

ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Jing Qin

Biased Sampling, Over-identified Parameter Problems and Beyond



ICSA Book Series in Statistics

Series editors

Jiahua Chen, Department of Statistics, University of British Columbia, Vancouver, Canada

Ding-Geng (Din) Chen, University of North Carolina, Chapel Hill, NC, USA

More information about this series at <http://www.springer.com/series/13402>

Jing Qin

Biased Sampling, Over-identified Parameter Problems and Beyond



Springer

Jing Qin
Biostatistics Research Branch
National Institute of Allergy and Infectious
Diseases
Bethesda, MD
USA

ISSN 2199-0980
ICSA Book Series in Statistics
ISBN 978-981-10-4854-8
DOI 10.1007/978-981-10-4856-2

ISSN 2199-0999 (electronic)
ISBN 978-981-10-4856-2 (eBook)

Library of Congress Control Number: 2017939898

© Springer Nature Singapore Pte Ltd. 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

When I was a graduate student more than twenty five years ago, I was struggling to read many statistical research papers. This is particularly true at the time when I had passed my Ph.D. qualification examination. The goal of this book is to make it easier for Ph.D. students and new researchers to embark in their research area. During the past 30 years, statistics has become more an applied and more diversified science. In response to this trend, I have tried to cover as many different topics as possible. My main research interest focuses on likelihood-based inferences, which includes parametric likelihood, biased sampling likelihood, semi-parametric likelihood, empirical likelihood and Godambe's estimating function theory.

This book is devoted to biased sampling problems (also called choice-based sampling in econometric parlance) and over-identified parameter estimation problems. When a proper randomization cannot be achieved, the observed sample will not be representative of the population of interest. This biased sampling problem appears frequently since in the real world, truly random sampling is not easily achievable or practically feasible. Biased sampling problems appear in many areas of research, including medicine, epidemiology and public health, social sciences and economics. As pointed out by Prof. James Heckman (1979), the 2000 Nobel Laureate in Economics, "Sample selection bias may arise in practice for two reasons. First, there may be self selection by the individuals or data units being investigated. Second, sample selection decisions by analysts or data processors operate in much the same fashion as self selection". This book would be of interest to those who work in the health, biological, social and physical sciences, as well as those who are interested in survey methodology and other areas of statistical science, among others.

Due to its convenience and cost effectiveness, one of the most efficient designs in health sciences research is the case-control design. Under this design, individuals (called cases) with the condition of interest (for example, cancer) are sampled. Their risk profiles for the condition are collected. Then some controls (do not satisfy the condition of interest, for example cancer free) are enrolled along with their risk profiles are also recorded. Since the numbers of cases and controls are fixed by

researchers, the ratio of cases and controls does not necessarily match the one in the entire target population. Consequently, this selection bias can lead to a result that is different from what we would have gotten if we had enrolled the entire population. In missing data problems, the likelihood function based on only those individuals with complete data is a biased version of the targeted one. In capture and recapture studies, each animal may be captured with a probability proportional to its size. In observational casual studies, the choice of treatment and control depends on the baseline covariates, which may lead to biased results if simply using the two sample comparison methods. In finite population problems the proportional to size design is very popular in survey, which is a special type of biased sampling. Meanwhile the length-biased sampling is one of the most naturally occurred types of biased sampling. Length-biased data are clearly encountered in applications of renewal processes, etiologic studies, genome-wide linkage studies, epidemiologic cohort studies, cancer prevention trials, and studies of labor economy. In observational studies, a prevalent cohort design that draws samples from individuals with a condition or disease at the time of enrollment is generally more efficient and practical. The recruited patients who have already experienced an initiating event are followed prospectively for the failure event (e.g. disease progression or death) or are right censored. Under this sampling design, individuals with longer survival times measured from the onset of the disease are more likely to be included in the cohort, whereas those with shorter survival times are unconsciously excluded. Length-biased sample thereby manifests in various data sets, because the “observed” time intervals from initiation to failure within the prevalent cohort tend to be longer than those arising from the underlying distribution of the general population. Finding appropriate adjustments for the potential selection bias in analyzing length-biased data or more general biased sampling problems has been a long-standing statistical problem. Although we use a prevalent cohort study in medical applications here to illustrate length-biased data, it is apparent that similar issues caused by biased sampling are common in many potential applications and sampling designs. For example, biased sampling problem occurs frequently in stereology and estimating dark matter distribution in astronomy. Stereology is the study of three-dimensional properties of objects or matter usually observed two-dimensionally.

It is worth mentioning that biased sampling problems may occur even if the sampling is unbiased. In fact if we model only the density ratios but leave the baseline density arbitrary in multiple sample problems, then we end up with a biased sampling problem since those populations other than the baseline one can be treated as a biased version of the baseline population, where the selection bias functions are the density ratios. Therefore methods developed in biased sampling problems can be borrowed to make robust inference instead of the full parametric approach in multiple sample problems.

When a model is defined through more estimating functions than the free parameters, it becomes an over-identified parameter problem. This problem occurs naturally if there exists auxiliary information. For example, in survey sampling, summarized information is available from published reports. Meta analysis is an

exciting area to combine similar studies to achieve a more precise analysis. An effective statistical method is to synthesize estimating functions. The advantage of using estimating function approach over the full parametric likelihood is that one does not need to specify the full parametric model. Therefore this approach is robust against possible model mis-specifications. Moreover the likelihood-based approach may be very cumbersome in some complex setup, such as in finance dependent data and time series data problems. In statistical and econometric literature there are a few efficient methods to combine over-identified estimating functions, among them, generalized method of moments (GMM), empirical likelihood method and estimating function theory are the most popular ones.

The importance of biased sampling problem and over-identified parameter problem has been well recognized in statistics and econometrics. Among many other important contributions, econometricians Heckman and McFadden (2000) were awarded the Nobel prizes in economics for their fundamental contribution to selection sampling problems. Heckman was cited “for the development of theory and methods for analyzing selective samples,” and McFadden was cited “for his development of theory and methods for analyzing discrete choice”. The econometrician Hansen (2013) won the same prize for his contribution on the ground break research on GMM. In the introduction of his award, it was cited for “....His seminal paper, ‘Large Sample Properties of Generalized-Methods of Moments Estimators’, (1982 *Econometrica*) importantly altered the landscape for how empirical research is done in finance and macroeconomics”.

Breslow (2003), one of the leading bio-statisticians in the world, argued that statisticians and epidemiologists have made similar contributions to medicine with their work on case-control studies, analysis of incomplete data and causal inference. In spite of repeated nominations of such eminent figures, the Nobel Prize in physiology and medicine has been never awarded for work in biostatistics or epidemiology. Nevertheless, this indicates that the biased sampling and over-identified parameter (GMM) problems are fundamentally important and have wide applications in different research disciplines.

In addition to biased sampling and over-identified parameter problems, I will discuss some popular models and methods in econometrics, epidemiology, statistics, and biostatistics, including, among others, the Manski’s (1975) maximum rank score, Han’s (1986) maximum rank correlation estimation, Cosslett’s (1983) maximum likelihood estimator of the binary choice model under monotonic constraints, Godambe’s optimal estimating function theory, Owen’s (1988) empirical likelihood, Kullback–Leibler likelihood and entropy family, semiparametric genetic mixture models, Kou and Ying’s (1997) i.i.d. representation of a non-central hypergeometric distribution, causal inference and missing data, Vardi (1985)’s multiplicative censoring model, multi-sample Wicksell corpuscle problem, capture and recapture models, case and control problems with prevalent cases, and inference under monotonic function constraints using a combination of the EM algorithm and pool adjacent violation algorithm.

In this book I will give a brief overview of parametric likelihood inference and survival analysis, which makes easier for graduate students to understand the recent

developments of nonparametric or semiparametric methods and survival analysis based on truncated data, length-biased sampling data or backward time data in prevalent cohort studies. There are many good textbooks on parametric inference, among others, for example, Cox and Hinkley (1974), Lehmann and Casella (1998), Lehmann and Romano (2005), Caselle and Berger (2008) and Shao (2003). Excellent survival analysis books, include, among others, Cox and Oakes (1984), Kalbfleisch and Prentice (2002), Lawless (2002), Lancaster (1992), Andersen, Borgan, Gill and Keiding (1993) and Fleming and Harrington (1994). Since this book is aimed towards graduate students from different areas, the emphasis of this book is on statistical reasoning but not on the detailed mathematical derivations. I have tried to make the theories as assessable as possible. The detailed more advanced mathematical theories such as modern empirical process theory and information bound calculations can be found, among others, such as Bickel, Klassen, Ritov and Wellner (1993), Van der Vaart and Wellner (2003), Kosorok (2008a, b) and Tsiatis (2006). Since biased sampling problems have a natural connection with the sampling probability proportional to size problems in finite population, we recommend excellent survey sampling books by Cochran (1977), Sarndal, Swensson and Wretman (1991) and Thompson (1997) to readers. Other than those statistical books, some outstanding econometric reference books include, among others, Amemiya (1985), Gourieroux and Monfort (1990) and Lancaster (1990). Klugman, Panjer and Willmot's (2004) book provides in-depth coverage of modeling techniques used throughout many branches of actuarial science. Sham's (1998) "Statistics in Human Genetics" is an excellent statistical genetic reference book to new researchers in human disease genetics.

For young researchers, finally, I hope this book can play a role called "Pao zhuan yinyu" in Chinese, which means "cast a brick to attract jade", or, to offer a few commonplace remarks by way of introduction so that others may come up with valuable opinions. It would be the greatest encouragement to me if students can apply some methods discussed in this book to solve their own practical problems.

Rockville, USA

Jing Qin

Acknowledgements

It is impossible for me to write this book without the encouragements and supports from a large number of people. First I would like to thank faculty members, past and present, at the Department of Statistics and Actuarial Science, University of Waterloo, where I received a solid training in the statistical theory. These individuals include Profs. V.P. Godambe, J. Chen, J. Kalbfleisch, J. Lawless, M. Thompson, D. Sprott, C. Small, D. McLeish, D. Matthews, V. Farewell, S. Brown, Jeff Wu, C.B. Wu, G. Yi, P.F. Li and many others. I have learned a lot from Prof. A.B. Owen on his empirical likelihood method. My collaborators over the years have helped me a lot to formulate different statistical problems, including Drs. J. Chen, D. Follmann, B. Zhang, D.H. Leung, C.Y. Huang, Y. Shen, J. Ning, J. Sun, H. Zhou, J. Shao, P.F. Li, Y.K. Liu, A. Yuan, Z.H. Hu, G.Q. Diao, R. Pfeiffer, B.J. Chen, J. Cheng, G. Chan, Z. Guan, M.C. Wang, S.X. Chen, G. Li, B. Kedem, D. Small and many others. Profs. Jixiang Zhou, Shisong Mao and Jinglong Wang at East China Normal University helped me a lot when I pursued my Master degree in Statistics 30 years ago.

Dr. Denis Leung deserves special thanks for smoothing English and proofreading the entire book. Drs. Jiahua Chen, C.-Y. Huang, Jing Ning, Baojiang Chen, Zhong Guan, Pengfei Li, and Hao Liu have also read some chapters and given corrections. I take responsibility for all remaining errors. I also would like to thank my colleagues at the National Institute of Allergy and Infectious Diseases. We have a very harmonious and supportive environment. Particularly I would like to thank Drs. M. Proschan for his help with latex problems, M. Fay for his help on R programs and M. Gezmu, Z.H. Hu and D. Follmann for their help over the past years.

I must apologize to individuals who inadvertently have been left off in my acknowledgement list. Due to page limitation, many important related references cannot be cited in this book. And to those authors: Please accept my sincere apology.

Finally, I thank my parents, my wife, Wen, and children, Nancy and Richard, for their patience and tolerance over the years during my endeavor on writing this book. It is the greatest regret that my father passed away three years ago and could not read this.

Contents

1 Examples and Basic Theories for Length Biased Sampling	
Problems	1
1.1 Length Biased Sampling Examples	2
1.2 Basic Properties of Length Biased Sampling Problems	5
1.3 Stochastic Ordering	5
1.4 Lorenz Curve	8
1.5 Characterization of Length Biased Distribution	9
2 Brief Introduction of Renewal Process	11
2.1 Basic Concepts	11
2.2 Forward and Backward Recurrence Times	14
2.3 Basic Results on Poisson Process	18
3 Heuristical Introduction of General Biased Sampling with Various Applications	23
3.1 Natural Selection Biased Sampling Problems	23
3.2 Modelling Based Selection Biased Sampling Problems	40
4 Brief Review of Parametric Likelihood Inferences	49
4.1 Kullback–Leibler Information and Entropy Concepts	50
4.2 Issues in Maximum Likelihood Estimation	53
4.3 Popular Inference Methods in the Presence of Nuisance Parameters	65
4.4 Quasi-likelihood Methods in Linear Regression Models	69
4.5 Composite Likelihoods and Corrected Likelihoods	72
4.6 Variable Selection and Akaike Criterion	76
4.7 Two Useful Maximization Algorithms	78
4.8 Likelihood Based Inference with Inequality Constraints	82

5 Optimal Estimating Function Theory	85
5.1 Godambe's Optimality Criterion	86
5.2 Applications of Godambe's Theory in Missing Covariate Problems	90
5.3 Godambe's Theory in Length Biased Sampling	
AFT Models	93
5.4 Ancillarity and Fisher Information with Nuisance Parameters	98
5.5 Projection Methods in Parametric Models	102
5.6 Reduce Sensitivity with Respect to Nuisance Parameters	106
6 Projection Methods in General Semiparametric Models	111
6.1 Projection Method for the Mean Estimation and Linear Regression Model	112
6.2 Information Contained in the Conditional Expectation Model	115
6.3 Projection Method in a Two Sample Density Ratio Model	119
6.4 Information Calculation in Over-identified Semiparametric Models	121
6.5 Information Calculation for Missing Data Problems	123
6.6 A Non-root n Consistent Estimator Example	126
7 Generalized Method of Moments	129
7.1 Basic Concepts on Generalized Method of Moments	129
7.2 An Optimal Result Based on an Embed Exponential Family	133
7.3 Applications of GMM	136
8 Empirical Likelihood with Applications	139
8.1 Definition of Empirical Likelihood and Basic Properties	140
8.2 General Theory of Empirical Likelihood in Estimating Equations	143
8.3 Miscellaneous Topics on Empirical Likelihood	153
8.4 Hybrid Likelihoods and Utilization Auxiliary Information	159
8.5 Combine Summarized Information: A More Flexible Method in Meta Analysis	167
9 Kullback–Leibler Likelihood and Entropy Family	171
9.1 Minimize Kullback–Leibler Divergence Subject to Moment Constraints	171
9.2 Entropy Family in the Presence of Covariates	176
9.3 Some Miscellaneous Results	177
9.4 Entropy Family with Fixed Margins in Discrete Case	180
9.5 Generalized Empirical Likelihoods	185
9.6 Inference for Exponential Family with Specified Mean Function	187

10 General Theory on Biased Sampling Problems	191
10.1 Maximum Likelihood Estimation for Length Biased Sampling Problems	191
10.2 Maximum Likelihood Estimation for Multiple Biased Sampling Problems	196
10.3 Weight Function Depends on the Underlying Distribution Problems	203
11 General Theory for Case-Control Studies	207
11.1 Semiparametric Inference for Logistic Regression Analysis Based on Case-Control Data	207
11.2 Optimality of the Maximum Semiparametric Likelihood Estimating Equations	216
11.3 Different Semiparametric MLE Methods for Case-Control Data	219
11.4 Family-Based Case-Control Studies	226
11.5 Genetic Liability Model or Probit Model	232
11.6 Miscellaneous Problems	234
12 Conditioning Approach for Discrete Outcome Problems	241
12.1 Eliminate Nuisance Parameters in Logistic Regression Models	241
12.2 Matched Case Control Study	243
12.3 Matched Case and Control Sampling for Survival Analysis	246
13 Discrete Data Models	249
13.1 Regression and Ordered Categorical Variables	249
13.2 Using Continuous Distributions to Construct Discrete Choice Models	252
14 Gene and Environment Independence and Secondary Outcome Analysis in Case-Control Study	259
14.1 Score Test for Gene and Environment Independence	260
14.2 Inference Under the Assumption of Independence Between Gene and Environment	264
14.3 Secondary Outcome Analysis	267
14.4 Use Covariate Specific Disease Prevalent Information	272
14.5 Case-Control Study for Haplotype Data	277
15 Outcome Dependent Sampling and Maximum Rank Estimation	281
15.1 Linear Regression for Outcome Dependent Sampling	282
15.2 Semiparametric Approach	286
15.3 Manski's Maximum Rank Score Method	288
15.4 Han's Maximum Rank Correlation Estimation	289

15.5	Triple and Quadruple Wise Rank Likelihood	291
15.6	Retrospective Sampling and Maximum Rank Estimation	292
16	Noncentral Hypergeometric Distribution and Poisson Binomial Distribution	297
16.1	I.I.D. Representation of the Hypergeometric Distribution	297
16.2	Inferences for Poisson Binomial Distributions	303
17	Inferences and Tests in Semiparametric Finite Mixture Models	307
17.1	Examples for Hypothesis Test in Semiparametric Mixture Models	307
17.2	A New Score Test Statistic	311
17.3	Inference in Three Samples Semiparametric Mixture Models	316
17.4	Inference in Upgraded Mixture Models	322
18	Connections Among Marginal Likelihood, Conditional Likelihood and Empirical Likelihood	331
18.1	Unodered Pairs	332
18.2	Two-Component Mixture Model with Multiple Samples	338
18.3	Genetic Linkage Mixture Models	341
18.4	Shannon’s Mutual Information for Nuisance Parameter Elimination	344
19	Causal Inference and Missing Data Problems	353
19.1	Definition of Three Types of Missing Data and Basic Concepts	354
19.2	Some Existing Methods in Casual Inferences	358
19.3	Inference for Average Treatment Effect for Treated	373
19.4	Regression Generalizations	379
19.5	Projection Methods in Missing Data Problems	382
19.6	Optimal Estimating Function Based Inferences	389
19.7	Parameter Estimation for the “Working Regression Model”	394
19.8	Instrument Variable Approach in Casual Inferences	396
20	Inference in Finite Populations	409
20.1	Basic Concepts in Finite Sampling	409
20.2	Poisson and Binomial Sampling	412
20.3	Inferences in Super-population and Survey Population	414
20.4	Utilizing Auxiliary Information	416
20.5	Pseudo Likelihoods Method in Finite Sampling Problems	424

21 Inference for Density Ratio Model with Continuous Covariates	427
21.1 Generalized Odds Ratio Model and Pairwise Conditional Likelihood.	427
21.2 Generalized Kendall's Tau for Testing Conditional Independence	439
21.3 Applications in Graphical Models and High Dimensional Parameter Problems	442
21.4 Profile Likelihood Approach for Continuous Covariate Density Ratio Model.	443
22 Non-ignorable Missing Data Problems.	447
22.1 Model Identifiability Problem	447
22.2 Semiparametric Approaches for Non-ignorable Missing Data Problems	451
22.3 Empirical Likelihood Method and Instrument Variable Approach	455
22.4 Maximum Likelihood Estimation in Call-Back Problem	460
22.5 Heckman's Sample Selection Model.	464
23 Maximum Likelihood Estimation in Capture-Recapture Models	467
23.1 Estimating the Number of Species in the Absence of Covariate	467
23.2 Binomial Detection Model	470
23.3 Inference in Capture and Re-capture Models	474
24 A Review of Survival Analysis	477
24.1 Kaplan–Meier Estimator	477
24.2 Nonparametric MLE for Left Truncated Data	482
24.3 Comparable Set Approach in Truncation Problems	490
24.4 A Review of Cox Regression Model in Survival Analysis	493
24.5 Inferences for AFT Model and Quantile Regression Model	502
24.6 Double Empirical Likelihoods Utilizing Auxiliary Information	507
24.7 Case-Cohort Study	510
25 Length Biased Sampling, Multiplicative Censoring and Survival Analysis	519
25.1 Vardi's Four Equivalent Problems	519
25.2 A New EM Algorithm	524
25.3 Cox Model with Length Biased Sampling Data	529
25.4 Composite Partial Likelihood Approach	533
25.5 Linear Rank Statistics with Cross-Sectional Data	536
25.6 Estimating Equations Derived from an Embedded Likelihood	540

25.7	Generalized Multiplicative Censoring and Truncation	542
25.8	The Multi-sample Wicksell Corpuscle Problem	547
25.9	Missing Information Principle.	549
25.10	Case and Control Study with Prevalent Cases	556
26	Applications of the Pool Adjacent Violation Algorithm (PAVA) in Statistical Inferences	559
26.1	Pool Adjacent Violation Algorithm (PAVA).	559
26.2	Applications of Pool Adjacent Violation Algorithm in Exponential Families	560
26.3	Estimating Monotonic Decreasing Density and Hazard Functions	570
26.4	Cosslett's Maximum Likelihood Estimation and Related Problems.	577
26.5	Maximum Binomial Likelihood Estimation in a Genetic Mixture Model	581
26.6	Maximum Likelihood Estimation Based on Current Status Data	587
26.7	Application in Receiver Operating Characteristic (ROC)	595
26.8	Panel Count Data and Simplex Constraints.	598
References		601
Index		623

Chapter 1

Examples and Basic Theories for Length Biased Sampling Problems

We begin with a discussion of length biased sampling problems and their applications in different fields. Biased sampling occurs when an investigator naturally collects samples from a population, but the sampling distribution is different from the target population. It happens because not every unit in the population has an equal chance to be sampled when the natural sampling plan is adopted. In the well known Wicksell corpuscle problem (Wicksell 1925, 1926), the maximum size of random spheres in a reference volume is to be predicted from the size distribution of circles that are planar sections of spheres cut by a plane. Biased sampling appears because spheres with larger radii are more likely to be sampled, and it is clear that the probability of a sphere being sampled is proportional to its radius. This phenomenon is also common in stereological and medical research. For example, in the data collection stage for estimating the proportion of individuals having certain rare genetic traits such as albinism, Fisher (1934) noticed the biased sampling problem, though he did not use the terminology of “biased sampling”. Instead, he referred to the data collection procedure for counting the number of albino children in a family as the method of ascertainment. Rao (1965) introduced the concept of weighted distributions for discrete data, which is a distorted version of the original distribution. Many more examples can be found in the papers by Rao (1965) and Patil and Rao (1978). Typically biased sampling takes place in studying wild life population in which a large cluster of animals is more easily seen. Length biased sampling problems were also discussed by Cox (1969), Vardi (1982a,b) and Zelen (2004). Length-biased sampling arises in renewal processes when the probability that an interval is selected is proportional to the length of the interval. Intuitively, longer periods are more likely to contain an event that is independent of the renewal process.

1.1 Length Biased Sampling Examples

Let X be a random variable recording the size of some group from a target population with probability function denoted as $p_k = P(X = k)$, $k = 1, 2, \dots$. Suppose that a group from this population is observed only when at least one of the individuals in the group is sighted and each individual has an independent probability of θ of being sighted. Then the probability that an observed group has $X = k$ individuals is

$$P(\text{A group is sighted} | \text{group size} = k) = 1 - (1 - \theta)^k =: w(k).$$

The observed group size has a distribution

$$P(X = k | \text{A group is sighted}) = \frac{w(k)p_k}{\sum_{k=1}^{\infty} w(k)p_k} =: p_k^w.$$

If $\theta \rightarrow 0$, then $w(k) \approx k\theta$. As a consequence

$$p_k^w \rightarrow \frac{kp_k}{\sum_{k=1}^{\infty} kp_k}.$$

This distribution is called the size-biased or length-biased distribution of X .

Next we give an example of length-biased distribution derived from a continuous distribution. Let Y be a continuous random variable taking values in $(0, \tau)$, with density $f(y)$. Let A be a uniformly distributed random variable in $(0, \tau)$ and independent of Y . Suppose that a unit Y of size y in the population is recorded only if $y > A$. Since $P(A < Y|Y = y) = y/\tau$, a unit with a larger y is more likely to be recorded. Hence, the density function of a recorded unit is given by

$$f^w(y) = \frac{y\tau^{-1}f(y)}{\int \tau^{-1}yf(y)dy} = \frac{yf(y)}{\int_0^\tau yf(y)dy}, \quad 0 < y \leq \tau.$$

Clearly as $\tau \rightarrow \infty$, this becomes

$$f^w(y) = \frac{yf(y)}{\int_0^\infty yf(y)dy}, \quad y \geq 0.$$

This is a length-biased distribution derived from $f(y)$.

There are many examples where the observed distribution differs from the target population in this length biased fashion. We give two examples in genetical studies.

In genetic mapping studies, a test statistic is computed at each locus. When the sizes of the loci are plotted on the genetic map, disease suspect loci may be identified as the peaks. Using the renewal theory, Terwilliger et al. (1997) found that peaks caused by disease loci are longer than peaks caused by random fluctuation. Hence

longer peaks are more likely to contain the gene of interest than shorter ones. This phenomenon may have the potential to aid in linkage mapping.

Another example is in RNA sequence study. Next generation sequencing technologies have been used extensively to quantify transcribed RNA sequences. The differential expression of longer transcripts is more likely to be identified than that of a shorter transcript with the same effect size. For more details, see, Gao et al. (2011).

Next we give a few examples on length biased sampling version for some commonly used distributions.

1. Length Biased Binomial Distribution

Suppose in a boy school, data are collected on the number of brothers and sisters in the family of each boy in the school. Clearly the collected data are a biased representation of the general population since this is a boy school and each family has at least one boy.

Suppose X has a binomial distribution. Denote its length biased version as X^* . Hence

$$P(X^* = k) = \frac{k P(X = k)}{\sum_{k=0}^n k P(X = k)} = k \binom{n}{k} \theta^k (1 - \theta)^{n-k} / (n\theta) = \binom{n-1}{k-1} \theta^{k-1} (1 - \theta)^{n-k},$$

or

$$X^* \sim 1 + B((n-1), k, \theta), k = 0, 1, 2, \dots, (n-1).$$

Intuitively, we already know that there is a boy in a family with n children, the total number of boys should equal to 1 plus the number of boys out of the other $n-1$ children, which has a binomial distribution with parameters $(n-1)$ and θ .

2. Length Biased Poisson Distribution

Suppose the original random variable X has a Poisson distribution

$$P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}, \quad x = 0, 1, 2, \dots,$$

Its length biased version is

$$P(X^* = x) = \frac{\lambda^{x-1} \exp(-\lambda)}{(x-1)!}, \quad x = 1, 2, \dots,$$

Again it can be shown that in distribution X^* can be written as

$$X^* = 1 + X.$$

More examples on biased sampling discrete distributions can be found in Rao (1965) and Patil and Rao (1978). Next we discuss length biased version for continuous distribution examples.

3. Length Biased Exponential Distribution

The exponential density is given by

$$f(y) = \lambda \exp(-\lambda y), \quad y > 0.$$

The length biased version is

$$f_w(y) = \frac{y\lambda \exp(-\lambda y)}{\int_0^\infty y\lambda \exp(-\lambda y)dy}, \quad y > 0.$$

This is a gamma density with shape parameter 2 and scale parameter λ .

4. Length Biased Weibull Distribution

A two-parameter Weibull density is given by

$$f(y) = \alpha\lambda(\lambda y)^{\alpha-1} \exp\{-(\lambda y)^\alpha\}, \quad y > 0.$$

The length biased version is

$$f_w(y) = \frac{\alpha(\lambda y)^\alpha \exp\{-(\lambda y)^\alpha\}}{\int \alpha(\lambda y)^\alpha \exp\{-(\lambda y)^\alpha\} dy}, \quad y > 0.$$

This density is also a Weibull density, with shape parameter $\alpha + 1$ and scale parameter λ .

5. General Weighted Distributions

Other than the length biased distributions, Rao (1965) also discussed general weighted distributions. If $X \sim f(x)$, and $w(x)$ is a positive function. Then, he defined a weighted version X^w of X as having the following density,

$$X^w \sim f^w(x) = \frac{w(x)f(x)}{\mu}, \quad \mu = \int w(x)f(x)dx,$$

where μ is assumed to be finite. A length biased sampling corresponds to the choice of $w(x) = x$. When $w(x) = x^2$ and x^3 , they are called, respectively, area biased sampling and volume biased sampling. Another commonly used $w(x)$ is $w(x) = P(D = 1|x)$, where $0 < P(D = 1|x) < 1$ is a probability function. This form is useful in missing data or non-response data problems. Given a X value, the response probability of an individual is given by $P(D = 1|x) = w(x)$. Then conditioning on $D = 1$,

$$X|D = 1 \sim \frac{P(D = 1|x)f(x)}{P(D = 1)} = \frac{w(x)f(x)}{\int w(x)f(x)dx}.$$

In other words, those responders have a probability density which is the weighted version of original one.

Other popular and important weight functions include x^α , $\alpha > 0$, $\{1 - (1 - \beta)^x\}$, $0 < \beta < 1$, $(x + 1)$, $x(x - 1) \cdots (x - r + 1)$, $r > 0$, ϕ^x , $0 < \phi < 1$, $\exp(\phi t)$, $\Phi(\alpha + x\beta)$, where Φ is the standard normal cumulative distribution function, and $\exp(\alpha + x\beta)/\{1 + \exp(\alpha + x\beta)\}$.

1.2 Basic Properties of Length Biased Sampling Problems

In this section we derive some basic properties of length biased sampling distributions.

Theorem 1.1 Suppose a non-negative random variable X has a pdf $f(x)$ with mean $E(X) < \infty$. Let X^* be the length biased version of X , i.e., with pdf $f^*(x) = xf(x)/\mu$, $\mu = \int xf(x)dx$. Then

$$E(X^*) - E(X) = \text{var}(X)/E(X),$$

and $E(X^*) \geq E(X)$.

Proof Note

$$E(X^*) - E(X) = \frac{\int y^2 f(y)dy}{\mu} - \mu = \frac{\int y^2 f(y)dy - \mu^2}{\mu} = \frac{E(X^2) - E^2(X)}{\mu} = \frac{\text{var}(X)}{E(X)}.$$

Therefore on the average the length of X^* is longer than that of X . Intuitively this makes sense since larger X is more likely to be sampled in a length biased sampling.

1.3 Stochastic Ordering

In statistical literature if two random variables T_1 and T_2 (with cumulative distribution functions F_1 and F_2 , respectively) satisfy

$$F_1(t) \leq F_2(t),$$

then T_1 is called stochastically larger than T_2 . Next we show that a random variable X and its length biased version X^* satisfy this stochastic ordering. First we present a well known lemma.

Chebyshev Correlation Inequality

Lemma 1.2 Suppose g_1 and g_2 are either both monotonic non-increasing or non-decreasing functions of a random variable X . Then $g_1(X)$ and $g_2(X)$ are non-negatively correlated:

$$E[g_1(X)g_2(X)] \geq E[g_1(X)]E[g_2(X)].$$

Proof Let X_1 and X_2 be i.i.d. copies of X . By the monotonic assumption

$$\{g_1(X_1) - g_1(X_2)\}\{g_2(X_1) - g_2(X_2)\} \geq 0.$$

Taking expectation on both sides we have

$$E[\{g_1(X_1) - g_1(X_2)\}\{g_2(X_1) - g_2(X_2)\}] \geq 0.$$

Expanding the left hand side, we have

$$E\{g_1(X_1)g_2(X_1)\} - E\{g_1(X_2)g_2(X_1)\} - E\{g_1(X_1)g_2(X_2)\} + E\{g_1(X_2)g_2(X_2)\} \geq 0.$$

By using the fact that X_1 and X_2 are independently and identically distributed, we have proved the Chebyshev correlation inequality.

Now we are ready to show the stochastic ordering between X and X^* . Noting

$$X^* \sim dF^*(t) = \frac{tdF(t)}{\int_0^\infty t dF(t)},$$

we have

$$\begin{aligned} P(X^* > x) &= \frac{\int_x^\infty t dF(t)}{E(X)} \\ &= \frac{E[XI(X > x)]}{E(X)} \\ &\geq \frac{E(X)E[I(X > x)]}{E(X)} = P(X > x). \end{aligned}$$

We have used the Chebyshev correlation inequality by taking $g_1(X) = X$ and $g_2(X) = I(X > x)$.

The stochastic ordering between X and Y does not extend to their length biased versions X^* and Y^* .

Exercise Let

$$P(X = 1) = P(X = 3) = 1/2, \quad P(Y = 2) = P(Y = 4) = 1/2.$$

Show that $\bar{F}(x) \leq \bar{G}(x)$.

The length biased versions are, respectively,

$$P(X^* = 1) = 1/4 = 1 - P(X^* = 3), \quad P(Y^* = 2) = 1/3 = 1 - P(Y^* = 4).$$

Show the stochastic ordering no longer applies in this case.

The hazard function of a random variable is defined as $\lambda(t) = f(t)/\bar{F}(t)$, where $f(t)$ and $\bar{F}(t)$ are, respectively, density and survival functions.

The hazard function for the length biased version X^* of X satisfies

$$\begin{aligned}\lambda_{X^*}(t) &= \frac{tf(t)}{\int_t^\infty xf(x)dx} \\ &\leq \frac{tf(t)}{t \int_t^\infty f(x)dx} = \frac{f(t)}{\bar{F}(t)} = \lambda(t).\end{aligned}$$

In other words the hazard function with length biased data is smaller than that with unbiased sampling data. In general this is called hazard ordering if two hazard functions satisfy the above inequality constraint. Since “healthier individuals” are more likely to be selected in medical studies due to the length biased phenomenon, naturally the hazard of death for individuals in length biased sampling should be smaller than those in unbiased sampling.

Another popular stochastic ordering is called the likelihood ratio ordering, defined as monotonic function of the ratio of two densities. Therefore the length biased version X^* and X satisfy the stochastic ordering, hazard ratio ordering and likelihood ratio ordering.

Next we show that in general the hazard ratio ordering is stronger than the stochastic ordering.

In fact if $\lambda_1(t) \leq \lambda_2(t)$, then

$$\Lambda_1(t) = \int_0^t \lambda_1(u)du \leq \int_0^t \lambda_2(u)du = \Lambda_2(t), \quad t > 0.$$

Therefore

$$\bar{F}_1(t) = \exp\{-\Lambda_1(t)\} \geq \exp\{-\Lambda_2(t)\} = \bar{F}_2(t),$$

i.e., the hazard ratio ordering is stronger than the stochastic ordering. Moreover, the hazard ratio ordering is equivalent to

$$\bar{F}_1(t) = \exp\{-\Lambda_2(t)\} \exp\{-\int_0^t (\lambda_1(u) - \lambda_2(u))du\} =: \bar{F}_2(t)\bar{F}_3(t),$$

where $\bar{F}_3(t)$ is a(n) (improper) survival function.

Exercise Show that the likelihood ratio ordering (not necessarily the length biased case) is stronger than the stochastic ordering and hazard ratio ordering.

1.4 Lorenz Curve

If $X \sim F(x)$ denotes the annual household income in a population, then a popular method in economic study to measure inequality of the wealth distribution is the Lorenz curve (Lancaster 1992), which is a graphical representation of the cumulative income distribution. The Lorenz curve shows the percentage, p_2 , of the total income for the bottom p_1 percent of the households. The percentage of households is plotted on the horizontal axis, the percentage of income on the vertical axis. The Lorenz curve has a natural connection with length biased sampling distribution. Mathematically it is a plot of $F(x)$ vs. the length biased version of F , i.e., $F^*(x) = \int_0^x u dF(u)/\mu$. Note that $\mu = \int_0^\infty u dF(u)$ is the average income and $\int_0^x u dF(u) = E[XI(X < x)]$ is the average income for those households with income less than x .

Since we have shown that $F^*(x) \leq F(x)$, the Lorenz line should fall below the 45 degree straight line. As a matter of fact, if $p_1 = p_2$, the Lorenz curve is a straight line which says for instance that 50% of the households have 50% of the total income. Thus the straight line represents perfect equality. Any departure from this 45° line represents inequality.

Easily we can show that the Lorenz curve is a plot of p vs. $L(p) = \int_0^p F^{-1}(u) du / \int_0^1 F^{-1}(u) du$ by using transformation $p = F(x)$.

Exercise 1 Show that $L(p)$ is an increasing convex function of p .

Exercise 2 If X has a uniform distribution on $(0, \tau)$, find the distribution of the length biased version, X^* , of X . Is it possible to have X^* and X such that they have the same distribution?

A useful measure for the Lorenz curve is its distance from the line of perfect equality in income, i.e. $p - L(p)$. The Gini index is defined as

$$G = 2 \int_0^1 \{p - L(p)\} dp = 1 - 2 \int_0^1 L(p) dp$$

It can be shown that

$$G = \frac{1}{\mu} \int F(x)\{1 - F(x)\} dx = 2\text{cov}(X, F(X)).$$

The Gini index can also be understood as the difference between two areas under the curves, $y = x$ and $y = L(x)$.

Next we present a theoretical characterization result for a length biased sampling variable.

1.5 Characterization of Length Biased Distribution

We say that a random variable X is infinitely divisible if for all n , X can be decomposed in distribution as the sum of n iid variables. That is, for all n there exists a distribution dF_n such that if $X_1^{(n)}, \dots, X_n^{(n)}$ are i.i.d with this distribution and X and $X_1^{(n)} + \dots + X_n^{(n)}$ have the same distribution. The commonly used infinitely divisible distributions include, Poisson, negative binomial, normal, exponential, gamma, etc. As we have shown that if X^* is a length biased version of X , then X^* is stochastically larger than X . In binomial and Poisson cases we have found that in fact $X^* = 1 + X$. In general if X is infinitely divisible, then there exists a random variable $Z \geq 0$ such that in distribution

$$X^* = X + Z,$$

where X and Z are independent of each other. For more details, see, Pakes et al. (1996) and Arratia et al. (2015). In other words the length based random variable X^* may be decomposed as the summation of the original unbiased random variable X and an independent non-negative random variable Z .

Chapter 2

Brief Introduction of Renewal Process

Length biased sampling is a fundamental problem in studying renewal processes. Suppose buses arrive at a stop after independent random intervals of T minutes, where T is uniformly distributed between 10 and 20. It is natural to wonder how long one is expected to wait from some random point in time t until next bus arrives. The next bus could arrive immediately, or one could be unlucky with time t just after the previous bus left and could wait as long as 20 minutes for the next bus. Interestingly this waiting time is no longer uniformly distributed. This is the so-called “inspection paradox”.

In order to understand this phenomenon we briefly introduce some basic concepts in the renewal process. More details can be found from renewal process textbooks, such as Cox and Isham (1980) and Feller (1965). Professor Sigman (2009)’s lecture notes are also very valuable. Some results discussed in this chapter will be used in Chap. 25.

2.1 Basic Concepts

Let X_1, X_2, \dots be i.i.d. positive random variables with a common distribution function F . Define a sequence of times by $T_0 = 0$, and $T_k = T_{k-1} + X_k$ for $k \geq 1$. A typical example is that X_i is the lifetime of a light bulb, and T_k is the time the k -th bulb burns out. Another example is T_k is the time of arrival of the k -th customer, or bus arrival time, etc. We assume $\mu = E(X_i)$ and $\sigma^2 = \text{Var}(X_i) < \infty$.

Let

$$N(t) = \max\{k : T_k \leq t\},$$

be the number of renewals in $(0, t]$. Immediately we have $N(t) \rightarrow \infty$ as $t \rightarrow \infty$, and

$$\{N(t) = n\} = \{T_n \leq t, T_{n+1} > t\}, \quad n \geq 1.$$

Let F_n be the distribution of T_n . Clearly we have a convolution relationship

$$F_{k+1}(x) = \int_0^x F_k(x-y)dF(y), \quad k \geq 1.$$

Also from the fact

$$\{N(t) = k\} = \{N(t) \geq k\} - \{N(t) \geq k+1\},$$

we have

$$P\{N(t) = k\} = F_k(t) - F_{k+1}(t).$$

The renewal function $m(t)$ is defined as $m(t) = E\{N(t)\}$. Noting

$$N(t) = \sum_{k=1}^{\infty} I_k, \quad I_k = I(T_k \leq t),$$

we have

$$m(t) = E[N(t)] = \sum_{k=1}^{\infty} E(I_k) = \sum_{k=1}^{\infty} F_k(t).$$

Conditioning on the first interval time X_1 ,

$$m(t) = E\{N(t)\} = E[E\{N(t)|X_1\}].$$

If $t < x$, then $E\{N(t)|X_1 = x\} = 0$ because the first arrival occurs after time t . On the other hand if $t \geq x$,

$$E\{N(t)|X_1 = x\} = 1 + E\{N(t-x)\}$$

since after the first arrival time x , the renewal process starts over again. Now we show the **renewal equation**

$$\begin{aligned} m(t) &= \int_0^{\infty} E\{N(t)|X_1 = x\}dF(x) \\ &= \int_0^t [1 + m(t-x)]dF(x) \\ &= F(t) + \int_0^t m(t-x)dF(x). \end{aligned}$$

Next we study the elementary renewal theorem.

Theorem 2.1 For the renewal process $N(t)$, $t \geq 0$ defined above, almost surely

$$\lim_{t \rightarrow \infty} \frac{N(t)}{t} = \frac{1}{E(X)}, \quad \lim_{t \rightarrow \infty} \frac{E\{N(t)\}}{t} = \frac{1}{E(X)}.$$

Proof Noting $T_{N(t)} \leq t < T_{N(t)+1}$ and $T_{N(t)} = X_1 + \dots + X_{N(t)}$, we have inequality

$$\frac{1}{N(t)} \sum_{j=1}^{N(t)} X_j \leq \frac{t}{N(t)} \leq \frac{1}{N(t) + 1} \left\{ \sum_{j=1}^{N(t)+1} X_j \right\} \{1 + 1/N(t)\}.$$

Using the Law of Large Numbers we can show that both sides converge to $E(X)$ as $t \rightarrow \infty$.

The second part of proof needs $N(t)/t$ to be uniformly integrable. We leave this to readers as an exercise.

Clearly the number of renewals in the interval $(0, t]$ is inversely proportional to the length of inter-arrival time. Indeed the elementary renewal theorem tells us the average number of renewals in $(0, t]$ is reciprocal of the average inter-arrival time.

Next we ask whether the central limit theorem holds true

$$\frac{N(t) - E\{N(t)\}}{\sqrt{\text{Var}\{N(t)\}}} \rightarrow N(0, 1)$$

as $t \rightarrow \infty$?

We already know that $E\{N(t)\} = m(t) \approx t/\mu$. How do we find the variance of $N(t)$?

For any real number x , denote r_t as the integer part of $t/\mu + x\sqrt{t\sigma^2/\mu^3}$. Then

$$r_t = t/\mu + x\sqrt{t\sigma^2/\mu^3} - \theta, \quad 0 \leq \theta < 1.$$

Note

$$\begin{aligned} P(T_{r_t} \geq t) &= P\left(\frac{T_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} \geq \frac{t - r_t\mu}{\sigma\sqrt{r_t}}\right) \\ &= P\left(\frac{T_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} \geq \frac{-x\sigma(t/\mu)^{1/2} + \mu\theta}{\sigma\sqrt{r_t}}\right). \end{aligned}$$

As $t \rightarrow \infty$,

$$\frac{r_t}{t/\mu} = \frac{t/\mu + x\sqrt{t\sigma^2/\mu^3} - \theta}{t/\mu} \rightarrow 1.$$

Using the Central Limit Theorem we can show that

$$P(N(t) \leq r_t) = P(T_{r_t} \geq t) \rightarrow 1 - \Phi(-x) = \Phi(x),$$

where Φ is the standard normal distribution function.

On the other hand

$$\begin{aligned} P(N(t) \leq r_t) &= P\left(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \leq \frac{r_t - t/\mu}{\sqrt{t\sigma^2/\mu^3}}\right) \\ &= P\left(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \leq \frac{x\sqrt{t\sigma^2/\mu^3} - \theta}{\sqrt{t\sigma^2/\mu^3}}\right) \\ &\approx P\left(\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \leq x\right), \end{aligned}$$

as $t \rightarrow \infty$. In conclusion we have proved the Central Limit Theorem for a renewal process.

