}\int_{J_i}\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}^2 I\left(\left|\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right|>\epsilon\right)\nu(t;\boldsymbol{\theta}_0)dt \tag{4.8}
$$

and (4.8) $\xrightarrow{\mathbf{P}}0$ as $n\to\infty$ provided that (B′) is true, because

$$
\frac{\delta_i}{n}\xrightarrow{\mathbf{P}}p_{ci}
$$

and

$$
\begin{aligned}
& \int_{J_i}\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}^2 I\left(\left|\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right|>\epsilon\right)\nu(t;\boldsymbol{\theta}_0)dt \\
\le & \int_0^\tau\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}^2 I\left(\left\{\frac{\partial}{\partial\theta_j}\log\nu(t;\boldsymbol{\theta}_0)\right\}^2>n\epsilon^2\right)\nu(t;\boldsymbol{\theta}_0)dt \\
\to & \ 0 \quad \text{by (B′).}
\end{aligned}
$$

That is, Condition (B′) implies Condition (C).

To verify Condition (D), it suffices to show that, for any constant vector $\boldsymbol{c} = (c_1, ..., c_q)' \neq \mathbf{0}$, $\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c} > 0$. We have

$$\begin{aligned} \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c} &= \sum_{i=1}^{h} p_{ci} \sum_{j,l=1}^{q} c_j c_l \sigma_{ijl} \\ &= \sum_{i=1}^{h} p_{ci} \int_{J_i} \left[\sum_{j=1}^{q} c_j \frac{\partial}{\partial \theta_j} \log \nu(t; \boldsymbol{\theta}_0) \right]^2 \nu(t; \boldsymbol{\theta}_0) dt \\ &> 0, \end{aligned}$$

provided that the following Condition (D′) is satisfied.

(D′)

$$\int_0^{\tau} \left[\sum_{j=1}^{q} c_j \frac{\partial}{\partial \theta_j} \log \nu(t; \boldsymbol{\theta}_0) \right]^2 \nu(t; \boldsymbol{\theta}_0) dt > 0$$

for any $\boldsymbol{c} = (c_1, ..., c_q)' \neq \mathbf{0}$.

To verify Condition (E), first note that (4.6) can be written as

$$\sum_{i=1}^{h} \frac{\delta_i}{n} \int_{J_i} \left\{ \frac{\partial^2}{\partial \theta_j \partial \theta_l} \log \nu(t; \boldsymbol{\theta}_0) \right\}^2 \nu(t; \boldsymbol{\theta}_0) dt \quad \xrightarrow{\mathbf{P}} \quad \sum_{i=1}^{h} p_{ci} \int_{J_i} \left\{ \frac{\partial^2}{\partial \theta_j \partial \theta_l} \log \nu(t; \boldsymbol{\theta}_0) \right\}^2 \nu(t; \boldsymbol{\theta}_0) dt$$

as $n \to \infty$. Thus (4.6) can be restated as

$$\int_0^{\tau} \left\{ \frac{\partial^2}{\partial \theta_j \partial \theta_l} \log \nu(t; \boldsymbol{\theta}_0) \right\}^2 \nu(t; \boldsymbol{\theta}_0) dt < \infty.$$

Now let

$$\Gamma(t, \boldsymbol{\Theta}_0) = \max_{1 \leq j,l,m \leq q} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \left| \frac{\partial^3}{\partial \theta_j \partial \theta_l \partial \theta_m} \nu(t; \boldsymbol{\theta}) \right| \tag{4.9}$$

and

$$\Delta(t, \boldsymbol{\Theta}_0) = \max_{1 \leq j,l,m \leq q} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \left| \frac{\partial^3}{\partial \theta_j \partial \theta_l \partial \theta_m} \log \nu(t; \boldsymbol{\theta}) \right|. \tag{4.10}$$

Then

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}\left|\frac{\partial^3}{\partial\theta_j\partial\theta_l\partial\theta_m}\lambda(t;\boldsymbol{\theta})\right| \le \Gamma(t,\boldsymbol{\Theta}_0)\delta.(t) \equiv G_n(t)$$

and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0}\left|\frac{\partial^3}{\partial\theta_j\partial\theta_l\partial\theta_m}\log\lambda(t;\boldsymbol{\theta})\right| \le \Delta(t,\boldsymbol{\Theta}_0) \equiv H_n(t).$$

Thus

$$\begin{aligned} a_n^{-2}\int_0^\tau G_n(t)dt &= \sum_{i=1}^h \frac{\delta_i}{n}\int_{J_i}\Gamma(t,\boldsymbol{\Theta}_0)dt \\ &\xrightarrow{\mathbf{P}} \sum_{i=1}^h p_{ci}\int_{J_i}\Gamma(t,\boldsymbol{\Theta}_0)dt < \infty \quad \text{as } n\to\infty, \end{aligned}$$

provided that

$$\int_0^\tau \Gamma(t,\boldsymbol{\Theta}_0)dt < \infty.$$

Also

$$\begin{aligned} a_n^{-2}\int_0^\tau H_n(t)\lambda(t;\boldsymbol{\theta}_0)dt &= \sum_{i=1}^h \frac{\delta_i}{n}\int_{J_i}\Delta(t,\boldsymbol{\Theta}_0)\nu(t;\boldsymbol{\theta}_0)dt \\ &\xrightarrow{\mathbf{P}} \sum_{i=1}^h p_{ci}\int_{J_i}\Delta(t,\boldsymbol{\Theta}_0)\nu(t;\boldsymbol{\theta}_0)dt < \infty \quad \text{as } n\to\infty, \end{aligned}$$

provided that

$$\int_0^\tau \Delta(t,\boldsymbol{\Theta}_0)\nu(t;\boldsymbol{\theta}_0)dt < \infty.$$