Theorem 2.2 *As $t \rightarrow \infty$, in distribution*

$$\frac{N(t) - t/\mu}{\sqrt{t\sigma^2/\mu^3}} \rightarrow N(0, 1). \quad (2.1.1)$$

From this with mild moment conditions we can find

$$\text{Var}\{N(t)\}/t \rightarrow \sigma^2/\mu^3.$$

2.2 Forward and Backward Recurrence Times

The forward recurrence time is a very important concept in renewal processes. It is the time between any given time t and the next epoch of the renewal process under consideration,

$$V(t) = T_{N(t)+1} - t, \quad t \geq 0.$$

It is also called residual lifetime or residual waiting time. If X_i has an exponential distribution with rate λ , then by the memoryless property of the exponential distribution, we know $V(t) \sim \exp(\lambda)$, $t \geq 0$. However for the general renewal process, the distribution of $V(t)$ is complicated and depends on the time t .

Theorem 2.3 When the process $\{N(t), t \geq 0\}$ has run a long, it reaches the equilibrium, i.e.,

$$\lim_{t \rightarrow \infty} \int_0^t V(s)ds = \frac{E(X^2)}{2E(X)},$$

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t I\{V(s) > x\}ds = \frac{E(X - x)^+}{\mu}, \quad \mu = E(X),$$

where $a^+ = \max(0, a)$.

The proof is not difficult. In fact from

$$T_{N(t)} \leq t < T_{N(t)+1},$$

we have

$$\int_0^t V(s)ds = \sum_{i=1}^{T_{N(t)}} \int_{t_i}^{t_{i+1}} (s - t_i)ds + \int_{T_{N(t)}}^t (s - T_{N(t)})ds.$$

Immediately we have the inequality

$$\frac{1}{t} \sum_{j=1}^{N(t)} X_j^2 / 2 \leq \frac{1}{t} \int_0^t V(s)ds \leq \frac{1}{t} \sum_{j=1}^{N(t)+1} X_j^2 / 2.$$

Note that $t^{-1} = N^{-1}(t)N(t)/t$. Using the elementary renewal theorem and Strong Law of Large Numbers, we have the conclusion.

Similarly we can show that

$$\begin{aligned} \int_0^t I\{V(s) > x\}ds &= \sum_{j=1}^{N(t)} \int_{t_{j-1}}^{t_j} I\{V(s) > x\}ds + \int_{N(t)}^t I\{V(s) > x\}ds \\ &= \sum_{j=1}^{N(t)} \int_{t_{j-1}}^{t_j} I\{x_j - x > s - t_{j-1}\}ds \\ &= \sum_{j=1}^{N(t)} (x_j - x)^+. \end{aligned}$$

Note that if $x_j - x > 0$, then the integral becomes $\int_{t_{j-1}}^{t_j - x} 1ds = x_j - x$. On the other hand if $x_j - x < 0$, the indicator function gives zero, and the integral becomes 0.

Note that

$$E(X - x)^+ = \int_x^\infty (t - x)dF(t) = \int_x^\infty \bar{F}(t)dt.$$

Intuitively the limiting value of $P(V(t) > x)$ as $t \rightarrow \infty$ should be equal to the average $t^{-1} \int_0^t P\{V(s) > x\}ds$. Therefore the equilibrium distribution as $t \rightarrow \infty$ has a distribution

$$\frac{1}{t} \int_0^t I\{V(s) > x\}ds = \frac{N(t)}{t} \frac{1}{N(t)} \sum_{j=1}^{N(t)} (x_j - x)^+ \rightarrow \int_x^\infty \bar{F}(t)dt/\mu, \quad x \geq 0.$$

This is also called a residual density

$$f_R(x) = \frac{\bar{F}(x)}{\mu}, \quad x \geq 0.$$

Note that $f_R(x)$ is always non-increasing.

Similarly we can define backward time. For any given time t , the backward time is the time between the last epoch of the renewal process to t ,

$$A(t) = t - T_{N(t)}, \quad t \geq 0.$$

It can be shown that the backward time A and forward time V have the same limiting distribution. Intuitively, when a renewal process reaches the equilibrium status, it becomes reversible, i.e., the statistical properties of this process are the same as the statistical properties for time-reversed data from the same process. The backward time can also be treated as the forward time if time periods are reversible.

The inter-arrival interval is defined as

$$Y(t) = T_{N(t)+1} - T_{N(t)} = A(t) + V(t).$$

Using the results from backward time and forward time, we immediately have

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t Y(s)ds = \frac{E(X^2)}{E(X)} \geq E(X).$$

This is the so called “inspection paradox”. When the renewal process reaches the equilibrium status, the mean inter-arrival time is longer than $E(X)$.

Moreover

$$\lim_{t \rightarrow \infty} \int_0^t I\{Y(s) > x\}ds = \frac{E\{XI(X > x)\}}{\mu}.$$

This can be shown by noting

$$\frac{1}{t} \int_0^t I\{Y(s) > x\}ds = \frac{1}{t} \sum_{j=1}^{N(t)} X_j I(X_j > x),$$

the length of time during the j -th inter-arrival time that $T(s) > x$ is precisely X_j if $X_j > x$ and 0 otherwise.

Observe

$$t^{-1} \int_0^t I\{V(s) > y, A(s) > x\} ds = t^{-1} \sum_{j=1}^{N(t)} \{X_j - (y + x)\}^+.$$

The length of time during the j -th inter-arrival time that $A(s) > y$ and $V(s) > x$ is precisely $\{X_j - (y + x)\}^+$.

Consider the first inter-arrival time, so that $A(s) = s$ and $V(s) = X_1 - s$, $s \in [0, X_1]$. Then

$$I\{A(s) > y, V(s) > x\} = I(s > y, X_1 - s > x) = I(y < s < X_1 - x).$$

If $X_1 - (y + x) > 0$, then this indicator is non-zero and

$$\int_0^{X_1} I(y < s < X_1 - x) ds = \int_y^{X_1-x} ds = X_1 - (y + x).$$

If $X_1 - (y + x) \leq 0$, then the indicator is 0 and

$$\int_0^{X_1} I(y < s < X_1 - x) ds = 0.$$

Therefore we obtain $\{X_1 - (y + x)\}^+$.

As $t \rightarrow \infty$, the joint density of the backward time $A(t)$ and forward time $V(t)$ has a survival function

$$t^{-1} \int_0^t I\{V(s) > y, A(s) > x\} ds = \frac{N(t)}{t} \frac{1}{N(t)} \sum_{j=1}^{N(t)} \{X_j - (y + x)\}^+ \rightarrow \frac{\int_{x+y}^{\infty} \bar{F}(u) du}{\mu}.$$

Thus we have shown that

Theorem 2.4 *At the equilibrium, the backward time A and forward time V have a joint density*

$$\frac{f(a+v)}{\mu}, \quad a, v > 0. \quad (2.2.2)$$

Let $Y = A + V$ be the inter-arrival interval, then

$$(Y, A) \sim \frac{f(y)}{\mu}, \quad y > a > 0. \quad (2.2.3)$$

The marginal density of Y is

$$\int_0^y f(y)da/\mu = yf(y)/\mu, \quad y > 0, \quad (2.2.4)$$

which is a length biased version of $f(y)$. The conditional density of A given Y is

$$\frac{f(y)/\mu}{yf(y)/\mu} = 1/y, \quad 0 \leq a \leq y. \quad (2.2.5)$$

In other words, given the length of the inter-arrival time, the backward time is a uniform distribution between $(0, y)$, irrelevant of the density f . By symmetry, this result also applies to the forward time. From a statistical inference point of view the inter-arrival length has captured all information on the underlying distribution F and the backward time or forward time is not informative for F any more as long as the inter-arrival length is given. However, the backward time indeed contains information about F if the inter-arrival length is not available, which is the case in cross sectional studies where the forward time is not available or partially available (with right censoring). We will use this result in Chap. 25 for length biased sampling data with right censoring.

In contrast to the length biased sampling problem, Rao (1965) discussed a reciprocal length biased sampling problem in aerial survey of low density traffic streams. Assume that vehicles enter the highway according to a nonhomogeneous Poisson process. The vehicles choose velocities at random from a distribution of $V \sim f(v)$. However the recorded velocity of a vehicle in the interval $[a, b]$ at time t is

$$f_w(v) = \frac{v^{-1} f(v)}{\int v^{-1} f(v) dv},$$

as $t \rightarrow \infty$, i.e., at the equilibrium. In this example, slower vehicles take longer time to transverse $[a, b]$. As a result, it would be easier to be recorded.

We conclude this chapter with some important results on the Poisson process which may be useful in later chapters.

2.3 Basic Results on Poisson Process

A special case of a renewal process is the Poisson process. It is defined as follows.

- (1) $N(0) = 0$.
- (2) The process has stationary and independent increments.
- (3) The number of events in an interval of length t is Poisson distributed with mean λt , i.e., for all $s, t > 0$,

$$P(N(t+s) - N(s) = k) = \frac{(\lambda t)^k \exp(-\lambda t)}{k!}, \quad k = 0, 1, 2, \dots,$$

Note that the first inter-arrival time $X_1 > t$ implies $N(t) = 0$.

$$P(X_1 > t) = P(N(t) = 0) = \exp(-\lambda t).$$

Hence X_1 has an exponential distribution with mean $1/\lambda$.

$$\begin{aligned} P(X_2 > t | X_1 = s) &= P\{0 \text{ event in } (s, s+t] | X_1 = s\} \\ &= P\{0 \text{ event in } (s, s+t]\} \\ &= \exp(-\lambda t), \end{aligned}$$

where the second equality has used the independent increments assumption and the third one has used the stationary increments argument.

Therefore the inter-arrival times X_1, X_2, \dots are i.i.d. exponential random variables.

Exercise Given $N(t) = n$, the n arrival times T_1, \dots, T_n have the same distribution as the order statistics corresponding to n independent variables uniformly distributed on the interval $(0, t)$.

Nonhomogeneous Poisson Process

In practical applications the requirement that the arrival rate at any time t is a constant λ is too strong. To relax this assumption, we consider a nonhomogeneous Poisson process.

The counting process $\{N(t), t \geq 0\}$ is said to be a non-stationary or nonhomogeneous Poisson process with intensity function $\lambda(t)$, $t \geq 0$ if

- (1) $N(t) = 0$.
- (2) $\{N(t), t \geq 0\}$ has independent increments.
- (3) $P\{N(t+\delta) - N(t) \geq 2\} = o(\delta)$, as $\delta \rightarrow 0$.
- (4) $\{N(t+\delta) - N(t) = 1\} = \lambda(t)\delta + o(\delta)$.

Denote

$$p_j(t) = P(j \text{ event before time } t),$$

then

$$\begin{aligned} p_j(t + \Delta t) &= P\{j \text{ events before } t \text{ and no events in } (t, t + \Delta)\} \\ &\quad + P\{(j-1) \text{ events before } t \text{ and no events in } (t, t + \Delta)\} + (\Delta t)^2, \end{aligned}$$

or

$$p_j(t + \Delta t) = p_j(t)(1 - \lambda(t)\Delta t) + p_{j-1}(t)\lambda(t)\Delta t + (\Delta t)^2.$$

As a consequence we have a differential equation

$$p'_0(t) = -\lambda(t)p_0(t).$$

Clearly

$$p_0(t) = \exp \left\{ - \int_0^t \lambda(u) du \right\}$$

is the solution. In general we can derive (exercise)

$$P\{N(t) - N(s) = k\} = \frac{1}{k!} \{\Lambda(t) - \Lambda(s)\}^k \exp[-\{\Lambda(t) - \Lambda(s)\}], \quad t > s > 0.$$

where

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Define a time transformed process

$$N^*(s) = N(\Lambda(t)),$$

$$P(T_1 > t) = P(N(t) = 0) = \exp\{-\Lambda(t)\},$$

$$P\{T_2 > t_2 | T_1 = t_1\} = P\{N(t_2) - N(t_1) = 0\} = \exp[-\{\Lambda(t_2) - \Lambda(t_1)\}].$$

Therefore

$$f_2(t_2|t_1) = \lambda(t_2) \exp[-\{\Lambda(t_2) - \Lambda(t_1)\}].$$

In general if a Poisson process is observed in the interval $(0, \tau]$ with event times at $t_1 < t_2 < \dots < t_n < \tau$, then the likelihood is

$$\begin{aligned} L &= f_1(t_1)f_2(t_2|t_1) \cdots f_n(t_n|t_{n-1}) P(T_{n+1} > \tau) \\ &= \lambda(t_1) \exp\{-\Lambda(t_1)\} \cdots \lambda(t_n) \exp[-\{\Lambda(t_n) - \Lambda(t_{n-1})\}] \exp[-\{\Lambda(\tau) - \Lambda(t_n)\}] \\ &= \prod_{i=1}^n \lambda(t_i) \exp\{-\Lambda(\tau)\}. \end{aligned} \tag{2.3.6}$$

Theorem 2.5 Given that $N(\tau) = n$, the pdf of the arrival times T_1, \dots, T_n is

$$f(t_1, \dots, t_n) = \frac{n! \prod_{i=1}^n \lambda(t_i)}{\Lambda^n(\tau)}. \tag{2.3.7}$$

Define

$$h(t) = \lambda(t)/\Lambda(\tau), \quad 0 < t < \tau, \quad (2.3.8)$$

then $h(t)$ is a density function in $[0, \tau]$. As a consequence, $f(t_1, \dots, t_n)$ is the joint density of n order statistics from g .

In medical applications, due to correlations among recurrent events, frequently the intensity is assumed to be a random variable given by $\xi\lambda(t)$, where ξ (called frailty) has a gamma distribution. Then conditional on ξ , $N(t)$ still has independent increments. However, it is not the case unconditionally.

The joint likelihood is

$$\int \xi^n \prod_{i=1}^n \lambda(t_i) \exp\{-\xi\Lambda(\tau)\} dG(\xi), \quad \xi \sim G(\cdot).$$

The conditional density

$$(T_1, \dots, T_n)|N(\tau) = n \sim \frac{n! \prod_{i=1}^n \xi \lambda(t_i)}{\xi^n \Lambda^n(\tau)} = \frac{n! \prod_{i=1}^n \lambda(t_i)}{\Lambda^n(\tau)}, \quad 0 \leq t_1, \dots, t_n \leq \tau,$$

however, is free of the frailty ξ . In other words, conditional on $N(\tau) = n$, again T_1, \dots, T_n can be treated as order statistics generated from $h(t)$ defined in (2.3.8). Note that the frailty ξ is cancelled out.

Chapter 3

Heuristical Introduction of General Biased Sampling with Various Applications

When individuals of a population are less likely or more likely to be included than others in a study, then sampling bias occurs. Sampling bias is a systematic error due to the non-random sampling from a population. As a consequence, the sampled subjects are neither equally balanced nor objectively representing the target population. This problem appears naturally in evidence based economic, epidemiological, and medical research in which observational studies are conducted. We call this type of biased sampling “natural selection biased sampling problems”. Biased sampling may also arise when one models the ratio between densities by the exponential tilting model, though the original sampling has nothing to do with the selection bias. This is another fascinating application. We call this type of biased sampling “modelling based biased sampling problems”.

One of the main goals of this book is to develop some valid statistical inference methods for biased sampling problems. In this chapter, we give many examples to show that biased sampling is a ubiquitous problem in many research disciplines.

3.1 Natural Selection Biased Sampling Problems

Example 1 Meta analysis

Meta analysis refers to the quantitative synthesis of evidence from a set of related studies. According to the Web of Science, there were 2,006 meta-analysis publications during the years of 1985–1994, 13,154 during 1995–2004, and more than 55,000 during 2005–2014. It has a tremendous impact on medical research and clinical practice. The naive method of directly combining different studies may be erroneous due to publication bias. Publication bias is the tendency of investigators or editors to base decisions regarding submission or acceptance of manuscripts for publication, on the strength of the investigator’s study findings. Weighted distributions are the natural

choice to model this phenomenon because the weight function is proportional to the probability that the observation enters the record. More details can be found, among others, in Hedges and Olkins (1985), Iyengar and Greenhouse (1988) and Begg and Mazumder (1994).

Let T_i be the statistic from i -th study used to assess the treatment effect, $i = 1, 2, \dots, n$. A common assumption is that

$$T_i | v_i \sim N(\delta, v_i),$$

where δ is the true treatment effect and v_i is the variance, typically it is related to the sample size in the i -th study. After t_i is generated, the study is selected for inclusion in the meta-analysis with probability determined by the appropriate selection model characterized by the weight function. Specifically, for the i -th of the k selected studies, the probability density function of the observed effect t_i is $g_i(t_i)$, where

$$g_i(t_i) = \frac{f_i(t_i)w_i(t_i)}{\int f_i(x)w_i(x)dx}. \quad (3.1.1)$$

In this formulation $f_i(t)$ is the underlying probability density of t_i prior to selection, and $w_i(t)$ is the weight function, i.e., the probability that the study is published. Typically,

$$w_i(t_i) = \exp\{-bp_i^a\}, \quad (3.1.2)$$

for some $a, b \geq 0$, where $p_i = \Phi(-t_i/\sqrt{v_i})$ (one-sided test) or $p_i = 2\Phi(-t_i/\sqrt{v_i})$ (two-sided test).

Let $y_i = t_i/\sqrt{v_i}$, then

$$y_i | v_i \sim g_i^*(y_i) = \frac{h_i(y_i)w(y_i)}{\int h_i(y)w(y)dy},$$

where

$$h_i(y) = \frac{1}{\sqrt{2\pi}} \exp\{-(y - \delta/\sqrt{v_i})^2/2\}.$$

Inference on the treatment effect has to depend on the observed density $g_i^*(y_i)$, which is a biased version of $h_i(y_i)$. Simply ignoring this may yield biased results on δ .

Clearly $a = 0$ or $b = 0$ corresponds to the case that there is no publication bias. A valid statistic method is needed to test this assumption.

Example 2 Case-control sampling in epidemiology studies and choice based sampling in econometric studies

Epidemiological study and econometric study are two areas that use statistical methods extensively. Even though the scientific problems may not be the same, the statistical methods, however, as Breslow (2003) pointed out, are very similar. The typical example contains the case-control sampling problems discussed in cancer

epidemiologies and the choice-based sampling problem in economic studies. The most important references for choice based sampling problems in economics, among others, include Cosslett (1981), Hsieh et al. (1985). In biostatistical literature, important references include, Prentice and Pyke (1979), Anderson (1979), Breslow (1996), and others. Some outstanding contributions for biased sampling problems in statistical literature can be found in Vardi (1982a,b, 1985, 1989).

Consider a parametric multinomial model

$$P(Y = j|x) = P(Y = j|x, \beta), j = 0, 1, 2, \dots, J,$$

where $P(Y = j|x, \beta)$ is a known probability distribution with unknown finitely many parameter β . The marginal density $f(x)$ of X , however, is not specified.

In prospective sampling, one samples X from the density $f(x)$ first, and then samples Y from the probability model $P(Y = j|x)$. However, in choice based sampling or case-control sampling, one first samples $Y^* = j$ based on a known probability mass function $P(Y^* = j) = H(j)$, $j = 0, 1, 2, \dots, J$, then X is sampled from the conditional density $P(X = x|Y = j)$. In this case the joint density for the observed (Y^*, X) is

$$\begin{aligned} P(Y^* = j)P(X = x|Y = j) &= H(j) \frac{P(Y = j, X = x)f(x)}{P(Y = j)} \\ &= H(j) \frac{P(Y = j|x, \beta)f(x)}{\int P(Y = j|x, \beta)f(x)dx}. \end{aligned}$$

Since in general $H(j)$ is known, only the second factor contains information for β . In medical studies, the popular case-control study corresponds to the case $J = 1$, and represents respectively, disease free $Y = 0$ or disease $Y = 1$ status. The underlying model $P(Y = j|x)$, $j = 0, 1$ is given by a logistic regression model.

Essentially, choice based sampling can be extended to deal with continuous response model, i.e., Y is a continuous variable. In this case, one may select $Y = y$ according to some specified density $g(y)$, followed by sampling X from $P(X = x|Y = y)$. It is not necessarily $g(y) = \int f(y|x, \beta)f(x)dx$. Practically, this method may not be easy since it would be hard to find an individual with exactly $Y = y$ when sampling X if Y is a continuous variable.

Outcome dependent sampling is another related retrospective sampling method. In this case, one first divides the range of Y into a few subgroups, say,

$$-\infty < c_1 < c_2 < \dots < c_J < \infty.$$

Within each subgroup one may sample X_{ij} , $i = 1, 2, \dots, n_j$, $j = 1, 2, \dots, J$ from $P(X = x|Y \in c_{j-1} < Y \leq c_j)$. This is also called stratified sampling. We will find out in Chap. 15 that inferences based on outcome dependent sampling and case-control sampling are closely related.

Example 3 Left truncation and right censored data problems

Analogous to the stratified sampling problem, the so called truncation problem has wide applications in lifetime study. In epidemiology studies, a patient is recruited if he or she is alive at the beginning of study and has experienced an initial event (disease onset). Let A be the time from disease onset to the beginning of the study, and T be the lifetime calculated from disease onset. Then A is called the truncation variable because only those individuals with $T > A$ are included in the cohort. Denote the density of A as $h(a)$. In the absence of truncation, we may assume that A and T are independent given covariates X . Moreover, we may postulate a parametric model $f(t|x) = f(t|x, \beta)$, where X is some baseline covariate. Due to left truncation, the observed data, however, follow

$$P(T = t, A = a|x, T > A) = \frac{P(T = t, A = a|x)}{P(T > A|x)} = \frac{h(a)f(t|x\beta)}{\int h(a)\bar{F}(t|x\beta)da},$$

where $\bar{F}(t|x, \beta)$ is the survival function corresponding to $f(t|x, \beta)$. Clearly, if A takes only discrete finitely many values, the truncation problem is equivalent to an outcome dependent sampling problem. A special case is when A follows a uniform distribution in $(0, \tau)$, where $\tau \rightarrow \infty$, then the left truncation problem becomes the so called “length biased” sampling problem. The density becomes

$$P(T = t, A = a|x, T > A) = \frac{f(t|x\beta)}{\int \bar{F}(a|x\beta)da}, \quad a < t < \tau.$$

The density for observed T is

$$P(T = t|x, T > A) = \frac{tf(t|x\beta)}{\int tf(t|x\beta)dt}.$$

In other words, the recruitment of an individual depends on his or her lifetime. A healthier individual is more likely to be sampled. Brookmeyer and Gail (1994) noticed the biased sampling problem in AIDS epidemiologic studies. Simon (1980) used length biased sampling methods in etiologic studies. Gail and Benichou (2000) also discussed related problems in their *Encyclopedia of Epidemiologic Method*.

A recent popular length biased sampling example is the Canadian Study of Health and Aging. This is a multi-center epidemiologic study of dementia, in which 14,025 subjects aged 65 years or older were randomly selected throughout Canada to receive an invitation for a health survey. A total of 10,263 subjects agreed to participate in the study (Wolfson et al. 2001). The backward times from the dementia onset time to the beginning of the study were ascertained. Then, patients were followed up to either death or the end of the study. After excluding subjects with missing onset date of disease or missing classification of dementia subtype, there were a total of 818 patients who were classified into the following three categories: 393 classified as “probable Alzheimer’s disease”, 252 as “possible Alzheimer’s disease”, and 173 as “vascular dementia”. All these patients had been then followed until end of 1996

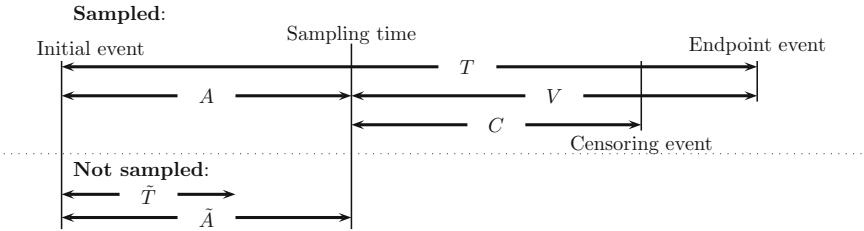


Fig. 3.1 Length biased sampling scheme, where those individuals died before sampling time ($\tilde{T} < \tilde{A}$) are not sampled

when the study terminated. At the time, 638 patients died, while others were right-censored.

In contrast to conventional survival analysis, in which the follow up time is calculated from disease onset time, in length biased sampling problems, the follow up starts from the beginning of the study. The observed data are $\{T_i = A_i + \min(V_i, C_i), \delta_i = I(V_i \leq C_i), i = 1, 2, \dots, n\}$, where A_i and V_i are, respectively, the backward time and forward time, and C is the censoring time. The sampling scheme is plotted in Fig. 3.1.

The observed data likelihood is

$$L = \prod_{i=1}^n \frac{f^{\delta_i}(t_i) \bar{F}^{1-\delta_i}(t_i)}{\mu}, \quad \mu = \int_0^\infty \bar{F}(u) du.$$

In spite of the independent assumption between C and (A, V) , the lifetime $T = A + V$ and censoring time $A + C$ are not independent of each other due to the fact that they share the common backward time. This becomes the dependent censoring problems in survival analysis. Simply ignoring these problems may lead to biased results. Wolfson et al. (2001) found that the unadjusted median survival was 6.6 years (the 95% confidence interval ranges from 6.2 to 7.1). After adjustment for length bias, the estimated median survival was 3.3 years (the 95% confidence interval ranges from 2.7 to 4.0). In other words, the unadjusted analysis severely overestimates the median survival time. The survival curves are plotted in Fig. 3.2 for three different estimators, (1) Vardi's nonparametric MLE for length biased sampling data, (2) Lynden-Bell's product limiting estimator, and (3) Kaplan–Meier estimator. Note that the first two are valid since the left truncation is taken into account but the Kaplan–Meier estimator is biased since it fails to adjust the left truncation.

The truncation problem also appears in astronomy applications. A current controversy in cosmology involves Hubble's Law and I.E. Segal's Chronometric Theory. They predicted different values for the slope β_0 of the straight line relating the magnitude (negative log of luminosity) and the log of velocity, measured by red shift for distant celestial objects. It is generally agreed that the residual V , which represents intrinsic luminosity, is independent of red shift. However, it is debated whether the distribution of V , which is of interest in itself, is of any special form or even has finite

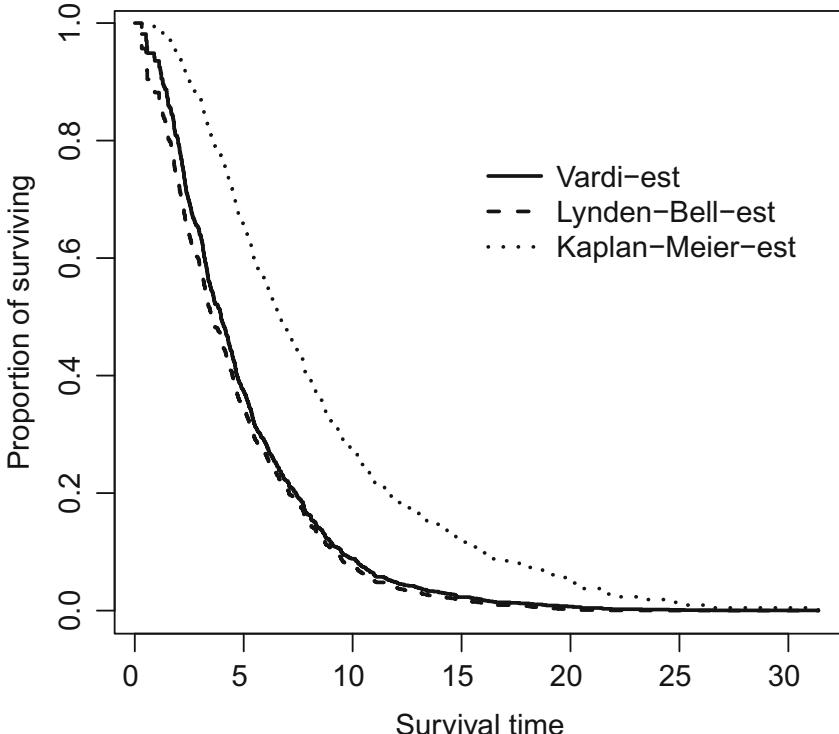


Fig. 3.2 Vardi estimator, Lynden-Bell estimator and Kaplan–Meier estimator based on the Canadian Study of Health and Aging data

second moment. Hence, nonparametric methods are sought to estimate the slope β and the distribution of V . The problem is complicated by the truncation due to the fact that objects extremely far away are not visible, and thus unobserved. Bhattacharya et al. (1983) proposed a nonparametric method based on Kendall’s tau for estimating the slope in a linear regression model. A further complication is the so called “double truncation” discussed by Efron and Petrosian (1999). They investigated quasar luminosities which were doubly truncated by some detection limits. Let $A \leq V$ be low and up detection limits, respectively. T is observed if and only if $A < T < V$. If $T \sim f(t)$ and $(A, V) \sim g(a, v)$, $0 < a \leq v < \infty$, then

$$T, A, V | A < T < V \sim \frac{f(t)g(a, v)I(a < t < v)}{\int f(t)g(a, v)I(a < t < v)dtdadav}.$$

The conditional density is

$$T | a < T < v \sim \frac{f(t)I(a < t < v)}{F(v) - F(a)}.$$

Efron and Petrosian (1999) discussed nonparametric MLE of F , without imposing any distributional assumptions.

Example 4 Secondary outcome analysis based on case-control data

It is well known that the case-control design is one of the most convenient and economic methods to make inferences on risk factors. Typically case-control studies involve the collection of data on a large number of outcomes. These data are of potentially great value to researchers, who, although not necessarily concerned with the disease that defined the case series in the original study, may want to use the available information for a regression analysis involving a secondary outcome. Because cases and controls are selected with unequal probabilities, regression analysis involving a secondary outcome generally must acknowledge the biased sampling feature. More specifically the underlying disease model is given by

$$P(D = 1|X = x, Y = y) = \frac{\exp(\alpha + \beta x + \gamma y + \xi xy)}{1 + \exp(\alpha + \beta x + \gamma y + \xi xy)} =: \pi(x, y).$$

In this model X, Y are covariates. Suppose we are further interested in examining the relationship between X and Y , we may postulate a model

$$f(y|x) = f(y|x\theta).$$

Since we do not have a direct observation from (Y, X) , we have to deal with biased sampling problems. The densities for the case and control data are, respectively

$$f(y, x|D = 1) = \frac{\pi(x, y)f(y, x)}{\int \pi(x, y)f(x, y)dxdy} = \frac{\pi(x, y)f(y|x\theta)g(x)}{\int \pi(x, y)f(y|x\theta)g(x)dxdy}$$

and

$$f(y, x|D = 0) = \frac{\{1 - \pi(x, y)\}f(y|x\theta)g(x)}{\int \{1 - \pi(x, y)\}f(y|x\theta)g(x)dxdy}.$$

However, in the absence of disease prevalent information, this model is almost not identifiable. In Lin and Zeng (2009)'s simulation study, two cases are considered. (1) the disease prevalent information is given by extra knowledge. (2) The disease is rare, for example, $P(D = 1) \leq 0.05$. In this situation $f(y, x) = P(D = 1)f(y, x|D = 1) + P(D = 0)f(y, x|D = 0) \approx f(y, x|D = 0)$. In other words, information in the general population is almost identical to that in the control group. As an alternative, Qin et al. (2016) have made a parametric assumption directly in the control population, i.e., $f(y|x, D = 0) = f(y|x, D = 0, \theta)$. It is shown that any information on θ can be used to improve odds ratio parameters β, γ and ξ . We will revisit this in Chap. 14.

Example 5 Casual inference in observational study

Based on the choice of either treatment (policy) or control, estimation of the average treatment effect on a scalar outcome is a basic goal of many empirical studies

in economics and epidemiology. In a randomized clinical trial, the assignment to the treatment is independent of potential outcomes and covariates, and the average treatment effect can be estimated by simply examining the mean difference. On the other hand, if the treatment assignment depends on the covariates, then simple two sample analysis may not be desirable and can even produce biased estimators. The main problem is that the observed complete data or treatment assignment is a biased version of the original data. When making causal inference in observational studies in many scientific disciplines the most common strategies are the adjustments to the observed confounding variables. It is found that the results based on regression adjustments can be sensitive to model specification when applied to the data in which the treatment and control groups differ substantially (in terms of their pre-treatment covariates). The seminal paper by Rosenbaum and Rubin (1983) introduced propensity score methods to address the fundamental problem by reducing the covariate imbalance between the two groups. They showed that under the assumption of no unmeasured confounding, adjusting for the propensity score, rather than potentially high-dimensional covariates, is sufficient for unbiased estimation of causal effects and this can be done by simple nonparametric methods such as matching and subclassification. Their paper is highly influential.

To be more concrete, let D be the treatment indicator, assuming 1 for treatment and 0 otherwise. Let Y be the response of interest and X be covariates of interest. If $D = 1$, then we only observe the treatment effect Y_1 , but the control effect Y_0 is missing. Similarly we can observe the control effect Y_0 if $D = 0$. In contrast to randomized experiments in which the treatment assignment $D = 1$ or 0 is independent of Y and X , in an observational study, the propensity of $D = 1$ or 0 depends on X . Denote

$$P(D = 1|X = x) = \pi(x).$$

Rosenbaum and Rubin (1983) called $\pi(x)$ the propensity score function. A popular model in practical applications for $\pi(x)$ is the logistic regression model.

In order to assess the treatment effect, one may compare those Y 's with $D = 1$ and $D = 0$. Unfortunately, this comparison may produce biased results since conditional on $D = 1$, the distribution of Y is a biased version of the original distribution of Y . Let $f_1(y|x)$ and $f_0(y|x)$ be, respectively, the conditional densities of treatment outcome Y_1 and control outcome Y_0 . Denote the marginal density of X as $g(x)$. Note that

$$\begin{aligned} P(Y = y|D = 1) &= \frac{P(Y = y, D = 1)}{P(D = 1)} = \frac{\int P(D = 1|x) f_1(y|x) g(x) dx}{\int \int \int P(D = 1|x) f_1(y|x) g(x) dx} \\ &= \frac{\int \pi(x) f_1(y|x) g(x) dx}{\int \int \pi(x) f_1(y|x) g(x) dy dx} \\ &\neq \int f_1(y|x) g(x) dx = f_1(y), \end{aligned}$$

where $f_1(y)$ is the marginal density of treatment outcome. Similarly,

$$P(Y = y|D = 0) = \frac{\int \{1 - \pi(x)\} f_0(y|x)g(x)dx}{\int \int \{1 - \pi(x)\} f_0(y|x)g(x)dydx} \neq \int f_0(y|x)g(x)dx = f_0(y),$$

again $f_0(y)$ is the marginal control outcome density. As a consequence, the **observed** treatment outcome Y_1 and control outcome Y_0 are biased versions of $f_1(y)$ and $f_0(y)$. In general, the treatment difference $\Delta = E(Y_1) - E(Y_0)$ is not identical to the observed difference $E(Y_1|D = 1) - E(Y_0|D = 0)$. Moreover, let $h_0(x|y)$ and $h_1(x|y)$ be the conditional densities of X given Y_0 and Y_1 , respectively, then we can easily see

$$P(Y = y|D = i) = \frac{w_i(y) f_i(y)}{\int w_i(y) f_i(y) dy}, \quad i = 0, 1,$$

where

$$w_i(y) = \int h_i(x|y) \pi(x) dx, \quad i = 0, 1$$

are weight functions. The main goal in causal inference is to estimate the average treatment by removing the selection bias and utilizing the baseline covariate effectively.

Example 6 Missing data problems

Missing data is a ubiquitous problem in medical and social science studies. For example, in survey sampling, one may collect incomplete data or missing data due to cost, lack of response, or other reasons. Based on different missing patterns, Little and Rubin (2002) classified missing data in three categories: (1) Missing completely at random (MCAR); (2) Missing at random (MAR); (3) Missing not at random or nonignorable missing (MNAR).

Missing completely at random indicates that there is no relationship between whether a data point is missing and values in the data set, missing or observed, namely the missing data are just a random subset of the data. Statistical inference about the original population based on complete data only is valid.

Missing at random occurs when missing data depend on the observed quantities. During the past few decades, a large volume of literature in statistics and economics have discussed these problems. The most popular methods, among others, include likelihood based inference, imputation or multiple imputation method, inverse weighting or augmented inverse weighting approach, and calibration method, etc.

When neither MCAR nor MAR hold, the data are called missing not at random or non-ignorable. These problems are the most difficult to handle because the missing probability function and density of response variable are not separable, unless strong, unverifiable assumptions are made. As a consequence, in many cases the underlying parameters may not be identifiable.

In missing at random problems, there are two different ways to modelling the data. Let $D = 1$ or 0 be the indicator of where Y is missing or not. A commonly used prospective modelling method is to assume two parametric models

$$P(D = 1|y, x) = \pi(x, \beta), \quad f(y|x) = f(y|x, \theta).$$

Alternatively, Little (1993) suggested a retrospective method. He called it a “pattern mixture model”. Essentially, he directly modelled

$$P(X = x, Y = y|D = 1) = f_1(x, y, \gamma_1),$$

and

$$P(X = x, Y = y|D = 0) = f_0(x, y, \gamma_0),$$

separately for those complete data ($D = 1$) and incomplete data ($D = 0$). If one assumes

$$f_1(x, y) = f_0(x, y) \exp(\alpha + x\beta + y\gamma),$$

then

$$\begin{aligned} P(D = 1|x, y) &= \frac{P(D = 1)f(x, y|D = 1)}{P(D = 1)f(x, y|D = 1) + P(D = 0)f(x, y|D = 0)} \\ &= \frac{P(D = 1)f_0(x, y) \exp(\alpha + x\beta + y\gamma)}{P(D = 1)f_0(x, y) \exp(\alpha + x\beta + y\gamma) + P(D = 0)f_0(x, y)} \\ &= \frac{\exp(\alpha^* + x\beta + y\gamma)}{1 + \exp(\alpha^* + x\beta + y\gamma)}. \end{aligned}$$

In other words, if we assume a density ratio model, then the missing probability becomes a logistic regression. This shows the two methods produce the same type of modelling.

Example 7 Heckman's selection biased sampling model

Sample selection bias as a specification error was proposed by Heckman (1979). Since then, it has become an extremely popular method in handling non-ignorable missing data. Many people believed that in addition to his many other important contributions in economics and statistics, Heckman got Nobel Prize award for his sample selection bias paper.

Suppose the observed covariate data are (X_i, Z_i) , $i = 1, 2, \dots, N$. The outcome variable Y_{1i} , however, may not be observable for all individuals. Whether Y_{1i} is observable or not depends on Y_{2i} , where Y_{1i} and Y_{2i} are linked by

$$Y_{1i} = \begin{cases} X_i\beta + \epsilon_{1i}, & \text{if } Y_{2i} \geq 0 \\ \text{Not observed} & \text{if } Y_{2i} < 0 \end{cases}, \quad Y_{2i} = Z_i\gamma + \epsilon_{2i}.$$

The first equation for Y_{1i} is the ordinary linear regression equation. Define a dummy indicator variable

$$D_{2i} = 1, \quad \text{if } Y_{2i} \geq 0, \quad D_{2i} = 0, \quad \text{if } Y_{2i} < 0.$$

It is used to indicate whether the variable Y_{1i} of interest is available. In general Y_{1i} and Y_{2i} are correlated bivariate normal variables and covariates X_i and Z_i may have some overlap components.

A simple way to analyze this data set is to use complete data only ($D_{2i} = 1$). Unfortunately, this may produce biased results. In fact

$$E[Y_1|X, Y_2 > 0] = X^T \beta + E[\epsilon_1|\epsilon_2 > -Z\gamma].$$

Clearly β can be estimated consistently if and only if ϵ_1 and ϵ_2 are independent. This is so called “ignorable missing data problem”. Denote the joint density of (ϵ_1, ϵ_2) as $f(\epsilon_1, \epsilon_2, \theta)$. Using Bayes formula we can show that

$$E[\epsilon_1|\epsilon_2 > -Z\gamma] = \frac{\int_{-\infty}^{\infty} \int_{-Z_i\gamma}^{\infty} \epsilon_2 f(\epsilon_1, \epsilon_2, \theta) d\epsilon_2 d\epsilon_1}{\int_{-\infty}^{\infty} \int_{-Z_i\gamma}^{\infty} f(\epsilon_1, \epsilon_2, \theta) d\epsilon_2 d\epsilon_1} := \lambda(z\gamma; \theta),$$

where $\lambda(\cdot)$ is the so-called “Mills ratio” function, or reciprocal of a normal hazard function. Since $\lambda(z\gamma; \theta) \neq 0$, this shows that a complete data only analysis is biased. We will revisit this problem in Chap. 22.

Example 8 Vardi’s multiplicative censoring problem

In his influential *Biometrika* paper, Vardi (1989) discussed four important statistical problems, including (1) multiplicative censoring problem, (2) renewal process with incomplete renewal, (3) de-convolution problem, and (4) estimation of a monotonic density. To be brief, we will only discuss his de-convolution problem here. The other three problems will be discussed in Chaps. 25 and 26.

Suppose we observe a sample from the distribution of the sum of two independent random variables X and Y , where X has an unknown distribution F and Y has a known distribution, for example, the standard normal distribution. The goal is to estimate F non-parametrically. This is a well-known de-convolution problem. In general it would be very difficult to do so Lindsay (1995). However if the density of Y is given by

$$g(y) = \exp(y), \quad -\infty < y < 0,$$

then the nonparametric maximum likelihood estimation of F would be much easier to obtain (Groeneboom and Jongbloed 2014).

Let $Z^* = X + Y$, then

$$Z = \exp(Z^*) = \exp(X) \exp(Y) = VU.$$

It can be shown that $U = \exp(Y)$ has a uniform distribution in $(0, 1)$. Denote $f(v)$ as the density of V .

$$Z|U \sim \frac{f(z/u)}{u},$$

$$Z \sim \int_0^1 u^{-1} f(z/u) du = \int_z^\infty t^{-1} dF(t).$$

Let

$$dH(z) = \frac{zdF(z)}{\mu}, \quad \mu = \int_0^\infty zdF(z)$$

be the length biased version of F , then

$$Z \sim \frac{\bar{H}(z)}{\mu}.$$

Clearly Z can be treated as the backward time discussed in Chap. 2. Note that

$$P(Z < t) = P(UV < t, V < t) + P(UV < t, V > t) = F(t) + \int_t^\infty (t/v) dF(v) \geq F(t).$$

In other words, Z is stochastically smaller than V . This can serve as a semiparametric stochastic ordering model in the future applications.

Vardi (1989) discussed nonparametric maximum likelihood estimation of H by using an EM algorithm. As a consequence, F can be estimated by using the inverse weighting of H .

Example 9 Window censored recurrence data

Many applications, particularly in the failure, repair, and replacement of industrial components, involve recurrent events. Interval censored sampling is a very popular method to extract data. In this type of sampling, only events that occurred during this particular interval are recorded (Vardi 1982b). Below we will discuss an application of window sampling in economics study introduced by McClean and Devine (1995).

The manpower problem in the nursing service had caused big concern for nursing managers in UK. Nursing staff may withdraw from active service and return to work after a period of time. The duration of spells of absence from the profession prior to returning to active service is of main interest. Around 50,000 nurses in the period 1977–1987 database were extracted on the duration of absence, which consisted of the career histories. Any nurse withdrawing from or returning to active service during this time was recorded. On the other hand, no information was recorded on the nurses who withdrew from active service prior to 1977 and had still not returned by the end of the end data collection in 1987. This study was conducted in Northern Ireland.

There are four types of data.

- (1) Complete data, in which the start of the duration and the end of duration both occur within the sampling window.
- (2) Right censored data, in which the start of the duration is somewhere within the observation window but has not ended by the end of the window.
- (3) Left censored forward time data, in which the start of duration before the observation window but ends before the end of the observation window.

(4) Right censored forward time data, in which the start of duration before the observation window but ends after the end of the observation window. This type of data is not observable in McClean and Devine (1995)'s example.