Furthermore, for all $\epsilon > 0$, the left side of (4.7) can be written as

$$\sum_{i=1}^h \frac{\delta_i}{n}\int_{J_i}\Delta(t,\boldsymbol{\Theta}_0)I\left\{\frac{1}{\sqrt{n}}\left[\Delta(t,\boldsymbol{\Theta}_0)\right]^{1/2} > \epsilon\right\}\nu(t;\boldsymbol{\theta}_0)dt, \tag{4.11}$$

and (4.11) $\xrightarrow{\mathbf{P}} 0$ as $n \to \infty$, because $\delta_i/n \xrightarrow{\mathbf{P}} p_{ci}$ and

$$\int_{J_i} \Delta(t, \boldsymbol{\Theta}_0) I\left\{\Delta(t, \boldsymbol{\Theta}_0) > n\epsilon^2\right\} \nu(t; \boldsymbol{\theta}_0) dt \to 0$$

as $n \to \infty$, provided that $\int_0^\tau \Delta(t, \boldsymbol{\Theta}_0)\nu(t; \boldsymbol{\theta}_0)dt < \infty$.

In summary, Condition (E) can be restated as follows.

(E′) For all j, l,

$$\int_0^\tau \left\{\frac{\partial^2}{\partial\theta_j \partial\theta_l} \log \nu(t; \boldsymbol{\theta}_0)\right\}^2 \nu(t; \boldsymbol{\theta}_0) dt,$$

$$\int_0^\tau \Gamma(t, \boldsymbol{\Theta}_0) dt, \tag{4.12}$$

and

$$\int_0^\tau \Delta(t, \boldsymbol{\Theta}_0)\nu(t; \boldsymbol{\theta}_0) dt$$

are finite quantities, where $\Gamma(t, \boldsymbol{\Theta}_0)$ and $\Delta(t, \boldsymbol{\Theta}_0)$ are defined in (4.9) and (4.10), respectively.

4.4.3.1 Asymptotic Properties of the Loglinear NHPP Estimator

For the loglinear NHPP model, $\nu(t; \gamma_0, \gamma_1) = \exp(\gamma_0 + \gamma_1 t)$, and $\log \nu(t; \gamma_0, \gamma_1) = \gamma_0 + \gamma_1 t$. Let $\boldsymbol{\theta}_0 = \left(\gamma_0^{(0)}, \gamma_1^{(0)}\right)'$ be the true value of $\boldsymbol{\theta} = (\gamma_0, \gamma_1)'$. Note that for any nonnegative integers α and β,

$$\frac{\partial^{\alpha+\beta}}{\partial\gamma_0^\alpha \partial\gamma_1^\beta} \nu(t; \gamma_0, \gamma_1) = t^\beta \nu(t; \gamma_0, \gamma_1),$$

$$\frac{\partial}{\partial\gamma_0}\log\nu(t;\gamma_0,\gamma_1)=1,$$

$$\frac{\partial}{\partial\gamma_1}\log\nu(t;\gamma_0,\gamma_1)=t,$$

and

$$\frac{\partial^{\alpha+\beta}}{\partial\gamma_0^{\alpha}\partial\gamma_1^{\beta}}\log\nu(t;\gamma_0,\gamma_1)=0\quad\text{for }\alpha+\beta>1. \tag{4.13}$$

These are all continuous functions in both $(\gamma_0,\gamma_1)'$ and in $t\in[0,\tau]$. Thus, Conditions (A′) and (B′) are satisfied.

To verify (D′), note that for any $\boldsymbol{c}=(c_0,c_1)'\neq 0$,

$$\int_0^{\tau}\left[c_0\frac{\partial}{\partial\gamma_0}\log\nu(t;\boldsymbol{\theta}_0)+c_1\frac{\partial}{\partial\gamma_1}\log\nu(t;\boldsymbol{\theta}_0)\right]^2\nu(t;\boldsymbol{\theta}_0)dt=\int_0^{\tau}(c_0+c_1t)^2\exp\left(\gamma_0^{(0)}+\gamma_1^{(0)}t\right)dt>0.$$

Thus (D′) is satisfied.

To verify (E′), because of (4.13), it suffices to show that (4.12) $<\infty$. This is again straightforward because

$$\Gamma(t,\boldsymbol{\Theta}_0)\leq(1+t)^3\exp\left[\left(\left|\gamma_0^{(0)}\right|+1\right)+\left(\left|\gamma_1^{(0)}\right|+1\right)t\right],$$

where $\boldsymbol{\Theta}_0=\{(\gamma_0,\gamma_1)':\left|\gamma_0-\gamma_0^{(0)}\right|<1,\left|\gamma_1-\gamma_1^{(0)}\right|<1\}$.

4.4.3.2 Asymptotic Properties of the Power Law NHPP Estimator

To verify Conditions (A′), (B′), (D′), and (E′) for the power law NHPP model, we note that

$$\nu(t;\beta,\eta)=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}$$

and

$$\log\nu(t;\beta,\eta)=\log(\beta)-\beta\log(\eta)+(\beta-1)\log(t)$$

are smooth functions of (t, β, η) for $t, \beta, \eta > 0$ and their partial derivatives of any order exist and are continuous. Thus Condition (A′) is satisfied. Furthermore, the partial derivatives with respect to β and η are polynomial functions of $t^{\beta-1}$ and $\log t$ that are linear in $t^{\beta-1}$. Thus to verify Conditions (B′), (D′), and (E′), we only need to show that for any $0 < a < 1$, and $j \geq 0$,

$$\int_0^a t^{\beta-1} \left|\log t\right|^j dt < \infty,$$

which is indeed true because $\int_0^a t^{\beta-1} \left|\log t\right|^j dt = \int_{-\log a}^{\infty} u^j \exp(-\beta u) du < \infty$.

4.5 Concluding Remarks

In this article, for the one-dimension counting process, we have established the asymptotic properties for both the nonparametric and the NHPP estimators of the MCF from window-observation recurrence data. With the assumptions in Section 4.2.3, the NP estimator of the MCF is shown to be uniformly consistent on compact intervals, and the NHPP estimators of the MCF has consistent solutions of the likelihood equations as long as Conditions (A′), (B′), (D′) and (E′) in Section 4.4.3 are satisfied. One advantage of these restated conditions is that they are on the recurrence rate function $\nu(t; \boldsymbol{\theta})$ of the NHPP model instead of the multiplicative intensity model $\lambda(t; \boldsymbol{\theta}) = \nu(t; \boldsymbol{\theta})\delta.(t)$, and illustrations with the loglinear and the power law NHPP models are given in Sections 4.4.3.1 and 4.4.3.2. The asymptotic normality for the NP and the NHPP MCF estimators is also established when the assumptions and the conditions (needed for the NHPP MCF estimators) are satisfied.

The results in this article for univariate MCF estimators, with those conditions and theorems from Andersen et al. (1993), can be extended to the multiple-dimension counting process. Examples of multiple-dimension counting processes include numbers of recurrences for multiple failure modes from automobile repair history.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Billingsley, P. (1986), *Probability and Measure*, New York: Wiley.

Cook, R. J., and Lawless, J. F. (2007), *The Statistical Analysis of Recurrent Events*, New York: Springer-Verlag.

Ghosh, D., and Lin, D. Y. (2000), "Nonparametric Analysis of Recurrent Events and Death," *Biometrics*, 56, 554-562.

Lawless, J. F., and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.

Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: Wiley.

Nelson, W. B. (2003), *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Philadelphia: ASA-SIAM.

Peña, E. A., Strawderman, R. L., and Hollander, M. (2001), "Nonparametric Estimation with Recurrent Event Data," *Journal of the American Statistical Association*, 96, 1299-1315.

Rigdon, S. E., and Basu, A. P. (2000), *Statistical Methods for the Reliability of Repairable Systems*, New York: Wiley.

Zuo, J., Meeker, W. Q., and Wu, H. (2008), "Analysis of Window-Observation Recurrence Data," *Technometrics*, 50, 128-143.

CHAPTER 5. CONCLUSIONS

Our research extends the existing statistical methods, both nonparametric and parametric, to estimate the MCF with window-observation recurrence data.

Chapter 2 provides descriptions of four MCF estimators, that is, the NP estimator, the NHPP estimator, the local hybrid estimator, and the NHPP hybrid estimator. Besides the technical details for the MCF estimators, their variance estimators, and needed assumptions, applications of these estimators are illustrated with two examples, along with some graphical tools used in analysis with the recurrence data. Chapter 2 also offers some suggestions on selection among the MCF estimators.

Chapter 2 includes brief descriptions of the procedures to construct CIs for the MCF estimators. It is of interest to know, however, how different CI procedures perform, especially when the amount of observed information in the data is small with window-observation recurrence data. Chapter 3 provides the details and the summary results of an extensive simulation study on five CI procedures for each of the four MCF estimators described in Chapter 2, and makes suggestions on the uses of those CI procedures.

Chapter 4 establishes the asymptotic properties for the NP and NHPP MCF estimators, and provides the assumptions and conditions that are needed for the MCF estimators to be consistent and asymptotically normal.

APPENDIX A. SUPPLEMENTAL MATERIAL

A.1 Calculation of $\widehat{\mathrm{Var}}[d_{\cdot}^{\dagger}(t)]$

Recall that $d_{\cdot}^{\dagger}(t)$ is the estimator of the sum of the increases in the MCF $\mu(t)$ over all RSSZ intervals from 0 to t. Here, using the power law NHPP model, we illustrate how to use the delta method to calculate an estimate of $\mathrm{Var}[d_{\cdot}^{\dagger}(t)]$.