Denote F as the distribution function for the duration of spells. The first two types of data can be treated prospective sampling data, i.e., the conventional survival with possible right censoring. The last two types of data can be treated as prevalent sampling, which suffers from selection bias. Denote the four types of data as y_{ij} , $j = 1, 2, 3, 4$, $i = 1, 2, \dots, n_j$. Then the likelihood contribution is

$$L = \left[\prod_{i=1}^{n_1} dF(y_{i1}) \right] \left[\prod_{i=1}^{n_2} \bar{F}(y_{i2}) \right] \left[\prod_{i=1}^{n_3} \frac{\bar{F}(y_{i3})}{\mu} \right] \left[\prod_{i=1}^{n_4} \frac{\int_{y_{i4}}^{\infty} \bar{F}(u) du}{\mu} \right],$$

where $\mu = \int_0^{\infty} \bar{F}(u) du$.

For McClean and Devine (1995)'s example, the type 4 data are not observable, which leads to a truncated version of type 3 data. Let l be the sampling window length, we need to consider the forward time V conditioning on $V \leq l$

$$V|V \leq l \sim \frac{\bar{F}(v)}{\mu^*}, \quad 0 < v \leq l, \quad \mu^* = \int_0^l \bar{F}(u) du.$$

Therefore the overall likelihood contribution is

$$L = \left[\prod_{i=1}^{n_1} dF(y_{i1}) \right] \left[\prod_{i=1}^{n_2} \bar{F}(y_{i2}) \right] \left[\prod_{i=1}^{n_3} \frac{\bar{F}(y_{i3})}{\mu^*} \right].$$

If one assumes a parametric model for F , then inference is straightforward. The challenge is to maximize this type of likelihood without specifying the form of F .

A more complicated example was discussed by Dewanji and Kalbfleisch (1987) for estimating the inter-occurrence distributions when n independent and identically distributed cyclic semi-Markov processes, each being ergodic, irreducible and in equilibrium, are observed over finite windows. Zhu and Wang (2012) also discussed window sampling problems with the truncation distribution not necessarily being uniform. Excellent discussions on related problems in economics applications can be found in Lancaster's (1992) book.

Example 10 Wicksell corpuscle problem

The Wicksell's corpuscle problem has a history of more than 80 years. Its name comes from the Swedish Mathematician Sven Wicksell, who constructed a mathematical model for an anatomical sample of tissues in many organs, for example, spleen, thymus and pancreas (Wicksell 1925). Since his pioneering works (1925, 1926), many applications of this model have been found in other science areas, such as astronomy, geology, metallurgy, material science, stereology, etc. Essentially, it is the study of three dimensional properties of objects or matter usually observed two-dimensionally or one-dimensionally.

Suppose that spherical particles of different radii are randomly distributed in R^3 , where the center of the spheres are distributed according to a stationary spatial Poisson process. We are interested in estimating the distribution function of the radii of the particles. Denote $f(R)$ as its density function. In practical applications, one may sample the spheres using a two-dimensional planar sample, and only the circular profiles of the spheres intersecting the plane are observed. Let r be the radii of the two-dimensional circular profiles. There are two complications. The first one is the biased sampling, and the second one is the indirect measurement. As we discussed before, spheres with larger radii are more likely to be sampled. The sampling probability is proportional to their radii. Therefore the sampled density of the spherical radii is

$$f^S(R) = \frac{Rf(R)}{\int_0^\infty Rf(R)dR}.$$

The chance of a plane will intersect the center of a sphere is small. As a consequence the second complication is that the radii of the two-dimensional circular profiles will always be smaller than the radii of the sphere being sampled.

Given a sphere with radius R , Wicksell (1925) showed that the radii r of the circular profiles has a density

$$h(r|R) = \frac{r}{R\sqrt{R^2 - r^2}}, \quad 0 \leq r \leq R.$$

As a consequence, the sampling density of r is given by

$$f^S(r) = \int_r^\infty \frac{rf^S(R)dR}{R\sqrt{R^2 - r^2}} = \frac{r}{\int_0^\infty Rf(R)dR} \int_r^\infty \frac{f(R)dR}{\sqrt{R^2 - r^2}}.$$

This is a complicated problem.

The statistical literature on estimation of the corpuscle problem is very rich. Comprehensive review of this problem is given by Chiu et al. (2013). More recent monograph on this problem can be found in Groeneboom and Jongbloed (2014). It seems very natural to solve the integral equation in $f^S(r)$ by replacing the left side with the empirical distribution function. However, this is a well known ill-posed Abel-type integral equation. Figure 3.3 gives a naive plug in estimator (a) and the maximum likelihood estimator (b) of the distribution function of spherical radii based on two-dimensional observations with sample size of $n = 50$. The true distribution of F is chosen as a uniform distribution. Clearly the behaviour of the plug in estimator is not very good. Recently, Chan and Qin (2016) discussed nonparametric maximum likelihood estimation of F when one-dimensional, two-dimensional, and three-dimensional observations are available. We will revisit this problem in Chap. 26.

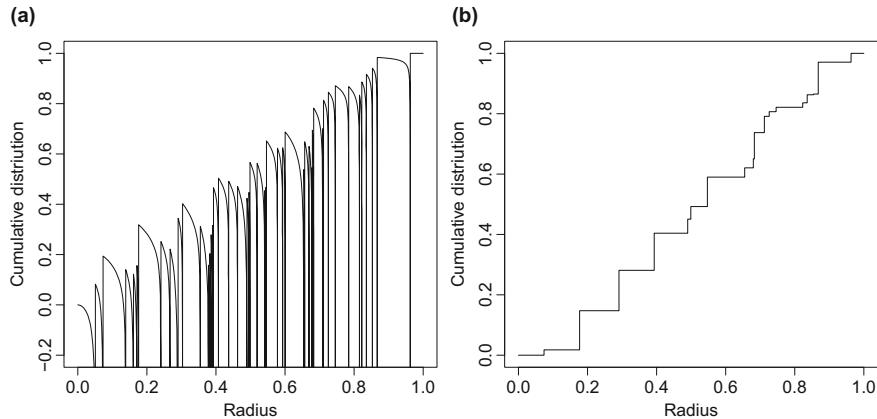


Fig. 3.3 Comparing a naive plug-in estimator (a) and the maximum likelihood estimator (b) of the distribution function of spherical radii based on two-dimensional observations with sample size 50. The sample is generated from a standard uniform distribution

Example 11 Capture-recapture

In many fields, such as biology, ecology, fishery etc., it is of great importance to know the abundance of a species or the size of a closed population (Borchers et al. 2002). Single-capture and mark-recapture experiments are widely used to collect necessary data. In mark-recapture experiments, individuals or animals from a population of interest are captured, marked, and then released. After the captured and un-captured animals are mixed, another sample is taken from the population at a later time (Chen and Lloyd 2000).

The application of the capture-recapture methods to estimation of population parameters for epidemiologic and demographic studies have been growing. For example, the U.S. Census Bureau uses the dual system to estimate the population size of the United States (Hogan 1993). The dual system estimation produces valid population estimates as long as certain assumptions based on the chosen model hold true. Boden and Ozonoff (2008) examined the reporting of nonfatal injury and illness reporting for the two most important data sources in the United States: workers' compensation data and the Bureau of Labor Statistics' annual Survey of Occupational Injuries and Illnesses. They used capture-recapture analysis to estimate the proportion of injuries reported. In software engineering, the capture-recapture model is used to estimate the number of defects in an inspected artifact. This estimate can be a valuable source of information for deciding whether or not the artifact requires a reinspection to improve the phase containment of defects.

Due to the fact that the classical estimator is known to be biased under population heterogeneity, Alho (1990a, b) introduced an estimator for the unknown population size in a dual registration process based on a logistic regression model. His model allows different capture probabilities across individuals and capture times. The probabilities are estimated from the observed data using the conditional maximum likelihood

method. One important feature of this type of data analysis is that the collected data are biased sampling of the original population since larger size animals are more likely to be captured.

Next we discuss the semiparametric approach proposed by Alho (1990a,b). For simplicity, we consider only one recapture case.

Suppose a species has N (unknown) individuals. Let

$$X_1, \dots, X_N \sim F(x),$$

where X_i is some characteristic of the i -individual, for example, animal length or weight, etc. Let D_j , $j = 1, 2$ be the capture history in the j -th capture process. $D_j = 1$ if a species is captured in the j -th occasion. Given $X_i = x$, $i = 1, \dots, N$ we may assume a probability model

$$P(D_1 = 1|x) = \pi_1(x\beta_1), \quad P(D_2 = 1|x) = \pi_2(x\beta_2),$$

where $\pi_1(\cdot)$ and $\pi_2(\cdot)$ are logistic regression models. Let

$$\pi(x) = P(D_1 + D_2 \geq 1|x) = 1 - P(D_1 + D_2 = 0|x) = 1 - \{1 - \pi_1(x)\}\{1 - \pi_2(x)\}$$

be the probability of at least one catch out of two occasions. We have made the assumption that the two capture processes are independent each other. Note that

$$P(D_1, D_2|D_1 + D_2 \geq 1, x) = \frac{\pi_1^{D_1}(x)\{1 - \pi_1(x)\}^{1-D_1}\pi_2^{D_2}(x)\{1 - \pi_2(x)\}^{1-D_2}}{\pi(x)}.$$

Denote the observed data as $(D_{i1}, D_{i2}, D_{i1}X_i, D_{i2}X_i)$, $i = 1, 2, \dots, N$. The likelihood is

$$L = \prod_{i=1}^N \left\{ \frac{\pi_1^{D_{i1}}(x_i)\{1 - \pi_1(x_i)\}^{1-D_{i1}}\pi_2^{D_{i2}}(x_i)\{1 - \pi_2(x_i)\}^{1-D_{i2}}}{\pi(x_i)} \right\}^{I\{D_{i1}+D_{i2} \geq 1\}}$$

$$\prod_{i=1}^N \{\pi(x_i)dF(x_i)\}^{I\{D_{i1}+D_{i2} = 0\}} \left[\int \{1 - \pi(x)\}dF(x) \right]^{I\{D_{i1}+D_{i2} = 0\}}.$$

If a parametric model for F is assumed, then it is habitual to perform a parametric maximum likelihood estimation. On the other hand, if the distribution form of F is left arbitrary, it would be more of a challenge (Alho 1990a,b). We discuss this case in detail in Chap. 23.

Example 12 Unequal weight sampling in survey

Unequal weight sampling in survey is a very popular method when there are heterogeneity in the population units. Comprehensive discussions on this type of sampling design can be found in many survey sampling books, for example, Cochran (1977), Sarndal et al. (1991) and Thompson (1997). Let Y_1, \dots, Y_N be a finite

population drawn from a super-population with distribution function $F(y)$. Denote $\mathcal{P} = \{Y_1, \dots, Y_N\}$. For a sampling design, a subpopulation $S \in \mathcal{P}$ is collected. The main interest is to estimate either the finite sample total $T = Y_1 + \dots + T_N$ or the mean. In general, likelihood inference in finite sample population is complicated and difficult. To illustrate this, we consider the following example.

In analyzing discovery data, the size-bias sampling occurs naturally. For example, in petroleum resource estimation, the size of a pool affects its chance of discovery. In software debugging problems, larger bugs are clearly more likely to be detected. Nair and Wang (1989) considered successive sampling discovery model in a finite population problem. The basic idea is to sample successively from a finite population with sampling probability proportional to the size and without replacement (Kaufman et al. 1975). Without loss of generality, we assume the first n individuals (y_1, \dots, y_n) are selected in the sample. The inclusion probability is

$$\frac{w(y_1)}{\sum_{i=1}^N w(y_i)} \frac{w(y_2)}{\sum_{i=1}^N w(y_i) - w(y_1)} \cdots \frac{w(y_n)}{\sum_{i=1}^N w(y_i) - \sum_{j=1}^{n-1} w(y_j)},$$

where $w(y)$ is a known function of y , in most applications $w(y) = y$. The overall likelihood contribution is

$$\frac{N!}{(N-n)!} \left\{ \prod_{i=1}^n \frac{f(y_i)w(y_i)}{b_i} \right\} E \left\{ \prod_{i=1}^n \frac{b_i}{b_i + w(Y_{n+1}) + \dots + w(Y_N)} \right\},$$

where $b_i = w(y_i) + \dots + w(y_n)$ and the expectation is with respect to Y_{n+1}, \dots, Y_N .

Even if the form of f is known up to some unknown parameter θ , maximum likelihood estimation is clumsy. We need to evaluate the expectation numerically, which is not easy. Nair and Wang (1989) used an EM algorithm to accomplish this maximization. On the other hand, Bickel et al. (1992) studied maximum likelihood estimation without making any parametric assumption on f .

Example 13 The case cohort design

Case cohort designs use a sub-sampling technique in survival data for estimating the relative risk of disease in a cohort study without collecting covariate data from the entire cohort. There were sporadic discussions on this problem in earlier works, followed by a systematic treatment given in Prentice's (1986) *Biometrika* paper. This type of study was originally designed for efficient analysis of studies in which the population size is too large to collect detailed data on all the participants, e.g., large survey studies. The basic idea is that the contribution to the likelihood for the earlier censored individuals is insignificant compared to the diseased individuals. Let $Y_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, X_i , $i = 1, 2, \dots, N$ be the possibly censored lifetime, censoring indicator and covariates. However, the covariate information is available only for those sampled individuals. At each failure point t , one may form a case and control sample as follows.

Let $\bar{F}(t|x)$ and $\bar{G}_C(t|x)$ be the survival functions of lifetime T and censoring variable C conditioning of covariate X , respectively. The conditional density of T given X is denoted as $f(t|x)$. Let $h(x)$ be the marginal density of X .

(1) Cases include all individuals died at time t .

$$X|\delta = 1, T = t \sim \frac{\bar{G}_C(t|x)f(t|x)h(x)}{\int \bar{G}_C(t|x)f(t|x)h(x)dx}.$$

(2) Controls include all individuals who are alive at least up to time t .

$$X|Y \geq t \sim \frac{\bar{G}_C(t|x)\bar{F}(t|x)h(x)}{\int \bar{G}_C(t|x)\bar{F}(t|x)h(x)dx}.$$

The case cohort design collects covariate information for all cases. However, the covariate information is collected only for randomly selected individuals. Biased sampling occurs since the probabilities of the including case covariate and control covariate are different. Define, respectively,

$$\pi_i = \begin{cases} 1, & \delta_i = 1 \\ p, & \delta_i = 0 \end{cases}$$

as the sampling weights, and sample indicator $V_i = 1$ if the i -th individual is selected and 0 otherwise.

The inverse weighting estimator

$$\sum_{i=1}^N \frac{V_i}{\pi_i} \exp(x_i \beta) I(y_i \geq t)$$

is an unbiased estimate of $\sum_{i=1}^N \exp(x_i \beta) I(y_i \geq t)$ in the Cox proportional hazards regression model, thus leading to an unbiased estimate for the hazard ratio parameter. Many adaptations have been developed since Prentice's (1986) work, leading to different versions of case cohort design. We discuss those problems in details in Chap. 24. Essentially this is a covariate missing at random problem in survival set up.

3.2 Modelling Based Selection Biased Sampling Problems

The examples discussed so far are related to the selection bias sampling. As mentioned before, biased sampling problems may occur if we use the density ratio model to link different densities together. Below are some examples. The basic idea comes from the case-control study with the logistic regression model assumption, in which

the case density and control density are bonded by the exponential tilting model (3.2.3) below.

Example 14 Biased sampling in one-way layout

Sometimes the sampling process has nothing to do with selection bias. However, biased sampling may occur if one makes a semiparametric modelling assumption.

In the classic one-way ANOVA analysis with $m = q + 1$ independent normal random samples,

$$X_{i1}, \dots, X_{in_i} \sim g_i(x), \quad i = 1, 2, \dots, m,$$

where $g_i(x)$ is the probability density of $N(\mu_i, \sigma^2)$, $i = 1, 2, \dots, m$. Holding $g_m(x)$ as a reference, the ratios are

$$\frac{g_i(x)}{g_m(x)} = \exp(\alpha_i + x\beta_i), \quad i = 1, 2, \dots, q,$$

where

$$\alpha_i = \frac{\mu_m^2 - \mu_i^2}{2\sigma^2}, \quad \beta_i = \frac{\mu_i - \mu_m}{\sigma^2}.$$

It follows that testing $H_0 : \mu_1 = \mu_2 = \dots = \mu_m$ is equivalent to testing $H'_0 : \beta_1 = \beta_2 = \dots = \beta_q = 0$.

In the conventional ANOVA analysis, the baseline $g_m(x)$ is a normal density. A direct generalization (Qin and Zhang 1997; Fokianos et al. 2001) is to assume

$$\frac{g_i(x)}{g_m(x)} = \exp(\alpha_i + h(x)\beta_i), \quad i = 1, 2, \dots, q, \quad (3.2.3)$$

where $h(x)$ is a given function of x , and the baseline density $g_m(x)$ is not specified. In this situation except for direct observations X_{m1}, \dots, X_{mn_m} from $g_m(x)$, all other X_{i1}, \dots, X_{in_i} , $i = 1, 2, \dots, q$, are indirect observations from $g_m(x)$, i.e., biased samples from $g_m(x)$ with different weight functions $\exp(\alpha_i + h(x)\beta_i)$.

Clearly, if two normal populations have different means but the same variance, then the exponential tilting model holds true with $h(x) = x$. On the other hand, if both the means and variances are different, a quadratic term will be necessary in the exponential tilting model. We can perform an ANOVA analysis without using the normal assumption. With some modifications, this method can be generalized to dealing with two-way ANOVA.

Multiple-sample exponential tilting density ratio models (3.2.3) have many applications. For example, Chen and Liu (2013) used this model to combine information for some similar conditions or experiments. In particular, they applied this density ratio model to Canadian lumber study for a better estimation of quantile functions. De Carvalho and Davison (2014) developed a semiparametric model for the situation where several multivariate extremal distributions are linked through the action of a covariate on an unspecified baseline distribution, through the density ratio model.

Due to its importance in risk assessment, the modeling of multivariate extremes has received increasing attention recently.

As an alternative approach, we may also assume

$$\frac{g_i(x)}{g_m(x)} = \psi(x\beta_i), \quad i = 1, 2, \dots, q,$$

where $g_m(x) = g_m(x, \theta)$ is a given parametric form, $\psi(\cdot)$'s are unknown monotonic nondecreasing function. In general the shape restricted inference is difficult.

Example 15 Biased sampling with application in mixture model

Finite mixture models have been used extensively in many areas. For simplicity, we assume a two components mixture model problem. Anderson (1979) first introduced the density ratio model in the mixture model below. Denote

$$X_1, \dots, X_{n_1} \sim f(x), \quad Y_1, \dots, Y_{n_2} \sim g(x), \quad Z_1, \dots, Z_{n_3} \sim \lambda f(x) + (1 - \lambda)g(x).$$

The traditional approach is to assume a parametric model for f and g and then perform maximum likelihood estimation and likelihood ratio test for λ or other parameters in f and g . It is well known that the full parametric likelihood approach is the most efficient one, though it may not be robust. A wrong conclusion may be drawn if the underlying model is misspecified.

Under the density ratio model or the exponential tilting assumption between f and g ,

$$g(x) = \exp(\alpha + x\beta)f(x),$$

where $f(x)$ is an unspecified density and α is a normalizing constant, the observed data have densities, respectively,

$$\begin{aligned} X_1, \dots, X_{n_1} &\sim f(x), \quad Y_1, \dots, Y_{n_2} \sim \exp(\alpha + x\beta)f(x), \quad Z_1, \dots, Z_{n_3} \\ &\sim \{\lambda + (1 - \lambda)\exp(\alpha + x\beta)\}f(x). \end{aligned}$$

In other words, we can treat Y_i 's and Z_i 's as biased sampling data from f with weighting functions $\exp(\alpha + x\beta)$ and $\{\lambda + (1 - \lambda)\exp(\alpha + x\beta)\}$, respectively.

As in the ANOVA example, the original problem has nothing to do with biased sampling, however, by using the exponential tilting model assumption, we end up with a biased sampling problem. In Anderson's example, X_i 's and Y_i 's are called "training samples", which are direct observations from the two components f and g , respectively. However, the Z_i 's are samples from the mixture density. Qin (1999) derived the theoretical results for Anderson's estimator. In particular, he showed that the semiparametric likelihood ratio statistic for λ has an asymptotic chi-squared distribution if the true value λ_0 satisfies $0 < \lambda_0 < 1$.

In many applications of mixture models, unfortunately it is very rare for "training samples" to be available. In this case, Anderson's (1979) approach does not work due to lack of identifiability based on one sample from the mixture model only.

Nevertheless, Leung and Qin (2006) found that Anderson's method still works if there are repeated observations. In particular, they considered a mixture model

$$(X_1, \dots, X_I) \sim \lambda \prod_{i=1}^I f(x_i) + (1 - \lambda) \prod_{i=1}^I g(x_i).$$

The joint density is

$$\left\{ \lambda + (1 - \lambda) \prod_{i=1}^I \exp(\alpha + x_i \beta) \right\} \prod_{i=1}^I f(x_i).$$

using the exponential tilting assumption. The underlying model is identifiable if $I \geq 2$. Moreover, it is possible to profile out $dF(x_i)$ to perform semiparametric maximum likelihood estimation. More related mixture model examples in genetic studies will be discussed in Chaps. 17 and 18.

Example 16 Skewed normal distribution

In probability theory and statistics, the skewed normal distribution is a continuous probability distribution that generalizes the normal distribution to allow for non-zero skewness. Azzalini (2013) published a monograph on skewed distributions. The basic idea is perturbation of a symmetric base symmetric density f_0 . The skewed density is given by

$$f(x) = 2f_0(x)G_0(w(x)),$$

where $w(-x) = -w(x)$ and G_0 is a symmetric distribution function $G_0(-x) = 1 - G_0(x)$.

By using the symmetry property, one can easily show that $f(x)$ is indeed a density. In fact, if $X \sim g_0(x)$, $Y \sim f_0(y)$ and they are independent each other, then

$$P\{X < w(Y)\} = P\{-X > -w(Y)\} = P\{-X > w(-Y)\} = P\{X > w(Y)\}.$$

Therefore,

$$1/2 = P\{X < w(Y)\} = E[G_0\{w(Y)\}] = \int G_0\{w(y)\}f_0(y)dy.$$

Denote $G(x) = G_0(w(x))$, then $G(x) \geq 0$, $G(-x) + G(x) = 1$. In the special case $w(x) = x$,

$$f(x) = 2f_0(x)G_0(x).$$

If $w(x) = 0$, then $G_0(x) = 1/2$, which implies $f(x) = f_0(x)$.

Ma et al. (2005, 2013) studied the semiparametric estimation by assuming either the form of weight function G_0 is known or the form of baseline density function f_0 is known.

It is interesting to point out that the skewed normal density can be treated as a truncated version of a normal density.

In fact, let X and Y be two independent random variables with symmetric densities $f_0(x)$ and $g_0(x)$, respectively. We observe

$$P(X = x | Y \leq X) = \frac{P(X = x, Y \leq X)}{P(Y \leq X)} = 2f_0(x)G_0(x).$$

Therefore we can apply an EM algorithm to compute the maximum likelihood estimators if parametric models for f_0 and g_0 are imposed.

There is a related selection bias sampling problem in the study of cometary orbits. It is helpful to find whether or not the directed normals to the orbits are uniformly distributed on the celestial sphere. Previous studies by statisticians have not taken the selection effects into account and have tended to reject uniformity. Jupp et al. (2003) proposed a plausible selection bias mechanism that gives rise to a one-parameter family of distributions on the sphere. Data on long-period comets are analyzed using this one-parameter family. A nonzero selection effect is detected, and its size is estimated. Subject to this selection bias effect, uniformity of the directed normals can no longer be ruled out.

Example 17 Importance sampling

When a probability density function involves a normalizing constant, theoretical integration is difficult to compute. Thus, Monte Carlo simulation is commonly used. Let

$$f(x) = \frac{\phi(x)}{c}, \quad c = \int \phi(y)dy,$$

where $\phi(x)$ is a given non-negative function. More generally we have observations

$$X_1, \dots, X_n \sim f(x, \theta) = \frac{\phi(x, \theta)}{\int \phi(x, \theta)dx} := \frac{\phi(x, \theta)}{c(\theta)},$$

and we are interested in estimating θ .

The basic concept of importance sampling is the fact that

$$\int \phi(x)dx = \int \frac{\phi(x)}{g(x)}g(x)dx = E_g[\phi(Y)/g(Y)],$$

where $Y \sim g(y)$, and g is some specified density function. Note that

$$\text{Var}[\phi(Y)/g(Y)] = \int \phi^2(y)/g(y)dy - c^2.$$

Using Cauchy's inequality

$$c^2 = \{\int \phi(y)dy\}^2 = \{\int \sqrt{g(y)}\phi(y)/\sqrt{g(y)}dy\}^2 \leq \{\int \phi^2(y)/g(y)dy\}(\int g(y)dy).$$

Therefore the optimal choice is $g(y) = f(y)$. Unfortunately this is not feasible.

Let us return to the estimation problem. A good strategy is to use an initial guess g as close to f as possible. We then generate a Monte Carlo sample

$$y_1, \dots, y_N \sim g(y).$$

We then have a two sample problem with density ratio being linked by

$$\begin{aligned} g(x) &= f(x)g(x)c(\theta)/\phi(x, \theta) = f(x)\exp\{\alpha + \log\psi(x, \theta)\}, \\ \psi(x, \theta) &= \log g(x) - \log\phi(x, \theta). \end{aligned}$$

Note that the observed data sample size n is fixed. However, the Monte Carlo sample size N can be made as large as possible. Suppose there exist some known functions $\mu(x)$ such that the integral $\int \mu(y)g(y)dy = E[\mu(Y)]$ can be easily evaluated, for example $E[\mu(Y)] = a$. This information can be used to improve estimation of θ by imposing this constraint in the stage of maximizing out $dF(x)$. We discuss this in Sect. 11.6.

Example 18 Density ratio model with continuous covariate

So far, we have studied exponential tilting models with multiple group data. A natural generalization is the regression exponential tilting model by allowing continuous covariates.

Consider a generalized linear model with conditional density function

$$f(y_i|x_i; \theta_i, \phi) = \exp[a(\phi)\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)], \quad (3.2.4)$$

where $\theta_i = U(\beta^T x_i)$, U is a known link function. This model was proposed by Nelder and Wedderburn (1972). It is well known that

$$E(Y_i|x_i) = b'(U(\beta^T x_i)), \quad Var(Y_i|x_i) = b''(U(\beta^T x_i))/a(\phi),$$

where $b'(\theta)$ and $b''(\theta)$ denote the first and second derivatives of b with respect to θ . With the assumption on the forms of a , b , c , Nelder and Wedderburn (1972) obtained the maximum likelihood estimates for β and ϕ . In the estimating function setup, forms for $E(Y_i|x_i)$ and $Var(Y_i|x_i)$ are assumed directly. Sometimes it is helpful to estimate β or the direction of β whereas the forms of $b(\cdot)$ and $c(\cdot)$ are of less interest. If the form of $\exp\{c(\cdot)\}$ (baseline density) is not specified, then the form of $b(\cdot)$ is unknown. It is clear that this is equivalent to the baseline density $f(y|x=0) = \text{const} * \exp\{c(y, \phi)\}$ being unknown. Therefore the conditional expectation is a unknown function of $\beta^T x$.

Even though a full parametric model is specified, the generalized linear model with unspecified baseline density may also arise due to non-ignorable missing data or selection biased sampling. Let $f(y|x, \theta)$ be a fully specified conditional density

of Y given X . Given Y , define an indicator random variable R . $R(y) = 1$ if Y is observed and 0 otherwise. Let

$$\pi(y) = P(R = 1|Y = y),$$

that is, the response Y is observable with probability $\pi(y)$. By using Bayes' formula, the probability density function of the observed response conditioning on covariate X is

$$f(y|x, R = 1) = \frac{\pi(y)f(y|x, \theta)}{\int \pi(y)f(y|x, \theta)dy}.$$

Comparing with the generalized linear, $\pi(y)$ and $\{\int \pi(y)f(y|x, \theta)dy\}^{-1}$ play the roles of $\exp\{c(y, \phi)\}$ and $\exp\{b(x)\}$ in (3.2.4), respectively.

Back to the partially specified generalized linear model (3.2.4), Kalbfleisch (1978) pointed out that it is possible to test the hypothesis that $\beta = 0$ without specifying the forms of $b(\cdot)$ and $c(\cdot)$. Specifically, suppose that data $(y_1, x_1), \dots, (y_n, x_n)$ are available from model (3.2.4). If Y is a continuous variable, the probability of the observing pairs (y_i, x_i) , $i = 1, 2, \dots, n$ given the order statistics $y_{(1)}, \dots, y_{(n)}$ and covariates (x_1, \dots, x_n) is easily seen to be

$$P(y|y_{(.)}; (x_1, \dots, x_n)) = \frac{\exp\{a(\phi) \sum_{l=1}^n y_l U(\beta^T x_l)\}}{\sum_{j \in P} \exp\{a(\phi) \sum_{l=1}^n y_{jl} U(\beta^T x_{jl})\}}, \quad (3.2.5)$$

where $j = (j_1, \dots, j_n)$ and P is the set of permutations of the integers $\{1, 2, \dots, n\}$. The functions $b(\cdot)$ and $c(\cdot)$ are cancelled out in the conditional likelihood of β given the order statistics. When Y is dichotomous, inferences should use $\Pr(y|y_+)$ where $y_+ = \sum_i y_i$. In the absence of knowledge on the form of c and in particular on b , it is essential to address the interpretation of β under this circumstance. With U known, note that for two individuals with X values of x_1 and x_0 , respectively,

$$\frac{f(y|x_1)/f(y^*|x_1)}{f(y|x_0)/f(y^*|x_0)} = \exp\{a(\phi)(y - y^*)(U(\beta^T x_1) - U(\beta^T x_0))\}, \quad (3.2.6)$$

where y and y^* are two different Y values. Thus β characterizes the effect of X through U on the “odds” of Y through its probability (density) function. This is very similar to the Cox proportional hazards model (Cox 1972, 1975) in which the hazards ratio has a known parametric form.

The well-known invariance property of the odds ratio (Prentice and Pyke 1979) implies that

$$\frac{f(y|x_1)/f(y^*|x_1)}{f(y|x_0)/f(y^*|x_0)} = \frac{f(x_1|y)/f(x_1|y^*)}{f(x_0|y)/f(x_0|y^*)}.$$

When Y is a categorical variable, this equivalence implies that knowledge of either $f(y|x)$, from a prospective study, or $f(x|y)$, from a case-control study, suffices to identify the odds ratio (Prentice and Pyke 1979). The conditional inference method

discussed above can be carried through by changing the roles of X and Y . When Y and some of the covariates X are continuous, complications regarding inference arises for this problem considered above due to the fact that only a small number of subjects are expected to share the same X values. Meanwhile, the conditional likelihood proposed by Kalbfleisch, which is applicable for either categorical or discrete X , becomes computationally prohibitive even when the sample size n is only moderate. For example, with $n = 10$, there are 3,628,800 terms appearing in the denominator. It seems impractical to use his approach even though it is possible to test $\beta = 0$ since in this case it can be simplified.

Qin and Liang (1999) and Liang and Qin (2000) directly modelled the general odds ratio $[f(y|x_1)/f(y^*|x_1)][f(y|x_0)/f(y^*|x_0)]^{-1} = R(y, y^*, x, x^*; \theta)$. They formed a pairwise conditional likelihood $L_{ij}(\theta)$ by applying Kalbfleisch's (1978) conditional likelihood to any two observations $(y_i, x_i), (y_j, x_j)$, $i \neq j$. The overall pairwise likelihood is $\prod_{i < j} L_{ij}(\theta)$.

Luo and Tsai (2012) and Diao et al. (2012) used the profile likelihood approach to estimate the baseline distribution function. Interestingly, the later paper found that the pairwise or triple-wise conditional likelihood approach and the profile likelihood method produced very similar results for β estimation.

So far we have listed many models where biased sample phenomenon exists in different fields. In the subsequent chapters we will develop statistical methodologies to solve those problems.

Chapter 4

Brief Review of Parametric Likelihood Inferences

Maximum likelihood estimation (MLE) under regular conditions can be found in most statistical books. In non-regular cases, however, it involves all kinds of problems, such as solution on the boundary of parameter space, multiple roots, non-existence, inconsistency in the presence of many incidental parameters, etc. Under strong regularity conditions, the consistency and asymptotic normality of the maximum likelihood estimate can be found in classical statistical inference books, for example, Lehmann and Casella (1998). The maximum likelihood estimator can be treated as a special case of M-estimators. The definition of M-estimators was motivated by robust statistics. Pioneering works on M-estimation were discussed by Huber (1967). Under weaker regularity conditions given by Huber's (1967), consistency and asymptotic normality were established for the general M-estimate derived by minimizing an objective function. In particular, in Huber's (1967) paper (i) the data distribution is not required to be a member of the parametric family used in the MLE, and (ii) the regularity conditions do not involve second and higher derivatives of the likelihood function. More discussions on M-estimation using advanced empirical process theories can be found in the books by Pollard (1984) and Van der Vaart and Wellner (1992). The works by Niemiro (1992) and Hjort and Pollard (1997) are instrumental for general M-estimators defined by minimizing a convex function that is not necessarily smooth.

Let Θ be a parameter space. Denote the observed data as

$$X_1, \dots, X_n \sim i.i.d. f(x, \theta_0), \quad \theta_0 \in \Theta,$$

where *i.i.d.* is the abbreviation of “independent and identically distributed”. The maximum likelihood estimator is defined as $\hat{\theta}$ such that

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n f(x_i, \theta) = \operatorname{argmax}_{\theta \in \Theta} \sum_{i=1}^n \log f(x_i, \theta).$$

The basic tool for the consistent proof of $\hat{\theta}$ is the Law of Large Numbers and the Kullback–Leibler divergence inequality.

4.1 Kullback–Leibler Information and Entropy Concepts

For two probability measures P and Q , the Kullback–Leibler divergence is defined as

$$KL(P, Q) = E_P \left[\log \frac{dQ}{dP} \right] = \int \left\{ \log \frac{dQ}{dP} \right\} dP.$$

Since $\log(x)$ is a convex function, by Jensen's inequality

$$KL(P, Q) \geq \log \int \frac{dQ}{dP} dP = \log 1 = 0,$$

with equality if and only if P and Q are the same. This is a fundamental result. It has been used to show the consistency of the maximum likelihood estimate not only in full parametric models, but also in semiparametric and non-parametric models.

Under the following regularity conditions, it can be shown that the maximum likelihood estimator is consistent.

Regularity Conditions

- (1) The parameter space Θ is compact.
- (2) $\log f(x, \theta)$ is continuous with respect to θ for all x .
- (3) There is an integrable function $\psi(x)$ such that $|\log f(x, \theta)| \leq \psi(x)$ for all $\theta \in \Theta$ and $x \in \Omega$.

In fact, let $\hat{\theta}_n$ be the MLE in Θ based on the observed data x_1, \dots, x_n . Since Θ is compact, there exists a subsequence $\hat{\theta}_{n_k}$ such that $\hat{\theta}_{n_k} \rightarrow \theta^* \in \Theta$. Due to the definition of maximum likelihood estimation,

$$n_k^{-1} \sum_{i=1}^{n_k} \log f(x_i, \hat{\theta}_{n_k}) \geq n_k^{-1} \sum_{i=1}^{n_k} \log f(x_i, \theta_0),$$

where θ_0 is the true value of θ . Under the specified regularity conditions above,

$$E_{\theta_0} \{ \log f(X, \theta^*) \} \geq E_{\theta_0} \{ \log f(X, \theta_0) \}.$$

Therefore the only possible case is $\theta^* = \theta_0$.

In general, the Kullback–Leibler divergence is not a distance function since is not symmetric. A commonly used symmetric metric is the Hellinger distance. For two densities f_0 and f_1 , the Hellinger distance is defined as

$$H(f_0, f_1) = \sqrt{\int \{f_0^{1/2}(x) - f_1^{1/2}(x)\}^2 dx}.$$

Using the fact that $1 - x \leq -\log(x)$, $x > 0$ and Jensen's inequality, easily one can show that

$$\begin{aligned} H^2(f_0, f_1) &= 2(1 - \int \sqrt{f_0/f_1} dx) \leq 2 \log \left[E_{f_1} \left\{ \sqrt{f_1/f_0} \right\} \right] \\ &= 2E_{f_1} [\log \sqrt{f_1/f_0}] = KL(f_0, f_1). \end{aligned}$$

Beran (1997) discussed minimum Hellinger distance estimate based on the robust consideration. In general, the maximum likelihood estimate and minimum Hellinger distance estimate are asymptotically equivalent. However, it is not so convenient to use the minimum Hellinger distance estimate since it involves using a kernel method to estimate the underlying density. The choice of window size is not easy.

It is well known that under some regularity conditions such as those given in Lehmann and Casella (1998)'s book, in distribution the maximum likelihood estimate satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, \sigma^2), \quad \sigma^2 = 1/E[\partial \log f / \partial \theta]^2,$$

where the variance is the inverse of Fisher information.

Entropy Definition

Entropy is a very important concept in probability and information theory. It is a measure of unpredictability of information, or randomness of a random variable or a probability distribution. The larger the entropy is, the less predictable the underlying random variable is. Consider a discrete probability distribution p on a countable set $\{x_1, x_2, \dots\}$ with probability masses $p_i = p(x_i)$, $i = 1, 2, \dots$. The entropy of p is defined as

$$H(p) = - \sum_{i=1}^{\infty} p_i \log p_i.$$

It is a convention to define $0 \log 0 = 0$. Since $0 \leq p_i \leq 1$, it can be shown that $H(p) \geq 0$.

For a continuous probability density function $p(x)$, the entropy is defined by

$$H(p) = - \int p(x) \log p(x) dx.$$

In contrast to the discrete variable entropy, the entropy for continuous random variable may be infinitely large, negative or positive.

Entropy of the Discrete Uniform Distribution

Suppose $p(x_i)$, $i = 1, 2, \dots, n$ is a discrete probability mass function on $\{x_1, \dots, x_n\}$. First we will show that the uniform probability mass function on $\{x_1, \dots, x_n\}$ has the maximum entropy, which turns out to be the least predictable distribution.

In fact, define

$$\psi = - \sum_{i=1}^n p_i \log p_i + \lambda(\sum_{i=1}^n p_i - 1).$$

From

$$\frac{\partial \psi}{\partial p_i} = -(\log p_i + 1) + \lambda = 0,$$

we have

$$p_i = \exp(\lambda - 1).$$

Using the fact $\sum_{i=1}^n p_i = 1$, we have $p_i = 1/n$, $i = 1, 2, \dots, n$. In other words, the uniform distribution has the maximum entropy. It is a distribution with the least informative prior or least predictable. Maximizing entropy minimizes the amount of prior information built into a distribution. Many physical systems tend to move towards maximal entropy configurations.

On the other hand, if $p_j = 1$ for some j , $j = 1, 2, \dots, n$, then the entropy is $-1 \log 1 = 0$. In other words, if the underlying distribution concentrates on one point, say, x_j , then it has the least amount of variation or least amount of entropy, or easiest predictable.

Entropy of the Normal Density

It is easy to find the entropy for the normal random variable $X \sim N(\mu, \sigma^2)$.

$$\begin{aligned} H(p) &= - \int \frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(x-\mu)^2/\sigma^2) (-0.5 \log(2\pi\sigma^2) - 0.5(x-\mu)^2/\sigma^2) \\ &= 0.5\{1 + \log(2\pi\sigma^2)\}. \end{aligned}$$

We have pointed out earlier that entropy measures the variation of a random variable or distribution. Indeed in the normal case, we know that σ^2 is the variance. The entropy of a normal distribution is an increasing function of σ^2 and is independent of the location of the underlying normal distribution. The larger the σ^2 is, the larger the entropy is. In general for a given parametric family, the maximum entropy could be unbounded if there is no restriction on the underlying parameters.

Sometimes, we have no knowledge on the underlying distribution except for its first or second moment. We are going to explore the maximum entropy family such that it achieves maximum under the first two moment constraints.

Maximum Entropy Under Moment Constraints

First we consider a discrete random variable example.

Suppose $P(X = i) = p_i$, $i = 1, 2, 3$. We have a prior information that the mean of X is 2.5. The Lagrangian is

$$L = \max_{p_i} [- \sum_{i=1}^3 p_i \log p_i - (\lambda_0 - 1)(\sum_{i=1}^3 p_i - 1) - \lambda_1(\sum_{i=1}^3 i p_i - 2.5)].$$

First order conditions yield

$$p_i = \exp(-\lambda_0 - i\lambda_1), \quad \lambda_0 = 2.987, \quad \lambda_1 = -0.834.$$

The maximum entropy distribution under the mean constraint is

$$p_1 = 0.116, \quad p_2 = 0.268, \quad p_3 = 0.616.$$

Next, we consider a continuous distribution case. Suppose $p(x)$ is a density for a continuous random variable in $(-\infty, \infty)$. For fixed μ and σ^2 , if $p(x)$ satisfies $\int xp(x)dx = \mu$ and $\int(x - \mu)^2 p(x)dx = \sigma^2$, then the normal density with mean μ and variance σ^2 achieves the maximum entropy.

In fact, let $\phi(x)$ be the normal density with mean μ variance σ^2 . Using Kullback–Leibler information inequality

$$\begin{aligned} - \int p(x) \log p(x) &\leq - \int p(x) \log \phi(x) \\ &= \int p(x)[0.5 \log(2\pi\sigma^2) + 0.5(x - \mu)^2/\sigma^2]dx = 0.5\{1 + \log(2\pi\sigma^2)\} \end{aligned}$$

since $p(x)$ has variance σ^2 . This inequality becomes equality if and only if $p(x) = \phi(x)$.

Exercise Use a Lagrange multiplier method to show the normal density achieves the maximum entropy under first and second moment constraints.

A direct application of this result is the heuristic understanding of the Central Limit Theorem. Let X_1, \dots, X_n be i.i.d. random variables with mean zero and variance 1. Then the Central Limit Theorem tells us that as $n \rightarrow \infty$, $Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \rightarrow N(0, 1)$ in distribution. Note that Y_n has mean 0 and variance 1, it gets the largest variation (or least predictable) due to the summation of a large number of random variables X_i 's when n increases. It is reasonable to assume that the limiting distribution should follow a standard normal distribution since it has the largest uncertainty or entropy under the first two moment constraints.

Exercise If $p(x)$ is a density on R^+ such that it has fixed first moment. Then the exponential distribution has the maximum entropy.

4.2 Issues in Maximum Likelihood Estimation

It is well known that maximum likelihood method is the most efficient one. However this method may not be robust when the postulated parametric model is incorrect. As an alternative, the estimating equation approach is robust. Below we discuss different issues in maximum likelihood estimation.