We start from a simple case with only one RSSZ interval before time t, say (t_{a1}, t_{b1}). Then the parametric adjustment is

$$d_{\cdot}^{\dagger}(t) = d_1^{\dagger}(t_{a1}, t_{b1}) = \int_{t_{a1}}^{t_{b1}} \hat{\nu}(x)dx = \int_{t_{a1}}^{t_{b1}} \nu(x; \hat{\beta}, \hat{\eta})dx = \left(\frac{t_{b1}}{\hat{\eta}}\right)^{\hat{\beta}} - \left(\frac{t_{a1}}{\hat{\eta}}\right)^{\hat{\beta}}.$$

Let $g_1(t; \beta, \eta) = \int_{t_{a1}}^{t_{b1}} \nu(x; \beta, \eta)dx$. Then taking first derivative with respect to the model parameters $\boldsymbol{\theta} = [\beta,\ \eta]'$, we have

$$\frac{\partial g_1(t; \beta, \eta)}{\partial \boldsymbol{\theta}} = \left[\left(\frac{t_{b1}}{\eta}\right)^{\beta} \ln\left(\frac{t_{b1}}{\eta}\right) - \left(\frac{t_{a1}}{\eta}\right)^{\beta} \ln\left(\frac{t_{a1}}{\eta}\right),\ -\frac{\beta}{\eta}\left(\left(\frac{t_{b1}}{\eta}\right)^{\beta} - \left(\frac{t_{a1}}{\eta}\right)^{\beta}\right)\right]'.$$

Applying the delta method, one can calculate an estimate for $\mathrm{Var}[d_{\cdot}^{\dagger}(t)]$ as

$$\widehat{\mathrm{Var}}[d_1^{\dagger}(t_{a1}, t_{b1})] = \left[\frac{\partial g_1(t; \beta, \eta)}{\partial \boldsymbol{\theta}}\right]'_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} \left[\frac{\partial g_1(t; \beta, \eta)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Now we extend to the more general cases, which allow more than one RSSZ intervals before time t as well as allow time t to be in one of the RSSZ intervals. In the more general cases, we first identify all the RSSZ intervals on or before time t, say $(t_{a1}, t_{b1}), (t_{a2}, t_{b2}), ..., (t_{ak}, t_{bk})$. If t is in an RSSZ interval (t_{ak}, t_{bk}), then the last interval is replaced by (t_{ak}, t). Without loss of generality, we take the last interval as (t_{ak}, t_{bk}). Then

$$d_{\cdot}^{\dagger}(t) = \sum_{i=1}^{k} d_i^{\dagger}(t_{ai}, t_{bi}),$$

where

$$d_i^{\dagger}(t_{ai}, t_{bi}) = \int_{t_{ai}}^{t_{bi}} \hat{\nu}(x)dx = \int_{t_{ai}}^{t_{bi}} \nu(x; \hat{\beta}, \hat{\eta})dx = \left(\frac{t_{bi}}{\hat{\eta}}\right)^{\hat{\beta}} - \left(\frac{t_{ai}}{\hat{\eta}}\right)^{\hat{\beta}}.$$

Let $g(t; \beta, \eta)$ be the row vector with element i given by $g_i(t; \beta, \eta) = \int_{t_{ai}}^{t_{bi}} \nu(x; \beta, \eta)dx$. Then by calculating

$$\left[\frac{\partial g(t; \beta, \eta)}{\partial \boldsymbol{\theta}}\right]'_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \widehat{\Sigma}_{\hat{\boldsymbol{\theta}}} \left[\frac{\partial g(t; \beta, \eta)}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

we obtain a $k \times k$ symmetric matrix, with the diagonal elements $\widehat{\text{Var}}[d_i^{\dagger}(t_{ai}, t_{bi})]$ and the off-diagonal elements $\widehat{\text{Cov}}[d_i^{\dagger}(t_{ai}, t_{bi}), d_j^{\dagger}(t_{aj}, t_{bj})]$, $i, j = 1, 2, ..., k$. It can be shown that the sum of all of the elements in this matrix is an estimator of $\text{Var}[d_{\cdot}^{\dagger}(t)]$.

A.2 A Conservative Estimator of $\text{Var}[\bar{d}(t_k)]$ When the Size of the Risk Set is 1

When the size of the risk set at t_k, $\delta_{\cdot}(t_k)$, is 1, the variance of $\bar{d}(t_k)$ in (2.2) is not estimable. The simple moment estimator in the following formula returns a value of zero (as in estimating the variance with a sample size of one):

$$\begin{aligned} \widehat{\text{Var}}[\bar{d}(t_k)] &= \frac{\widehat{\text{Var}}[d_1(t_k)]}{\delta_{\cdot}(t_k)} \\ &= \frac{\left\{\left[d_i(t_k) - \frac{d_i(t_k)}{\delta_{\cdot}(t_k)}\right]^2 + [\delta_{\cdot}(t_k) - 1]\left[0 - \frac{d_i(t_k)}{\delta_{\cdot}(t_k)}\right]^2\right\} / \delta_{\cdot}(t_k)}{\delta_{\cdot}(t_k)}. \end{aligned} \tag{A.1}$$

Equation (A.1) assumes that all events occur at distinct times. The term $\left[d_i(t_k) - \frac{d_i(t_k)}{\delta_{\cdot}(t_k)}\right]^2$ is for the unit with an event at time t_k, and the term $[\delta_{\cdot}(t_k) - 1]\left[0 - \frac{d_i(t_k)}{\delta_{\cdot}(t_k)}\right]^2$ is for the other $\delta_{\cdot}(t_k) - 1$ units that had no event at the time.