Robust Confidence Intervals for Maximum Likelihood Estimators

Let $g(X, \theta)$ be a given function of random variable X and an unknown parameter θ . An estimating equation $g(X, \theta)$ is called unbiased if it has zero mean, i.e.,

$$E[g(X, \theta)] = 0.$$

Under some regularity conditions, the asymptotic variance of the estimator, solving the empirical version $n^{-1} \sum_{i=1}^n g(x_i, \theta) = 0$, using n iid observations x_1, \dots, x_n , is

$$E^{-1}[\partial g/\partial\theta]E[gg^T]E^{-1}[\partial g/\partial\theta].$$

In general, one needs to estimate both expectations for the variance estimation. If $g = \partial \log f(x, \theta)/\partial\theta$ is the score, however, the asymptotic variance can be simplified as $[E\{\partial \log f(x, \theta)/\partial\theta\}^2]^{-1}$ due to the Fisher information identity

$$E[\partial \log f(x, \theta)/\partial\theta]^2 = -E[\partial^2 \log f(x, \theta)/\partial\theta^2].$$

Nevertheless, Royall (1986) suggested that even if g is the score estimating equation, it would be more robust to use the sandwich variance estimate

$$\hat{V}_2 = n \sum_{i=1}^n \{\partial \log f(x_i, \hat{\theta})/\partial\theta\}^2 / \hat{I}^2(\hat{\theta}), \quad \hat{I}(\hat{\theta}) = - \sum_{i=1}^n \frac{\partial^2 \log f(x_i, \theta)}{\partial\theta^2} \Big|_{\theta=\hat{\theta}}.$$

Example 1 If y_1, \dots, y_n are iid observations from a binomial model

$$P(Y = y) = \binom{k}{y} \theta^y (1-\theta)^{k-y}, \quad y = 0, 1, 2, \dots, k.$$

The maximum likelihood estimator is $\hat{\theta} = \bar{y}/k$, where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. Two variance estimators of $n^{1/2}(\hat{\theta} - \theta)$ are

$$\hat{V}_1 = n/I(\hat{\theta}) = (\bar{y}/k)[1 - \bar{y}/k]/k, \quad \hat{V}_2 = \sum_{i=1}^n (y_i - \bar{y})^2/nk^2.$$

If the binomial model is wrong and Y is actually hypergeometric, say, with

$$P(Y = y) = \binom{N\theta}{y} \binom{N(1-\theta)}{k-y} / \binom{N}{k}.$$

Then with probability one, $\hat{V}_2 \rightarrow [\theta(1-\theta)/k](N-k)/(N-1)$, the correct value, but $I(\hat{\theta}) \rightarrow \theta(1-\theta)/k$.

Non-regular Examples

Under some regularity conditions, the log-likelihood is well behaved and produces a root- n consistent estimator. Moreover the MLE achieves the information lower bound. However, in non-regular cases, the MLE is associated with problems such as multiple roots, boundary problem, inconsistency, etc. Below we give a few examples.

Example 1 The multiple roots problem for the Cauchy location family.

Let x_1, \dots, x_n be i.i.d. observations from a Cauchy distribution with density

$$f(x, \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \quad -\infty < x < \infty.$$

The score equation is

$$2 \sum_{i=1}^n \frac{x_i - \theta}{1 + (x_i - \theta)^2} = 0,$$

or

$$0 = \sum_{i=1}^n (x_i - \theta) \prod_{j \neq i} \{1 + (x_j - \theta)^2\}.$$

This is a polynomial in θ of degree $2(n-1) + 1 = 2n-1$, which can have as many as $2n-1$ roots. Let $\tilde{\theta}$ be the median of $x_i, i = 1, 2, \dots, n$. Clearly it is a consistent estimator. In fact

$$\sqrt{n}(\tilde{\theta} - \theta) \rightarrow N(0, \pi^2/4).$$

However this estimator may not be efficient. The one-step adjusted estimator is

$$\begin{aligned} \hat{\theta} &= \tilde{\theta} + I^{-1}\{\tilde{\theta}\}(n^{-1} \sum_{i=1}^n \partial \log f(x_i, \tilde{\theta}) / \partial \theta) \\ &= \tilde{\theta} + 4n^{-1} \sum_{i=1}^n \frac{x_i - \tilde{\theta}}{1 + (x_i - \tilde{\theta})^2}. \end{aligned}$$

Exercise Show that the one-step adjusted estimator is efficient.

The Cauchy density with location and scale parameters is given by

$$f(x, \theta, \sigma) = \frac{1}{\pi \sigma} \frac{1}{1 + (x - \theta)^2 / \sigma^2}, \quad -\infty < x < \infty.$$

Copas (1975) found an interesting result on the roots of score estimating equations. Although the likelihood for the median of a Cauchy distribution with known scale is often multimodal, the joint likelihood for both location and scale parameters has exactly one point of maximum and at most one stationary point. The maximized likelihood function is therefore unimodal.

Next, we discuss two applications of the **information identity test**.

1. A criterion for the global maximum likelihood estimate

How do we find the global maximum likelihood estimator if the maximum likelihood estimating equation has multiple roots? Gan and Jiang (1999) gave simple necessary and sufficient conditions for the consistency and asymptotic optimality for the root of a score equation. The conclusion is that a global maximizer should satisfy

$$E \left\{ \left(\frac{\partial \ell}{\partial \theta} \right)^2 + \frac{\partial^2 \ell}{\partial \theta^2} \right\} = 0. \quad (4.2.1)$$

This is a well known information identity for the maximum likelihood score. In general, the quasi-likelihoods discussed later do not inherit this property.

Since only a genuine log-likelihood posses the information identity property, it has been used widely in econometric literature for model mis-specification test, see, for example, White (1982). This test statistic can also be used in stratified data analyses where one is interested in testing whether the regression coefficients are homogeneous across different strata.

2. Test homogeneous

To test the homogeneity of nuisance parameters from many strata in a mixed model setting, Liang (1987) studied a locally most powerful test. More specifically he considered a random effects model

$$Y_i \sim \int f_i(y_i, \beta, \alpha + \theta^{1/2} v) dG(v),$$

where the form of f_i is known but the distribution $G(v)$ of random effects is not specified except for the 0 mean constraint, and i is a stratum index. Clearly if $\theta = 0$, then the density is homogeneous across strata. Let

$$\ell = \sum_{i=1}^n \log \int f_i(y_i, \beta, \alpha + \theta^{1/2} v) dG(v)$$

be the log-likelihood. One is interested in testing $H_0 : \theta = 0$, i.e., there is no random effects. Using L'Hospital's rule, one has the score statistic

$$S = \sum_{i=1}^n \frac{\partial f_i / \partial \theta^2}{f_i^2} = \sum_{i=1}^n \left(\frac{\partial \log f_i}{\partial \theta} \right)^2 + \left(\frac{\partial^2 \log f_i}{\partial \theta^2} \right).$$

This is exactly the information matrix test.

Example 2 Inference for a uniform distribution.

Suppose the observe data come from a uniform distribution,

$$X_1, \dots, X_n \sim U(0, \theta).$$

The likelihood is

$$L = \left(\frac{1}{\theta}\right)^n, \quad 0 < \theta \leq \max(X_1, \dots, X_n).$$

Clearly the maximum likelihood estimate is $\hat{\theta} = \max(X_1, \dots, X_n)$. Note that $\hat{\theta} \leq \theta$. For any $t > 0$

$$P\{n(\hat{\theta} - \theta) < -t\} = \left(1 - \frac{t}{n\theta}\right)^n \rightarrow \exp(-t/\theta).$$

In this example, the convergence rate is $O(1/n)$ and the limiting distribution is the exponential distribution instead of the conventional normal distribution. This example shows that the boundary parameter problem may have a faster convergence rate than the conventional root- n rate.

Example 3 Inference in a mixture model.

Suppose the observed data come from a mixture model

$$X_1, \dots, X_n \sim \lambda f_1(x) + (1 - \lambda) f_2(x),$$

where f_1 and f_2 are two known densities. We are interested in testing $H_0 : \lambda = 1$. The log-likelihood is

$$\ell = \sum_{i=1}^n \log[\lambda f_1(x_i) + (1 - \lambda) f_2(x_i)].$$

The first derivative of this log-likelihood is

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^n \frac{f_1(x_i) - f_2(x_i)}{\lambda f_1(x_i) + (1 - \lambda) f_2(x_i)}.$$

The second derivative satisfies

$$\frac{\partial^2 \ell}{\partial \lambda^2} = - \sum_{i=1}^n \frac{\{f_1(x_i) - f_2(x_i)\}^2}{\lambda \{f_1(x_i) + (1 - \lambda) f_2(x_i)\}^2} < 0.$$

Therefore $\partial \ell / \partial \lambda$ is a decreasing function in $(0, 1)$. If $\partial \ell(1) / \partial \lambda \geq 0$, then $\partial \ell / \partial \lambda \geq 0$ in $(0, 1)$. As a result, $\ell(\lambda)$ is increasing. The MLE is $\lambda = 1$. Otherwise the MLE is attained in the interior of $(0, 1)$. Note that

$$\frac{1}{\sqrt{n}} \frac{\partial \ell(1)}{\partial \lambda} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (1 - f_2(x_i)/f_1(x_i)) \rightarrow N(0, \sigma^2),$$

and

$$P(\partial\ell(1)/\partial\lambda > 0) \rightarrow 1/2.$$

It can be shown that the likelihood ratio statistic is

$$R(1) = 2 \max_{0 \leq \lambda \leq 1} [\ell(\lambda) - \ell(1)] \rightarrow 0.5\chi^2(0) + 0.5\chi^2(1),$$

which is different from the regular parameter case.

Example 4 Consider another mixture model

$$X \sim \pi N(0, 1) + (1 - \pi)N(\mu, 1),$$

where both π and μ are unknown. We are interested in testing $X \sim N(0, 1)$. This is equivalent to testing either $\pi = 1$ or $\mu = 0$. Note that if $\pi = 1$, then μ disappears in the likelihood. Similarly, if $\mu = 0$, then π disappears in the likelihood again. Hartigan (1985) showed that the likelihood ratio statistic for testing $H_0 : \pi = 1$ or $\mu = 0$ fails to converge.

In a more general case, if

$$X_1, \dots, X_n \sim \pi N(\mu_1, \sigma_1^2) + (1 - \pi)N(\mu_2, \sigma_2^2),$$

it can be shown that if one chooses $\hat{\mu}_1 = X_1$ then the likelihood is unbounded when $\sigma_1 \rightarrow 0$. Thus the likelihood function is unbounded and there are many maximum likelihood estimators that are clearly not consistent. However, there exists a consistent and asymptotically efficient sequence of roots of the score equations. One may use the modified one step score estimator to achieve efficiency.

Interested readers may refer to Chen (2016) on the proofs of consistency of the MLE under mixture models.

Exercise Suppose we only observe $X_i = \min(X_{1i}, X_{2i})$, $i = 1, 2, \dots, n$, where X_{1i} and X_{2i} are independent and

$$X_{1i} \sim N(\mu_1, \sigma_1^2), \quad X_{2i} \sim N(\mu_2, \sigma_2^2).$$

In this example the global MLE is not consistent. The likelihood is

$$L = \prod_{i=1}^n \left[\sigma_1^{-1} \phi \left(\frac{x_{1i} - \mu_1}{\sigma_1} \right) \left\{ 1 - \Phi \left(\frac{x_{1i} - \mu_2}{\sigma_2} \right) \right\} + \sigma_2^{-1} \phi \left(\frac{x_{2i} - \mu_2}{\sigma_2} \right) \left\{ 1 - \Phi \left(\frac{x_{2i} - \mu_1}{\sigma_1} \right) \right\} \right].$$

Using the facts

$$\sigma^{-1} \phi \left(\frac{x - \mu}{\sigma} \right) = \begin{cases} 0, & \text{if } \mu = x \\ \infty & \text{if } \mu \neq x \end{cases}, \quad \Phi \left(\frac{x - \mu}{\sigma} \right) = \begin{cases} 0, & \text{if } x < \mu \\ 1/2 & \text{if } x = \mu \\ 1, & \text{if } x > \mu. \end{cases}$$

show that the likelihood function diverges to ∞ when μ_2 and σ_2 are fixed, $\hat{\mu}_1 = \max_{1 \leq i \leq n} x_i$ and σ_1 converges to zero. Therefore the global maximum likelihood estimator is inconsistent.

Intuitively if μ_1 is much smaller than μ_2 , then the observed X_i 's are most likely from X_{1i} 's. Therefore there is almost no information on μ_2 based on observed X_i 's. In Chap. 18 we will discuss unordered pairs where only $\min(X_{1i}, X_{2i})$, $\max(X_{1i}, X_{2i})$, $i = 1, 2, \dots, n$ are available. We show that it is possible to estimate the underlying distribution functions even without the normal distribution assumption.

Parameter Transformation

Suppose X_1, \dots, X_n are iid observations from $f(x, \theta)$. We are interested in constructing a confidence interval for $g(\theta)$. Denote $\hat{\theta}$ as the maximum likelihood estimate. Using Taylor's expansion, we have

$$g(\hat{\theta}) - g(\theta) = g'(\theta)(\hat{\theta} - \theta) + o_p(n^{-1/2}).$$

Moreover from

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \sigma^2),$$

where σ^2 is the inverse Fisher information, it can be shown

$$\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} \rightarrow N(0, \sigma_g^2), \quad \sigma_g^2 = (g'(\theta))^2 \sigma^2.$$

A 95% confidence interval for $g(\theta)$ is $(g(\hat{\theta}) - 1.96\hat{\sigma}_g, g(\hat{\theta}) + 1.96\hat{\sigma}_g)$.

Consider an example $g(\theta) = \exp(\theta)$. Then a 95% confidence interval for $\exp(\theta)$ is

$$(\exp(\hat{\theta}) - 1.96 \exp(\hat{\theta})\hat{\sigma}, \exp(\hat{\theta}) + 1.96 \exp(\hat{\theta})\hat{\sigma}).$$

As an alternative, a better confidence interval is $(\exp(\hat{\theta} - 1.96\hat{\sigma}), \exp(\hat{\theta} + 1.96\hat{\sigma}))$.

To achieve the accuracy, Sprott (1973) approximated the likelihood ratio statistic using re-parametrization of the underlying parameter.

Consider a simple example for constructing a confidence interval for a binomial mean. The log likelihood is

$$\ell = k \log \theta + (n - k) \log(1 - \theta).$$

Let $\hat{\theta} = k/n$ be the maximum likelihood estimator. The likelihood ratio statistic $R(\theta) = 2[\ell(\theta) - \ell(\hat{\theta})]$ can be expanded up to the quadratic term to get a chi-squared approximation. The accuracy of this approximation depends on how quickly the third and higher order terms go to zero. The first three derivatives are

$$\frac{\partial \ell}{\partial \theta} = \frac{k}{\theta} - \frac{n - k}{1 - \theta}, \quad \frac{\partial^2 \ell}{\partial \theta^2} = -\frac{k}{\theta^2} - \frac{n - k}{(1 - \theta)^2}, \quad \frac{\partial^3 \ell}{\partial \theta^3} = \frac{2k}{\theta^3} - \frac{2(n - k)}{(1 - \theta)^3}.$$

Let

$$I = -E[\partial^2 \ell / \partial \theta^2] = n / \{\theta(1-\theta)\}$$

be the Fisher information.

$$E\left[\frac{\partial^3 \ell}{\partial \theta^3}\right] = 2n \frac{1-2\theta}{\theta^2(1-\theta)^2}.$$

Sprott (1973) used $|\Delta(\theta)|$ to measure the deviation from normality, where $\Delta(\theta)$ is given by

$$\Delta(\theta) = \frac{E[\partial^3 \ell / \partial \theta^3]}{I^{3/2}}.$$

Easily we can find

$$|\Delta(\theta)| = 2|1-2\theta|\{\theta(1-\theta)\}^{-1/2}n^{-1/2}.$$

If θ is close to either 0 or 1, then this quantity gets larger.

Next we consider the logit transformation

$$\gamma = \log \frac{1-\theta}{\theta}, \quad \theta = \frac{1}{1+\exp(\gamma)},$$

the likelihood becomes

$$\ell = (n-k)\gamma - n \log\{1 + \exp(\gamma)\}.$$

The derivatives are

$$\frac{\partial \ell}{\partial \gamma} = (n-k) - n \frac{\exp(\gamma)}{1 + \exp(\gamma)}, \quad \frac{\partial^2 \ell}{\partial \gamma^2} = -n \frac{\exp(\gamma)}{(1 + \exp(\gamma))^2}, \quad \frac{\partial^3 \ell}{\partial \gamma^3} = -n \frac{\exp(\gamma)\{1 - \exp(\gamma)\}}{(1 + \exp(\gamma))^3}.$$

It can be shown

$$|\Delta(\gamma)| = |1-2\theta|\{\theta(1-\theta)\}^{1/2}n^{-1/2},$$

$$|\Delta(\gamma)|/|\Delta(\theta)| = \theta(1-\theta) \leq 1/4.$$

Exercise Find Δ for transformations

$$\psi = \sin^{-1}(\theta^{1/2}), \quad \phi = \int_0^\theta \frac{dt}{\{t(1-t)\}^{2/3}}.$$

Conclude which transformation is better?

Neyman–Scott Problems

Statistical inference in the presence of many nuisance parameters is stimulated from the well known Neyman–Scott problem. Neyman and Scott (1948) discussed

inference problem based on independent random variables whose probability densities involve parameters of two types. The first type appears in the density of every random variable; the second type appears in the density of only a finite number, possibly one. In Neyman and Scott's terminology, parameters of the first and second types are called, respectively, structural and incidental parameters. More specifically, they considered independent random variables with distributions

$$X_{i1}, X_{i2} \sim N(\mu_i, \sigma^2), i = 1, 2, \dots, n.$$

The log-likelihood is

$$\ell = -0.5 \sum_{i=1}^n \{(x_{i1} - \mu_i)^2 + (x_{i2} - \mu_i)^2\}/\sigma^2 - n \log \sigma^2 + \text{constant}.$$

$$\frac{\partial \ell}{\partial \mu_j} = -(x_{j1} - \mu_j) - (x_{j2} - \mu_j) = 0,$$

$$\hat{\mu}_j = \frac{x_{j1} + x_{j2}}{2},$$

$$\frac{\partial \ell}{\partial \sigma^2} = 0.5 \sum_{i=1}^n \{(x_{i1} - \mu_i)^2 + (x_{i2} - \mu_i)^2\}/(\sigma^2)^2 - n \frac{1}{\sigma^2} = 0,$$

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^n \{(x_{i1} - \hat{\mu}_i)^2 + (x_{i2} - \hat{\mu}_i)^2\} = \frac{1}{4n} \sum_{i=1}^n (x_{i1} - x_{i2})^2 \rightarrow \sigma^2/2,$$

since $X_{i1} - X_{i2} \sim N(0, 2\sigma^2)$.

In this example $\mu_i, i = 1, 2, \dots, n$ are incidental parameters. When $\mu_i, i = 1, 2, \dots, n$ are unknown constants, the MLE of the structural parameter σ^2 does not need to be consistent. Note that there are only finitely many observations that involve a particular μ_i . It is in general impossible to estimate μ_i consistently even if $n \rightarrow \infty$. Because the dimension of the parameter space $(\sigma^2, \mu_1, \dots, \mu_n)$ increases with n , the traditional large sample theory does not apply.

Exercise 1 Clearly $T_i = X_{i1} + X_{i2}$ is a sufficient statistic for μ_i . Show the maximum conditional likelihood (conditional on T_i) produces consistent estimates of σ^2 .

Exercise 2 Let $X_{ij} \sim N(\mu, \sigma_i^2)$, $i = 1, 2, \dots, n; j = 1, 2, \dots, n_i$. In this case μ is the structural parameter whereas $\sigma_i^2, i = 1, 2, \dots, n$ are incidental parameters.

(1) Show the score estimating equation for μ is

$$\sum_{i=1}^n \frac{n_i^2(\bar{x}_i - \mu)}{(n_i - 1)s_i^2 + n_i(\bar{x}_i - \mu)^2} = 0,$$

where

$$\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}, \quad s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2.$$

(2) Neyman and Scott (1948) found an estimating equation

$$\sum_{i=1}^n \frac{n_i(n_i - 2)(\bar{x}_i - \mu)}{(n_i - 1)s_i^2 + n_i(\bar{x}_i - \mu)^2} = 0.$$

Show that Neyman and Scott's (1948) estimate has smaller asymptotic variance than the maximum likelihood estimate as $n \rightarrow \infty$.

(3) Note that for any fixed μ , $T_i = \sum_{j=1}^{n_i} (x_{ij} - \mu)^2$ is a sufficient statistic for σ_i^2 . Derive the maximum conditional likelihood estimate of μ by conditioning on T_1, \dots, T_n . Compare above three different estimators.

Nuisance Parameter Problem and Parameter Orthogonality

In statistical literature parameter orthogonalizing methods are used to minimize the impact of nuisance parameters. Cox and Reid (1987) systematically discussed this method to separate the nuisance parameter and parameters of interest. The intuition is as follows. As the sample size goes to ∞ , the MLE of the nuisance parameter η and parameter of interest β are normally distributed with covariance equal to the inverse Fisher information matrix. If this matrix is block diagonal then $\hat{\eta}$ and $\hat{\beta}$ are distributed independently and approximately separate inference is achieved. The orthogonality between $\hat{\eta}$ and $\hat{\beta}$ implies the dependence of $\hat{\beta}$ on $\hat{\eta}$ is minimum. The basic procedure is as follows:

Denote the density as $f(x) = f(x, \eta, \beta)$. Let

$$\eta = h(\eta^*, \beta),$$

where the new parameter η^* is to be chosen to be orthogonal to β , i.e.,

$$E \left[\frac{\partial^2 \log f(x, \eta^*, \beta)}{\partial \eta^* \partial \beta} \right] = 0.$$

Using the chain rule, we have

$$\frac{\partial h}{\partial \beta} E \left\{ \frac{\partial^2 \log f}{\partial \eta^2} \right\} + E \left\{ \frac{\partial \log f}{\partial \eta \partial \beta} \right\} = 0.$$

We need to solve this differential equation to determine the transformation function h . In general this may not be easy.

Exercise 1 Lawless (1987) and Lancaster (2000) considered a panel Poisson count model with count data Y_{ij} follow a Poisson distribution with rate $\eta_i \exp(x_{ij}\beta)$, $i = 1, 2, \dots, n$; $j = 1, 2, \dots, n_i$. The likelihood contribution from the i -th individual is

$$L(\eta_i, \beta) = \exp[-\eta_i \sum_j \exp(x_{ij}\beta)] \eta_i^{\sum_j y_{ij}} \exp(\sum_j y_{ij}x_{ij}\beta).$$

Let

$$\eta_i^* = \eta_i \sum_j \exp(x_{ij}\beta),$$

then the likelihood can be written as

$$L(\eta_i^*, \beta) \propto \exp(-\eta_i^*) (\eta_i^*)^{\sum_j y_{ij}} \prod_i \left\{ \frac{\exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} \right\}^{y_{ij}}.$$

Clearly

$$\frac{\partial^2 \log L}{\partial \eta^* \partial \beta} = 0.$$

In fact we can observe that

$$Y_{i+} = \sum_{j=1}^{n_i} Y_{ij} \sim \text{Poisson}(\eta_i^*),$$

and

$$(Y_{i1}, \dots, Y_{in_i})|Y_{i+} \propto \frac{\eta_i^{y_{i+}} \exp(\sum_j y_{ij}x_{ij}\beta) \exp\{-\eta_i \sum_j \exp(x_{ij}\beta)\}}{\exp(-\eta_i^*)(\eta_i \sum_j \exp(x_{ij}\beta))^{y_{i+}}} = \prod_i \left\{ \frac{\exp(x_{ij}\beta)}{\sum_j \exp(x_{ij}\beta)} \right\}^{y_{ij}}.$$

The incident parameters η_1, \dots, η_n are eliminated in the conditional likelihood.

Exercise 2 Incidental parameter problems in a panel probit model

Consider panel data with a Probit model

$$P(D_{ij} = 1|\alpha_i, x_{ij}) = \Phi(\alpha_i + x_{ij}\beta), \quad i = 1, 2, \dots, n; j = 1, 2, \dots, n_i,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution. The existing conditional likelihood method cannot be used to eliminate α_i . Lancaster (2000) proposed an orthogonal reparametrization

$$\alpha_i^* = \sum_{j=1}^{n_i} \int_{-\infty}^{\alpha_i + x_{ij}\beta} h(u) du, \quad h(u) = \frac{\phi^2(u)}{\Phi(u)\{1 - \Phi(u)\}},$$

where $\phi(u) = d\Phi(u)/du$. However, it is not clear whether this method can produce a consistent estimator of β . In order to estimate β consistently, Kiefer and Wolfowitz (1956) assumed that $\alpha_i, i = 1, 2, \dots, n$ are random variables with a common distribution.

Next, we give another example due to Machado (2004) in the logistic regression model when there are many incidental intercepts.

Inconsistent for Panel Data

Consider a logistic regression model

$$P(D_{ij} = 1|x_{ij}) = \frac{\exp(\alpha_i + x_{ij}\beta)}{1 + \exp(\alpha_i + x_{ij}\beta)}, \quad j = 1, 2; \quad i = 1, 2, \dots, n.$$

If we directly maximize the logistic likelihood with respect to α_i , $i = 1, 2, \dots, n$ and β , we can show the MLE is inconsistent for estimating β .

More specifically, let us consider a simpler case, where

$$P(D_i = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}, \quad i = 1, 2, \dots, I,$$

$$P(\Delta_{ij} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, \quad i = 1, 2, \dots, I; \quad j = 1, 2.$$

The log-likelihood is

$$\ell = \sum_{i=1}^I [D_i(\alpha_i + \beta) - \log\{1 + \exp(\alpha_i + \beta)\}] + \sum_{i=1}^I [\sum_{j=1}^2 \Delta_{ij} \alpha_i - 2 \log\{1 + \exp(\alpha_i)\}].$$

The score estimating equations for α_i are

$$\frac{\partial \ell}{\partial \alpha_i} = D_i + \sum_{j=1}^2 \Delta_{ij} - \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)} - 2 \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} = 0,$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^I D_i - \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)} = 0.$$

The possible value of $K_i = D_i + \sum_{j=1}^2 \Delta_{ij}$ are $\{0, 1, 2, 3\}$. Clearly if $K_i = 3$ or 0, then $\alpha_i = \infty$ or $\alpha_i = -\infty$, respectively. If $K_i = 1$, the solution for α_i satisfies

$$\exp(\alpha_i) = \frac{-1 + \sqrt{1 + 8 \exp(\beta)}}{4 \exp(\beta)}.$$

If $K_i = 2$, then

$$\exp(\alpha_i) = \frac{\exp(\beta) + \sqrt{\exp(2\beta) + 8 \exp(\beta)}}{2 \exp(\beta)}.$$

Replacing these solutions in the score equation for β gives

$$\sum_{i=1}^I D_i = n_1 \left(\frac{-1 + \sqrt{1 + 8 \exp(\beta)}}{3 + \sqrt{1 + 8 \exp(\beta)}} \right) + n_2 \left(\frac{\exp(\beta) + \sqrt{\exp(2\beta) + 8 \exp(\beta)}}{2 + \exp(\beta) + \sqrt{\exp(2\beta) + 8 \exp(\beta)}} \right) + n_3,$$

where $n_j = \sum_{i=1}^I I(K_i = j)$, $j = 1, 2, 3$.

Exercise Show that the maximum likelihood estimation of β is inconsistent as $I \rightarrow \infty$ unless the true value of β is 0. Moreover the maximum likelihood estimator underestimates β if $\beta_0 < 0$, and overestimates β if $\beta_0 > 0$. Furthermore $1 < \hat{\beta}/\beta_0 < 2$.

Exercise Is it possible to find a consistent estimator of β by using a conditional likelihood approach?

4.3 Popular Inference Methods in the Presence of Nuisance Parameters

To find consistent estimates for the structural parameter in the presence of infinitely many incidental parameters has become an extremely important topic since the landmark work by Neyman and Scott (1948). The most popular method is the conditional approach. This approach conditions on the sufficient statistic X_{i+} for the incidental parameter μ_i as in the Neyman–Scott problem. Among others, Kalbfleisch and Sprott (1970)'s work is very influential. Theoretical results on the maximum conditional likelihood estimation can be found in Andersen (1970).

We summarize three popular methods for obtaining consistent estimates of structural parameters in the presence of many incidental parameters. There are also other methods, such as the Bayes' approach etc., which are not covered here.

(1) Conditional approach.

Suppose the density $f(x, y; \theta, \eta)$ can be decomposed as

$$f(x, y; \theta, \eta) = f(x; \theta|y)g(y; \theta, \eta),$$

where the conditional density $f(x|y)$ only depends on the structural parameter θ , then the conditional likelihood $L_c = \prod_{i=1}^n f(x_i; \theta|y_i)$ may be used to make inferences on θ . We have used this approach in previous examples.

Example Let T and A be two independent random variables, and $T \sim f(t, \theta)$ and $A \sim h(a, \eta)$. Instead of observing (T, A) , we only observe those pairs such that $T > A$. Denote the observed data as (T_i, A_i) , $i = 1, 2, \dots, n$, where $T_i > A_i$.

$$(T, A)|T > A \sim \frac{f(t, \theta)h(a, \eta)}{\int \bar{F}(a, \theta)h(a, \eta)da}, \quad t > a.$$

Note that the conditional likelihood $T|(T > A, A = a)$ is $f(t, \theta)/\bar{F}(a, \theta)$ which is independent of η . Therefore, we can use

$$L_c(\theta) = \prod_{i=1}^n \frac{f(t_i, \theta)}{\bar{F}(a_i, \theta)}$$

to make inference on θ . This is a well known truncation problem. We will discuss it in Chap. 24 on the nonparametric inference for F .

(2) Marginal likelihood approach.

Suppose the density $f(x, y; \theta, \eta)$ can be decomposed as

$$f(x, y; \theta, \eta) = f(x, y; \theta, \eta|y)g(y; \theta),$$

where the marginal density $g(y)$ only depends on the structural parameter θ , then the marginal likelihood $L_M = \prod_{i=1}^n g(y_i; \theta)$ may be used to make inference on θ .

Example Holt and Prentice (1974) and Wild (1983) conducted survival analyses in twin studies with matched pair experiments. Suppose the observed failure time data are t_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2$. The corresponding covariate for the (i, j) -th pair is denoted by z_{ij} . The hazard function can be specified as one of the following four models:

Model (1)

$$\lambda_i(t|z_{ij}) = \lambda \eta t^{\eta-1} \exp(z_{ij}\beta).$$

Model (2)

$$\lambda_i(t|z_{ij}) = \lambda_i \eta t^{\eta-1} \exp(z_{ij}\beta), \quad \lambda_i \sim \Gamma(a, b),$$

where $\Gamma(a, b)$ denotes a gamma distribution with shape parameter a and scale parameter b .

Model (3)

$$\lambda_i(t|z_{ij}) = \lambda_i \eta t^{\eta-1} \exp(z_{ij}\beta),$$

where λ_i 's are unknown parameters.

Model (4)

$$\lambda_i(t|z_{ij}) = \lambda_{i0}(t) \exp(z_{ij}\beta),$$

where $\lambda_{i0}(t)$ is unspecified baseline hazard function.

Inference based on model 1 is straightforward. However, this model may lose the feature of matched pairs since it treats λ as a constant for all i 's.

In model 2, inference on β and η can be based on the marginal likelihood, i.e., integrating out the frailty λ_i with respect to the gamma density.

In model 3, the marginal likelihood can be constructed through a transformation. Denote

$$\bar{t}_i = \frac{t_{i1} + t_{i2}}{2}, \quad w_i = t_{i1}/t_{i2}.$$

The induced homomorphic group acting on the parameter space is transitive on λ_i and leaves β and η invariant. The quantities $w_i, i = 1, 2, \dots, m$ are then marginally sufficient for β . The marginal likelihood based on the w_i 's is

$$L_{M3} = \prod_{i=1}^n \frac{\eta w_i^{\eta-1} \exp\{(z_{i1} - z_{i2})\beta\}}{[1 + w_i^\eta \exp\{(z_{i1} - z_{i2})\beta\}]^2}.$$

Finally, a different marginal likelihood can be constructed in model 4 through

$$P(T_{i2} < T_{i1}) = \frac{1}{1 + \exp\{(z_{i1} - z_{i2})\beta\}}.$$

The corresponding marginal likelihood is

$$L_{M4} = \prod_{i=1}^n \frac{1}{1 + \exp\{\epsilon_i(z_{i1} - z_{i2})\beta\}},$$

where $\epsilon_i = 1$ if $t_{i2} < t_{i1}$ and $\epsilon_i = -1$ otherwise.

Wild (1983) conducted theoretical and numerical comparisons for above four models. Readers may read his paper for details.

(3) Profile likelihood approach.

Suppose the likelihood $L(\beta, \eta)$ contains a parameter of interest β and a nuisance parameter η . If both β and η are of finite dimensions, then, for fixed β , η can be profiled out to give the profile likelihood

$$L(\beta, \hat{\eta}(\beta)) = \max_{\eta} L(\beta, \eta).$$

Inference on β can be then based on $L(\beta, \hat{\eta}(\beta))$. In general $L(\beta, \hat{\beta}(\eta))$ is well behaved if η is of finite dimension. On the other hand, it may lead to inconsistent estimator of β if the dimension of η goes to infinity.

In the Neyman and Scott (1948) problem, Kiefer and Wolfowitz (1956) showed that σ^2 can be consistently estimated if the means $\mu_i, i = 1, 2, \dots, n$ are treated as i.i.d. random variables with a common distribution G_0 . If the form of G_0 is known, then this may be treated as a mixture model and the EM algorithm can be used to estimate the structural parameters. On the other hand if the form of G_0 is unknown, then it would be difficult to estimate G_0 non-parametrically. Lindsay (1995)'s monograph is a good reference on this result. In the survival analysis Chap. 24, we will show that even if the dimension of η goes to infinity, the profile likelihood $L(\beta, \hat{\eta}(\beta))$ still produces consistent and efficient estimator of β under the Cox proportional hazards model.

(4) Rank based inference and tests

Consider a transformation linear model

$$h(y_i) = x_i\beta + \epsilon_i,$$

where h is an unknown monotonic non-decreasing function, and the error ϵ has a density $\epsilon_i \sim f(\epsilon)$ which is known. The observed data (Y_1, \dots, Y_n) can be decomposed as order statistics $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ and rank statistics $R = [(1), (2), \dots, (n)]$, where (i) is the rank statistic refers to the label attached to the i -th element of the order statistic $y_{(i)}$. Without loss of generality we denote the corresponding covariates as $x_{(1)}, \dots, x_{(n)}$. Under a monotonic transformation, the rank statistics R are invariant.

Note that

$$\begin{aligned} P\{R = [(1), (2), \dots, (n)] | (x_1, \dots, x_n)\} &= \int_{y_{(1)} < \dots < y_{(n)}} \prod_{i=1}^n f(y_{(i)} - x_{(i)}\beta) dy_{(i)} \\ &= \int_{y_{(1)} < \dots < y_{(n)}} \prod_{i=1}^n \frac{f(y_{(i)} - x_{(i)}\beta)}{f(y_{(i)})} \prod_{i=1}^n f(y_{(i)}) dy_{(i)} \\ &= \frac{1}{n!} E \left[\prod_{i=1}^n \frac{f(V_{(i)} - x_{(i)}\beta)}{f(V_{(i)})} \right], \end{aligned}$$

where $V_{(1)} \leq V_{(2)} \leq \dots \leq V_{(n)}$ are order statistics from $f(v)$. This is the well known Hoeffding (1951) formula.

In general, there is no closed form for evaluating this expectation. Doksum (1987) proposed using Monte Carlo method to evaluate this integral. Unfortunately, if n is moderately large, computational burden becomes an issue. The expectation has a closed form only in the special case that f has an extreme value density (corresponding to the Cox regression model, for example, Kalbfleisch and Prentice (1973)). We will return to this case in Chap. 24.

(5) A challenging problem

Instead of the Neyman–Scott location shift model, Small and Murdoch (1993) discussed a stratified exponential tilting model

$$(X_{k1}, X_{k2}) \sim f_k(x_1)f_k(x_2) \propto \exp(\beta_{k1}x_1)\exp(\beta_{k2}x_2)f(x_1)f(x_2), k = 1, 2, \dots, K,$$

where $f(x)$ is the common baseline density for X_1, X_2 . The main interest is to estimate the cumulative distribution function F by treating $\beta_{k1}, \beta_{k2}, k = 1, 2, \dots, K$ as nuisance parameters. It is possible to eliminate the nuisance parameters by using the conditional likelihood argument. In fact let $T_k = X_{k1} + X_{k2}$, then

$$(X_{k1}, X_{k2})|T_k = t \sim \frac{f(x_1)f(x_2)}{\int f(t - x_2)f(x_2)dx_2}.$$

It is straightforward to estimate the underlying parameters in f if a parametric form for f is assumed. However it is not clear how to find the non-parametric maximum conditional likelihood estimator of $dF(x)$.

4.4 Quasi-likelihood Methods in Linear Regression Models

Suppose Y has mean μ_0 and finite variance σ^2 . The underlying density of Y is denoted by $f_0(y)$, which is unknown. We are interested in estimating μ . We consider a “working normal model”

$$f(y, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y - \mu)^2/(2\sigma^2)\}.$$

The maximum likelihood estimator of μ can be attained by minimizing $\sum_{i=1}^n (y_i - \mu)^2$ with respect to μ , where $y_i, i = 1, 2, \dots, n$ are the observed data. The population version is

$$\min_{\mu} E_0(Y - \mu)^2 = \min_{\mu} \left\{ \int y^2 f_0(y) dy - 2\mu\mu_0 + \mu^2 \right\} = \min_{\mu} \left\{ \int y^2 f_0(y) dy + (\mu - \mu_0)^2 - \mu_0^2 \right\},$$

where f_0 and μ_0 are the true density and mean, respectively. Clearly $\mu = E_0(Y) = \mu_0$ is the minimizer. In other words, the least squares method leads to a consistent estimator of μ_0 whether the true density is from a normal distribution. However if the “working density” is not normal but something else, then in general, the maximum likelihood estimator may not be consistent.

Essentially, the least squares method is equivalent to using a normal density to approximate the true density under the Kullback–Leiber divergence. We can also use a t -distribution or other distributions to do so too. In the linear regression model

$$Y = \alpha + x^T \beta + \sigma \epsilon, \quad \epsilon \sim f_0(\epsilon).$$

Gould and Lawless (1988) showed the maximum likelihood method by arbitrarily specifying a “working density” is robust for the slope parameter β but not for the intercept parameter α . Next we use the Kullback–Leibler information to interpret this phenomenon.

For a given “working kernel” f , the Kullback–Leibler divergence is given by

$$\begin{aligned} & \int \sigma_0^{-1} f_0((y - \alpha_0 - x\beta_0)/\sigma_0) \log f((y - \alpha - x\beta)/\sigma) dy - \log \sigma \\ &= \int f_0(t) \log f\{\sigma_0\sigma^{-1}t + (\alpha_0 - \alpha)/\sigma + (\beta_0 - \beta)x/\sigma\} dt - \log \sigma. \end{aligned}$$

Let α^*, σ^* maximize

$$\int f_0(t) \log f(\alpha + (\sigma_0/\sigma)t) dt - \log \sigma.$$

Clearly

$$\begin{aligned} & \int f_0(t) \log f\{\sigma_0\sigma^{-1}t + (\alpha_0 - \alpha)/\sigma + (\beta_0 - \beta)x/\sigma\} dt - \log \sigma \\ & \leq \int f_0(t) \log f\{(\sigma_0/\sigma^*t + \alpha^*)\} dt - \log \sigma^*. \end{aligned}$$

Therefore the maximum value is achieved if

$$(\alpha_0 - \alpha)/\sigma + (\beta_0 - \beta)x/\sigma = \alpha^*, \quad \sigma_0/\sigma = \sigma_0/\sigma^*$$

for all x . If x has at least two distinct values, this can happen only if $\beta = \beta_0$ and $\alpha = \sigma^*\alpha_0 - \alpha^*$. Thus we have shown that the slope parameter β can be consistently estimated.

Next we derive large sample results. The log-likelihood is

$$\ell = \sum_{i=1}^n \log f((y_i - \alpha - x_i\beta)/\sigma) - n \log \sigma.$$

The score estimating equations are

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \sum_{i=1}^n \frac{f'((y_i - \alpha - x_i\beta)/\sigma)}{f((y_i - \alpha - x_i\beta)/\sigma)} \sigma^{-1} = 0, \quad \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n x_i \frac{f'((y_i - \alpha - x_i\beta)/\sigma)}{f((y_i - \alpha - x_i\beta)/\sigma)} \sigma^{-1} = 0, \\ \frac{\partial \ell}{\partial \sigma} &= \sum_{i=1}^n \left[\frac{-f'((y_i - \alpha - x_i\beta)/\sigma) \frac{y_i - \alpha - x_i\beta}{\sigma^2}}{f((y_i - \alpha - x_i\beta)/\sigma)} - \frac{1}{\sigma} \right]. \end{aligned}$$

Let $\epsilon_i = (y_i - \alpha - x_i\beta)/\sigma$, and $\psi'(\epsilon) = f'(\epsilon)/f(\epsilon)$. Equivalently the score equations are

$$S_1 = \sum_{i=1}^n x_i \psi'(\epsilon_i), \quad S_2 = \sum_{i=1}^n \psi'(\epsilon_i), \quad S_3 = \sum_{i=1}^n \psi'(\epsilon_i)\epsilon_i - 1.$$

Without loss of generality we assume that $E(X) = 0$. Let $\eta = (\beta, \alpha, \sigma)$. Easily we can show

$$\frac{\partial S_1}{\partial \alpha} = \sum_{i=1}^n x_i \psi''(\epsilon_i) \sigma^{-1}, \quad \frac{\partial S_1}{\partial \beta} = \sum_{i=1}^n x_i x_i^T \psi''(\epsilon_i) \sigma^{-1}, \quad \frac{\partial S_1}{\partial \sigma} = - \sum_{i=1}^n x_i \psi''(\epsilon_i) \sigma^{-2},$$

$$\frac{\partial S_2}{\partial \alpha} = \sum_{i=1}^n \psi''(\epsilon_i) \sigma^{-1}, \quad \frac{\partial S_2}{\partial \beta} = \sum_{i=1}^n x_i \psi''(\epsilon_i) \sigma^{-1}, \quad \frac{\partial S_2}{\partial \sigma} = - \sum_{i=1}^n \psi''(\epsilon_i) \sigma^{-2},$$

$$\frac{\partial S_3}{\partial \alpha} = \sum_{i=1}^n \sigma^{-1} [\psi''(\epsilon) \epsilon_i + \psi'(\epsilon_i)], \quad \frac{\partial S_3}{\partial \beta} = \sum_{i=1}^n x_i \sigma^{-1} [\psi''(\epsilon) \epsilon_i + \psi'(\epsilon_i)],$$

and

$$\frac{\partial S_3}{\partial \sigma} = - \sum_{i=1}^n \sigma^{-1} [\psi''(\epsilon) \epsilon_i^2 + \psi'(\epsilon) \epsilon_i].$$

We can show in distribution

$$\frac{1}{\sqrt{n}} \begin{pmatrix} S_1 \\ S_2 \\ S_3 \end{pmatrix} \rightarrow N(0, A), \quad A = \begin{pmatrix} E(XX^T)E(\psi'(\epsilon)^2) & 0 & 0 \\ 0 & E(\psi'(\epsilon)^2) & E\{\psi'(\epsilon)^2 \epsilon\} \\ 0 & E\{\psi'(\epsilon)^2 \epsilon\} & E\{\psi'(\epsilon)\epsilon - 1\}^2 \end{pmatrix},$$

and in probability

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial S}{\partial \eta} \rightarrow \begin{pmatrix} E(XX^T)E(\psi''(\epsilon))\sigma^{-1} & 0 & 0 \\ 0 & E(\psi''(\epsilon)) & -E(\psi'')\sigma^{-1} \\ 0 & \sigma^{-1}E(\psi''\epsilon) & -\sigma^{-1}E\{\psi''\epsilon^2 + \psi'\epsilon\} \end{pmatrix}.$$

Finally we can show that the asymptotic variance and covariance of $\hat{\beta}$ is

$$\sigma^2 E(XX^T) \frac{E(\psi')^2}{E(\psi'')}.$$

Since $E(XX^T)$ is fixed by design, based on the Godambe's optimal theory (will be discussed in Chap. 5), the second factor is minimized if the true density is used. In practice, however, the true density is unknown. In that case, we may specify a "working model"

$$\psi'(\epsilon) = \psi'(\epsilon, \xi).$$

Then we can minimize

$$\frac{E(\psi')^2}{E(\psi'')}$$

with respect to ξ ! This is exactly Godambe's optimal criterion in which the functional form of ψ is given but with finitely many unknown parameters ξ . However when ψ is unknown, this problem becomes more difficult. Theoretically, a nonparametric kernel method can be used to estimate ψ . Bickel (1982) proposed an adaptive estimation method.