Equation (A.1) is a function of $\delta_{\cdot}(t_k)$, because $d_i(t_k) = 1$. It can be shown that 2 is the positive integer that maximizes $\widehat{\text{Var}}[d_1(t_k)]$ in (A.1). Therefore, for events recorded in the intervals with the size of the risk set at 1, we use (A.1) with $\delta_{\cdot}(t_k)$ at the value of 2. Then simplifying (A.1), we obtain a conservative estimator for $\text{Var}[\bar{d}(t_k)]$ as 1/8, in intervals with risk set size having a value of 1.

A.3 Method to Estimate $\widehat{\text{Cov}}\left[d_{\cdot}^{\dagger}(t), \bar{d}_{\cdot}(t)\right]$ for the NHPP Hybrid Estimator

This section outlines the method we use to estimate $\text{Cov}\left[d_{\cdot}^{\dagger}(t), \bar{d}_{\cdot}(t)\right]$ for the NHPP hybrid estimator, using the power law NHPP model as an example. Note that $d_{\cdot}^{\dagger}(t)$ is a function of the model parameter estimators $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\eta})'$ and can be expressed as

$$d_{\cdot}^{\dagger}(t) = \begin{cases} \sum_{m:t_{mU} \le t} d_m^{\dagger} & \text{if } t \text{ is not in an RSSZ interval} \\ \sum_{m:t_{mU} \le t} d_m^{\dagger} + \int_{t_{wL}}^{t} \hat{\nu}(x)dx & \text{if } t \text{ is in the } w\text{th RSSZ interval } (t_{wL} < t \le t_{wU}), \end{cases}$$

where

$$d_m^{\dagger} \;=\; \int_{t_{mL}}^{t_{mU}} \hat{\nu}(t)dt = \hat{\eta}^{-\hat{\beta}}(t_{mU}^{\hat{\beta}} - t_{mL}^{\hat{\beta}}).$$

Also, $\bar{d}_{\cdot}(t)$ is a function of recurrence times t_{ij} and the size of the risk set at the recurrence time $\delta_{\cdot}(t_{ij})$ and can be expressed as

$$\bar{d}_{\cdot}(t) \;=\; \sum_{i=1}^{n} \left[\sum_{j=1}^{r_i} \frac{I(t_{ij} < t)}{\delta_{\cdot}(t_{ij})} \right], \tag{A.2}$$

where $\delta_{\cdot}(t_{ij}) = \#\left\{k : 1 \le k \le n, t_{ij} \in \cup_{l=1}^{p_k}(t_{klL}, t_{klU}]\right\}$ is the size of the risk set at t_{ij}. Note that (A.2) is a different (but numerically equivalent) representation of $\widehat{\mu}_{\text{NP}}$ in (2.2), but for the convenience of calculations below uses the same notation as that for $\widehat{\mu}_{\text{NHPP}}$.

Let $w = \{\max(m) : 1 \le m \le q, t \ge t_{mL}\}$. Then by estimating the parts of $\text{Cov}\left[d_m^{\dagger}, \bar{d}_{\cdot}(t)\right]$, $m = 1, ..., w$, we will be able to estimate $\text{Cov}\left[d_{\cdot}^{\dagger}(t), \bar{d}_{\cdot}(t)\right]$. Therefore, from now on, we focus on the estimation of $\text{Cov}\left[d_m^{\dagger}, \bar{d}_{\cdot}(t)\right]$. Let

$$\mu_m^{\dagger} \;=\; \int_{t_{mL}}^{t_{mU}} \nu(t)dt = \eta^{-\beta}(t_{mU}^{\beta} - t_{mL}^{\beta}).$$

Then by the delta method, we have

$$\text{Cov}\left[d_m^\dagger, \bar{d}_\cdot(t)\right] \approx \frac{\partial \mu_m^\dagger}{\partial \beta}\text{Cov}\left[\hat{\beta}, \bar{d}_\cdot(t)\right] + \frac{\partial \mu_m^\dagger}{\partial \eta}\text{Cov}\left[\hat{\eta}, \bar{d}_\cdot(t)\right]$$

and an estimator is

$$\begin{aligned}\widehat{\text{Cov}}\left[d_m^\dagger, \bar{d}_\cdot(t)\right] &= \frac{\partial \mu_m^\dagger}{\partial \beta}_{|_{\eta=\hat{\eta}}^{\beta=\hat{\beta}}}\widehat{\text{Cov}}\left[\hat{\beta}, \bar{d}_\cdot(t)\right] + \frac{\partial \mu_m^\dagger}{\partial \eta}_{|_{\eta=\hat{\eta}}^{\beta=\hat{\beta}}}\widehat{\text{Cov}}\left[\hat{\eta}, \bar{d}_\cdot(t)\right] \\ &= \left\{\hat{\eta}^{-\hat{\beta}}\left[t_{mU}^{\hat{\beta}}\log(t_{mU}) - t_{mL}^{\hat{\beta}}\log(t_{mL})\right] - \hat{\eta}^{-\hat{\beta}}\log(\hat{\eta})(t_{mU}^{\hat{\beta}} - t_{mL}^{\hat{\beta}})\right\}\widehat{\text{Cov}}\left[\hat{\beta}, \bar{d}_\cdot(t)\right] \\ &\quad -\hat{\beta}\hat{\eta}^{-(\hat{\beta}+1)}(t_{mU}^{\hat{\beta}} - t_{mL}^{\hat{\beta}})\widehat{\text{Cov}}\left[\hat{\eta}, \bar{d}_\cdot(t)\right]. \end{aligned} \tag{A.3}$$

To compute $\widehat{\text{Cov}}\left[\hat{\beta}, \bar{d}_\cdot(t)\right]$ and $\widehat{\text{Cov}}\left[\hat{\eta}, \bar{d}_\cdot(t)\right]$ in (A.3), we define a set of estimating equations. The log likelihood function for the power law NHPP model is

$$\ell(\beta, \eta) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{r_i}\left[\log(\beta) - \beta\log(\eta) + (\beta-1)\log(t_{ij})\right] - \eta^{-\beta}\sum_{k=1}^{p_i}(t_{ikU}^{\beta} - t_{ikL}^{\beta})\right\}.$$

Simplifying the likelihood equations $\frac{\partial \ell(\beta,\eta)}{\partial \beta} = 0$ and $\frac{\partial \ell(\beta,\eta)}{\partial \eta} = 0$, we obtain the following estimating equations for β and η

$$F_1(\beta, \eta) = \frac{1}{n}\sum_{i=1}^{n} g_{i1}(\beta, \eta) = 0 \tag{A.4}$$

$$F_2(\beta, \eta) = \frac{1}{n}\sum_{i=1}^{n} g_{i2}(\beta, \eta) = 0 \tag{A.5}$$

where

$$\begin{aligned} g_{i1}(\beta, \eta) &= r_i + \beta\sum_{j=1}^{r_i}\log(t_{ij}) - \beta\eta^{-\beta}\sum_{k=1}^{p_i}\left[t_{ikU}^{\beta}\log(t_{ikU}) - t_{ikL}^{\beta}\log(t_{ikL})\right] \\ g_{i2}(\beta, \eta) &= r_i - \eta^{-\beta}\sum_{k=1}^{p_i}(t_{ikU}^{\beta} - t_{ikL}^{\beta}). \end{aligned}$$

By the delta method, we have

$$\begin{pmatrix} \widehat{\text{Cov}}\left[\hat{\beta}, \bar{d}_\cdot(t)\right] \\ \widehat{\text{Cov}}\left[\hat{\eta}, \bar{d}_\cdot(t)\right] \end{pmatrix} = -\begin{pmatrix} \frac{\partial F_1}{\partial \beta} & \frac{\partial F_1}{\partial \eta} \\ \frac{\partial F_2}{\partial \beta} & \frac{\partial F_2}{\partial \eta} \end{pmatrix}^{-1}_{|_{\eta=\hat{\eta}}^{\beta=\hat{\beta}}} \begin{pmatrix} \text{Cov}\left[F_1, \bar{d}_\cdot(t)\right] \\ \text{Cov}\left[F_2, \bar{d}_\cdot(t)\right] \end{pmatrix}_{|_{\eta=\hat{\eta}}^{\beta=\hat{\beta}}} \tag{A.6}$$

and using (A.4) and (A.5), it is easy to calculate the elements in the matrix on the right side of (A.6). For the vector on the right side of (A.6), because recurrences among units in the data are independent from each other, we have

$$\mathrm{Cov}\left[F_l, \bar{d}_\cdot(t)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}\left[g_{il}(\beta,\eta), \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right], \quad l=1,2.$$

More specifically,

$$\begin{aligned}
\mathrm{Cov}\left[g_{i1}, \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] &= \mathrm{Cov}\left[r_i + \beta\sum_{j=1}^{r_i}\log(t_{ij}), \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] \\
&= \mathrm{Cov}\left[r_i, \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] + \beta\mathrm{Cov}\left[\sum_{j=1}^{r_i}\log(t_{ij}), \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] \\
\mathrm{Cov}\left[g_{i2}, \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] &= \mathrm{Cov}\left[r_i, \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right]
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{Cov}\left[r_i, \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] &= \mathrm{E}(r_i)\mathrm{E}\left[\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\middle| r_i\right] \\
&= \beta\eta^{-\beta}\sum_{k=1}^{p_i}\int_{t_{ikL}}^{t_{ikU}} x^{\beta-1}\frac{I(x<t)}{\delta_\cdot(x)}dx \qquad \text{(A.7)}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}\left[\sum_{j=1}^{r_i}\log(t_{ij}), \sum_{j=1}^{r_i}\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\right] &= \mathrm{E}(r_i)\mathrm{E}\left[\log(t_{ij})\frac{I(t_{ij}<t)}{\delta_\cdot(t_{ij})}\middle| r_i\right] \\
&= \beta\eta^{-\beta}\sum_{k=1}^{p_i}\int_{t_{ikL}}^{t_{ikU}} x^{\beta-1}\log(x)\frac{I(x<t)}{\delta_\cdot(x)}dx. \qquad \text{(A.8)}
\end{aligned}$$

The following are results used in the derivation of (A.7) and (A.8).