In applications we have found the following three-stage approach is useful.

(1) Use a least squares method to find $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$.

(2) Define $\hat{\epsilon}_i = (y_i - \hat{\alpha} - x_i^T \hat{\beta})/\hat{\sigma}$, $i = 1, 2, \dots, n$. Minimize

$$\frac{[n^{-1} \sum_{i=1}^n \psi'(\hat{\epsilon}_i, \xi)]^2}{n^{-1} \sum_{i=1}^n \psi''(\hat{\epsilon}_i, \xi)}$$

with respect to ξ . Denote the minimizer as $\hat{\xi}$.

(3) Maximize

$$\prod_{i=1}^n \frac{1}{\sigma} f((y_i - \alpha - x_i \beta)/\sigma, \hat{\xi}).$$

with respect to (α, β, σ) .

Gould and Lawless (1988) made asymptotic variance comparison between the maximum likelihood estimate and the pseudo likelihood estimate. In general the asymptotic efficiency depends only on the choice of f but not on $E(XX^T)$.

If one is interested in finding a consistent estimator for the intercept α , then the least squares method can be applied again

$$\min_{\alpha} (y_i - \alpha - x_i \hat{\beta})^2,$$

where $\hat{\beta}$ is the estimator discussed above.

Exercise Are the results for the accelerated failure time model generalizable to the heteroscedastic variance case?

Next we briefly introduce some commonly used pseudo likelihood methods in statistical inference.

4.5 Composite Likelihoods and Corrected Likelihoods

1. Composite likelihoods

The main goals for using composite likelihoods are the considerations of model robustness and computational feasibility. The journal of *Statistica Sinica* published a special 2011 issue devoted to composite likelihood methods. There are many variations of composite likelihoods.

Consider a m -dimensional random variable Y , with probability density function $f(y, \theta)$ for some unknown p -dimensional parameter vector $\theta \in \Theta$. Denote by $\{A_1, \dots, A_K\}$ a set of marginal or conditional events with associated likelihoods $L_k(y, \theta) = f(y \in A_k; \theta)$. Following Lindsay (1988) a composite log-likelihood is the weighted summation

$$\ell_C(y, \theta) = \sum_{k=1}^K w_k \ell_k(y, \theta),$$

where w_k 's are non-negative weights to be chosen. If the weights are all equal then they can be ignored.

In statistical inference for spatial processes, Besag (1974; 1975) defined a pseudo likelihood as the product of the conditional densities of a single observation given its neighbours,

$$L_C(y, \theta) = \prod_{j=1}^m f(y_j | \{y_i : y_i \text{ is neighbour of } y_j\}, \theta).$$

This method also applies to time series problems where the composite likelihood is constructed by conditioning on the past history one time lag or two time lags ahead.

If two conditional densities

$$f(y|x) = f_1(y|x, \beta), \quad f(x|y) = f_2(x|y, \beta)$$

are easily obtained, Arnold et al. (2007) suggested maximizing the product of the two conditional densities. In general, two conditional densities can determine the joint density $f(x, y)$ of (X, Y) uniquely. However, the joint density may not have a closed form due to the need of the integration.

An immediate application of the composite likelihood method is for the transformation linear model,

$$h(y_i) = x_i \beta + \epsilon_i,$$

where h is an unknown monotonic non-decreasing function, and the error density $\epsilon_i \sim f(\epsilon)$ is known. For example, for any $i \neq j \neq k$, we may find $P(Y_i < Y_j | x_i, x_j) = p_{ij}(x_i, x_j, \beta)$ or $P(Y_i < Y_j < Y_k | x_i, x_j, x_k) = p_{ijk}(x_i, x_j, x_k, \beta)$. Then we can formulate either pairwise or triplet-wise composite rank likelihoods. Those likelihoods would be much easier to calculate than the full rank likelihood method when the error has a non extreme distribution (The extreme value distribution is corresponding to the Cox proportional hazards model). Moreover the transformation function $h(\cdot)$ does not play a role for estimating β in the rank based methods. More details can be found in Thas and De Neve (2012).

More comprehensive discussions on different composite likelihoods can be found in the excellent review paper by Varin et al. (2011).

2. Corrected likelihoods or scores for measurement error problems

When independent variables or covariates in models are subject to measurement errors, the simple maximum likelihood estimation is inconsistent. Nakamura (1990) proposed the corrected likelihood or score function method for errors-in-variables models. In contrast to imputation methods used in statistical literature, the corrected score function method somehow goes backward.

Suppose the true covariate Z is measured with error

$$X = Z + \epsilon,$$

where X is the observed version. Assume the true conditional density of Y given Z is given by a parametric model

$$f(y|z) = f(y|z, \beta).$$

If $z_i, i = 1, 2, \dots, n$ are available, the log-likelihood and score are, respectively,

$$\ell(y, z, \beta) = \sum_{i=1}^n \log f(y_i|z_i, \beta), \quad U(y, z, \beta) = \frac{\partial \ell(y, z, \beta)}{\partial \beta}.$$

A function $\ell^*(y, x, \beta)$ is called a “corrected log-likelihood” if

$$E[\ell^*(Y, X, \beta)|Y = y, Z = z] = \ell(y, z, \beta)$$

for any β . Similarly the corrected score $U^*(y, x, \beta)$ satisfies

$$E\{U^*(Y, X, \beta)|Y = y, Z = z\} = U(y, z, \beta).$$

Using iterated expectations, we have $E[U^*(Y, X, \beta)] = 0$. Therefore the corrected score is unbiased. In general we can only show that the estimating equation $\sum_{i=1}^n U^*(x_i, y_i, \beta) = 0$ has a consistent root. However, it may have multiple roots. It would be interesting to study whether the information identity principle developed by Gan and Jiang (1999) can be used to select consistent roots. Huang (2014b) proposed to use empirical likelihood to deal with the multiple roots problem.

Consider a generalized linear model

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\},$$

where

$$\theta = z\beta, \quad x = z + \epsilon, \quad \epsilon \sim N(0, \Lambda).$$

Denote $\xi = \beta^T \Lambda \beta / 2$. By the normal moment generating function property,

$$E[\exp(\beta^T X)|z] = \exp(\beta^T z + \xi), \quad E[X \exp(\beta^T X)|z] = (z + \Lambda \beta) \exp(\beta^T z + \xi),$$

$$E[XX^T \exp(\beta^T X)|z] = [\Lambda + (z + \Lambda \beta)(z + \Lambda \beta)^T] \exp(\beta^T z + \xi).$$

The following examples are special cases.

Example 1 Consider

$$Y_k = z_k \beta + \epsilon_k, \quad \epsilon \sim N(0, \sigma^2).$$

The log-likelihood is

$$\ell(y, z, \beta) = -0.5n \log(2\pi) - n \log \sigma - 0.5\sigma^{-2} \sum_{k=1}^n (y_k - z_k \beta)^2.$$

A corrected log-likelihood is

$$\ell^*(y, x, \beta) = -0.5n \log(2\pi) - n \log \sigma - 0.5\sigma^{-2} \sum_{i=1}^n \{(y_k - x_k \beta)^2 - \beta^T \Lambda \beta\}.$$

Example 2 Suppose Y follows the Poisson distribution with mean $\exp(\beta^T z)$. A corrected log-likelihood is

$$\ell^*(x, y, \beta) = \sum_{k=1}^n \{-\exp(x_k \beta - \xi) + y_k x_k \beta - \log y_k!\}.$$

Example 1 Consider a measurement problem in the gamma regression model

$$f(y; \theta, \phi) = \theta(\theta y)^{\phi-1} \exp(-\theta y) / \Gamma(\phi), \quad \theta = \exp(z \beta).$$

Show

$$\ell^*(\beta) = \sum_{i=1}^n [(\phi - 1) \log y_i + \phi x_i \beta - y_i \exp(x_i \beta) - \xi] - \log \Gamma(\phi)$$

is the corrected likelihood.

Example 2

$$P(Y = 1|z) = \frac{\exp(\beta^T z)}{1 + \exp(\beta^T z)},$$

$$X = z + \epsilon.$$

Use the inversion theorem of the Laplace transformation to show that a corrected score function does not exist. (Stefanski 1989).

A thorough introduction of statistical inference on measurement errors in nonlinear regression models can be found in Carroll et al. (2006). A recent monograph on measurement error problems can be found in Yi (2016).

4.6 Variable Selection and Akaike Criterion

Variable selection has long been of interest to statisticians. Recently, it has been an active area of research since the arrival of big data problems created with the recent advance of technology, where thousands of variables are available in data analysis. For example, researchers can conduct genome-wide association studies by sequencing the DNA of human subjects, enabling them to statistically correlate specific genes to particular diseases. Social scientists and survey researchers are also faced with integrating data from multiple data sources and expanding their activities beyond experiments and surveys, especially with large volume of data that can be collected through the world wide web. Variable selection is intended to select the “best” subset of predictors in a regression model. The main motivations of variable selection include (1) To explain the data in a parsimonious way; (2) To eliminate unnecessary predictors since they will only add noise to the estimation of other quantities of interest; (3) To stabilize the model caused by collinearity; (4) To achieve cost effectiveness such that irrelevant or redundant predictors can be omitted from future data collection.

Denote

$$X_1, \dots, X_n \sim i.i.d. f_0(x).$$

Let $f(x) = f(x, \theta)$ be the postulated parametric family which may or may not include the true density $f_0(x)$. Model selection can be approached by minimizing the Kullback–Leibler information

$$I_{KL} = \inf_{\theta} \left\{ \int f_0(x) \log f_0(x) dx - \int f_0(x) \log f(x, \theta) dx \right\}.$$

Since the first term is a constant, we only need to minimize the second term with respect to θ . Since $F_0(x)$ is unknown, we may replace it by the empirical distribution $F_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$. When comparing two competing models, minimizing I_{KL} is equivalent to maximizing over different competing models the maximum of the log-likelihood

$$\ell = \sum_{i=1}^n \log f(x_i, \theta)$$

with respect to θ . Clearly such a criterion is not ideal. It encourages selection of the larger model even if a subset of it is the true model. For a fixed positive integer p , denote θ_p as p -dimensional parameters. Let $\hat{\theta}_p$ maximize the log-likelihood $\ell(\theta_p) = \sum_{i=1}^n \log f(x_i, \theta_p)$. Akaike (1973) proposed the Akaike information criterion (AIC)

$$\min_p \{\ell(\hat{\theta}_p) - p\}. \quad (4.6.2)$$

In other words, the AIC is the maximum log-likelihood penalized by a quantity equal to the number of parameters.

Denote the expected log-likelihood as

$$\eta(F_0) = \int f_0(x) \log f(x, \hat{\theta}) dx,$$

where $\hat{\theta} = \operatorname{argmax} \int \log f(x, \theta) dF_n(x)$. Next we introduce Konishi and Kitagawa's (1996) arguments.

Replacing F_0 by the empirical distribution F_n , we have

$$\eta(F_n) = n^{-1} \sum_{i=1}^n \log f(x_i, \hat{\theta}).$$

Note that $\eta(F_n)$ overestimates $\eta(F_0)$ since F_n corresponds more closely to $\hat{\theta}$ than the true F_0 . Write the bias as

$$b(F_0) = E_0[\eta(F_n) - \eta(F_0)].$$

The information criterion based on the bias-corrected log-likelihood is to maximize

$$GIC = n^{-1} \sum_{i=1}^n \log f(x_i, \hat{\theta}) - b(F_n).$$

In general, let $\hat{\theta}$ be an estimator of θ (not necessarily be the maximum likelihood estimator and not necessarily consistent). Suppose

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^n T(x_i) + o_p(n^{-1}).$$

Using Taylor's expansion, we can show that

$$b(F_0) = E_0[\eta(F_n) - \eta(F_0)] = \frac{1}{n} b_1(F_0) + o_p(1/n),$$

where

$$b_1(F_0) = \operatorname{tr} \left\{ \int T(x, F_0) \frac{\partial \log f(x_i, \theta)}{\partial \theta^T} |_{T(F_0)} dF_0(x) \right\},$$

and tr is the trace of a matrix. The bias corrected log-likelihood is

$$GIC = - \sum_{i=1}^n \log f(x_i, \hat{\theta}) + \frac{1}{n} \sum_{i=1}^n \operatorname{tr} \left\{ \int T(x, F) \frac{\partial \log f(x_i, \theta)}{\partial \theta^T} |_{\hat{\theta}} dF_0(x) \right\}.$$

In the special case that $\hat{\theta}$ is the MLE, then

$$T(x_i) = J(F_0)^{-1} \frac{\partial \log f(x_i, \theta_0)}{\partial \theta}, \quad J = - \int \frac{\partial^2 \log f(x, \theta_0)}{\partial \theta \partial \theta^T} dF_0(x).$$

The biased corrected log-likelihood becomes

$$GIC = - \sum_{i=1}^n \log f(x_i, \hat{\theta}) + \frac{1}{n} \sum_{i=1}^n \text{tr}\{J^{-1}(F_n)I(F_n)\}. \quad (4.6.3)$$

If the postulated parametric family contains the true density, it becomes the AIC (4.6.2) since the second term is p by the well known information identity.

In classical linear regression models, the ridge regression minimizes

$$\sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2,$$

where λ is a tuning parameter, which can be determined by, for example, cross validation. In big data problems where the number of explanatory variables is comparable to the sample size, many of those β_i are expected to be zero. Tibshirani (1996) considered the Lasso regression by minimizing

$$\sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|,$$

where again λ is a tuning parameter. In contrast to ridge regression, Lasso regression forces the $\hat{\beta}_j$'s associated with the un-important variables to be exactly 0.

4.7 Two Useful Maximization Algorithms

Next we discuss two popular algorithms for finding maximum likelihood estimates. The first one is the EM algorithm for missing data and incomplete observation problems. The second one is called alternating minimum algorithm, or coordinator descending algorithm. This algorithm is mainly used in penalized likelihood inference when there exist many parameters.

1. The EM algorithm

Orchard and Woodbury (1972) provided theoretical foundation of the missing information principle. The expectation and maximization (EM) algorithm was formally proposed by Dempster et al. (1977) for handling missing data problems in a parametric model set up. However, recent research shows this method is also useful

for finding maximum nonparametric likelihood or semiparametric likelihood estimates. We begin with the parametric cases and then move on to nonparametric and semiparametric problems in the later Chapters.

Consider a parametric model

$$f(x, y, \theta) = f(y|x, \theta) f(x, \theta),$$

where the form of f is known but θ is an unknown parameter. Due to missing data, we only observe X but Y is missing. The log-likelihood based on X can be written as

$$\log f(x, \theta) = \log f(x, y, \theta) - \log f(y|x, \theta).$$

Taking conditional expectation $Y|X$ at $\theta = \theta_0$ on both sides, we have

$$\log f(x, \theta) = Q(\theta, \theta_0) - K_{Y|X}(\theta, \theta_0),$$

where

$$Q(\theta, \theta_0) = \int f(y|x, \theta_0) \log f(x, y, \theta) dy, \quad K_{Y|X}(\theta, \theta_0) = \int f(y|x, \theta_0) \log f(y|x, \theta) dy.$$

Note that $K_{Y|X}$ is the conditional Kullback–Leibler information. Now we seek θ_1 such that

$$Q(\theta_1, \theta_0) = \max_{\theta} Q(\theta, \theta_0).$$

We repeat this process by successively replacing θ_j with θ_{j+1} , $j = 0, 1, \dots$, such that

$$\theta_0, \theta_1, \dots, \theta_n,$$

satisfy

$$Q(\theta_i, \theta_i) \leq Q(\theta_{i+1}, \theta_i), \quad i = 0, 1, 2, \dots$$

Since

$$\log f(x, \theta_{i+1}) - \log f(x, \theta_i) = Q(\theta_{i+1}, \theta_i) - Q(\theta_i, \theta_i) + \{K_{Y|X}(\theta_i, \theta_i) - K_{Y|X}(\theta_{i+1}, \theta_i)\} \geq 0,$$

using the Kullback–Leibler information inequality, we have a nondecreasing sequence

$$\log f(x, \theta_0) \leq \log f(x, \theta_1) \leq \dots \leq \log f(x, \theta_n).$$

If $f(x, \theta)$ is a convex function, then convergence is guaranteed. When $f(x, \theta)$ is not convex, this algorithm may not converge to the global maximum likelihood estimate. In practice, this process should be repeated using different initial values θ_0 . Wu (1983) discussed theoretical convergence aspects of the EM algorithm. In general, the convergence rate of EM algorithm is slow. In statistical literature there are many discussions on how to speed up the EM algorithm.

Differentiating the equation

$$-\log f(x, \theta) = -\log f(x, y, \theta) + \log f(y|x, \theta)$$

twice with respect to θ , we have

$$-\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2} = -\frac{\partial^2 \log f(x, y, \theta)}{\partial \theta^2} + \frac{\partial^2 \log f(y|x, \theta)}{\partial \theta^2}.$$

Taking conditional expectation given $X = x$ leads to

$$-\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2} = -E \left\{ \frac{\partial^2 \log f(x, Y, \theta)}{\partial \theta^2} | x \right\} + E \left\{ \frac{\partial^2 \log f(Y|x, \theta)}{\partial \theta^2} | x \right\}.$$

The left hand side is the observed-data information matrix. The first term on the right hand side is the conditional information matrix for the complete data. The second term is minus the missing information matrix. This is the “missing information principle” (Orchard and Woodbury 1972; Louis 1982), i.e.,

$$\text{Observed Information} = \text{Complete Information} - \text{Missing Information}.$$

As an alternative, the Louis’s formula is given by

$$-\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2} = -E \left\{ \frac{\partial^2 \log f(x, Y, \theta)}{\partial \theta^2} | x \right\} - \text{var} \left\{ \frac{\partial \log f(Y|x, \theta)}{\partial \theta} | x \right\}.$$

This is a direct result of the information identity.

More discussions on the EM algorithm can be found in Tanner (1996) and Lange (2010).

Exercise Consider the Heckman selection model

$$Y_1 = x_1\beta_1 + \epsilon_1, \quad Y_2 = x_2\beta_2 + \epsilon_2,$$

where Y_1 is observable if and only if $Y_2 > 0$. X_1, X_2 are available, but the only information on Y_2 is $Y_2 > 0$ or $Y_2 < 0$. Assume the errors (ϵ_1, ϵ_2) have a bivariate normal distribution with mean zero and covariate matrix (σ_{ij}) , $i = 1, 2$; $j = 1, 2$, where $\sigma_{22} = 1$. Use the EM algorithm to find the maximum likelihood estimate.

2. Alternating minimization

A more general algorithm is the so called alternating minimization algorithm. Csiszar and Tusnády (1984) derived convergence properties of this algorithm.

Let P and Q be two probability measures from \mathcal{P} and \mathcal{Q} , respectively, where \mathcal{P} and \mathcal{Q} are two given probability measure spaces. Let the distance between P and Q be $d(P, Q)$, which is a mapping from \mathcal{P} and \mathcal{Q} to R^+ . The sequences $\{P_k, Q_k\}_{k=0}^\infty$ are obtained by alternating minimization for $k = 0, 1, 2, \dots$,

$$P_{k+1} = \operatorname{argmin}_{P \in \mathcal{P}} d(P, Q_k), \quad Q_{k+1} = \operatorname{argmin}_{Q \in \mathcal{Q}} d(P_{k+1}, Q),$$

where P_0 is an initial point in \mathcal{P} . We may use notation

$$P_0 \rightarrow Q_0 \rightarrow P_1 \rightarrow Q_1 \cdots$$

Csiszar and Tusnády (1984) showed that if \mathcal{P} and \mathcal{Q} are convex measures and $d(P, Q)$ is the Kullback–Leiber information divergence, then the alternating minimization algorithm converges monotonically to a global minimum. If a likelihood function involves two unknown probability measures P and Q , on the other hand, the popular method for finding maximum likelihood estimates is the profile likelihood approach. The disadvantage of the profile likelihood method is that the structure of $d(P, \hat{Q}(P))$ is destroyed after profiling out Q . However the alternative minimization (or maximization) algorithm preserves this structure. We will discuss this point in details in Sect. 24.2.

A closely related minimization algorithm is called coordinate descent or coordinate wise minimization algorithm. This method circumvents the difficulty of the process of simultaneously maximizing over all coordinates by maximizing over them one by one.

Given a convex, differentiable function $f : R^n \rightarrow R$, if we are at a point x such that $f(x)$ is minimized along each coordinate axis, have we found a global minimizer? The answer is yes! A necessary condition is f is differentiable. If $f(x)$ itself is not differentiable but can be written as

$$f(x) = g(x) + \sum_{i=1}^n h_i(x_i),$$

where g is convex, differentiable and each h_i is convex. Then the answer is yes again! The non-smooth parts $h_i(x_i)$ are called separable. Therefore we can define $x_i^{(k)}$, $i = 1, 2, \dots, n$, $k = 1, 2, \dots$, sequentially

$$x_1^{(k)} = \operatorname{argmin}_{x_1} f(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \dots, x_n^{(k-1)}), \quad x_2^{(k)} = \operatorname{argmin}_{x_2} f(x_1^{(k)}, x_2, x_3^{(k-1)}, \dots, x_n^{(k-1)}),$$

$$x_3^{(k)} = \operatorname{argmin}_{x_3} f(x_1^{(k)}, x_2^{(k)}, x_3, \dots, x_n^{(k-1)}), \quad \dots, \quad x_n^{(k)} = \operatorname{argmin}_{x_n} f(x_1^{(k)}, x_2^{(k)}, \dots, x_{n-1}^{(k)}, x_n).$$

Tseng (2001) proved that for such f (provided f is continuous on compact set $\{x : f(x) \leq f(x^{(0)})\}$ and f attains its minimum), any limit point of $x^{(k)}$, $k = 1, 2, 3, \dots$ is a minimizer of f .

Remarks

(1) Order of cycle through coordinates is arbitrary, any permutation of $\{1, 2, \dots, n\}$ can be used.

(2) Individual coordinates can be replaced by blocks of coordinates anywhere.

(3) “One-at-a-time” update scheme is critical, and “all-at-once” scheme does not necessarily converge.

The coordinate decent minimization algorithm is mainly used in variable selection problems where the number of explanatory variables may be much larger than the sample size, as in high dimensional data problems, see example, Friedman et al. (2007). We will show in Chap. 24 that this algorithm can also be used for maximum likelihood estimation in a nonparametric or semiparametric set up where the number of parameters grows with the sample size. A typical example is the truncation problem which will be discussed in Chap. 24.

4.8 Likelihood Based Inference with Inequality Constraints

Natural physical or biological phenomenon often lead to inequality constraints for the underlying parameters. Suppose there are iid data

$$X_1, \dots, X_n \sim f(x, \theta),$$

where the underlying parameter θ satisfies the following inequality constraints

$$g_1(\theta) \geq 0, \dots, g_r(\theta) \geq 0.$$

By introducing Lagrange multipliers, the objective function is

$$L = \ell(\theta) + \sum_{j=1}^r \lambda_j g_j(\theta).$$

We need to check Khun–Tucker conditions (Boyd and Vandenberghe 2004) for the maximum likelihood estimator under the inequality constraints.

The first order condition is

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial \ell(\theta)}{\partial \theta} + \sum_{j=1}^r \frac{\partial g_j(\theta)}{\partial \theta} \lambda_j = 0$$

with sign constraints

$$g_j(\theta) \geq 0, \lambda_j \geq 0, j = 1, 2, \dots, r,$$

and exclusion conditions

$$\lambda_j g_j(\theta) = 0, j = 1, 2, \dots, r.$$

More details on the theoretical developments can be found in the work by Gourieroux et al. (1982). Interested readers may also read the excellent book by Robertson et al. (1988). In general the likelihood ratio statistic does not converge to a standard chi-squared distribution under the inequality constraints due to boundary conditions.

Chapter 5

Optimal Estimating Function Theory

Classical statistical inference emphasizes unbiased estimators rather than unbiased estimating equations. Suppose random variable X has a density with a known parametric form $f(x, \theta)$, where θ is a $p \times 1$ unknown parameter. An estimator of θ , denoted as $\hat{\theta}$, is called unbiased if the expectation of $\hat{\theta}$ equals to θ , i.e., $E(\hat{\theta}) = \theta$. For example if X has a normal density with mean θ and variance σ^2 , then the sample mean $\bar{X} = n^{-1} \sum_{i=1}^n x_i$ is an unbiased estimator of θ . On the other hand the maximum likelihood estimator of variance $n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is biased for σ^2 . The adjusted variance estimator $(n - 1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is unbiased. In general the requirement of unbiasedness of the point estimator is strong. Even the maximum likelihood estimator may not satisfy this requirement. Sometimes unbiased estimator does not exist.

Example Suppose $X \sim N(\theta, 1)$, and we are interested in estimating $g(\theta) = |\theta|$. If $\hat{g}(x)$ is an unbiased estimator of $g(\theta)$, then for any $-\infty < \theta < \infty$ we have

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{g}(x) \exp\{-0.5(x - \theta)^2\} dx = |\theta|.$$

Note that the left side in above formula is continuously differentiable with respect to θ everywhere. However the right side is not differentiable at $\theta = 0$. Thus we conclude that it is impossible to find an unbiased estimator of $|\theta|$.

In many situations there are many unbiased estimators. Then naturally we seek among these unbiased estimators the one with the smallest variance. In statistical literature a uniformly minimum-variance unbiased estimator or minimum-variance unbiased estimator (UMVUE or MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter. The well known Cramer–Rao information lower bound states that under some regularity conditions, no unbiased estimator has a variance that is lower than the inverse of the Fisher information lower bound. In general the UMVUE is very hard to find unless the underlying distribution comes from exponential families.

5.1 Godambe's Optimality Criterion

Instead of seeking unbiased estimators, in his original work, Godambe (1960) studied the unbiased estimating functions. Any measurable function $g(x, \theta)$ of x and θ is called unbiased if it has expectation 0. In practice it is very easy to construct unbiased estimating functions, for example, by using the method of moments. The moments based estimation method is to match the sample moments to the population moments. The population moments are defined as

$$\mu_r = E(X^r) = \int x^r f(x, \theta) = \psi_r(\theta), \quad r = 1, 2, \dots$$

The sample moments are the empirical versions, i.e., $n^{-1} \sum_{i=1}^n x_i^r$. The moment estimating equations are defined by

$$n^{-1} \sum_{i=1}^n x_i^r - \psi_r(\theta) = 0, \quad r = 1, 2, \dots, p.$$

If we denote

$$g_r(x, \theta) = x^r - \psi_r(\theta),$$

then $E[g_r(X, \theta)] = 0$. Clearly if the density of X belongs to a specified parametric family, we can construct as many unbiased estimating functions as possible. A natural question is, which one is optimal? Godambe (1960) gave a stringent definition of regular estimating functions, which satisfy the following:

- (1) $E_\theta g(X, \theta) = 0$, for $\theta \in \Theta$, where Θ is the parameter space.
- (2) $\partial g / \partial \theta$ exists for $\theta \in \Theta$.
- (3) $\int g(x, \theta) f(x, \theta) dx$ is differentiable under the integral sign.
- (4) $[E\{\partial g / \partial \theta\}]^2 > 0$ for $\theta \in \Theta$.

Let $\mathcal{G} = \{g : g \text{ satisfy 1) - 4)}\}$ be regular estimating functions.

For the time being we assume $p = 1$, i.e., θ is a scalar parameter.

Definition A $g^* \in \mathcal{G}$ is said to be an optimal estimating function if

$$\frac{E_\theta(g^{*2})}{[E_\theta\{\partial g^* / \partial \theta\}]^2} \leq \frac{E_\theta(g^2)}{[E_\theta\{\partial g / \partial \theta\}]^2}$$

for any $g \in \mathcal{G}$ and $\theta \in \Theta$.

In this definition one would like the estimating function g to cluster around 0 as much as possible, i.e., $E_\theta(g^2)$ should be as small as possible. At the same time, it is desirable for a “good” estimating function to be sensitive to the underlying parameter, i.e., $E[g(X, \theta + d\theta)]$ should be as away from 0 as possible when the parameter departs away from the true one. Since the asymptotic variance of the estimator derived from

the given unbiased estimating function g is $E_\theta(g^2)/[E_\theta\{\partial g/\partial\theta\}]^2$, we need to search for a g such that the resulting estimator has asymptotic minimum variance.

Theorem 5.1 For any $g \in \mathcal{G}$,

$$\frac{E_\theta g^2}{[E_\theta \partial g / \partial \theta]^2} \geq \frac{1}{E_\theta [\partial \log f / \partial \theta]^2}.$$

Proof Note that

$$Eg(x, \theta) = 0 \text{ or } \int g(x, \theta) f(x, \theta) dx = 0.$$

Differentiating the above equation with respect to θ , we have

$$\int \frac{\partial g(x, \theta)}{\partial \theta} f(x, \theta) dx + \int g(x, \theta) \frac{\partial f}{\partial \theta} dx = 0,$$

or

$$E \left(\frac{\partial g(X, \theta)}{\partial \theta} \right) + E \left(g(X, \theta) \frac{\partial \log f}{\partial \theta} \right) = 0.$$

By using Cauchy–Schwarz's inequality,

$$\left[E \left(\frac{\partial g}{\partial \theta} \right) \right]^2 = \left[E \left(g \frac{\partial \log f}{\partial \theta} \right) \right] \leq E(g^2) E \left(\frac{\partial \log f}{\partial \theta} \right)^2.$$

Corollary 1 If $g = \partial \log f(x, \theta) / \partial \theta$, then the following information equality holds:

$$E \left[\frac{\partial^2 \log f}{\partial \theta^2} \right] = - \left[E \left(\frac{\partial \log f}{\partial \theta} \right) \right]^2.$$

Corollary 2 Let $\phi(x)$ be an unbiased estimator of $\phi(\theta)$. Then the Cramer–Rao inequality starts from

$$\begin{aligned} & E \left[\phi(X) \frac{\partial \log f(X, \theta)}{\partial \theta} \right] \\ &= \int \phi(x) \frac{\partial f(x, \theta)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int \phi(x) f(x, \theta) dx = \frac{\partial \phi(\theta)}{\partial \theta}. \end{aligned}$$

Using Jensen's inequality, the Cramer–Rao bound is

$$\text{Var}(\phi(X)) \geq [\phi'(\theta)]^2 / E[\partial \log f(X, \theta) / \partial \theta]^2.$$

Exercise Generalize Godambe's approach to the vector parameter case.

The foundations of finite sample estimation in stochastic processes

Instead of independent and identically distributed data problems, the optimal estimating function theory was extended by Godambe (1985) to dependent data problems. Let \mathcal{F} be a class of probability distributions F on R^n and $\theta = \theta(F)$ ($F \in \mathcal{F}$) be a real parameter. Let

$$h_i = h_i(y_1, \dots, y_i, \theta), i = 1, 2, \dots, n$$

be some specified functions. We require conditional unbiasedness, i.e.,

$$E_{i-1}h_i = 0, \quad E_{i-1} = E(\cdot | y_1, \dots, y_{i-1}), \quad E_0 = E_F.$$

Clearly conditional unbiasedness implies overall unbiasedness since $E(h_i) = E[E(h_i | y_1, \dots, y_{i-1})] = E[0] = 0$. Moreover it can be shown that $E(h_i h_j) = 0, i \neq j$.

Define a class of estimating functions

$$\mathcal{L} = \{g | g = \sum_{i=1}^n h_i a_{i-1}, \quad a_{i-1} = a_{i-1}(y_1, \dots, y_{i-1}, \theta)\}.$$

For any $g \in \mathcal{L}$, $Eg = 0$. We are interested in finding a_i 's such that the resulting estimating function is optimal.

In fact if we let

$$g^* = \sum_{i=1}^n h_i a_{i-1}^*, \quad a_{i-1}^* = \frac{E_{i-1}(\partial h_i / \partial \theta)}{E_{i-1}(h_i^2)}, \quad (5.1.1)$$

then g^* is the optimal estimating function in the class \mathcal{L} . We can prove this by the following steps.

Note that

$$g^2 = \sum_{i=1}^n a_{i-1}^* h_i^2 + 2 \sum_{i < j} h_i a_{i-1} h_j a_{j-1}.$$

Let

$$A^2 = \sum_{i=1}^n a_{i-1}^2 E_{i-1} h_i^2.$$

By using the fact that h_i only depends on y_1, \dots, y_i , a_{i-1} only depends on y_1, \dots, y_{i-1} and $E_{i-1}h_i = 0$, we have $E(g^2) = E(A^2)$. Also

$$\frac{\partial g}{\partial \theta} = \sum_{i=1}^n a_{i-1} \frac{\partial h_i}{\partial \theta} + \frac{\partial a_{i-1}}{\partial \theta} h_i.$$

Let

$$B = \sum_{i=1}^n a_{i-1} E_{i-1}(\partial h_i / \partial \theta).$$

Easily we can show $E(\partial g / \partial \theta) = E(B)$. Therefore

$$\frac{E(\partial g / \partial \theta)^2}{E(g^2)} = \frac{[E(B)]^2}{E(A^2)} \leq E(B^2 / A^2)$$

by the Cauchy–Schwarz's inequality.

Note that if

$$a_{i-1} = a_{i-1}^* = \frac{E_{i-1}(\partial h_i / \partial \theta)}{E_{i-1} h_i^2},$$

then

$$A^2 = \sum_{i=1}^n \left(\frac{E_{i-1}(\partial h_i / \partial \theta)}{E_{i-1} h_i^2} \right)^2 E_{i-1} h_i^2 = \sum_{i=1}^n \left(\frac{[E_{i-1}(\partial h_i / \partial \theta)]^2}{E_{i-1} h_i^2} \right) = B.$$

Therefore we have shown the optimality of (5.1.1).

Exercise An alternative proof is given by Professor Mary Thompson at University of Waterloo. It has the following steps.

1. If $EX_1 = EX_2 = 0$, and $E[\{X_1/c_1 - X_2/c_2\}X_2] = 0$ for constants c_1 and c_2 , then $V[X_1/c_1] \geq V[X_2/c_2]$.
2. Show that

$$E \left[\left\{ \frac{g}{E(\partial g / \partial \theta)} - \frac{g^*}{E(\partial g^* / \partial \theta)} \right\} g^* \right] = 0$$

for any $g \in \mathcal{L}$.

3. Show g^* is optimal in \mathcal{L} .

$$E(gg^*) / \{E(\partial g / \partial \theta)\} = E\{\sum_{i=1}^n a_{i-1} a_{i-1}^* E_{i-1}(h_i^2)\} / E\{\sum_{i=1}^n a_{i-1} E_{i-1}(\partial h_i / \partial \theta)\} = 1.$$

Similarly we have

$$E(g^{*2}) / \{E(\partial g^* / \partial \theta)\} = 1.$$

Example Conditional Least Squares with Applications in Branching Processes

Godambe's general theory has a nice application in branching processes. Branching processes are used to model reproduction. For example, in the study of the growth

of bacteria, individuals might correspond to bacteria, each of which generates finite many offsprings with some probability in a single time unit. Let y_0, \dots, y_n be a branching process with $y_0 = 1$, where y_i is the number of individuals at generation i , then y_i can be written as

$$y_i = \sum_{k=1}^{y_{i-1}} x_{ik},$$

where in the $(i - 1)$ -th generation each individual produces a random number of offsprings, x_{ik} , $k = 1, 2, \dots, y_{i-1}$ in the next generation, independently of other individuals. Note that conditioning on y_{i-1} , x_{ik} , $k = 1, 2, \dots, y_{i-1}$ are iid random variables with the same distribution as y_1 . Assume $E(Y_1) = \theta$ and $\text{Var}(Y_1) = \sigma^2$. Then

$$E_{i-1}(Y_i) = \theta y_{i-1}.$$

Let

$$h_i = y_i - \theta y_{i-1},$$

then $E_{i-1}h_i = 0$, $E_{i-1}(\partial h_i / \partial \theta) = -y_{i-1}$ and $E_{i-1}h_i^2 = \sigma_i^2 = y_{i-1}\sigma^2$. Easily we can find

$$a_{i-1}^* = \frac{E_{i-1}(\partial h_i / \partial \theta)}{E_{i-1}(h_i^2)} = \frac{-y_{i-1}}{\sigma^2 y_{i-1}} = -\frac{1}{\sigma^2}.$$

Therefore the optimal estimating function is equivalent to $g^* = \sum_{i=1}^n h_i$.

As alternative estimators, we can define

$$g_1 = \sum_{i=1}^n \{y_i/y_{i-1} - \theta\}, \quad g_2 = y_n/y_{n-1} - \theta.$$

Clearly both have the form $\sum_{i=1}^n h_i a_{i-1}$. Therefore they are inferior to g^* .

Exercise Define

$$h_i = (y_i - \theta y_{i-1})^2 - \sigma^2 y_{i-1}.$$

Find the optimal estimating equation for σ^2 based on the estimating functions $\sum_{i=1}^n h_i a_{i-1}$.

5.2 Applications of Godambe's Theory in Missing Covariate Problems

Next we present the results by Qin and Zhang (2011) on the application of Godambe's estimating function theory to the missing data problems.

The past decade has witnessed a surge of interest in statistical methods for studying missing data. This is mainly due to the increasing interest in the social sciences and

medical community, where missing data is a ubiquitous problem. In this section we consider the covariate-missing data problem under the assumption missingness only depends on the observable variables. This is the so-called missing at random problem. Other forms of missingness will be discussed in Chap. 19. The main purpose here is to illustrate the application of Godambe's optimal estimating function theory.

Let Y denote the response variable and let X and Z represent covariates. Suppose Y and X are always observed, but Z may be missing. Let δ be the missing indicator, where $\delta = 1$ when Z is observed and $\delta = 0$ otherwise. Without loss of generality, we denote the observed data as

$$(\delta_1 = 1, y_1, x_1, z_1), \dots, (\delta_m = 1, y_m, x_m, z_m), (\delta_{m+1} = 0, y_{m+1}, x_{m+1}, ?), \dots, (\delta_n = 0, y_n, x_n, ?).$$

The missing probability function $P(\delta = 0|y, x, z) = P(\delta = 0|y, x)$ may be completely known, as in the survey sampling problem, or may be modeled by a parametric model

$$P(\delta = 0|y, x, z) = P(\delta = 0|y, x) = 1 - \pi(y, x, \eta).$$

In the latter case, the unknown parameter η can be estimated by $\hat{\eta}$, by maximizing the log binomial likelihood

$$\ell_B(\eta) = \sum_{i=1}^n [\delta_i \log \pi(y_i, x_i, \eta) + (1 - \delta_i) \log\{1 - \pi(y_i, x_i, \eta)\}].$$

We assume that the conditional distribution of Y given (X, Z) is modeled by a parametric model

$$f(y|x, z) = f(y|x, z, \beta),$$

where β is a $p \times 1$ vector parameter of interest. The marginal distribution $g(x, z)$ of (X, Z) is left unspecified. Let $S(y, x, z, \beta) = \partial \log f(y|x, z, \beta) / \partial \beta$ denote the conditional score function. In the absence of missing data, it follows from Godambe's optimal estimating function theory that $\sum_{i=1}^n S(y_i, x_i, z_i, \beta)$ is the optimal unbiased estimating function for β . In the presence of missing data, a popular choice of an unbiased estimating function for β is the Horvitz–Thompson-type inverse weighted estimating function given by

$$h_{1n}(\beta) = \sum_{i=1}^n \frac{\delta_i S(y_i, x_i, z_i, \beta)}{\pi(y_i, x_i, \eta_0)},$$

where η_0 is the true value of η and is assumed to be known for the time being. This type of inverse weighted estimating functions has been discussed extensively in statistical literature; for example, Kalbfleisch and Lawless (1988) considered likelihood-based and estimating functions-based approaches in multi-state models. In a survey sampling review paper, Godambe and Thompson (1986) defined a class of finite population estimating functions which are unbiased estimates of the super-population

estimating equations. Analogously, we can define a class of estimating equations

$$\mathbf{H} = \left\{ h : \sum_{i=1}^n E\{h(\delta_i, Y_i, X_i, Z_i, \beta) | y_i, x_i, z_i\} = \sum_{i=1}^n S(y_i, x_i, z_i, \beta) \right\},$$

so that conditioning on $\{(y_i, x_i, z_i), i = 1, \dots, n\}$, any estimating function based on the observed complete data should be an unbiased estimator of the score function in the absence of missing data. It can be showed that the estimating function $h_{1n}(\beta)$ is optimal in the class \mathbf{H} . It is of interest to determine whether $h_{1n}(\beta)$ is also optimal in the following class of estimating functions

$$\Psi_1 = \left\{ \sum_{i=1}^n \frac{\delta_i \psi(y_i, x_i, z_i, \beta)}{\pi(y_i, x_i, \eta_0)} : E\{\psi(Y, X, Z, \beta) | x, z\} = 0 \right\},$$

which is a larger class than \mathbf{H} .

To address this issue, define $\pi^*(x, z, \beta, \eta_0) = P(\delta = 1 | x, z)$. It is seen that $\pi^*(x, z)$ can be written as

$$\pi^*(x, z, \beta, \eta_0) = E\{P(\delta = 1 | X, Y, Z) | x, z\} = E\{\pi(X, Y, \eta_0) | x, z\} = \int \pi(y, x, \eta_0) f(y | x, z, \beta) dy.$$

The conditional density of y given $(\delta = 1, x, z)$ is given by

$$f^*(y | \delta = 1, x, z) = f^*(y, x, z, \beta, \eta_0) = \frac{f(y | x, z, \beta) \pi(y, x, \eta_0)}{\pi^*(x, z, \beta, \eta_0)}.$$

Furthermore, conditional on $(\delta = 1, x, z)$, the conditional likelihood is given by

$$L_C = \prod_{i=1}^m \frac{\pi(y_i, x_i, \eta_0) f(y_i | x_i, z_i, \beta)}{\pi^*(x_i, z_i, \beta, \eta_0)}.$$

Let

$$S^*(y, x, z, \beta, \eta_0) = \frac{\partial \log f^*(y, x, z, \beta, \eta_0)}{\partial \beta} = \frac{\partial \log\{f(y | x, z, \beta) / \pi^*(x, z, \beta, \eta_0)\}}{\partial \beta}$$

denote the conditional score of β (conditioning on $(\delta = 1, x, z)$). According to Godambe (1960) optimal estimating function theory, $S^*(y, x, z, \beta, \eta_0)$ is the optimal estimating function among all regular estimating functions using data $(\delta_i y_i, \delta_i = 1, \delta_i x_i, \delta_i z_i), i = 1, \dots, n$. Thus, we need to show that $\sum_{i=1}^n \delta_i S^*(y_i, x_i, z_i, \beta, \eta_0) \in \Psi_1$. Since

$$\delta S^*(y, x, z, \beta, \eta_0) = \delta S^*(y, x, z, \beta, \eta_0) \pi(y, x, \eta_0) / \pi(y, x, \eta_0),$$

we only need to show that $E[S^*(Y, X, Z, \beta, \theta_0) \pi(y, x, \eta_0) | x, z] = 0$. Noting that

$$S^*(y, x, z, \beta, \eta_0)\pi(y, x, \eta_0) = E\{\delta S^*(Y, X, Z, \beta, \eta_0)|Y = y, X = x, Z = z\},$$

we have

$$\begin{aligned} E\{S^*(Y, X, Z, \beta, \eta_0)\pi(Y, X, \eta_0)|X, Z\} &= E\{\delta S^*(Y, X, Z, \beta, \eta_0)|X = x, Z = z\} \\ &= P(\delta = 1|X, Z)E\{S^*(Y, X, Z, \beta, \eta_0)|\delta = 1, X = x, Z = z\} = 0, \end{aligned}$$

which implies that $\sum_{i=1}^n \delta_i S^*(y_i, x_i, z_i, \beta, \eta_0)$ is optimal in class Ψ_1 .

Remark 1 Note that the inverse weighted estimating function $\sum_{i=1}^n \delta_i \psi(y_i, x_i, z_i, \beta)/\pi(y_i, x_i)$ does not use those observed data $y_i, x_i, \delta_i = 0$, which can be further improved by the so-called “augmented inverse probability weighted” estimating function (Robins et al. 1994). We discuss this in detail in Chap. 19.

Remark 2 If the true η_0 is unknown, we can replace it by its maximum binomial likelihood estimator $\hat{\eta}$. It can be shown that the impact on β is small by using estimated $\hat{\eta}$. This is due to the fact that the estimating function for η is ancillary to β (Small and McLeish 1989, see definition of ancillarity in next section).

If we replace $\partial \log f(y|x, x, \beta)/\partial \beta$ by some estimating equation $S(y, x, z, \beta)$, clearly the inverse weighted estimating function $\delta S(y, x, z, \beta)/\pi(x, y)$ is still valid. As an alternative we need to seek a new estimating equation by defining

$$\begin{aligned} S^*(y, x, z, \beta) &= S(y, x, z, \beta) - E^*[S(y, x, z, \beta)|x, z] = S(y, x, z, \beta) \\ &\quad - \frac{\int \pi(x, y)S(y, x, z)f(y|x, z)dy}{\int \pi(x, y)f(y|x, z)dy}. \end{aligned}$$

However this method may not be feasible since it needs full knowledge of $f(y|x)$ in order to evaluate the expectation. In the next section we find that this new approach may be substantially more efficient than the inverse weighting method for a special location shift model.

5.3 Godambe's Theory in Length Biased Sampling AFT Models

As demonstrated in Chaps. 1 and 3 that length biased sampling problems occur frequently. In this section we consider the accelerated failure time (AFT) model that is used widely in survival analysis in medical research and reliability studies. This model assumes that log survival time follows a linear model with an unspecified error distribution. We only illustrate this problem in the absence of right censoring case. This problem will be discussed more thoroughly in Chaps. 24 and 25 for handling right censoring problems. Chen (2010) used hazard-based estimation methods originally developed for censored observations, whereas Mandel and Ritov (2010) proposed a simple linear regression model on a log scale.

Suppose that in the absence of length bias, the conditional density of the positive random variable Y given $X = x$ is modeled by a parametric model

$$Y|X = x \sim f(y|x, \beta).$$

The corresponding score function is $S(y, x, \beta) = \partial \log f(y|x, \beta) / \partial \beta$. In the presence of length bias, the conditional density of Y given $X = x$ is given by

$$f^*(y, x, \beta) = \frac{yf(y|x, \beta)}{\mu(x, \beta)},$$

where $\mu(x, \beta) = \int yf(y|x, \beta)dy$.

Define a class of estimating functions as

$$\Psi_2 = \left\{ \sum_{i=1}^n \frac{\psi(y_i, x_i, \beta)}{y_i} : E_f\{\psi(Y, X, \beta)|X = x\} = 0 \right\}.$$

Let $S(y, x, \beta) = \partial \log f(y|x, \beta) / \partial \beta$ be the score function. Clearly, the conventional inverse weighted estimating function $\psi_{1n}(\beta) = \sum_{i=1}^n S(y_i, x_i, \beta)/y_i \in \Psi_2$ since $E^*[S(Y, X\beta)|X] = 0$, where the expectation is with respect to the length biased version. However it is not the optimal one in the class Ψ_2 !

In fact, according to Godambe (1960) optimal estimation function theory, the optimal estimation function in the regular class of estimating functions is the score function based on the length biased likelihood $\psi_2(\beta) = \sum_{i=1}^n \partial \log f^*(y_i, x_i, \beta) / \partial \beta$. We only need to show that $\psi_2(\beta) \in \Psi_2$ or $E^*[Y \partial \log f_l(Y, X, \beta) / \partial \beta | X] = 0$; this follows by observing that

$$\begin{aligned} \int y \frac{\partial \log f^*(y, x, \beta)}{\partial \beta} f(y|x, \beta) dy &= \int y \frac{\partial f(y|x, \beta)}{\partial \beta} dy - \frac{1}{\mu(x, \beta)} \int y \frac{\partial f(y|x, \beta)}{\partial \beta} dy \\ &\quad \int yf(y|x, \beta) dy = 0. \end{aligned}$$

In general, if we can replace the score function by some unbiased estimating function $U(y, x, \beta)$ satisfying $E\{U(Y, X, \beta)|X = x\} = 0$ in the absence of length bias, then $\sum_{i=1}^n U(y_i, x_i, \beta)/y_i$ is an unbiased estimating function for β in the presence of length bias.

As an alternative approach one may consider estimating function

$$\psi = U(Y, x, \beta) - E^*\{U(Y, x, \beta)\}.$$

However

$$E^*\{U(Y, x, \beta)\} = \frac{\int yU(y, x, \beta)f(y|x, \beta)dy}{\int yf(y|x, \beta)dy}$$

involves integrals which may not be feasible without fully knowledge of $f(y|x, \beta)$. Fortunately in the AFT model case we can circumvent this problem.

We now discuss a type of robustness of the optimal estimating function $\psi_{2n}(\beta)$ in the context of an accelerated failure time model. Assume that

$$\log Y = \alpha + x\beta + \epsilon,$$

where the density $f(\epsilon)$ of ϵ is not specified. It is easy to see that the conditional density of Y given $X = x$ is given by

$$f(y|x, \alpha, \beta) = \frac{1}{y} f(\log y - \alpha - \beta x), \quad y > 0.$$

With length biased data, the conditional density of the observed Y given $X = x$ is given by

$$f^*(y, x, \alpha, \beta) = \frac{f(\log y - \alpha - \beta x)}{\int_0^\infty f(\log y - \alpha - \beta x) dy} = \frac{f(\log y - \alpha - \beta x)}{\int_{-\infty}^\infty f(t) \exp(t + \alpha + \beta x) dt}, \quad y > 0.$$

As a result, the conditional density of the observed $T = \log Y$ given $X = x$ is equal to

$$f_T(t, x, \alpha, \beta) = \frac{f(t - \alpha - \beta x) \exp(t - \alpha - \beta x)}{\int_{-\infty}^\infty \exp(t) f(t) dt}, \quad -\infty < t < \infty.$$

The vector score function based on a single T is given by

$$S_T(t, x, \alpha, \beta) = - \begin{pmatrix} 1 \\ x \end{pmatrix} \left\{ \frac{f'(t - \alpha - \beta x)}{f(t - \alpha - \beta x)} + 1 \right\}.$$

If the true conditional density of the observed $T = \log Y$ given $X = x$ is

$$f_0(y, x, \gamma_0, \beta) = \frac{1}{\mu_0} g(t - \gamma_0 - \beta x) \exp(t - \gamma_0 - \beta x),$$

where $\mu_0 = \int \exp(t) g(t) dt$, then we have

$$\begin{aligned} E_{f_0}\{S_T(T, X, \alpha, \beta)|X=x\} &= -\frac{1}{\mu_0} \begin{pmatrix} 1 \\ x \end{pmatrix} \int \left\{ \frac{f'(t - \alpha - \beta x)}{f(t - \alpha - \beta x)} + 1 \right\} g(t - \gamma_0 - \beta x) \exp(t - \gamma_0 - \beta x) dt \\ &= -\frac{1}{\mu_0} \begin{pmatrix} 1 \\ x \end{pmatrix} \int \left\{ \frac{f'(t - \alpha)}{f(t - \alpha)} + 1 \right\} g(t - \gamma_0) \exp(t - \gamma_0) dt. \end{aligned}$$

If we can choose $\alpha = \alpha^*$ satisfying

$$\int \left\{ \frac{f'(t - \alpha^*)}{f(t - \alpha^*)} + 1 \right\} g(t - \gamma_0) \exp(t - \gamma_0) dt = 0,$$

then

$$E_{f_0} \left[\left\{ \frac{f'(T - \alpha^* - \beta X)}{f(T - \alpha^* - \beta X)} + 1 \right\} \middle| X = x \right] = 0$$

even when f is not the correct conditional density of the error term ϵ given $X = x$. Consequently, the optimal estimating equations

$$\sum_{i=1}^n \binom{1}{x_i} \left\{ \frac{f'(t_i - \alpha - \beta x_i)}{f(t_i - \alpha - \beta x_i)} + 1 \right\} = 0$$

can produce consistent estimation for the slope parameter β even when f is misspecified. This type of robustness is analogous to that of Gould and Lawless (1988) under a location shift model in the absence of length bias. This is only valid in the absence of right censoring.

Next we conduct some numerical comparisons with the inverse weighting estimating function approach. The regression model is assumed to be

$$\log Y = \alpha + \beta x + \epsilon.$$

Under the normal assumption for the distribution of error ϵ , the score equation of (α, β) based on the length-biased likelihood is

$$\sum_{i=1}^n \binom{1}{x_i} (\log y_i - \alpha - \beta x_i + 1) = 0,$$

whose solution is denoted as $(\hat{\alpha}, \hat{\beta})$. The inverse weighted score equation is given by

$$\sum_{i=1}^n \frac{1}{y_i} \binom{1}{x_i} (\log y_i - \alpha - \beta x_i) = 0$$

and their solution is denoted as $(\hat{\alpha}_w, \hat{\beta}_w)$.

If the variance of ϵ is σ^2 , then $\alpha = \alpha^* - \sigma^2$.

We consider two cases: (1) the true error distribution is $N(0, 1)$, and (2) the true error distribution is $0.2 \times t(6)$, where $t(6)$ stands for a student t distribution with six degrees of freedom. The first case corresponds to a correctly specified underlying distribution. Therefore, we expect that the maximum length-biased likelihood method would produce consistent estimators for both intercept and slope parameters. The second case corresponds to a misspecified underlying distribution. As a result, the maximum length-biased likelihood can only produce consistent estimators for the slope parameter. We generated 1000 random samples of the length biased sampling data of sample size $n = 100$. The estimated mean and sample variance are reported, respectively, in Tables 5.1 and 5.2 for cases (1) and (2).

Table 5.1 Means and variances of estimators of α and β using the maximum likelihood estimation (MLE) and the inverse weighted score estimating equation method with a correct regression model: $\log Y = \alpha + \beta X + \epsilon$, $X \sim N(0, 1)$, $\epsilon \sim N(0, 1)$. The sample size is $n = 100$ with 1000 repetitions

α	β	MLE				Inverse weighted score			
		Mean		Var		Mean		Var	
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}_w$	$\hat{\beta}_w$	$\hat{\alpha}_w$	$\hat{\beta}_w$
1	1	0.99120	1.00154	0.02077	0.01025	1.08569	0.92243	0.07772	0.06391
1	2	1.06662	1.95040	0.05496	0.01079	1.30091	1.78523	0.25613	0.15666
0	-1	0.00427	-0.99552	0.02016	0.01078	0.08824	-0.91599	0.07526	0.06139
2	2	2.04633	1.96108	0.05205	0.01117	2.27915	1.78254	0.27343	0.17560

Table 5.2 Means and variances of estimators of α and β using the maximum likelihood estimation (MLE) and the inverse weighted score estimating equation method with a misspecified regression model: $\log Y = \alpha + \beta X + \epsilon$, $X \sim N(0, 1)$, $\epsilon \sim 0.2 * t(6)$. The sample size is $n = 100$ with 1000 repetitions

α	β	MLE				Inverse weighted score			
		Mean		Var		Mean		Var	
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}_w$	$\hat{\beta}_w$	$\hat{\alpha}_w$	$\hat{\beta}_w$
1	1	0.06062	0.99995	0.00134	0.00066	1.00338	0.99430	0.00166	0.00218
1	2	0.06736	1.99636	0.00332	0.00068	1.02367	1.97857	0.01046	0.00911
0	-1	-0.93785	-0.99951	0.00147	0.00070	0.00044	-1.00083	0.00193	0.00253
2	2	1.06704	1.99729	0.00346	0.00072	2.01469	1.98944	0.01115	0.00958

We observe from Table 5.1 that the maximum length-biased sampling likelihood estimates for both the intercept and slope parameters have much smaller variances than those using the inverse weighted score estimating equations. For $(\alpha, \beta) = (2, 2)$, the variance based on the inverse weighted score equation for estimation of the slope parameter can be 17 times larger than that based on the maximum length-biased sampling likelihood estimation. Moreover, there is some bias based on the inverse weighted score method when the value of β gets larger. For model (2), similar results are observed in Table 5.2 for the slope parameter even though the maximum length-biased sampling likelihood estimate for the intercept parameter α is biased.

This example shows that in general the inverse probability weighted estimating equation may incur significant loss in terms of efficiency, though it gives unbiased estimation. For the accelerated failure time model, it is best to use the maximum likelihood estimate even if the underlying model is misspecified. When the slope parameter is consistently estimated, one can easily estimate the intercept by using the method of least squares. Unfortunately this method is not valid for right censored data. It is a challenging task to extend current approach to the right censored data problems.

5.4 Ancillarity and Fisher Information with Nuisance Parameters

So far we have focused on one parameter problems. Inference in the presence nuisance parameters is one of the richest research topics. Denote $\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is a nuisance parameter.

In the presence of nuisance parameter θ_2 , the most commonly used method to make inference for θ_1 is the profile likelihood approach,

$$L(\theta_1, \hat{\theta}_2(\theta_1)) = \max_{\theta_2} L(\theta_1, \theta_2),$$

where $\hat{\theta}_2(\theta_1)$ maximizes the likelihood function when θ_1 is fixed. In most cases the profile likelihood method works nicely. However, the weakness of this approach is that it may have large bias if the available data are sparse for θ_2 . Moreover, this method may produce inconsistent estimates if there are many incidental parameters or nuisance parameters, for example, in the well known Neyman and Scott (1948) examples discussed in last Chapter.

In the special case where the underlying estimating function is independent of the nuisance parameters, Godambe (1984) established the following optimal estimating function theory.

A real function $g(x, \theta_1)$ of x and θ_1 is a regular unbiased estimating function if

$$E_\theta[g(X, \theta_1)] = 0, \quad \theta = (\theta_1, \theta_2) \in \Theta.$$

Differentiating

$$\int g(x, \theta_1) f(x, \theta_1, \theta_2) dx = 0$$

with respect to θ_1 , we have

$$E_\theta[(\partial \log f(X, \theta_1, \theta_2)/\partial \theta_1)g(X, \theta_1)] = -E_\theta(\partial g(X, \theta_1)/\partial \theta_1).$$

Similarly differentiating with respect to θ_2 k times

$$\int (\partial^k f/\partial \theta_2^k) g dx = 0, \quad \text{or} \quad E \left\{ \frac{\partial^k f/\partial \theta_2^k}{f} g \right\} = 0.$$

Moreover for any constant c_1, \dots, c_k

$$E_\theta[g\{(\partial \log f/\partial \theta_1) - \sum_{k=1}^K c_k(1/f)(\partial^k f/\partial \theta_2^k)\}] = -E_\theta(\partial g/\partial \theta_1).$$

Using Cauchy–Schwarz's inequality we have

$$E_\theta \left[\{g/E_\theta(\partial g/\partial\theta_1)\}^2 \right] \geq \left[E_\theta \left\{ (\partial \log f/\partial\theta_1) - \sum_{k=1}^K c_k(1/f)(\partial^k f/\partial\theta_2^k) \right\}^2 \right]^{-1}.$$

Since this is true for any c_1, \dots, c_k , the information lower bound for the estimator $\hat{\theta}$ satisfies

$$E_\theta \left[\{g/E_\theta(\partial g/\partial\theta_1)\}^2 \right] \geq \max_{c_1, \dots, c_K} \left[E_\theta \left\{ (\partial \log f/\partial\theta_1) - \sum_{k=1}^K c_k(1/f)(\partial^k f/\partial\theta_2^k) \right\}^2 \right]^{-1}.$$

In the special case $k = 1$, we need to find c_1 to minimize

$$E[\partial \log f/\partial\theta_1 - c_1 \partial \log f/\partial\theta_2]^2.$$

This was discussed by Bartlett (1953), Lindsay (1980) and Liang (1983).

In general, define an estimating function space $\mathcal{G} = \{g(x, \theta_1) | E_F[g(x, \theta_1)] = 0\}$. Let $\mathcal{U} = \mathcal{G}^C$ be the complement space of \mathcal{G} , i.e., for any $u(x, \theta) \in \mathcal{U}$ and $g(x, \theta_1) \in \mathcal{G}$,

$$E[g(X, \theta_1)u(X, \theta)] = 0.$$

Note that g only depends on X and θ_1 but u depends on X, θ_1 and θ_2 . Clearly

$$\frac{\partial^k f/\partial\theta_2^k}{f} \in \mathcal{U}, \quad k = 1, 2, \dots.$$

Also the linear combination

$$\sum_{k=1}^K c_k(\theta) \frac{\partial^k f/\partial\theta_2^k}{f} \in \mathcal{U}.$$

Therefore

$$\inf_{c_1, \dots, c_K} E_\theta \{ (\partial \log f/\partial\theta_1) - \sum_{k=1}^K c_k(1/f)(\partial^k f/\partial\theta_2^k) \}^2 \geq \inf_{u \in \mathcal{U}} E_\theta \{ (\partial \log f/\partial\theta_1) - u \}^2.$$

The Fisher information $I(\theta_1)$ in the presence of nuisance parameter θ_2 is defined (Godambe 1980) as

$$I(\theta_1) = \inf_{u \in \mathcal{U}} E_\theta \{ \partial \log f/\partial\theta_1 - u \}^2.$$

A sufficient condition for a function $u^* \in \mathcal{U}$ to achieve the lower bound is $\{\partial \log f/\partial\theta_1 - u^*\} \in \mathcal{G}$, i.e., it is independent of θ_2 . In fact

$$E[\{\partial \log f/\partial\theta_1 - u\}^2] = E[\{\partial \log f/\partial\theta_1 - u^*\}]^2 + E[u - u^*]^2 + 2E[(u^* - u)\{\partial \log f/\partial\theta_1 - u^*\}].$$

If $u, u^* \in \mathcal{U}$, so does $u^* - u \in \mathcal{U}$. Therefore the last term vanishes. In general it would be very difficult to find u^* except for a few special cases in Godambe (1984).

Godambe (1976) called a statistic $t(x)$ of x an ancillary for θ_1 in the presence of the nuisance parameter θ_2 if

$$f(x, \theta_1, \theta_2) = f(x|t, \theta_1)h(t, \theta_1, \theta_2).$$

The class of distributions of t obtained by fixing θ_1 and letting θ_2 vary is complete for each value of θ_1 . In other words, if $E_\theta\{g(T, \theta_1)\} = 0$ for any $\theta \in \Theta$, then $P_\theta\{g(T, \theta_1) = 0\} = 1$.

Godambe (1984) showed that the conditional score

$$g^* = \partial \log f(x|t, \theta_1) / \partial \theta_1$$

is the optimal estimating equation in the class of unbiased estimating equations which are independent of the nuisance parameter θ_2 .

In fact, for any estimating function $g(x, \theta_1)$,

$$0 = E[g(X, \theta_1)] = E[E\{g(X, \theta_1)|t\}].$$

By the assumption of completeness, we have

$$E\{g(X, \theta_1)|t\} = 0.$$

Since h only depends on x through $t(x)$, we have

$$E[g(X, \theta_1)\partial \log h(t, \theta_1, \theta_2)/\partial \theta_1] = E[E\{g(X, \theta_1)|t\}\log h(t, \theta_1, \theta_2)/\partial \theta_1] = 0.$$

Note that

$$\partial \log f(x, \theta_1, \theta_2)/\partial \theta_1 = \partial \log f(x|t, \theta_1)/\partial \theta_1 + \partial \log h(t, \theta_1, \theta_2)/\partial \theta_1$$

we have

$$\begin{aligned} E(g^*g) &= E\left[\frac{\partial \log f(X|t, \theta_1, \theta_2)}{\partial \theta_1} g(X, \theta_1)\right] \\ &= E\left[\frac{\partial \log f(X, \theta_1, \theta_2)}{\partial \theta_1} g(X, \theta_1)\right] \\ &= -E[\partial g/\partial \theta_1]. \end{aligned}$$

Since $f(x|t, \theta_1)$ is a conditional density, it satisfies the information identity, i.e.,

$$E[\partial g^*/\partial \theta_1] = -\text{Var}(g^*).$$

From the inequality

$$0 \leq \text{Var}[E^{-1}(\partial g^*/\partial\theta_1)g^* - E^{-1}(\partial g/\partial\theta_1)g].$$

We can show that

$$E^{-1}(\partial g/\partial\theta_1)\text{Var}(g)E^{-1}(\partial g/\partial\theta_1) \geq E^{-1}(\partial g^*/\partial\theta_1)\text{Var}(g^*)E^{-1}(\partial g^*/\partial\theta_1) = \text{Var}(g^*).$$

This showed that the conditional score g^* (which is independent of θ_2) is optimal.

Example 1 Suppose X_i and $Y_i, i = 1, 2, \dots, n$ are independent of each other.

$$X_i \sim N(\theta_1, 1), \quad Y_i \sim N(\theta_1 + \theta_{2i}, 1), \quad i = 1, 2, \dots, n.$$

Clearly Y_i 's do not carry information for θ_1 . We may take $t = (y_1, \dots, y_n)$.

$$X_i|Y_i \sim X_i \sim N(\theta_1, 1).$$

Therefore $g^* = \bar{X} - \theta_1$ is the optimal estimating equation for θ_1 .

Exercise 1

$$X_i \sim N(\theta_1 + \theta_{2i}, 1), \quad Y_i \sim N(\theta_{2i}, 1), \quad i = 1, 2, \dots, n.$$

Show the optimal estimating equations is given by $g^* = \bar{X} - \bar{Y} - \theta_1$.

Exercise 2

$$X_i \sim N(\theta_i, \sigma^2), \quad Y_i \sim N(\theta_i, \sigma^2) \quad i = 1, 2, \dots, n$$

Let

$$T_i = X_i + Y_i, \quad i = 1, 2, \dots, n$$

Show the optimal estimating equation for σ in the presence of nuisance θ_i is

$$g^* = \frac{\partial \log f(x|t)}{\partial \theta_1} = -0.5n/\sigma + 0.5 \sum_{i=1}^n (x_i - y_i)^2 / \sigma^2.$$

Exercise 3 Godambe and Thompson (1974) defined

$$g^*(x, \theta_1, \theta_2) = C_1(\theta_1, \theta_2) \frac{\partial \log f}{\partial \theta_1} + C_2(\theta_1, \theta_2) \left[\left(\frac{\partial \log f}{\partial \theta_2} \right)^2 + \frac{\partial^2 \log f}{\partial \theta_2^2} \right].$$

If g^* is independent of θ_2 , then g^* is the optimal estimating equation for θ_1 in the class of all unbiased estimating functions such that they depend only on x and θ_1 .

A Truncation Example

Truncation is a common problem in epidemiologic and economics study. The observed pairs (Y, A) have joint density

$$(Y, A)|Y > A \sim \frac{f(y, \beta)}{\bar{F}(a, \beta)} \frac{\bar{F}(a, \beta)h(a, \gamma)}{\int \bar{F}(a, \beta)h(a, \gamma)}, \quad y > a.$$

In this case the conditional likelihood $f(Y|Y > A = a)$ is independent of the nuisance parameter γ . The resulting conditional score estimating function is optimal for β in the class of unbiased estimating functions not involving γ . However it may lose information when compared with the profile maximum likelihood estimate.

5.5 Projection Methods in Parametric Models

Godambe's basic ideas were further developed by Small and McLeish in a series of works in later of 80's and early 90's. Below we briefly describe their approaches.

Small and McLeish (1989) defined the centred likelihood ratio as

$$\psi_\eta(\theta) = \frac{L(\eta)}{L(\theta)} - 1,$$

where $L(\theta)$ and $L(\eta)$ are the likelihoods evaluated at the true parameter θ and a candidate parameter η (close to θ), respectively. We expand the centred likelihood ratio as

$$\psi_\eta(\theta) = \frac{L(\eta)}{L(\theta)} - 1 = \frac{\partial L(\theta)/\partial\theta}{L(\theta)}(\eta - \theta) + \frac{1}{2} \frac{\partial^2 L(\theta)/\partial\theta^2}{L(\theta)}(\eta - \theta)^2 + \dots$$

The k -th order E -sufficient subspace is defined as the linear space spanned by the

$$L^{(j)}/L = \frac{\partial^j L(\theta)/\partial\theta^j}{L(\theta)}, \quad j = 1, 2, \dots, k.$$

Naturally the E -sufficient space is defined as the closure of the k -th order E -sufficient subspace as $k \rightarrow \infty$. An unbiased estimating function $\phi(x, \theta)$ is called E -ancillary if $E_\eta \phi(X, \theta) = 0$ for any $\eta, \theta \in \Theta$, where Θ is the parameter space, and the expectation is taken with respect η . Clearly an E -ancillary estimating function is less sensitive to the true parameter. In practical applications we should avoid it. An unbiased estimating function is called k -th-order E -ancillary if

$$E_\eta[\phi(X, \theta)] = o(\eta - \theta)^k$$

for all θ and η such that $\eta \rightarrow \theta$. The k -th-order E -ancillary space is spanned by all k -th-order E -ancillary estimating functions.

From

$$\begin{aligned} o(\eta - \theta)^k &= E_\eta[\phi(X, \theta)] = \int \phi(x, \theta) L(x, \eta) dx \\ &= \int \phi(x, \theta) \left(\frac{L(\eta)}{L(\theta)} - 1 \right) L(x, \theta) dx \\ &= \sum_{j=1}^k E\{\phi(x, \theta) L^{(j)}/L(\theta)\} (\eta - \theta)^j / j! + \dots \end{aligned}$$

we conclude $E\{\phi(x, \theta) L^{(j)}/L(\theta)\} = 0$, $j = 1, 2, \dots, k$. Therefore the k -th order E -sufficient subspace and k -th-order E -ancillary space are orthogonal to each other.

Example 1 Suppose $X_1, \dots, X_n \sim N(\theta, 1)$. Let $L(\theta) = (2\pi)^{1/2} \exp\{-(x - \theta)^2/2\}$

$$\frac{\partial L}{\partial \theta} = (x - \theta)L, \quad \frac{\partial^2 L}{\partial \theta^2} = -L(\theta) + (x - \theta)^2 L(\theta), \quad \frac{\partial^3 L}{\partial \theta^3} = -3(x - \theta)L(\theta) + (x - \theta)^3 L(\theta), \dots$$

Therefore the centred likelihood ratio space is spanned by

$$\psi = (x - \theta), \quad \psi_1 = (x - \theta)^2 - 1, \quad \psi_2 = (x - \theta)^3 - 3(x - \theta), \dots$$

Easily we can show

$$E_\eta\{\psi(X, \theta)\} = \eta - \theta,$$

$$E_\eta[\psi_1(X, \theta)] = E_\eta[(X - \eta + \eta - \theta)^2 - 1] = E_\eta(X - \eta)^2 + (\eta - \theta)^2 - 1 = (\eta - \theta)^2 = o(\eta - \theta).$$

Therefore ψ_1 is a first order E -ancillary. Similarly

$$E_\eta[\psi_2(X, \theta)] = (\eta - \theta)^3 = o((\eta - \theta)^2),$$

which shows ψ_2 is a second order E -ancillary.

Since the k -th order ($k \geq 1$) E -ancillary is insensitive to the parameter θ , a good estimating function should be orthogonal to it. Easily we can show that

$$E[\psi\psi_1] = 0, \quad E[\psi\psi_2] = 0, \dots$$

Therefore $\psi(x, \theta) = x - \theta$ is a good estimating function for θ in this example.

Example 2 Suppose $X = (X_1, \dots, X_n)$ comes from a one-parameter exponential family

$$f(x; \theta) = a(x)b(\theta) \exp(\theta t),$$

where $T = t(X)$ is a complete sufficient statistic for θ . Then

$$\frac{\partial L(\theta)/\partial\theta}{L(\theta)} = \frac{a(x)b'(\theta)\exp(\theta t) + a(x)b(\theta)t\exp(\theta t)}{L(\theta)} = \frac{\partial b(\theta)}{\partial\theta} + t.$$

It can be shown that the k -th order E -sufficient subspace is spanned by t, t^2, \dots, t^k . Let $\psi(x, \theta)$ be an unbiased estimating function, it does not need to lie within the E -sufficient subspace. A good estimating function should fall in the E -sufficient subspace.

Let $\psi^*(\theta, T) = E[\psi(X, \theta)|T]$. It can be shown that ψ^* is the projection of estimating function ψ on the E -sufficient space. In fact we only need to show the residual $\psi - \psi^*$ is orthogonal to the E -sufficient space. For any measurable function $h(T)$

$$E[\{\psi(X, \theta) - \psi^*(T, \theta)\}h(T)] = E[\psi(X, \theta)h(T) - E\{\psi(X, \theta)h(T)|T\}] = 0.$$

The generalizes the Rao-Blackwellization of an estimator.

Example 3 Mixture distribution example. The likelihood is

$$L(\theta) = \prod_{i=1}^n \{\theta f(x_i) + (1-\theta)g(x_i)\},$$

where f and g are two known densities. It can be shown that

$$L(\eta) = L(\theta) \prod_{i=1}^n \{1 + (\eta - \theta)S_i(\theta)\}, \quad S_i(\theta) = \frac{f(x_i) - g(x_i)}{\theta f(x_i) + (1-\theta)g(x_i)}.$$

Furthermore we can show that

$$L(\eta) = L(\theta)\{1 + \sum_{i=1}^n k_i(\theta)T_i\},$$

where

$$T_k = \sum_{j_1 < j_2 < \dots < j_k} S_{j_1}(\theta)S_{j_2}(\theta)\dots S_{j_k}(\theta), \quad k_j(\theta) = (\eta - \theta)^j.$$

In other words $T_1(\theta), \dots, T_n(\theta)$ are the bases of the subspace spanned by

$$L^{(j)}/L = \frac{\partial^j L(\theta)/\partial\theta^j}{L(\theta)}, \quad j = 1, 2, \dots$$

Suppose we are interested in projecting $\psi(\theta)$ onto the space generated by T_1, \dots, T_n , we can write

$$\psi_S = \sum_{i=1}^n b_i(\theta)T_i(\theta).$$

We require the residual $\psi(\theta) - \psi_S(\theta)$ be insensitive to the parameter θ , i.e., $E_\eta[\psi(\theta) - \psi_S(\theta)] = o((\eta - \theta)^n)$. Note that

$$E_\eta[\psi_S(\theta)] = \sum_{i=1}^n b_i(\theta) E_\eta[T_i(\theta)] = \sum_{i=1}^n b_i(\theta) \binom{n}{i} [E_\eta S_1(\theta)]^i,$$

where

$$S_1 = \frac{f(x_1) - g(x_1)}{\theta f(x_1) + (1 - \theta)g(x_1)}.$$

Note that

$$\frac{\eta f(x_1) + (1 - \eta)g(x_1)}{\theta f(x_1) + (1 - \theta)g(x_1)} = 1 + \frac{f(x_1) - g(x_1)}{\theta f(x_1) + (1 - \theta)g(x_1)}(\eta - \theta) = 1 + S_1(\eta - \theta).$$

$$\begin{aligned} E_\eta[S_1(\theta)] &= E_\theta[S_1(\theta)L(\eta)/L(\theta)] \\ &= E_\theta S_1^2(\theta)(\eta - \theta) \\ &= J(\eta - \theta), \quad J = E_\theta[S_1^2(\theta)]. \end{aligned}$$

On the other hand

$$E_\eta[\psi(\theta)] = E_\theta[\psi(\theta)L(\eta)/L(\theta)] = \sum_{j=1}^n E_\theta\{T_j\psi(\theta)\}(\eta - \theta)^j.$$

Therefore we need to choose the coefficients of the projection as

$$b_i(\theta) = \binom{n}{j}^{-1} \frac{E_\theta\{T_j\psi(\theta)\}}{J^j(\theta)}.$$

As an example suppose we are interested in testing $\theta = \theta_0$ vs. $\theta > \theta_0$. By the Neyman–Pearson lemma, it is well known that the likelihood ratio statistic is the most powerful test for the local alternative. In general if the underlying parametric family is not from the exponential family, then a two sided most powerful test does not exist. A practical method is to symmetrize the score statistic $S(\theta) = \sum_{i=1}^n S_i$. Let $\psi(\theta) = S^2(\theta) - E\{S^2(\theta)\}$. The projection is (exercise)

$$\psi_P = 2T_2(\theta) + c(\theta)T_1(\theta), \quad c(\theta) = E_\theta[S_1^3(\theta)]/J(\theta).$$

5.6 Reduce Sensitivity with Respect to Nuisance Parameters

To handle the nuisance parameters, Godambe (1984) required his estimating function to be independent of the nuisance parameters, which is a very strong requirement. Let θ be the parameters of interest and ξ be the nuisance parameter. The commonly used profile maximum likelihood method is sensitive to nuisance parameters. A popular method used to repair the maximum likelihood estimating equations is the projected score approach (Small and McLeish 1989; Waterman and Lindsay 1996). The aim of this procedure is to reduce the impact of the nuisance parameters on the estimating equation for the primary parameter. This is achieved by subtracting from the primary-score function from its projection onto the space V_t , where V_t is the closed linear space spanned by the functions $V_k = \partial^k f / \partial \xi^k / f, k = 1, 2, \dots$. Under regularity conditions V_t approaches the E -sufficient subspace for the nuisance parameter as $k \rightarrow \infty$ (Small and McLeish 1989). The basis formed by $V_k, k = 1, 2, \dots$ is referred to as the Bhattacharyya basis. An unbiased estimating function ϕ can be projected onto the k -th order E -sufficient space spanned by $E_\xi^{(k)} = \{V_i, i = 0, .1, 2, \dots, k\}$, denoted the projection as ϕ_k . Since the residual $\phi - \phi_k$ is orthogonal to $E_\xi^{(k)}$, it is less sensitive to the nuisance parameter ξ . In practical applications the second order ($k = 2$) projection is often sufficient.

A Nice Property of the Second Order E -ancillary Estimating Function

For simplicity we only discuss second order E -ancillary estimating functions. Let $\psi(\theta, \xi) = \sum_{i=1}^n \psi(x_i, \theta, \xi)$ be an estimating function for θ . In general ξ may be replaced by a consistent estimator $\hat{\xi}_\theta$. Suppose it is root- n consistent, i.e., $\hat{\xi}_\theta - \xi = O_p(n^{-1/2})$, for example $\hat{\xi}_\theta$ can be the MLE for fixed θ . By using Taylor's expansion, we have

$$\sum_{i=1}^n [\psi_i(\theta, \hat{\xi}(\theta)) - \psi_i(\theta, \xi)] = \sum_{i=1}^n \left[\frac{\partial \psi_i}{\partial \xi} (\hat{\xi} - \xi) + \frac{1}{2} \frac{\partial^2 \psi_i}{\partial \xi^2} (\hat{\xi} - \xi)^2 + \dots \right]. \quad (5.6.2)$$

If $E[\partial \psi / \partial \xi] \neq 0$ and $E[\partial^2 \psi / \partial \xi^2] \neq 0$, then

$$\psi(\theta, \hat{\xi}_\theta) - \psi(\theta, \xi) = O_p(n) O_p(n^{-1/2}) + O_p(n) O_p(n^{-1}) = O_p(n^{1/2}).$$

However for second order E -ancillary with respect to nuisance ξ , this difference becomes $O_p(1)$.

An estimating equation $\psi(x, \theta, \xi)$ is called second order E -ancillary with respect to nuisance ξ if it satisfies

$$\int \psi(x, \theta, \xi) \partial^i f(x, \theta, \xi) / \partial \xi^i dx = 0, \quad i = 0, 1, 2.$$

From

$$0 = \frac{\partial}{\partial \xi} \int \psi f dx = \int \left\{ \frac{\partial \psi}{\partial \xi} f(x) dx + \psi \frac{\partial f}{\partial \xi} dx \right\},$$

we obtain

$$\int \frac{\partial \psi}{\partial \xi} f(x) dx = 0, \text{ or } E[\partial \psi / \partial \xi] = 0.$$

Again the second order E -ancillary property

$$0 = \frac{\partial}{\partial \xi} \int \psi \frac{\partial f}{\partial \xi} dx = \int \left\{ \frac{\partial \psi}{\partial \xi} \frac{\partial f(x)}{\partial \xi} dx + \psi \frac{\partial^2 f}{\partial \xi^2} dx \right\},$$

gives

$$\int \frac{\partial \psi}{\partial \xi} \frac{\partial f(x)}{\partial \xi} dx = 0$$

Finally from

$$0 = \frac{\partial}{\partial \xi} \int \frac{\partial \psi}{\partial \xi} f(x) dx = \int \left\{ \frac{\partial^2 \psi}{\partial \xi^2} f(x) dx + \frac{\partial \psi}{\partial \xi} \frac{\partial f(x)}{\partial \xi} dx \right\},$$

we arrived at

$$E[\partial^2 \psi / \partial \xi^2] = 0.$$

As a consequence in the Taylor expansion (5.6.2), the difference has an order

$$O_p(n^{1/2}) O_p(n^{-1/2}) + O_p(n^{1/2}) O_p(n^{-1}) = O_p(1).$$

This shows that the second order E -ancillary estimating equation is less sensitive for the nuisance parameter ξ . Moreover assuming the uniform integrability, from (5.6.2), $E\{\psi(\theta, \xi_\theta) - \psi(\theta, \xi)\} = o(1)$.

Any given estimating function ψ can be projected onto the second E -ancillary nuisance space, i.e., choose $c_i(\theta, \eta)$, $i = 1, 2$ such that

$$\psi^* = \psi - c_1(\theta, \xi)V_1 - c_2(\theta, \xi)V_2, \quad V_1 = \frac{\partial f / \partial \xi}{f}, \quad V_2 = \frac{\partial^2 f / \partial \xi^2}{f},$$

where c_1 and c_2 are chosen such that

$$E[\psi^* V_1] = E[\psi^* V_2] = 0.$$

This leads to the projection estimating function ψ^* to be less sensitive with respect to ξ , when compared with ψ .

Example 1 Suppose we have a parametric model $f(x, \theta, \xi)$, where θ is the parameter of interest and ξ is a nuisance parameter. The scores for θ and ξ , are, respectively

$$\psi_1(x) = \frac{\partial \log f(x, \theta, \xi)}{\partial \theta}, \quad \psi_2(x) = \frac{\partial \log f(x, \theta, \xi)}{\partial \xi}.$$

Consider the two cases.

- (1) If $E[\psi_1 \psi_2] = 0$, then the space for the parameter of interest and the nuisance parameter space are orthogonal to each other. We can use $\sum_{i=1}^n \psi_1(x_i, \theta, \xi)$ and $\sum_{i=1}^n \psi_2(x_i, \theta, \xi) = 0$ to estimate θ .
- (2) If $E[\psi_1 \psi_2] \neq 0$, then based on the general theory developed above we should use $\psi_1 - c\psi_2$ as estimating function for θ , where $c = E(\psi_1 \psi_2)/E(\psi_2^2)$ is chosen such that

$$E[(\psi_1 - c\psi_2)\psi_2] = 0.$$

On the other hand the profile maximum likelihood method amounts to solving

$$\sum_{i=1}^n \psi_2(x_i, \theta, \xi) = 0$$

for fixed θ . Denote the solution as $\xi = \xi(\theta)$. Differentiating both sides of the estimating equation above with respect to θ , we have

$$0 = \sum_{i=1}^n \frac{\partial \psi_2(x_i, \theta, \xi)}{\partial \theta} + \sum_{i=1}^n \frac{\partial \psi_2(x_i, \theta, \xi)}{\partial \xi} \frac{\partial \xi}{\partial \theta} = 0,$$

or

$$\frac{\partial \xi}{\partial \theta} = - \left(\sum_{i=1}^n \frac{\partial \psi_2(x_i, \theta, \xi)}{\partial \xi} \right)^{-1} \left(\sum_{i=1}^n \frac{\partial \psi_2(x_i, \theta, \xi)}{\partial \theta} \right).$$

Differentiating with respect to θ in the profile log-likelihood $\ell_p(\theta) = \sum_{i=1}^n \log f(x_i, \theta, \xi(\theta))$, we have

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \sum_{i=1}^n \frac{\partial \log f(x_i, \theta, \xi(\theta))}{\partial \theta} + \sum_{i=1}^n \frac{\partial \log f(x_i, \theta, \xi(\theta))}{\partial \xi} \frac{\partial \xi(\theta)}{\partial \theta} \\ &= \sum_{i=1}^n \left[\psi_1(x_i, \theta, \xi(\theta)) + \psi_2(x_i, \theta, \xi(\theta)) \frac{\partial \xi(\theta)}{\partial \theta} \right]. \end{aligned} \quad (5.6.3)$$

Since ψ_2 is the score for ξ , it satisfies

$$E(\partial \psi_2 / \partial \xi) = E \left(\frac{\partial^2 \log f(x, \theta, \xi)}{\partial \xi^2} \right) = -E \left(\frac{\partial \log f(x, \theta, \xi)}{\partial \xi} \right)^2 = -E(\psi_2^2).$$

Differentiating $E(\partial \log f / \partial \theta) = 0$ with respect to ξ , we have

$$0 = \int \frac{\partial^2 \log f}{\partial \theta \partial \xi} f dx + \int \frac{\partial \log f}{\partial \theta} \frac{\partial f}{\partial \xi} dx,$$

or

$$E \left(\frac{\partial^2 \log f}{\partial \theta \partial \xi} \right) + E \left(\frac{\partial \log f}{\partial \theta} \frac{\partial \log f}{\partial \xi} \right) = 0, \quad E(\partial \psi_2 / \partial \theta) + E(\psi_1 \psi_2) = 0.$$

Therefore asymptotically

$$\frac{\partial \xi}{\partial \theta} = -\{-E(\psi_2^2)\}^{-1}\{-E(\psi_1 \psi_2)\} = -E(\psi_1 \psi_2)/E(\psi_2^2).$$

Equation (5.6.3) implies that asymptotically the score obtained from the profile likelihood is orthogonal to ψ_2 . In other words the projection method and the profile likelihood method are asymptotically equivalent. However, this is true only if both ψ_1 and ψ_2 are genuine scores for θ and ξ , respectively. The estimating equation based approach does not need to be true.

Example 1

$$X_1, \dots, X_n \sim h(x, \theta, \xi) = \xi f(x, \theta) + (1 - \xi)g(x, \theta).$$

We wish to estimate θ and treat ξ as a nuisance parameter. The first component of the score vector is

$$S_1(\theta, \xi) = \sum_{i=1}^n \frac{\partial \log h(x_i, \theta, \xi)}{\partial \theta} = \sum_{i=1}^n \frac{\xi \partial f(x_i, \theta) / \partial \theta + (1 - \xi) \partial g(x_i, \theta) / \partial \theta}{h(x_i, \theta, \xi)}.$$

Denote the second component of the score vector as

$$S_2(\theta, \xi) = \sum_{i=1}^n \frac{\partial \log h(x_i, \theta, \xi)}{\partial \xi} = \sum_{i=1}^n \frac{f(x_i, \theta) - g(x_i, \theta)}{h(x_i, \theta, \xi)}.$$

As shown in Example 2, the nuisance E -sufficient space spanned by $\partial^j L(\theta, \xi) / \partial \xi^j / L$, $j = 1, 2, \dots$ has bases

$$T_k(\theta, \eta) = \sum_{j_1 < j_2 < \dots < j_k} S_2(x_{j_1}, \theta, \xi) S_2(x_{j_2}(\theta, \xi) \dots S_2(x_{j_k}, \theta, \eta)), \quad k = 1, 2, \dots,$$

Note that

$$E[S_1(\theta, \xi) T_k(\theta, \xi)] = 0, \quad k > 1.$$

Therefore we only need to project S_1 onto the space spanned by $\sum_{i=1}^n S_2(x_i, \theta, \xi)$, i.e., we need to find a constant $k(\theta, \xi)$ such that

$$E[(S_1 - kS_2)S_2] = 0.$$

Clearly the solution for k is

$$k(\theta, \xi) = \text{cov}(S_1, S_2) / \text{Var}(S_2).$$

Exercise

$$(X_{1i}, X_{2i}), i = 1, 2, \dots, n \sim \xi f(x_{1i}, \theta) f(x_{2i}, \theta) + (1 - \xi) g(x_{1i}, \theta) g(x_{2i}, \theta).$$

Project the score with respect to θ onto the nuisance space generated by $\partial^j L(\theta, \xi) / \partial \xi^j$, $j = 1, 2, \dots$

Chapter 6

Projection Methods in General Semiparametric Models

The projection method can be used not only in finitely many parameter problems but also in nuisance function or infinite many nuisance parameters cases. First we introduce the concept of regular estimator and influence function.