- For unit i, $r_i \sim POI(\sum_{k=1}^{p_i}\int_{t_{ikL}}^{t_{ikU}}\nu(t)dt)$, and thus

$$\mathrm{E}(r_i) = \mathrm{Var}(r_i) = \sum_{k=1}^{p_i}\eta^{-\beta}(t_{ikU}^{\beta} - t_{ikL}^{\beta}),$$

- Conditional on a particular value of r_i, the $t_{i1}, ..., t_{ir_i}$ are identically and independently distributed with the following density function:

$$f_i(t) = \begin{cases} \frac{\beta\eta^{-\beta}t^{\beta-1}}{\sum_{k=1}^{p_i}\eta^{-\beta}(t_{ikU}^{\beta}-t_{ikL}^{\beta})} & t \in \cup_{k=1}^{p_i}(t_{ikL}, t_{ikU}] \\ 0 & \texttt{otherwise} \end{cases}$$

  This follows from a straightforward generalization of the result on page 59 and 60 of Rigdon and Basu (2000), who give a similar result for units that have only one observation window starting at time 0.

A.4 Properties of the Complete Data Simulated from the Power Law NHPP Model

A.4.1 Distribution of the Number of Recurrences

Let X_1, X_2, ..., X_n be the number of observed recurrences for the n units from a Complete data set (i.e., no gaps), simulated from a power law NHPP model with a given value of $\mathrm{E}(r)$. Then X_1, X_2, ..., X_n are independent and identically distributed (*iid*) from a Poisson $(\mathrm{E}(r)/n)$ distribution. From this, the probability mass function (pmf) of (X_1, X_2, ..., X_n) is

$$\Pr\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n\right) = \prod_{i=1}^{n}\left\{\exp\left(-\mathrm{E}(r)/n\right) \times \frac{[\mathrm{E}(r)/n]^{x_i}}{x_i!}\right\}. \tag{A.9}$$

Because the sum of independent Poisson random variables is also a Poisson random variable, $\sum_{i=1}^{n} X_i$ follows a Poisson distribution with $\lambda = \sum_{i=1}^{n}\lambda_i = \mathrm{E}(r)$, where λ_i is the parameter for X_i. Therefore, the probability of a simulated data set with x recurrences is

$$\Pr\left(\sum_{i=1}^{n} X_i = x\right) = \exp\left(-\mathrm{E}(r)\right) \times \frac{[\mathrm{E}(r)]^{x}}{x!}, \tag{A.10}$$

which is a function of $\mathrm{E}(r)$ and x only. For the four values of $\mathrm{E}(r)$ in our simulation study, 10, 20, 50, and 100, the values of $\Pr\left(\sum_{i=1}^{n} X_i \leq 4\right)$ are 0.0293, 1.69×10^{-5}, 5.45×10^{-17}, and 1.61×10^{-37}, respectively.

Note that (A.9) and (A.10) depend only on E(r) and n, and thus results in this subsection apply not only to the power law NHPP model, but also to NHPP model with other forms of the recurrence functions.

A.4.2 NP MCF Estimator

One interesting property of the NP method for the Complete data case is that both $\widehat{MCF}_{NP}(t_{endobs})$ and $\widehat{\text{Var}}[\widehat{MCF}_{NP}(t_{endobs})]$, which are needed to construct the normal-approximation-based CIs, are functions of X_1, X_2, ..., X_n and n, and do not depend on the observed times of the recurrences. Then, for a given set of $(x_1, x_2, ..., x_n)$, we can construct the NORMA and the LNORMA CIs and find out whether the CIs capture the true MCF at E(r)/n. By (A.9), we can also calculate $\Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$. Therefore, if we can identify that part of the sample space of $(x_1, x_2, ..., x_n)$ for which the constructed CIs capture the true MCF, and then the sum of the corresponding probabilities is the actual CP of the CI procedure. Because x_i can take the value of any non-negative integer, for $i = 1, ..., n$, the complete sample space for $(x_1, x_2, ..., x_n)$ is countable, but infinite. As a result, it is not possible to identify the complete set of $(x_1, x_2, ..., x_n)$ that we want to identify to calculate the actual CP. However, a close approximation of the actual CP can be obtained by using the following approach.

1. Select a finite number of combinations of $(x_1, x_2, ..., x_n)$ from the complete sample space such that the sum of the corresponding probabilities is close to 1.

2. For each set of values of $(x_1, x_2, ..., x_n)$ in the selected sample space, calculate $\widehat{MCF}_{NP}(t_{endobs})$ and $\widehat{\text{Var}}[\widehat{MCF}_{NP}(t_{endobs})]$, and construct a CI. If the CI captures the true MCF, then add the corresponding probability of the set to CP.