An estimator $\hat{\theta}$ is called asymptotically linear if it is asymptotically equivalent to a simple average, i.e. there is a function $\psi(x)$ of a single observation such that

$$\sqrt{n}(\hat{\theta} - \theta) = \sum_{i=1}^n \psi(x_i)/\sqrt{n} + o_p(1),$$

where $E\{\psi(X)\} = 0$ and $E\{\psi(X)\psi^T(X)\}$ is finite and non-singular. The function $\psi(x)$ is called influence function.

One can imagine that the data are generated by a parametric model that satisfies the semiparametric assumptions and contains the truth. Such a model is called as a parametric submodel. One can obtain the classical Cramer–Rao bound for a parametric submodel. Any regular semiparametric estimator, i.e., one that is consistent and asymptotically normal under the semiparametric assumption, has an asymptotic variance that is comparable to the Cramer–Rao bound of a semiparametric model. As a consequence the asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer–Rao bounds for all parametric submodels.

Next we discuss a well-known non-regular estimator example given by Hodges (see Bickel et al. 1993 and Tsiatis 2006).

Superefficient Estimator Example

Let \bar{X} be the sample mean from an i.i.d. $N(\mu, 1)$ sample. It is the maximum likelihood estimator of μ . Define a new estimator

$$\hat{\mu} = \begin{cases} \bar{X}, & \text{if } |\bar{X}| > n^{-1/4} \\ 0 & \text{if } |\bar{X}| \leq n^{-1/4}. \end{cases}$$

Note that $\hat{\mu}$ is equal to \bar{X} with probability approaching one if $\mu \neq 0$. On the other hand $\hat{\mu}$ is equal to zero with probability approaching one if $\mu = 0$. Therefore the asymptotic distribution of $\hat{\mu}$ is the same as the sample mean if $\mu \neq 0$ but has asymptotic variance zero if $\mu = 0$. In other words this estimator is super-efficient at $\mu = 0$.

Unfortunately, this estimator can behave badly in a neighborhood of zero. It is not difficult to show that if the true value of μ_n is $n^{-1/3}$, then $\sqrt{n}(\hat{\mu} - \mu_n)$ diverges to $-\infty$.

Modern approaches in statistics to this problem use an asymptotic minimax criteria for evaluating efficiency by restricting attention to the class of estimators satisfying uniform conditions that rules out superefficient estimators.

Below we give some semiparametric model examples where the projection method and the information bound calculation are easily derived. However, in many semiparametric models, we may not be able to find an estimator that achieves the lower bound even if we can find the bound. As a thumb of rule, in general, the profile maximum likelihood estimate can achieve this lower bound. Unfortunately the profile maximum likelihood estimate may not be always available. Sometimes, it may also be inconsistent, see, Wellner's (2015) Le Came lecture in the joint statistical meeting. For more rigorous approaches, we refer readers to advanced works, for example, Begun et al. (1983), Chamberlain (1987), Newey (1990), Bickel et al. (1993) and Tsiatis (2006).

6.1 Projection Method for the Mean Estimation and Linear Regression Model

1. Mean estimation

Suppose we are interested in estimating the population mean $\mu = E(X)$ and let the distribution of X be unrestricted except for regularity conditions such as $\text{Var}(X) < \infty$. Naturally we can expect the sample mean to be the most efficient one. In fact the normal distribution family can be considered as a submodel. The maximum likelihood estimator is the sample mean under such a submodel. The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer–Rao bounds for all parametric submodels, including the normal. Therefore, sample mean should be the most efficient.

2. Mean estimation in the conditional model

Suppose we have a parametric model $f(y|x) = f(y|x, \theta)$, where the marginal density $g(x)$ of X is not specified. We would like to show that $n^{-1} \sum_{i=1}^n \mu(x_i, \hat{\theta})$ is the optimal estimate of μ , where $\mu(x, \theta) = \int y f(y|x, \theta) dy$ and $\hat{\theta}$ is the MLE based on the model $f(y|x, \theta)$.

Intuitively this is obvious since $\hat{\theta}$ is the MLE and $\hat{G}(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$ is the nonparametric MLE for G . As a consequence

$$\hat{\mu}_{MLE} = \int \mu(x, \hat{\theta}) d\hat{G}(x)$$

should be optimal.

Easily, we can show that

$$\sqrt{n}(\hat{\mu}_{MLE} - \mu) \rightarrow N(0, \sigma^2),$$

$$\sigma^2 = E(\partial\mu/\partial\theta)^T I^{-1}(\theta)E(\partial\mu/\partial\theta) + \text{Var}\{\mu(X)\}, \quad I(\theta) = -E\left\{\frac{\partial^2 \log f(Y|X, \theta)}{\partial\theta^T \partial\theta}\right\}.$$

Next, we calculate the information bound. For any submodel $g(x) = g(x, \xi)$, the log-likelihood is

$$\ell = \log f(y|x, \theta) + \log g(x, \xi).$$

The tangent space is spanned by

$$\frac{\partial\ell}{\partial\theta} = \frac{\partial \log f(y|x, \theta)}{\partial\theta} := s(y|x),$$

and

$$\frac{\partial\ell}{\partial\xi} = \frac{\partial \log g(x, \xi)}{\partial\xi} := \alpha(x).$$

Note that $E[s(Y|x)|x] = 0$ and $E[\alpha(X)] = 0$.

Let $\hat{\theta}$ and $\hat{\xi}$ be the MLEs. Then μ can be estimated by $\hat{\mu} = \mu(\hat{\theta}, \hat{\xi})$. Using Taylor's expansion,

$$\begin{aligned} \hat{\mu} &= \int yf(y|x, \hat{\theta})g(x, \hat{\xi})dydx \\ &= \mu(\theta, \xi) + \int y \frac{\partial \log f(y|x, \theta)}{\partial\theta} f(y|x, \theta)g(x, \xi)dydx(\hat{\theta} - \theta) \\ &\quad + \int yf(y|x, \theta) \frac{\partial \log g(x, \xi)}{\partial\xi} g(x, \xi)dydx(\hat{\xi} - \xi) + \dots \end{aligned}$$

Therefore

$$\sqrt{n}(\hat{\mu} - \mu) = E\left[\frac{\partial\mu(X, \theta)}{\partial\theta}\right]\sqrt{n}(\hat{\theta} - \theta) + E\left[\{\mu(X, \theta) - \mu\}\frac{\partial \log g(X, \xi)}{\partial\xi}\right]\sqrt{n}(\hat{\xi} - \xi) + \dots$$

The first term is the contribution from estimating θ based on the parametric model $f(y|x, \theta)$, which is independent of the choice of the marginal density $g(x, \xi)$. Note that

$$\sqrt{n}(\hat{\xi} - \xi) \rightarrow N(0, J^{-1}), \quad J = E\left[\frac{\partial \log g(x, \xi)}{\partial\xi}\right]^2.$$

Specifically, taking

$$\frac{\partial \log g(x, \xi)}{\partial \xi} = \mu(x, \theta) - \mu,$$

we can show that the asymptotic variance for the second term is no smaller than $E\{\mu(x, \theta) - \mu\}^2$. However we already observed that $\hat{\mu}_{MLE} = n^{-1} \sum_{i=1}^n \mu(x_i, \hat{\theta})$ achieves this lower bound.

3. Linear model

Consider a linear regression model in which

$$Y = x\beta + \epsilon, \quad \epsilon \sim f(\epsilon),$$

where $E(\epsilon) = 0$ and X and ϵ are independent each other. The log-likelihood with a single observation is

$$\ell = \log f(y - x\beta) + \log g(x).$$

Consider a submodel such that $f(\epsilon) = f(\epsilon, \eta)$ and $g(x) = g(x, \eta)$. Under this model,

$$\ell = \log f(y - x\beta, \eta) + \log g(x, \eta).$$

The score is

$$\frac{\partial \ell}{\partial \beta} = -x \frac{\partial \log f(\epsilon, \eta)}{\partial \epsilon} \Big|_{\epsilon=y-x\beta} = -xS(y - x\beta),$$

where

$$S(\epsilon) = \frac{\partial \log f(\epsilon)}{\partial \epsilon}.$$

The score for the nuisance parameter η is

$$\frac{\partial \ell}{\partial \eta} = \frac{\partial \log f(y - x\beta, \eta)}{\partial \eta} + \frac{\partial \log g(x, \eta)}{\partial \eta}.$$

We can decompose

$$-xS(y - x\beta) = -\{x - E(X)\}S(y - x\beta) - E(X)S(y - x\beta).$$

Note that

$$E \left[\{x - E(X)\}S(y - x\beta) \left\{ \frac{\partial \log f(y - x\beta, \eta)}{\partial \eta} + \frac{\partial \log g(x, \eta)}{\partial \eta} \right\} \right] = 0.$$

In other words $-\{x - E(X)\}S(y - x\beta)$ is orthogonal to the functional space generated by the nuisance parameter score. Therefore, we should use

$$\sum_{i=1}^n \{x_i - E(X)\} S(y_i - x_i \beta) = 0$$

as the estimating equation for β . We may replace $E(X)$ by $n^{-1} \sum_{i=1}^n x_i$. In general, $S = \partial \log f(\epsilon)/\partial \epsilon$ is unknown. Fortunately, even if $S(y - x\beta)$ is misspecified, this estimating equation still has mean 0, which implies that it still produces a consistent estimator of β , though with some loss of efficient for the mis-specified score function S .

Next we consider a parametric linear regression $f(y - x\beta) = f_0(y - \alpha - x\beta)$, where the form of f_0 is known. The log-likelihood is

$$\ell = \sum_{i=1}^n \log f_0(y_i - \alpha - x_i \beta).$$

$$S_\beta = \frac{\partial \ell}{\partial \beta} = -x S(\epsilon), \quad S(\epsilon) = \frac{\partial \log f_0(\epsilon)}{\partial \epsilon}.$$

$$S_\alpha = \frac{\partial \ell}{\partial \alpha} = -S(\epsilon).$$

The projection of S_β onto S_α is

$$E[S_\beta S_\alpha][E S_\alpha^2]^{-1} S_\alpha.$$

It can be shown that

$$S_\beta - E[S_\beta S_\alpha][E S_\alpha^2]^{-1} S_\alpha = -\{X - E(X)\} S(\epsilon).$$

Therefore, the projection estimating function in the full parametric model and semi-parametric model are coincident. In fact, Bickel (1982) used adaptive method to estimate β if f is symmetric and unknown. He proved that the adaptive estimator achieves full efficiency for β , i.e., the asymptotic variance for the estimate of β is the same whether f is known.

6.2 Information Contained in the Conditional Expectation Model

In this Section, we calculate the information contained in a conditional expectation model. Consider a semiparametric model, that for some given $\psi(y, x, \beta)$,

$$\mathcal{P} = [P : E_P\{\psi(Y, X, \beta)|X = x\} = 0].$$

Clearly, there are infinitely many probability distributions satisfying this model. We consider a specific conditional density

$$f(y|x) = f(y|x, \beta, \lambda),$$

where f is a known density with parameter of interest β and nuisance parameter λ .

The information matrix for β and λ is

$$I(\beta, \lambda) = -E_x \begin{pmatrix} \frac{\partial \log f}{\partial \beta \partial \beta^T} & \frac{\partial \log f}{\partial \beta \partial \lambda^T} \\ \frac{\partial \log f}{\partial \lambda \partial \beta^T} & \frac{\partial \log f}{\partial \lambda \partial \lambda^T} \end{pmatrix} =: \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where E_x is the conditional expectation for given x . For notation simplicity, we will use E_x or $E(\cdot|x)$ and V_x or $V(\cdot|x)$, exchangeably. The information for β is

$$(I^{11})^{-1} = I_{11} - I_{12}I_{11}^{-1}I_{21}.$$

Since this argument applies for any submodel $f(y|x, \beta, \lambda)$, the information bound is

$$\min_f (I_{11} - I_{12}I_{11}^{-1}I_{21}).$$

Next, we evaluate this bound in the conditional expectation model. For any x ,

$$\int \psi(y, x, \beta) f(y|x, \beta, \lambda) dy = 0.$$

Differentiating with respect to β and λ , respectively, we have

$$\int \frac{\partial \psi(y, x, \beta)}{\partial \beta} f(y|x, \beta, \lambda) dy + \int \psi(y, x, \beta) \frac{\partial f(y|x, \beta, \lambda)}{\partial \beta} = 0,$$

and

$$\int \psi(y, x, \beta) \frac{\partial f(y|x, \beta, \lambda)}{\partial \lambda} = 0.$$

Equivalently

$$E[\partial \psi / \partial \beta | x] + E[\psi \partial f / \partial \beta | x] = 0,$$

and

$$E[\psi \partial \log f / \partial \lambda | x] = 0.$$

Using the fact that

$$V_x[h - E(h^T g)V^{-1}(g)g] \geq 0,$$

or

$$V_x(h) - 2E_x(h^T g)V_x^{-1}(g)E_x(hg^T) + E_x(h^T g)V_x^{-1}(g)V_x(g)V_x^{-1}(g)E_x(gh^T) \geq 0,$$

we have

$$V_x(h) \geq E_x(h^T g)V_x^{-1}(g)E_x(hg^T).$$

Letting

$$h = \begin{pmatrix} \partial \log f / \partial \beta \\ \partial \log f / \partial \lambda \end{pmatrix}, \quad g = \psi(y, x, \beta),$$

we can show

$$V \begin{pmatrix} \partial \log f / \partial \beta & |x \\ \partial \log f / \partial \lambda & \end{pmatrix} \geq \begin{pmatrix} E(\partial \psi / \partial \beta | x) \\ 0 \end{pmatrix} V^{-1}(\psi | x) (E(\partial \psi / \partial \beta^T | x), 0).$$

Next we need to use the following fact for a positive matrix:

If

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22,1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \geq \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix},$$

then

$$\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22,1}^{-1} \end{pmatrix} \geq \begin{pmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{pmatrix} = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix},$$

i.e.

$$A_{11}^{-1} \geq C.$$

Since we have

$$\begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}^{-1} \geq \begin{pmatrix} E[E(\partial \psi / \partial \beta | x)V^{-1}(\psi | x)E(\partial \psi / \partial \beta^T | x)] & 0 \\ 0 & 0 \end{pmatrix},$$

therefore

$$(I^{11})^{-1} = I_{11} - I_{12}I_{11}^{-1}I_{21} \geq E[E(\partial \psi / \partial \beta | x)V^{-1}(\psi | x)E(\partial \psi / \partial \beta^T | x)].$$

The bound (Chamberlain 1987) is achievable if

$$\frac{\partial \log f}{\partial \beta} = -E(\partial \psi / \partial \beta | x)V^{-1}(\psi | x)\psi. \quad (6.2.1)$$

In fact this is the optimal function defined in the linear class of estimating functions $\{a(x)\psi(y, x, \beta)\}$, where $a(x)$ is measurable. Note that

$$E[a(x)\psi(Y, x, \beta)|x] = a(x)E[\psi(Y, x, \beta)|x] = 0.$$

Godambe's (1960) estimating function theory gives the optimal choice

$$a(x) = E[\partial\psi(Y, x, \beta)/\partial\beta|x]V^{-1}[\psi(Y, x, \beta)|x].$$

In general the class of regular estimating functions considered by Chamberlain (1987) is larger than the class of linear estimating functions defined above. However, the optimal one is the same in both classes, i.e., given in (6.2.1).

Next, we project the true unknown score or likelihood ratio into the linear subspace spanned by a given class of estimating functions.

1. Projection of score function

Let $S(x, \theta) = \partial \log f(x, \theta)/\partial\theta$ be the true score function which is unknown in general, where θ is a $p \times 1$ unknown vector parameter. For a given vector $r \times 1$ estimating equation $g(X, \theta)$ with $r \geq p$, we are interested in projecting $S(X, \theta)$ into the linear space \mathcal{G} ,

$$\mathcal{G} = \{\psi(x, \theta) | \psi = zg(x, \theta)\},$$

where $z_{p \times r}$ is a non-random allowed matrix. Denote the projection as $z_P g(x, \theta)$. Again the residual

$$S(X, \theta) - z_P g(x, \theta)$$

should be orthogonal to \mathcal{G} . Hence we must have

$$z_P = \text{Cov}(S, g)[E\{gg^T\}]^{-1}.$$

By differentiating $E_\theta g(X, \theta) = 0$, we find

$$\begin{aligned} \text{Cov}(S, g) &= E[Sg] = \int \partial f(x, \theta)/\partial\theta g(x, \theta) dx = - \int f(x, \theta) \partial g(x, \theta)/\partial\theta dx \\ &= E[\partial g^T(x, \theta)/\partial\theta]. \end{aligned}$$

Therefore the projection of score onto the space \mathcal{P} is

$$E[\partial g(X, \theta)/\partial\theta]^T [Eg(X, \theta)g^T(X, \theta)]^{-1} g(X, \theta).$$

2. Projection of Likelihood Ratio

Next, we project the likelihood ratio $f(x, \eta)/f(x, \theta)$ in the neighbour of θ into the space spanned by g , i.e., \mathcal{G} . We need to find constants c_1 and c_2 (which may depend on θ) such that

(1)

$$E_\theta \left\{ \frac{f(X, \eta)}{f(X, \theta)} - c_1 - c_2 g^T(X, \theta) \right\} = 0,$$

and the residual is orthogonal to \mathcal{G} , i.e.,

$$(2) \quad E_\theta \left[\left\{ \frac{f(X, \eta)}{f(X, \theta)} - c_1 - c_2 g^T(X, \theta) \right\} g(X, \theta) \right] = 0.$$

Clearly, from (1) the solution of c_1 is 1, and from (2) the solution of c_2 is given by

$$c_2 = E_\theta[f(X, \eta)g(X, \theta)/f(X, \theta)][E_\theta\{g^T(X, \theta)g(X, \theta)\}]^{-1}.$$

Expanding $f(x, \eta)$ at $\eta = \theta$, we have

$$f(x, \eta) = f(x, \theta) + \partial f(x, \theta)/\partial \theta(\eta - \theta) + \dots.$$

Therefore

$$E_\theta[f(X, \eta)g(X, \theta)/f(X, \theta)] \approx \int [\{\partial f(x, \theta)/\partial \theta\}g(x, \theta)]dx(\eta - \theta) = -E_\theta[\partial g(X, \theta)/\partial \theta](\eta - \theta).$$

Finally in a neighbourhood of θ ,

$$f(x, \eta)/f(x, \theta) = 1 - E_\theta\{\partial g(X, \theta)/\partial \theta\}\{E_\theta g^T(X, \theta)g(X, \theta)\}^{-1}g(x, \theta)(\eta - \theta) + \dots.$$

The projection of likelihood ratio is

$$L_P(\eta, \theta) = \prod_{i=1}^n [1 + E_\theta\{\partial g(x, \theta)/\partial \theta\}\{E_\theta g^T(X, \theta)g(X, \theta)\}^{-1}g(x, \theta)(\theta - \eta)].$$

We may use this projection to make inference for θ . For example, we may study point estimation and likelihood ratio limiting distribution, etc. The current version is a slight generalization of McLeish and Small (1992)'s results. Moreover, the projection of likelihood ratio very much resembles the empirical likelihood ratio, which will be discussed in Chap. 8.

Exercise Study the large sample property of the projection of log-likelihood ratio statistic $\log L_P(\hat{\theta}, \theta_0)$.

6.3 Projection Method in a Two Sample Density Ratio Model

Suppose we have a two-sample density ratio model given by

$$X_1, \dots, X_m \sim f(x), \quad Z_1, \dots, Z_n \sim g(x) = \frac{f(x) \exp(x\beta)}{\int f(x) \exp(x\beta) dx} =: \exp(\alpha + x\beta)f(x),$$

where α is a normalizing constant. The baseline density $f(x)$ is unspecified. Readers may find a natural connection between this model and the logistic regression model based on the case and control study in Chap. 11. Our main interest is the parameter β . We may consider a submodel $f(x) = f(x, \eta)$. The log-likelihood is

$$\ell = \sum_{i=1}^N [\log f(t_i, \eta) + \delta_i t_i \beta - \delta_i \log \int f(x, \eta) \exp(x\beta) dx],$$

where $\delta_i = 0, i = 1, 2, \dots, m$; and $\delta_j = 1, j = m+1, \dots, N = m+n$, and t_1, \dots, t_N are pooled data. The score for β is

$$S_\beta = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^N \delta_i \{t_i - a(\beta, \eta)\}, \quad a = E_1(t),$$

where E_1 denotes expectation with respect to g . The score for η is

$$S_\eta = \frac{\partial \ell}{\partial \eta} = \sum_{i=1}^N \left[\frac{\partial \log f(t_i, \eta)}{\partial \eta} - \delta_i \frac{\int \partial f / \partial \eta \exp(x\beta) dx}{\int \exp(x\beta) f(x) dx} \right].$$

Let $h(t) = \partial \log f(t) / \partial \eta$. Then

$$S_\eta = \sum_{i=1}^N [h(t_i) - \delta_i E_1\{h(T)\}].$$

$$E[S_\eta] = \int [m + n \exp(\alpha + t\beta)] h(t) f(t) dt - n \int \exp(\alpha + t\beta) h(t) dt = m \int h(t) f(t) dt = 0.$$

The nuisance parameter space is spanned by all measurable functions $h(t)$. Define

$$S_A = \sum_{i=1}^N t_i \{\delta_i - \pi(t_i)\}, \quad (6.3.2)$$

where

$$\pi(t_i) = \frac{\exp(\alpha + t_i \beta)}{1 + \rho \exp(\alpha + t_i \beta)}, \quad \rho = n/m, \quad \exp(-\alpha) = \int \exp(x\beta) f(x) dx.$$

Note that S_β can be decomposed as

$$S_\beta = \sum_{i=1}^n \delta_i \{t_i - a(\beta, \eta)\} = \sum_{i=1}^n t_i \{\delta_i - \pi(t_i)\} + \sum_{i=1}^n \{t_i \pi(t_i) - \delta_i a(\beta, \eta)\}.$$

Moreover

$$\begin{aligned} E[S_A S_\eta] &= m E_0[Th(T) \exp(\alpha + T\beta)] - E_0[\{m + n \exp(\alpha + T\beta)\} Th(T) \pi(T)] \\ &= m E_0[Th(T) \exp(\alpha + T\beta)] - m E_0[Th(T) \exp(\alpha + T\beta)] = 0, \end{aligned}$$

where E_0 is the expectation with respect to f . In other words S_A is orthogonal to the nuisance parameter space. Therefore, we can use $S_A = 0$ as estimating equations for (α, β) . Fortunately, they do not depend on the nuisance density f . We will discuss this model in details in the biased sampling problem in Chap. 11. In fact, S_A is precisely the score estimating function for β after profiling out the baseline density f . We will use an elementary method in Sect. 11.2 to show the optimality of S_A .

6.4 Information Calculation in Over-identified Semiparametric Models

We now establish the asymptotic optimality for the projection method in an over-identified parameter problem given by $E[g(X, \theta)] = 0$, where g is a r vector function and θ is a $p \leq r$ unknown parameter. This result was given in Qin and Lawless (1994).

Without loss of generality we write

$$g = (g_A, g_B)^T, \quad g_A = (g_1(x, \theta), \dots, g_p(x, \theta))^T, \quad g_B = (g_{p+1}(x, \theta), \dots, g_r(x, \theta))^T,$$

where the inverse of matrix $E[\partial g_A / \partial \theta]$ is assumed to be finite. Using Taylor's expansion, we can show that the estimator derived from $\sum_{i=1}^n g_A(x_i, \theta) = 0$ is asymptotically equivalent to having the influence function

$$IF = -[E\{\partial g_A / \partial \theta\}]^{-1} g_A.$$

Now we need to find the nuisance parameter space. Consider a parametric submodel $F(x) = F(x, \eta)$. From $\int g_A(x, \theta(F_\eta)) dF_\eta(x) = 0$, we have

$$\int \frac{\partial g_A(x, \theta(F_\eta))}{\partial \theta} \frac{\partial \theta(F_\eta)}{\partial \eta} dF_\eta(x) + \int g_A(x, \theta(F_\eta)) \frac{\partial \log f_\eta}{\partial \eta} dF_\eta(x) = 0,$$

or

$$\frac{\partial \theta(F_\eta)}{\partial \eta} = -[E(\partial g_A / \partial \theta)]^{-1} E[g_A(x, \theta) h], \quad h = \frac{\partial \log f_\eta}{\partial \eta}.$$

Similarly from $\int g_B(x, \theta(F_\eta)) dF_\eta = 0$ we have

$$\frac{\partial \theta(F_\eta)}{\partial \eta} E[\partial g_B / \partial \eta] + E[g_B h] = 0.$$

Substituting the expression of $\partial\theta(F_\eta)/\partial\eta$ in above equation, we have

$$E[\{g_B - E(\partial g_B/\partial\theta)(E(\partial_A/\partial\theta))^{-1}g_A\}h] = 0.$$

Therefore the nuisance parameter space \mathcal{S}_η is spanned by all functions $h(x)$ with $E[h(x)] = 0$ and $E[\beta(x)h(x)] = 0$, where

$$\beta(x) = g_B - E(\partial g_B/\partial\theta)(E(\partial g_A/\partial\theta))^{-1}g_A.$$

The influence function can be decomposed as

$$IF = S_{22.1}^{-1}S_{21}S_{11}^{-1}g - [E(\partial g_A/\partial\theta)^{-1}g_A + S_{22.1}^{-1}S_{21}S_{11}^{-1}g],$$

where

$$S_{22.1} = E(\partial g/\partial\theta)^T(Egg^T)^{-1}E(\partial g/\partial\theta), \quad S_{12} = S_{21}^T = E(\partial g/\partial\theta), \quad S_{11} = -E(gg^T).$$

We need to show

- (1) $\alpha := \{E(\partial g_A/\partial\theta)\}^{-1}g_A + S_{22.1}^{-1}S_{21}S_{11}^{-1}g \in \mathcal{S}_\eta$.
- (2) $S_{22.1}^{-1}S_{21}S_{11}^{-1}g$ is orthogonal to \mathcal{S}_η .

To show (1), we observe

$$\begin{aligned} & E[\{g_B - E(\partial g_B/\partial\theta)(E\partial g_A/\partial\theta)^{-1}g_A\}\{g^TS_{11}^{-1}S_{12}S_{22.1}^{-1}\}] \\ &= E[\{(0, I)g - (0, I)E(\partial g/\partial\theta)(E(\partial g_A/\partial\theta)^{-1}(I, 0)g\}\{g^TS_{11}^{-1}S_{12}S_{22.1}^{-1}\}] \\ &= (0, I)[Egg^T - E(\partial g/\partial\theta)(E\partial g_A/\partial\theta)^{-1}(I, 0)E(gg^T)][S_{11}^{-1}S_{12}S_{22.1}^{-1}] \\ &= (0, I)[-S_{11} + S_{12}(E\partial g_A/\partial\theta)^{-1}(I, 0)S_{11}][S_{11}^{-1}S_{12}S_{22.1}^{-1}] \\ &= -(0, I)[S_{12} + S_{12}(E\partial g_A/\partial\theta)^{-1}(I, 0)S_{12}]S_{22.1}^{-1} \\ &= -(0, I)S_{12}(E(\partial g_A/\partial\theta)^{-1}[E(\partial g/\partial\theta) - (I, 0)S_{12}]S_{22.1}^{-1}) = 0. \end{aligned}$$

To show (2), we have

$$\begin{aligned} \alpha &= S_{22.1}^{-1}[S_{22.1}(E\partial g_A/\partial\theta)^{-1}g_A + S_{21}S_{11}^{-1}g] \\ &= S_{22.1}^{-1}S_{21}S_{11}^{-1}[-E(\partial g/\partial\theta)(E\partial g_A/\partial\theta)^{-1}g_A + g] \\ &= S_{22.1}^{-1}S_{21}S_{11}^{-1}\left\{\begin{pmatrix} -E(\partial g_A/\partial\theta) \\ -E(\partial g_B/\partial\theta) \end{pmatrix}(E\partial g_A/\partial\theta)^{-1}g_A + g\right\} \\ &= S_{22.1}^{-1}S_{21}S_{11}^{-1}\left\{\begin{pmatrix} -E(\partial g_A/\partial\theta) \\ -E(\partial g_B/\partial\theta)(E\partial g_A/\partial\theta)^{-1}g_A \end{pmatrix} + \begin{pmatrix} g_A \\ g_B \end{pmatrix}\right\} \\ &= S_{22.1}^{-1}S_{21}S_{11}^{-1}\begin{pmatrix} 0 \\ g_B - E(\partial g_B/\partial\theta)(E\partial g_A/\partial\theta)^{-1}g_A \end{pmatrix} \in \mathcal{S}_\eta. \end{aligned}$$

Since the two terms in the decomposition of IF are orthogonal to each other, therefore,

$$\begin{aligned}\text{Var}(IF) &= \text{Var}[S_{22.1}^{-1} S_{21} S_{11}^{-1} g] + \text{Var}[\alpha] \geq \text{Var}[S_{22.1}^{-1} S_{21} S_{11}^{-1} g] \\ &= [E(\partial g / \partial \theta)^T (Egg^T)^{-1} E(\partial g / \partial \theta)]^{-1}.\end{aligned}$$

Thus we have finished the proof.

In Chaps. 7 and 8 we will show that Godambe's optimal estimating method, the method of moments (Hansen 1982) and the semiparametric empirical likelihood method (Qin and Lawless 1994) all achieve this information lower bound.

6.5 Information Calculation for Missing Data Problems

To further illustrate the information lower bound calculation, we consider the mean response estimation problem in the presence of missing at random data. This result is due to Hahn (1998). We will revisit this problem in Chap. 19 on missing data and causal inference problems. Some related results can be found in Robins et al. (1994).

Let $(Y_1, X_1, D_1), \dots, (Y_n, X_n, D_n)$ be the observed responses, covariates, and missing data indicators, respectively, where Y_i is missing if $D_i = 0$. Under the assumption of missing at random, the propensity score satisfies

$$P(D_i = 1|x_i, y_i) = P(D_i = 1|x_i) = \pi(x_i).$$

For simplicity we assume $\pi(x)$ is a known function. We are interested in estimating $\mu = E(Y)$. Consider a submodel

$$Y|X \sim f(y|x, \theta), \quad X \sim g(x, \theta).$$

The log-likelihood based on a generic observation and the regular parametric submodel is

$$\ell = d[\log f(y|x, \theta) + \log \pi(x)] + (1-d)\log\{1 - \pi(x)\} + \log g(x, \theta).$$

The derivative is

$$\frac{\partial \ell}{\partial \theta} = d \frac{\partial \log f(y|x, \theta)}{\partial \theta} + \frac{\partial \log g(x, \theta)}{\partial \theta}.$$

As a consequence, the tangent space for θ is

$$\mathcal{T} = \{T : T = DS(Y|X) + \alpha(X)\},$$

where

$$S(Y|X) = \frac{\partial \log f(Y|X, \theta)}{\partial \theta}, \quad \alpha(X) = \frac{\partial \log g(X, \theta)}{\partial \theta},$$

and they satisfy $E[S(Y|X)|x] = 0$ and $E[\alpha(X)] = 0$. Let

$$B = \frac{D}{\pi(X)}\{Y - \mu(X)\} + \mu(X) - \mu, \quad \mu(x) = E(Y|X = x).$$

Then $B \in \mathcal{T}$. The popular Horvitz and Thompson (1952) inverse probability weighted estimator of μ is $DY/\pi(X)$, which can be decomposed as

$$\frac{DY}{\pi(X)} = B + \frac{DY}{\pi(X)} - B = B + \frac{D - \pi(X)}{\pi(X)}\mu(X).$$

Note that

$$E\left\{\frac{D - \pi(X)}{\pi(X)}\mu(X)\right\} = 0$$

no matter what the underlying models $f(y|x)$ and $g(x)$ are. In other words, $\mu(X)\{D - \pi(X)\}/\pi(X)$ can be treated as an E -ancillary estimating function when we estimate μ . A good estimating function should be orthogonal to it. Moreover we can easily show that

$$E\left[B\left\{\frac{D - \pi(X)}{\pi(X)}\mu(X)\right\}\right] = 0.$$

Therefore we should use B as the estimating function for μ since $E(B) = 0$. From the fact

$$\begin{aligned}\text{Var}\left\{\frac{DY}{\pi(X)}\right\} &= \text{Var}(B) + \text{Var}\left\{\frac{D - \pi(X)}{\pi(X)}\mu(X)\right\} \geq \text{Var}(B) \\ &= E\left[\frac{\{Y - \mu(X)\}^2}{\pi(X)}\right] + E\{\mu(X) - \mu\}^2,\end{aligned}$$

we can claim that B is superior to the Horvitz and Thompson (1952) estimator.

Furthermore from

$$\mu = \int \int yf(t|x, \theta)g(x, \theta)dydx,$$

we have

$$\begin{aligned}\frac{\partial \mu}{\partial \theta} &= \int \int y \frac{\partial \log f(y|x, \theta)}{\partial \theta} f(y|x, \theta)g(x, \theta)dydx \\ &\quad + \int \int yf(y|x, \theta) \frac{\partial \log g(x, \theta)}{\partial \theta} g(x, \theta)dydx \\ &= E[YS(Y|X)] + E[Y\alpha(X)] \\ &= E[YS(Y|X) + \mu(X)\alpha(X)]\end{aligned}$$

Noting

$$T = DS(Y|X) + \alpha(X), \quad B = \frac{D}{\pi(X)}\{Y - \mu(X)\} + \mu(X) - \mu,$$

and the fact that $E\{S(Y|X)|X\} = 0$, $E\{Y - \mu(X)|X\} = 0$ and Y and D are independent of each other conditional on X , we can easily show that

$$\frac{\partial \mu}{\partial \theta} = E(BT).$$

Let $\hat{\theta}$ be the MLE based on this parametric submodel. Using Taylor's expansion, we have

$$\mu(\hat{\theta}) = \mu(\theta) + \frac{\partial \mu}{\partial \theta}(\hat{\theta} - \theta) + \dots$$

Since the maximum likelihood estimate satisfies

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}}[\partial^2 \ell / \partial \theta^2]^{-1} \partial \ell / \partial \theta + o_p(1) \rightarrow N(0, \{E(T^2)\}^{-1}),$$

easily we can show

$$\sqrt{n}\{\mu(\hat{\theta}) - \mu(\theta)\} = \frac{\partial \mu}{\partial \theta} \sqrt{n}(\hat{\theta} - \theta) + \dots \rightarrow N(0, \sigma^2),$$

where

$$\sigma^2 = E(\partial \mu / \partial \theta)[E(T^2)]^{-1}E(\partial \mu / \partial \theta) = E(BT)[E(T^2)]^{-1}E(BT).$$

Since this is true for all regular submodels, and $B \in \mathcal{T}$, specially if we let $T = B$, then

$$\sigma^2 \geq E[B^2].$$

In fact the variance is the expected square of the projection of B onto the tangent space \mathcal{T} . Since $B \in \mathcal{T}$, therefore the projection is itself.

Note that B involves the conditional mean of Y , i.e., $E(Y|x) = \mu(x)$, which is unknown in practical applications. B can be rewritten as

$$B = \frac{DY}{\pi(X)} - \frac{D - \pi(X)}{\pi(X)}\mu(x) - \mu.$$

Fortunately, we can still use B as the estimating function for the marginal mean of Y to get a consistent estimator even if we replace $\mu(x)$ by any measurable function of x since $E(B) = 0$ if $\pi(X)$ is correctly specified. We will further discuss this problem in Chap. 19.

6.6 A Non-root n Consistent Estimator Example

Sometimes, the available data do not contain enough information for a root- n consistent estimate of the underlying parameter. In that case the efficient score contains zero information.

Consider the binary choice model given by Cosslett (1983)

$$P(Y = 1|x) = F(x\beta),$$

where F is an unknown distribution function. The log likelihood for a generic data set is

$$\ell = Y \log F(X\beta) + (1 - Y) \log\{1 - F(X\beta)\}.$$

For a submodel $F(X\beta, \eta)$, the scores for β and η are, respectively,

$$S_\beta = \frac{\partial \ell}{\partial \beta} = X\{Y - F(X\beta)\}\sigma^{-2}(X\beta)F_\epsilon(X\beta), \quad \sigma^2(X\beta) = F(X\beta)\{1 - F(X\beta)\},$$

and

$$S_\eta = \frac{\partial \ell}{\partial \eta} = \{Y - F(X\beta)\}\sigma^{-2}(X\beta)F_\eta(X\beta),$$

where

$$F_\epsilon = \partial F(\epsilon, \eta)/\partial \epsilon, \quad F_\eta = \partial F(\epsilon, \eta)/\partial \eta.$$

Without restrictions on F_η , the nuisance space for η is generated by

$$\mathcal{J} = [S_\eta : S_\eta = \{Y - F(X\beta)\}D(X\beta)],$$

where $D(x\beta)$ is any measurable function.

Suppose we can find $J \in \mathcal{J}$ such that

$$m(Y, X) = S_\beta - J$$

is orthogonal to \mathcal{J} , and m is non-singular, i.e., $E(mm^T) \neq 0$. From $E[mS_\eta] = 0$, we have

$$\begin{aligned} 0 &= E[mS_\eta] = E[E[mS_\eta|x\beta]] \\ &= E[E[m(Y, X)\{Y - F(x\beta)\}|x\beta]D(x\beta)]. \end{aligned}$$

Since this is true for any measurable function $D(x\beta)$, it implies $E[m(Y, X)\{Y - F(x\beta)\}|x\beta] = 0$. Moreover from

$$\begin{aligned}
0 &= E[m(Y, X)\{Y - F(x\beta)\}|x\beta] = E[m(Y, X)E\{Y - F(x\beta)\}|X|x\beta] \\
&= E[m(1, X)\{1 - F(x\beta)\}F(x\beta) + m(0, X)\{-F(x\beta)\}\{1 - F(x\beta)\}|x\beta] \\
&= E[m(1, X) - m(0, X)|x\beta]\{1 - F(x\beta)\}F(x\beta).
\end{aligned}$$

We can always choose $F(x\beta)$ such that $0 < F(x\beta) < 1$, this implies

$$E[m(1, X) - m(0, X)|x\beta] = 0.$$

As a consequence we have $m(1, X) = m(0, X)$ almost surely. In other words $m(Y, X) = m(X)$, which is independent of Y .

Finally

$$\begin{aligned}
E[mm^T] &= E[m\{S_\beta - J\}^T] \\
&= E[m(X)S_\beta(Y, X)] = E[m(X)E\{S_\beta(Y, X)|X\}] \\
&= 0.
\end{aligned}$$

This contradicts the non-singular assumption on $m(Y, X)$.

Chamberlain (1987) showed that the information for β in this model is zero. It implies that there does not exist a root- n consistent estimator.

So far, we have discussed different projection methods for many commonly used models. In the survival analysis Chap. 24, we will use this method to find the efficient score for the Cox regression model. In many situations, unfortunately, the optimal projection estimating function involves the unknown baseline density, which may not be easily implemented. For more systematic approaches and examples we refer readers to the books by Bickel et al. (1993) and Tsiatis (2006). More examples on the information bound calculation for econometric models can be found in Newey (1990).

Chapter 7

Generalized Method of Moments

Generalized method of moments (Hansen 1982) is one of the most popular methods in econometric literature. Due to this ground-break work, Hansen was awarded Nobel prize in 2013. The generalized method of moments (henceforth GMM) has become an important unifying framework for inference in econometrics during the last thirty years. It is well known that maximum likelihood estimate (MLE) has the smallest variance in the class of consistent and asymptotic normally distributed estimators. However the MLE approach needs a full description and correct specification of the underlying distribution. Unlike MLE, GMM does not require complete knowledge of the distribution of observed data. Rather, only specified moments derived from an underlying model are needed for GMM estimation. In some cases, even if the distribution of observed data is known, the MLE can be computationally very burdensome, whereas GMM is easy. Among others, the log-normal stochastic volatility model used widely in economic finance is one of those examples. In over-identified models, in which there are more moment conditions than model parameters, the GMM estimation provides a straightforward way to combine estimating equations and to test the specification of the proposed models. This important feature is unique to GMM estimation.

7.1 Basic Concepts on Generalized Method of Moments

First we give a few examples of over-identified parameter problems.

Example 1 Suppose X_1, \dots, X_n are i.i.d. observations from a Poisson distribution with rate λ . The first and second moments satisfy

$$E(X) = \text{Var}(X) = \lambda.$$

Obviously we have two estimating equations for λ ,

$$E(X - \lambda) = 0, \quad E(X^2) - \lambda - \lambda^2 = 0.$$

On the other hand, if we remove the Poisson assumption but keep the two unbiased estimating equations, how do we estimate λ ? We need to find a way to combine them effectively.

Example 2 Consider an instrumental variable problem in econometrics.

In a linear regression model

$$Y_i = X_i\beta + \epsilon_i,$$

the covariate X and error term ϵ are assumed to be independent each other. In that case, the least squares method leads to a linear estimating equation

$$\sum_{i=1}^n x_i(y_i - x_i\beta) = 0,$$

which has mean 0. On the other hand, if X and ϵ are correlated each other, this estimating equation in general is biased. In econometric literature, a variable is called an “instrumental variable” if it is correlated with the covariate X but is uncorrelated to the error variable ϵ . Let $Z = Z_{m \times 1}$ be such an instrumental variable. Then the estimating function $\sum_{i=1}^n a(z_i)(y_i - x_i\beta) = 0$ is unbiased, where $a(z)$ is a given function of z . Using Godambe’s optimal estimating function theory, we can show that the optimal choice of $a(z)$ is

$$a(z) = E(X|z)E[(Y - X\beta)^2|z].$$

In general, the optimal form of $a(z)$ is unknown. To identify the underlying parameter β and to increase efficiency we can choose some function of z , say, $a_{m \times p}(z)$, where p is the dimension of β and $m \geq p$. Note that the instrumental variable approach is often associated with low precision. A large m is desirable.

A more specific example is discussed by Angrist and Krueger (2001). They were interested in investigating the number of years spent in education and the subsequent earning potentials of individuals (w). They used the model

$$\log(w_i) = \beta_0 + \text{edu}_i\beta_1 + \text{controls}_i + \epsilon_i$$

to examine the impact of compulsory schooling laws in the US, where controls are other variables that may affect wage. The parameter of interest was β_1 , the semi-elasticity of wage with respect to education. Estimating the parameters by ordinary least squares may be biased as edu_i is probably correlated with unobservable factors represented by the regression error term ϵ_i (such as individual costs and potential benefits of schooling or other options outside the schooling system). Using

the structure of compulsory school attendance laws at that time in the US they argued that (in addition to the controls) dummy variables indicating the quarter of birth for each individual could be used as instrument for the years spent in education.

Example 3 Consider a random variable X with a symmetric stable distribution with index α , location parameter θ and scale parameter c . The characteristic function of X is given by

$$\phi(t; \theta, c) = \exp(i\theta t - |ct|^\alpha).$$

The special cases $\alpha = 1$ and $\alpha = 2$ correspond to the Cauchy and normal distributions respectively. However, with the exception of these two cases, the symmetric stable law probability density functions are not expressible in closed forms suitable for likelihood methods. For simplicity, we assume α is known and $\alpha \neq 1, 2$. Let

$$h(X, t_1, t_2) = \left[\sin \left\{ t_1 \frac{(X - \theta)}{c} \right\}, \cos \left\{ t_2 \frac{(X - \theta)}{c} \right\} \right].$$

We can easily show that

$$E[h(X, t_1, t_2)] = (0, \exp(-|t_2|^\alpha)).$$

For different choices of t_1, t_2 , we have more estimating equations than the unknown parameters. Naturally, GMM can be employed to combine them. McLeish and Small (1992) used this example to project the likelihood function into the space spanned by h . More examples on over-identified models will be discussed in next chapter.