The complement of the selected sample space provides a bound on the error for the actual CP. Because the probability of getting large values of x_i is very small, the sample space we chose is $[0, k] \times [0, k] \times ... \times [0, k]$, where k is an integer large enough to keep the error bound small.

Table A.1 Probability in CI and Error Bound for Complete Data at $n = 10$

	NP Estimator			Power Law NHPP Estimator		
$E(r)/n$	NA	LOGNA	Error Bound	NA	LOGNA	Error Bound
1	0.89415	0.91038	1.11×10^{-6}	0.92573	0.96262	7.98×10^{-8}
2	0.89250	0.90384	2.07×10^{-6}	0.94751	0.94428	4.83×10^{-9}
5	0.89983	0.90550	1.40×10^{-5}	0.94878	0.94440	1.57×10^{-10}
10	0.90224	0.90510	7.65×10^{-6}	0.94503	0.94912	7.08×10^{-11}

We now derive the expressions for $\widehat{MCF}_{NP}(t_{endobs})$ and $\widehat{\text{Var}}[\widehat{MCF}_{NP}(t_{endobs})]$. For Complete data, having a constant risk set size of n, (3.1) leads to $\widehat{MCF}_{NP}(t_{endobs}) = \sum_{i=1}^{n} X_i/n$. Lawless and Nadeau (1995) present the following formula (here with our notation), for data with all units having the same end-of-observation times,

$$\widehat{\text{Var}}[\widehat{MCF}_{NP}(t_{endobs})] = \frac{1}{n^2} \sum_{i=1}^{n} \left[X_i - \frac{\sum_{i=1}^{n} X_i}{n} \right]^2 .$$

Table A.1 shows results for the approximate CP, which is the sum of the probabilities for the sets of $(x_1, x_2, ..., x_n)$ that have the calculated CI capturing the true MCF, as well as the error bound, for the NP estimator with $n = 10$. Two CI procedures are used, the NORMA and the LNORMA. The approximate probability of not capturing the true MCF equal (1 − Prob. in CI − error bound). Note that, the nominal coverage probability is 0.95, yet the probabilities with true MCF in the CI for the two procedures are well below this nominal value. This shows that the normal approximation procedures do not work well for the NP estimator when the sample size is small, and increasing the expected number of observed recurrences for each unit does not help improve the CP. The LNORMA CI procedure performs better than the NORMA procedure, because the number of recurrences is non-negative.

A.4.3 The Power Law NHPP Estimator

Rigdon and Basu (2000, Section 5.4) present the likelihood function of n independent systems from the same power law NHPP model, and point out that when all n systems are observed from time zero to the same end-of-observation time t_{endobs}, the Complete data scenario in our simulations, there is an explicit solution for estimating the two model parameters,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n}\sum_{j=1}^{X_i} \log(t_{endobs}/t_{ij})} \tag{A.11}$$

$$\hat{\eta} = \frac{n^{1/\hat{\beta}} t_{endobs}}{(\sum_{i=1}^{n} X_i)^{1/\hat{\beta}}}, \tag{A.12}$$

where t_{ij} denotes the j^{th} recurrence time for the i^{th} system. With the power law recurrence rate function (3.3), (3.2) can be simplified to

$$\widehat{MCF}_{NHPP}(t_{endobs}) = (t_{endobs}/\hat{\eta})^{\hat{\beta}} = \sum_{i=1}^{n} X_i/n.$$

This shows that $\widehat{MCF}_{NHPP}(t_{endobs}) = \widehat{MCF}_{NP}(t_{endobs})$ for the Complete data case and these estimators only depend on the number of observed recurrences and the number of observational units.

It can also be shown that,

$$\widehat{\text{Var}}[\widehat{MCF}_{NHPP}(t_{endobs})] = \sum_{i=1}^{n} X_i/n^2 = \widehat{MCF}_{NHPP}(t_{endobs})/n.$$

This result is not surprising, because for the Poisson distribution, the variance and the mean are the same, and here we have n units to estimate the variance. However, this simple form is only available for the Complete data case when all units are observed to t_{endobs}. The main steps to derive this result are listed below.

1. Derive the Hessian matrix by taking second derivatives of the likelihood function with β and η.

2. Obtain variance-covariance matrix of $\hat{\beta}$ and $\hat{\eta}$ by evaluating the inverse negative Hessian matrix at the MLEs $\hat{\beta}$ and $\hat{\eta}$.

3. Apply the delta method to get $\widehat{\text{Var}}[\widehat{MCF}_{NHPP}(t_{endobs})]$.

We used the same approach described in A.4.2 for the NP estimator, and obtained the approximate CP values and the error bound for the power law NHPP estimator in Table

A.1. Because both $\widehat{MCF}_{NHPP}(t_{endobs})$ and $\widehat{\text{Var}}[\widehat{MCF}_{NHPP}(t_{endobs})]$ depend only on the total number of observed recurrences and the number of units in the data, we used (A.10) instead of (A.9) to simplify the calculation. Compared to the NP estimator, the NORMA and the LNORMA procedures for the power law NHPP estimator have CPs that are very close to the nominal value at 0.95.