Three Versions of GMM

For a given vector estimating function $g(X, \theta)$, where g is a $r \times 1$ vector functions and θ is $p \times 1$ unknown parameters, it is clear if $r < p$, then g cannot determine θ uniquely. On the other hand, if $r > p$, then this is called an over-identified problem since only p estimating equations (out of r) are needed to identify θ . One may use any p out of r estimating functions for estimating θ . If that is the case, however, there may be a loss of information since not all components of $g(X, \theta)$ are used. A challenging problem is figuring out how to use all r estimating functions effectively for a better estimator. The GMM method proposed by Hansen (1982) specifically targets this problem.

There are three versions (Hansen et al. 1996) for the generalized method of moments.

(1) Two-step estimator.

(a) Find an initial estimator of $\hat{\theta}^0$. This can be done by choosing any p estimating functions from the r estimating functions $g(x, \theta)$.

(b) Minimize the quadratic form

$$\left[\sum_{i=1}^n g^T(x_i, \theta) \right] [E\{g(X, \hat{\theta}^0)g^T(X, \hat{\theta}^0)\}]^{-1} \left[\sum_{i=1}^n g(x_i, \theta) \right]$$

with respect to θ . Note that the covariate matrix can be replaced by the sample version $n^{-1} \sum_{i=1}^n g(x_i, \hat{\theta}^0)g^T(x_i, \hat{\theta}^0)$ if it is unknown.

(2) Iterative estimator.

Let $\hat{\theta}^1$ be the estimator obtained from the two-step method. Replace $\hat{\theta}^0$ by $\hat{\theta}^1$ in the covariate matrix and obtain a second GMM estimator, denoted as $\hat{\theta}^2$. Repeat this process until convergence.

(3) Direct estimator.

Minimize

$$\left[\sum_{i=1}^n g^T(x_i, \theta) \right] [E\{g(X, \theta)g^T(X, \theta)\}]^{-1} \left[\sum_{i=1}^n g(x_i, \theta) \right]$$

with respect to θ . In this case, θ appears in both estimating equations and covariate matrix. Again the covariate matrix can be replaced by its sample version if it is unknown.

Even though asymptotic results are equivalent for above three versions, small sample results may show different behaviours. In general, version 1 is more stable. An R program on calculating the generalized method of moments is available through <https://cran.r-project.org/web/packages/gmm/gmm.pdf>.

We will show that the GMM, Godambe's optimal estimating equations, the projection of likelihood ratio, and the empirical likelihood method for estimating equations developed by Qin and Lawless (1994) are all asymptotically first order equivalent. Moreover Chamberlain (1987) and Qin and Lawless (1994) showed that this class of estimators achieves the semiparametric efficient lower bound for a given set of moment restrictions.

Clearly minimizing the quadratic function defined in GMM with respect to θ is asymptotically equivalent to solving estimating equation

$$E[\partial g / \partial \theta] V^{-1}(g) \sum_{i=1}^n g(x_i, \theta) = 0.$$

Let $A(\theta)$ be a $p \times r$ vector. Define a class of estimating equations

$$\Psi = \{\psi(x, \theta) | \psi(x, \theta) = A(\theta)g(x, \theta)\}. \quad (7.1.1)$$

Then the GMM method is equivalent to choosing

$$A_0(\theta) = E[\partial g / \partial \theta] V^{-1}(g). \quad (7.1.2)$$

We will show indeed A_0 is the optimal choice in the class Ψ .

In fact under some regularity conditions, the solution $\hat{\theta}$ of the estimating equation $\sum_{i=1}^n A g(x_i, \theta) = 0$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) = [AE(\partial g/\partial\theta)]^{-1}A\frac{1}{\sqrt{n}}\sum_{i=1}^n g(x_i, \theta) + o_p(1).$$

Define $B_0 = [A_0 E(\partial g/\partial\theta)]^{-1} A_0 g$ and $B = [AE(\partial g/\partial\theta)]^{-1} Ag$. To demonstrate the optimality of using A_0 in (7.1.2), we need to show $V(B_0) - V(B)$ is a non-positive definite matrix for all A . Since

$$\text{Cov}(B_0, B) = \text{Var}(B_0) = [E(\partial g/\partial\theta^T)\text{Var}^{-1}(g)E(\partial g/\partial\theta)]^{-1},$$

therefore,

$$0 \leq \text{Var}(B_0 - B) = \text{Var}(B_0) + \text{Var}(B) - 2\text{Cov}(B_0, B) = \text{Var}(B) - \text{Var}(B_0).$$

We have shown that the GMM is equivalent to the optimal estimating equation in the linear estimating function space Ψ . Recall in Chap. 6 we also found that $A_0 g$ is the projection of score function in the linear estimating function space spanned by g .

7.2 An Optimal Result Based on an Embed Exponential Family

Hansen's (1982) GMM work is very influential in econometrics and statistics. It has received ten of thousands of citations from google scholar. Back and Brown (1992) argued that even if one can construct an exponential family which satisfies the moment constraints, the asymptotic variance derived from it is the same as the one obtained from GMM. Due to its theoretical importance, we will discuss their approach in details below.

Suppose $X \in R^m$. Denote the parameter space as $\Theta \in R^p$. Let g be a map from $R^m \times \Theta$ to R^r ($r \geq p$). Moreover let \mathcal{P} be the class of probability distributions P on R^m for which there is some $\theta \in \Theta$ satisfying

$$\int g(x, \theta) P(dx) = 0. \quad (7.2.3)$$

Denote P_0 as a particular element of \mathcal{P} . Suppose the solution θ to above equation is unique when $P = P_0$. The i.i.d. observations of X are

$$X_1, \dots, X_n \sim P_0.$$

We will show that the optimal GMM estimator is asymptotically equivalent under P_0 to a certain maximum likelihood estimator.

Define an exponential family of distribution \mathcal{P}' through the Radon–Nikodym derivatives

$$\frac{dP_\theta}{dP_0} = L(X, \theta) = \exp\{Z^T(\theta)g(X, \theta_0) - a(\theta)\}, \quad (7.2.4)$$

where the functions a and Z are determined implicitly by the following conditions:

$$Z(\theta_0) = 0, \quad a(\theta_0) = 0 \quad (7.2.5)$$

$$L(x, \theta_0) = 1, \quad P_{\theta_0} = P_0, \quad \int L(x, \theta)P_0(dx) = \int P_\theta(dx) = 1, \quad (7.2.6)$$

$$\int g(x, \theta)L(x, \theta)P_0(dx) = \int g(x, \theta)P_\theta(dx) = 0. \quad (7.2.7)$$

Note that (7.2.5) implies $P_0 \in \mathcal{P}'$, (7.2.6) implies P_θ is a probability measure and (7.2.7) implies $P_\theta \in \mathcal{P}'$.

We cannot calculate the MLE for the family \mathcal{P}' because it is defined in terms of θ_0 and P_0 , both of which are unknown. However, we can show that even if we could calculate it it would not be any better asymptotically than the GMM.

Define

$$A = \int \partial g(x, \theta_0)/\partial \theta P_0(dx), \quad B = \int g(x, \theta_0)g^T(x, \theta_0)P_0(dx). \quad (7.2.8)$$

Assume A has full column rank and B is positive definite.

Theorem 7.1 *There is a neighbourhood U of θ_0 on which (7.2.5)–(7.2.7) uniquely define $a(\theta)$ and $Z(\theta)$ in a neighbourhood of $a = 0$ and $Z = 0$. The function a and Z are continuously differentiable on U and*

$$\frac{\partial Z(\theta_0)}{\partial \theta} = -B^{-1}A, \quad (7.2.9)$$

$$\left(\frac{\partial a(\theta)}{\partial \theta}\right)^T = \left(\frac{\partial Z(\theta)}{\partial \theta}\right)^T \int g(x, \theta_0)P_\theta(dx). \quad (7.2.10)$$

Proof Differentiating $\int L(x, \theta)P_0(dx) = 1$ with respect to θ , we have

$$\int \left[\left(\frac{\partial Z(\theta)}{\partial \theta}\right)^T g(x, \theta_0) - \left(\frac{\partial a(\theta)}{\partial \theta}\right)^T \right] L(x, \theta)P_0(dx) = 0,$$

or

$$\left(\frac{\partial Z(\theta)}{\partial \theta}\right)^T \int g(x, \theta_0)L(x, \theta)P_0(dx) = \left(\frac{\partial a(\theta)}{\partial \theta}\right)^T.$$

This implies (7.2.9). Differentiating $\int g(x, \theta) L(x, \theta) P_0(dx) = 0$ with respect to θ , we have

$$\begin{aligned} & \int \frac{\partial g(x, \theta)}{\partial \theta} L(x, \theta) P_0(dx) \\ & + \int g(x, \theta) \left[\left(\frac{\partial Z(\theta)}{\partial \theta} \right)^T g(x, \theta_0) - \left(\frac{\partial a(\theta)}{\partial \theta} \right)^T \right] L(x, \theta) P_0(dx) = 0. \end{aligned}$$

Noting that $\int g(x, \theta_0) P_0(dx) = 0$ and $L(\theta_0) = 1$, we have

$$\begin{aligned} & \int \frac{\partial g(x, \theta_0)}{\partial \theta} P_0(dx) \\ & + \left(\frac{\partial Z(\theta_0)}{\partial \theta} \right)^T \int g(x, \theta_0) g^T(x, \theta_0) - \left(\frac{\partial a(\theta)}{\partial \theta} \right)^T \int g(x, \theta_0) P_0(dx) = 0. \end{aligned}$$

This implies (7.2.10).

Based on the observed data X_1, \dots, X_n from P_θ , the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^n [Z^T(\theta) g(x_i, \theta_0) - a(\theta)].$$

The corresponding score is

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \sum_{i=1}^n \left[\frac{\partial Z^T(\theta)}{\partial \theta} g(x_i, \theta_0) - \frac{\partial a(\theta)}{\partial \theta} \right] \\ &= \frac{\partial Z^T(\theta)}{\partial \theta} \sum_{i=1}^n \left[g(x_i, \theta_0) - \int g(x, \theta_0) P_\theta(dx) \right] \\ &=: \frac{\partial Z^T(\theta)}{\partial \theta} \sum_{i=1}^n \tilde{g}(x_i, \theta). \end{aligned}$$

Denote the MLE as $\hat{\theta}$. Clearly $\hat{\theta} \rightarrow \theta_0$ in probability. By Taylor's expansion,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(n^{-1} \frac{\partial \ell(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell(\theta_0)}{\partial \theta} + o_p(1).$$

It follows easily that, in distribution,

$$\frac{1}{\sqrt{n}} \frac{\partial \ell(\theta_0)}{\partial \theta} \rightarrow N(0, A^T B^{-1} A),$$

and in probability

$$n^{-1} \frac{\partial \ell(\theta_0)}{\partial \theta \partial \theta^T} \rightarrow -A^T B^{-1} A.$$

As a result, in distribution

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, (A^T B^{-1} A)^{-1}),$$

where A and B are defined in (7.2.8).

Remark In fact Back and Brown's (1992) approach is the entropy family method subject to the moment constraints. A detailed discussion will be given in Chap. 9.

7.3 Applications of GMM

In this section we give a few examples of GMM.

Example 1 For $i = 1, 2$, suppose we can specify two marginal densities

$$f(y_{ij}|x_{ij}) = f(y_{ij}|x_{ij}, \beta), i = 1, 2, j = 1, 2, \dots, n,$$

where the joint density of (Y_1, Y_2) given (x_1, x_2) is not specified. The marginal scores are

$$S_{ij}(y_{ij}, x_{ij}) = \frac{\partial \log f(y_{ij}|x_{ij}, \beta)}{\partial \beta}, \quad i = 1, 2.$$

Let

$$S_j = (S_{1j}, S_{2j}), j = 1, 2, \dots, n.$$

If we can specify the joint covariate matrix $\Sigma = \text{Var}(S)$, then the GMM estimate can be obtained by minimizing

$$S(\beta) \Sigma^{-1} S^T(\beta)$$

with respect to β . It is well known even if Σ^{-1} is misspecified, the GMM still produces consistent estimate. We may also replace Σ by its sample version.

Exercise On the other hand, we can construct more estimating equations, for example

$$S_{ij}^* = \frac{\partial^2 f(y_{ij}|x_{ij}, \beta)/\partial \beta^2}{f(y_{ij}|x_{ij}, \beta)}.$$

Do we get extra information by using GMM method to combine $S_j = (S_{1j}, S_{2j}, S_{1j}^*, S_{2j}^*)$?

Example 2 Generalized linear model for longitudinal data

In medical research and economic research longitudinal and panel data are frequently available. Consider a longitudinal study, where y_{ij} is the response measured at the j -th time point for the i -th individual, and x_{ij} is the corresponding covariate. Suppose

$$E(Y_{ij}|x_{ij}) = \mu(x_{ij}^T \beta), \quad \text{Var}(Y_{ij}|x_{ij}) = v_i(\mu_{ij}),$$

where

$$\mathbf{y}_i = (y_{i1}, \dots, y_{iT}^T).$$

Let $A_i = \text{diag}(v_i(\mu_{i1}), \dots, v_i(\mu_{ik}))$ be a diagonal matrix. Liang and Zeger (1986) proposed to use

$$U(\beta, R) = \sum_{i=1}^n U_i(\beta, R) = \sum_{i=1}^n D_i^T A_i^{-1/2} R^{-1}(\alpha) A_i^{-1/2} (\mathbf{y}_i - \mu_i) = 0$$

as estimating equation, where R is a given correlation coefficient matrix. A nice feature of this method is that R does not need to be correct for estimating β , even though a correct specification may increase the estimation efficiency. It is an interesting topic to find efficient estimate for the unknown parameter α in $R(\alpha)$.

Suppose R_j , $j = 1, 2, \dots, J$ are possible candidates for the working correlation matrices. Qu et al. (2000) used GMM method to combine estimating functions

$$(S^1(\beta), \dots, S^J(\beta)),$$

where

$$S_i^j(\beta) = D_i^T A_i^{-1/2} R_j^{-1} A_j^{-1/2} (\mathbf{y}_i - \mu_i).$$

Another interesting alternative approach for estimating β is studied by Xu et al. (2012). Define

$$\epsilon_i = A_i^{-1/2} (\mathbf{y}_i - \mu_i) | L = l \sim N(0, R_l), \quad l = 1, 2, \dots, J.$$

They treated ϵ_i as having a normal mixture distribution with mean 0 but possibly different covariate matrices. The log mixture likelihood

$$\ell = \sum_{i=1}^n \log \left\{ \sum_{l=1}^J \pi_l \phi_l(\epsilon_i, R_l) \right\}.$$

can be used to estimate the underlying parameters, where the mixing proportions π_l , $l = 1, 2, \dots, L$ can be treated as either known or unknown.

Example 3 Optimizing GMM for sparse data

Consider an estimating function $g_{r \times 1}(X, \beta)$ again. When the observed data are sparse, i.e., r is large and the sample size n is small, the inverse of a high dimensional covariate matrix based on its sample version may not be numerically stable. We may postulate a “working” inverse $r \times r$ covariate matrix $\Sigma^{-1}(\alpha)$ indexed by α , For fixed α , then we can apply the GMM method to minimize

$$\left[\sum_{i=1}^n g^T(x_i, \beta) \right] \Sigma^{-1}(\alpha) \left[\sum_{i=1}^n g(x_i, \beta) \right],$$

with respect to β , where $x_i, i = 1, 2, \dots, n$ are the observed data. For each fixed α , denote the GMM estimate of β , as $\hat{\beta}(\alpha)$. Then we can minimize the trace of the asymptotic variance of $\hat{\beta}(\alpha)$ with respect to α to find the optimal α^* .

As mentioned at the beginning of this chapter, the generalized method of moments has become almost indispensable in econometric applications. Interested readers can read the book by Hall (2005) on the applications GMM in financial and economic models. This method has also been used in survival analysis setup to incorporate auxiliary information, for example, Li and Yin (2009) and Huang and Qin (2013).

Chapter 8

Empirical Likelihood with Applications

The maximum likelihood method for regular parametric models has many optimality properties. As a result, it is one of the most popular methods in statistical inference. However, model mis-specification is a big concern since a misspecified model may lead to bias results. When the underlying distribution is multinomial, Hartley and Rao (1968) proposed a mean constrained estimator for the population total in survey sampling problems. To mimic the parametric likelihood but with robust properties, Owen (1988, 1990) proposed the empirical likelihood method, which is a natural generalization of the multinomial likelihood when the number of categories is the same as the sample size. To compute the profile likelihood of this general multinomial distribution which has atoms at data points, we can find the empirical likelihood method for constructing confidence regions has sampling properties similar to bootstrap, but the bootstrap uses re-sampling. As Owen pointed out, a version of this technique dates back at least to Thomas and Grunkemeier (1975), in the context of likelihood based confidence intervals related to the Kaplan and Meier estimator of the survivor function. By means of linear and quadratic expansions of the log-likelihood function, Cox and Oakes (1984 pp 51–52) were able to show the connection between Greenwood's estimator of the variance of the Kaplan–Meier estimator and the confidence intervals based on the likelihood ratio statistic. Owen (1988) obtained nonparametric likelihood-based interval estimates of the mean of a distribution, and also proved that, in this case, the theorem of Wilks (1938) on the asymptotic distribution of the likelihood ratio has a nonparametric analogue. Owen (1990) discussed the empirical likelihood as an alternative to likelihood type bootstrap methods. Hall and La Scala (1990) argued that empirical likelihood is a strong competitor with contemporary methods such as the bootstrap, and deserves a prominent place in the modern statisticians armoury of computer-intensive tools. They gave four main advantages of the empirical likelihood over the bootstrap.

(1) The shape of empirical likelihood confidence regions automatically reflects the observed data set. The regions tend to be concentrated in places where the density of the parameter estimator is greatest.

(2) Empirical likelihood regions are Bartlett correctable (DiCiccio et al. 1991). That is a simple correction for the mean of the empirical likelihood ratio reduces the coverage error from order n^{-1} to order n^{-2} , where n denotes sample size. Signed root adjustment is also possible for empirical likelihood regions (DiCiccio and Romano 1989).

(3) Empirical likelihood regions do not require estimation of the scale or skewness parameter. It allows the data to “speak for themselves” and thus may be more robust against model mis-specification.

(4) Empirical likelihood regions are range preserving and transformation respecting. Empirical likelihood yields confidence regions that are very close to those of likelihood based statistics.

More comprehensive discussions on the empirical likelihood are given in Owen’s (2001) monograph. It was highly recommended by Professor Peter Hall for researchers interested in the empirical likelihood. In a book review, he wrote “....A great amount of thought and care has gone into preparing this fascinating monograph..... If we look at statistics from the vantage point of empirical likelihood, we can see a long way; Owen shows us how, and how far....” The latest monograph by Zhou (2015) provides outstanding applications of the empirical likelihood method in survival analyses.

8.1 Definition of Empirical Likelihood and Basic Properties

Suppose i.i.d. observations X_1, \dots, X_n are obtained from a distribution F . Without loss of generality we assume there are no ties. Let

$$dF(x_i) = P(X = x_i), \quad i = 1, 2, \dots, n,$$

be the jumps at the observed data points. The nonparametric likelihood is

$$L_n(F) = \prod_{i=1}^n p_i, \quad p_i = dF(x_i), \quad i = 1, 2, \dots, n.$$

Clearly $p_i > 0$, $i = 1, 2, \dots, n$, otherwise $L_n(F) = 0$.

If we maximize this log-likelihood $\ell(F) = \sum_{i=1}^n \log p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0,$$

then we end up with

$$p_i = 1/n, \quad i = 1, 2, \dots, n.$$

Therefore we can estimate the underlying distribution F by

$$F_n(x) = \sum_{i=1}^n p_i I(x_i \leq x) = n^{-1} \sum_{i=1}^n I(x_i \leq x).$$

This is the reason that the empirical distribution is called the nonparametric maximum likelihood estimate.

Suppose we are interested in constructing a confidence interval for the mean of X , $\mu = E(X) = \int x dF(x)$. Since we have discretized F at each of the observed data points, the integral becomes $\mu = \sum_{i=1}^n p_i x_i$. Next we maximize the log nonparametric likelihood subject to one extra constraint

$$\sum_{i=1}^n p_i (x_i - \mu) = 0.$$

To implement this maximization, we can apply the Lagrange multiplier method. Assume that μ is an interior point of the convex hull formed by the n observations. Define

$$h(\lambda_1, \lambda) = \sum_{i=1}^n \log p_i + \lambda_1 \left(\sum_{i=1}^n p_i - 1 \right) - n\lambda \sum_{i=1}^n p_i (x_i - \mu).$$

Taking derivative with respect to the p_i 's, we can easily show that

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(x_i - \mu)},$$

where λ satisfies the constraint equation

$$\sum_{i=1}^n \frac{x_i - \mu}{1 + \lambda(x_i - \mu)} = 0. \quad (8.1.1)$$

Since all $0 < p_i < 1$, in the univariate case, we can find that

$$\frac{1 - n^{-1}}{\mu - x_{(n)}} < \lambda < \frac{1 - n^{-1}}{\mu - x_{(1)}},$$

where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum observed values, respectively. In addition, the left hand side of the constraint equation (8.1.1) is monotone in λ . Hence the solution to (8.1.1) can be found numerically by the bisection method. When μ is a vector, the constraint equation can be shown to be the derivative of a convex

objective function. A modified Newton's method can be used to solve the equation. Once the value of λ is obtained, the profile log-likelihood is

$$\ell_n(\mu) = - \sum_{i=1}^n \log[1 + \lambda(x_i - \mu)] - n \log n. \quad (8.1.2)$$

This can be treated as a parametric likelihood with the parameter μ . Clearly if we maximize $\ell_n(\mu)$ with respect to μ , the maximum empirical likelihood estimate of μ is $\hat{\mu} = \bar{X} = n^{-1} \sum_{i=1}^n x_i$, which is exactly the sample mean. The likelihood ratio statistic is defined by

$$R(\mu) = 2[\max_{\mu} \ell_n(\mu) - \ell(\mu)] = 2[\ell_n(\bar{X}) - \ell(\mu)].$$

Theorem 8.1 *Under the regularity conditions specified in Owen (1988, 1990), in distribution,*

$$R(\mu_0) \rightarrow \chi^2(p),$$

where p is the dimension of μ .

Remark In this theorem, the p_i 's can be treated as incidental parameters and μ as a structural parameter. In contrast to the Neyman–Scott problem, the empirical likelihood based estimator for the structural parameter has nice large sample properties.

We will show that the empirical likelihood confidence region is always an interval in the univariate case.

In fact, if μ_1 and μ_2 are in the empirical likelihood confidence interval, there exist $p_{1i}, p_{2i}, i = 1, 2, \dots, n$ such that

$$\sum_{i=1}^n p_{ji} x_i = \mu_j, \quad j = 1, 2,$$

and $-2 \sum_{i=1}^n \log(np_{ji}) \leq \chi_1^2(\alpha)$. Let $0 < \tau < 1$, consider $\sum_{i=1}^n \{\tau p_{1i} + (1 - \tau)p_{2i}\}$ $x_i = \tau \mu_1 + (1 - \tau)\mu_2$. Using the fact that $-2 \log(x)$ is a convex function, we have

$$\begin{aligned} -2 \sum_{i=1}^n \log\{\tau p_{1i} + (1 - \tau)p_{2i}\} &\leq -2\tau \sum_{i=1}^n \log(np_{1i}) - 2(1 - \tau) \sum_{i=1}^n \log(np_{2i}) \\ &\leq \tau \chi_1^2(\alpha) + (1 - \tau) \chi_1^2(\alpha) = \chi_1^2(\alpha). \end{aligned}$$

Therefore $\tau \mu_1 + (1 - \tau)\mu_2$, $0 \leq \tau \leq 1$, the segment connecting any two of interior points also falls in the confidence interval.

8.2 General Theory of Empirical Likelihood in Estimating Equations

As discussed in Chap. 7, the generalized method of moments can be used to combine information for over-identified parameter problems, Qin and Lawless (1994) demonstrated that empirical likelihood can also be used naturally to solve the same problem.

Suppose we have a vector estimating equation

$$E[g(X, \theta)] = 0,$$

where g is a $r \times 1$ vector function and θ is a $p \times 1$ parameter. We consider the over-identified parameter case $r \geq p$.

For fixed θ we need to maximize the nonparametric likelihood

$$\prod_{i=1}^n p_i, \quad dF(x_i) = p_i$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i g(x_i, \theta) = 0.$$

After discretizing F , we see that the expectation constraint becomes the summation constraint shown above. Using the same approach as before, up to a constant term, the profile log empirical likelihood is

$$\ell(\theta) = - \sum_{i=1}^n \log[1 + \lambda^T g(x_i, \theta)], \quad (8.2.3)$$

where λ is the Lagrange multiplier determined by

$$\sum_{i=1}^n \frac{g(x_i, \theta)}{1 + \lambda^T g(x_i, \theta)} = 0. \quad (8.2.4)$$

We then maximize $\ell(\theta)$ with respect to θ . In the following we use $\|\cdot\|$ to denote Euclidean norm. We wish to show that for any fixed $\theta \in \|\theta - \theta_0\| \leq d_n = O_p(n^{-1/3-\delta})$, $\delta > 0$, the constraint equation has a solution $\lambda = \lambda(\theta) = O_p(d_n)$. The following lemma is useful to show the existence of roots in the constraint estimating equations.

Aitchison and Silvey (1958) Lemma

If $h(\lambda)$ is a continuous function mapping from R^r onto itself, with the property that for every λ such that $\|\lambda\| = 1$, $\lambda^T h(\lambda) < 0$, then there exists a point $\hat{\lambda}$ such that $\|\hat{\lambda}\| < 1$ and $g(\hat{\lambda}) = 0$.

Next we give a proof for the existence of roots of the Eq. (8.2.4).

For any $\theta \in \|\theta - \theta_0\| \leq d_n$, define

$$h(\lambda) = \frac{1}{n} \sum_{i=1}^n \frac{g(x_i, \theta)}{1 + n^{-1/3-\delta/2} \lambda^T g(x_i, \theta)}.$$

Note that $E\|g(x_i, \theta)\|^3 < \infty$ implies $\max_{1 \leq i \leq n} |g_k(x_i, \theta)| = o_p(n^{1/3})$, $k = 1, 2, \dots, p$ (**Exercise**). Thus $h(\lambda)$ is well defined and almost surely continuous in the unit ball $\lambda \in \{\lambda : \|\lambda\| \leq 1\}$. When $\|\lambda\| = 1$, we have

$$\begin{aligned} \lambda^T h(\lambda) &= \lambda^T \frac{1}{n} \sum_{i=1}^n \{1 - n^{-1/3-\delta/2} \lambda^T g(x_i, \theta) + O_p(n - 2/3 - \delta)\} g(x_i, \theta) \\ &= \lambda^T n^{-1} \sum_{i=1}^n g(x_i, \theta) - n^{-1/3-\delta/2} \lambda^T n^{-1} \sum_{i=1}^n g(x_i, \theta) g^T(x_i, \theta) + O_p(n^{-2/3-\delta}) \\ &\leq O_p(n^{-1/3-\delta}) - cn^{-1/3-\delta/2} + O_p(n^{-2/3-\delta}) \\ &< 0, \end{aligned}$$

where c is the smallest eigenvalue of $E\{g(x_i, \theta) g^T(x_i, \theta)\}$. By using Aitchison and Silvey (1958)'s Lemma, we conclude that there exists a $\hat{\lambda}$ such that $\|\hat{\lambda}\| < 1$ and $h(\hat{\lambda}) = 0$.

Now we are ready to show the main result.

Theorem 8.2 *If $E\{g(X, \theta_0) g^T(X, \theta_0)\}$ is positive definite and $\partial g(x, \theta)/\partial\theta$ is continuous in a neighbour of θ_0 . Moreover $\|\partial g/\partial\theta\|$ and $\|g(x, \theta)\|^3$ can be bounded by some integrable function $G(x)$ in this neighbour. Also assume the rank of $E[\partial g(X, \theta_0)/\partial\theta]$ is p . Then in distribution*

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, V),$$

where

$$V = \left[E \left(\frac{\partial g}{\partial \theta} \right)^T (E g g^T)^{-1} E \left(\frac{\partial g}{\partial \theta} \right) \right]^{-1}.$$

Proof

First we show that $\ell(\theta)$ achieves its maximum in the interior of the ball $\|\theta - \theta_0\| \leq n^{-1/3}$.

In fact for $\theta = \theta_0 + un^{-1/3}$, where $\|u\| = 1$, using Taylor's expansion, we can show that

$$\ell(\theta) \geq (c - \epsilon)n^{1/3},$$

where c is the smallest eigenvalue of

$$E(\partial g(x, \theta_0)/\partial\theta^T)[Eg(x, \theta_0)g^T(x, \theta_0)]^{-1}E(\partial g(x, \theta_0)/\partial\theta).$$

Similarly it can be shown that

$$\ell(\theta_0) = O_p(\log \log n).$$

Since $\ell(\theta)$ is a continuous and differentiable function of θ in the ball $||\theta - \theta_0|| \leq n^{-1/3}$, $\ell(\theta)$ achieves its maximum value in the interior of this ball. Then we must have

$$\frac{\partial \ell(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0.$$

As a note, if we choose a ball $||\theta - \theta_0|| \leq n^{-1/2}$ instead of the ball $||\theta - \theta_0|| \leq n^{-1/3}$, then the maximum may achieve on the surface of this ball. As a result, we are not sure whether $\partial \ell(\theta)/\partial \theta|_{\theta=\hat{\theta}} = 0$ since $\hat{\theta}$ may fall on the boundary of the parameter space.

Denote

$$Q_{1n}(\theta, \lambda) = n^{-1} \sum_{i=1}^n \frac{1}{1 + \lambda^T g(x_i, \theta)} g(x_i, \theta),$$

and

$$Q_{2n}(\theta, \lambda) = n^{-1} \sum_{i=1}^n \frac{1}{1 + \lambda^T g(x_i, \theta)} \left(\frac{\partial g(x_i, \theta)}{\partial \theta^T} \right) \lambda.$$

Note that

$$\begin{aligned} \frac{\partial Q_{1n}(\theta, 0)}{\partial \theta} &= n^{-1} \sum_{i=1}^n \frac{\partial g(x_i, \theta)}{\partial \theta}, \quad \frac{\partial Q_{1n}(\theta, 0)}{\partial \lambda} = -n^{-1} \sum_{i=1}^n g(x_i; \theta) g^T(x_i; \theta), \\ \frac{\partial Q_{2n}(\theta, 0)}{\partial \theta} &= 0, \quad \frac{\partial Q_{2n}(\theta, 0)}{\partial \lambda} = -n^{-1} \sum_{i=1}^n \frac{\partial g(x_i; \theta)}{\partial \theta^T}. \end{aligned}$$

Expanding $Q_{in}(\hat{\theta}, \hat{\lambda})$ at $(\theta_0, 0)$, we have

$$0 = Q_{in}(\hat{\theta}, \hat{\lambda}) = \frac{\partial Q_{in}(\theta_0, 0)}{\partial \theta}(\hat{\theta} - \theta_0) + \frac{\partial Q_{in}(\theta_0, 0)}{\partial \lambda^T} \hat{\lambda} + o_p(\delta_n), \quad i = 1, 2,$$

where $o_p(\delta_n) = ||\hat{\theta} - \theta_0|| + ||\hat{\lambda}||$. In matrix form we have

$$\begin{pmatrix} -Q_{1n}(\theta_0, 0) \\ 0 \end{pmatrix} + o_p(\delta_n) = \begin{pmatrix} \frac{\partial Q_{1n}}{\partial \lambda^T} & \frac{\partial Q_{1n}}{\partial \theta} \\ \frac{\partial Q_{2n}}{\partial \lambda^T} & 0 \end{pmatrix}_{(\theta_0, 0)} \begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta_0 \end{pmatrix}.$$

Denote

$$S_n = \begin{pmatrix} \frac{\partial Q_{1n}}{\partial \lambda^T} & \frac{\partial Q_{1n}}{\partial \theta} \\ \frac{\partial Q_{2n}}{\partial \lambda^T} & 0 \end{pmatrix}_{(\theta_0, 0)}.$$

It can be shown that

$$\frac{1}{n} S_n \rightarrow S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{pmatrix}$$

in probability. We have

$$\begin{aligned} \hat{\theta} - \theta_0 &= (0, I) \begin{pmatrix} \hat{\lambda} \\ \hat{\theta} - \theta_0 \end{pmatrix} \\ &= (0, I) S_n^{-1} \begin{pmatrix} -Q_{1n}(\theta_0, 0) \\ 0 \end{pmatrix} + o_p(\delta_n) \\ &= S_{22,1}^{-1} S_{n21} S_{11}^{-1} Q_{1n}(\theta_0, 0) + o_p(\delta_n). \end{aligned}$$

By the matrix algebra we can show that, in distribution,

$$\sqrt{n}(\hat{\theta} - \theta_0) = S_{22,1}^{-1} S_{21} S_{11}^{-1} \sqrt{n} Q_{1n}(\theta_0, 0) + o_p(1) \rightarrow N(0, V),$$

where

$$S_{22,1} = S_{22} - S_{21} S_{11}^{-1} S_{12} = E(\partial g / \partial \theta)^T (E g g^T)^{-1} E(\partial g / \partial \theta) = V^{-1}.$$

Next we present a few interesting Corollaries from this Theorem.

Intuitively, information about the underlying parameters should increase with the number of estimating functions.

Corollary 1 Suppose $r - 1 \geq p$, then the asymptotic variance V_r of $\sqrt{n}(\hat{\theta} - \theta)$ is a decreasing matrix of r , i.e.,

$$V_{r-1} - V_r \geq 0,$$

a non-negative definite matrix.

In fact, denote

$$D_r(\theta) = (\partial g_1 / \partial \theta, \dots, \partial g_{r-1} / \partial \theta, \partial g_r / \partial \theta)^T = (D_{r-1}^T(\theta), \partial g_r / \partial \theta),$$

and

$$C_r(\theta) = E(gg^T) = \begin{pmatrix} C_{11}(\theta) & C_{12}(\theta) \\ C_{21}(\theta) & C_{22}(\theta) \end{pmatrix},$$

where $C_{11}(\theta)$ is a $(r - 1) \times (r - 1)$ matrix. Then

$$\begin{aligned} V_r^{-1} &= E(\partial g / \partial \theta)(Egg^T)^{-1}E(\partial g / \partial \theta) \\ &= D_r^T(\theta)C_r^{-1}(\theta)D_r(\theta) \\ &\geq (D_{r-1}^T(\theta), \partial g_r / \partial \theta^T) \begin{pmatrix} C_{11}^{-1}(\theta) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} D_{r-1}^T(\theta) \\ \partial g_r / \partial \theta \end{pmatrix} \\ &= D_{r-1}^T(\theta)C_{r-1}^{-1}(\theta)D_{r-1}(\theta) = V_{r-1}^{-1}. \end{aligned}$$

As discussed in Chap. 5, when the number of estimating equations and number of unknown parameters are the same, the score estimating functions are optimal. In the next Corollary we will show that if the first p estimating equations among r estimating functions are the score, increasing r does not lead to increase in information for β .

Corollary 2 Suppose the first p estimating functions are score estimating functions i.e.,

$$g_{1,\dots,p} = \frac{\partial \log f(x, \theta)}{\partial \theta},$$

then the information for β from the first p estimating function is equivalent to the information from all estimating functions.

Proof Let

$$h_1 = (g_1, \dots, g_p) = \frac{\partial \log f(x, \theta)}{\partial \theta},$$

$$h_2 = (g_{p+1}, \dots, g_r).$$

$$\begin{aligned} V_r^{-1} &= (E(\partial h_1 / \partial \theta), E(\partial h_2 / \partial \theta)) \begin{pmatrix} E(h_1 h_1^T) & E(h_1 h_2^T) \\ E(h_2 h_1^T) & E(h_2 h_2^T) \end{pmatrix}^{-1} \begin{pmatrix} E(\partial h_1 / \partial \theta) \\ E(\partial h_2 / \partial \theta) \end{pmatrix} \\ &= (E(\partial h_1 / \partial \theta), E(\partial h_2 / \partial \theta)) \begin{pmatrix} I & -(Eh_1 h_1^T)^{-1} E(h_1 h_2^T) \\ 0 & I \end{pmatrix} \begin{pmatrix} (Eh_1 h_1^T)^{-1} & 0 \\ 0 & A_{22,1}^{-1} \end{pmatrix} \\ &\quad \begin{pmatrix} I & 0 \\ -(Eh_1 h_1^T)^{-1} E(h_2 h_1^T) & I \end{pmatrix} \begin{pmatrix} E(\partial h_1 / \partial \theta) \\ E(\partial h_2 / \partial \theta) \end{pmatrix} \\ &= \begin{pmatrix} E(\partial h_1 / \partial \theta) \\ -E(h_2 h_1^T)(Eh_1 h_1^T)E(\partial h_1 / \partial \theta) + E(\partial h_2 / \partial \theta) \end{pmatrix}. \end{aligned}$$

Since h_1 is the score, $E(h_1 h_1^T) = -E(\partial h_1 / \partial \theta)$. Also from $E(h_2) = 0$, we have

$$\int \frac{\partial h_2}{\partial \theta} f(x, \theta) dx + \int h_2 \frac{\partial f(x, \theta)}{\partial \theta} dx = 0,$$

or

$$E(\partial h_2 / \partial \theta) + E(h_2 h_1^T) = 0.$$

Thus

$$\begin{aligned} V_r^{-1} &= (E(\partial h_1 / \partial \theta^T), 0) \begin{pmatrix} (Eh_1 h_1^T)^{-1} & 0 \\ 0 & A_{22.1}^{-1} \end{pmatrix} \begin{pmatrix} E(\partial h_1 / \partial \theta) \\ 0 \end{pmatrix} \\ &= E(\partial h_1 / \partial \theta^T) (Eh_1 h_1^T)^{-1} E(\partial h_1 / \partial \theta) = V_p^{-1}. \end{aligned}$$

Next we show the Wilks' likelihood ratio behavior holds true for the profile empirical likelihood $\ell(\theta)$.

Theorem 8.3 *The empirical likelihood ratio test is defined as*

$$R(\theta) = 2[\max_{\hat{\theta}} \ell(\hat{\theta}) - \ell(\theta)].$$

Then as $n \rightarrow \infty$, in distribution, $R(\theta_0)$ converges to a chi-squared distribution with degrees of freedom of p , where θ_0 is the true value of θ .

Proof Expanding the log-empirical likelihood ℓ at $(\theta_0, 0)$, we have

$$\begin{aligned} \ell(\hat{\theta}, \hat{\lambda}) &= \sum_{i=1}^n \log[1 + \hat{\lambda}^T g(x_i, \hat{\theta})] \\ &= \sum_{i=1}^n \lambda^T g(x_i, \hat{\theta}) - 0.5 \hat{\lambda}^T \sum_{i=1}^n g(x_i, \hat{\theta}) g^T(x_i, \hat{\theta}) \hat{\lambda} + o_p(1) \\ &= -0.5n Q_{1n}^T(\theta_0, 0) A Q_{1n}(\theta_0, 0) + o_p(1). \end{aligned}$$

Also under H_0 , from

$$n^{-1} \sum_{i=1}^n \frac{1}{1 + \lambda_0^T g(x_i, \theta_0)} g(x_i, \theta_0) = 0,$$

we have

$$\lambda_0 = -S_{11}^{-1} Q_{1n}(\theta_0, 0) + o_p(1),$$

$$\sum_{i=1}^n \log[1 + \lambda_0^T g(x_i, \theta_0)] = -0.5n Q_{1n}^T(\theta_0, 0) S_{11}^{-1} Q_{1n}(\theta_0, 0) + o_p(1).$$

Therefore the likelihood ratio statistic is

$$\begin{aligned} R &= n Q_{1n}^T(\theta_0, 0) (A - S_{11}^{-1}) Q_{1n}(\theta_0, 0) + o_p(1) \\ &= n Q_{1n}^T(\theta_0, 0) S_{11}^{-1} S_{12} S_{22.1}^{-1} S_{21} S_{11}^{-1} Q_{1n}(\theta_0, 0) + o_p(1) \\ &= [(-S_{11})^{-1/2} \sqrt{n} Q_{1n}(\theta_0, 0)]^T [(-S_{11})^{-1/2} S_{12} S_{22.1}^{-1} S_{21} (-S_{11})^{-1/2}] \\ &\quad [(-S_{11})^{-1/2} \sqrt{n} Q_{1n}(\theta_0, 0)] + o_p(1). \end{aligned}$$

Note that $(-S_{11})^{-1/2}\sqrt{n}Q_{1n}(\theta_0, 0)$ converges to a standard multivariate normal distribution and $(-S_{11})^{-1/2}S_{12}S_{22,1}^{-1}S_{21}(-S_{11})^{-1/2}$ is symmetric and idempotent with trace equal to p . Therefore the empirical likelihood ratio statistic R converges to χ_p^2 in distribution.

Corollary 3 Let $\theta^T = (\theta_1, \theta_2)^T$, where θ_1 and θ_2 are $q \times 1$ and $(p - q) \times 1$ vectors, respectively. Under $H_0 : \theta_1 = \theta_1^0$, the profile empirical likelihood ratio test statistic

$$R_2 = 2[\max_{\theta_2} \ell(\theta_1^0, \theta_2) - \max_{\theta} \ell(\theta_1, \theta_2)]$$

converges to a chi-squared distribution with q degrees of freedom. We leave the proof of this corollary to readers.

We calculated the information lower bound for θ in Chap. 6 for the over-identified model $E[g(x, \theta)] = 0$, where θ is a $p \times 1$ vector parameter and g is a $r \times 1$ vector function, $r \geq p$. Clearly the empirical likelihood based method and generalized method of moments both achieve the lower bound for estimation of θ . One advantage of the empirical likelihood method over the GMM method is that the empirical likelihood method provides a natural estimator for the underlying distribution function

$$\hat{F}(x) = \sum_{i=1}^n \hat{p}_i I(x_i \leq x).$$

Next we discuss its the optimality property.

Optimal Property for the Distribution Estimation

It is not difficult to show that, in distribution,

$$\sqrt{n}\{\hat{F}(x) - F(x)\} \rightarrow N(0, \sigma^2(x)),$$

where

$$\sigma^2(x) = F(x)\{1 - F(x)\} - BUB^T, \quad B = E[g(X, \theta_0)I(X \leq x)],$$

and U is the asymptotic variance of $\sqrt{n}\hat{\lambda}$. In other words, the constrained estimator of F has an asymptotic variance smaller than that of the conventional empirical distribution estimator.

The empirical distribution has influence function

$$\sum_{i=1}^n \{I(x_i \in C) - P(C)\},$$

where C is any measurable set. It can be decomposed as

$$I(x_i \in C) - P(C) = I(x \in C) - P(C) + \text{cov}(I(X \in C), g)Wg - \text{cov}(I(X \in C), g)Wg,$$

where $W = S_{11}^{-1} + S_{11}^{-1}S_{12}S_{22,1}^{-1}S_{21}S_{11}^{-1}$.

We can show

(1)

$$I(x \in C) - P(C) + \text{cov}(I(X \in C), g)Wg \in \mathcal{S}_\eta,$$

where \mathcal{S}_η is the space spanned by (See Sect. 6.4 in Chap. 6)

$$g_B - E(\partial g_B / \partial \theta)(E\partial g_A / \partial \theta)^{-1}g_A, \quad g = (g_A, g_B)^T.$$

(2) $\text{cov}(I(x \in C), g)Wg$ is orthogonal to \mathcal{S}_η .

Exercise 1 Show the above two properties.

As a result, the asymptotic variance based on the empirical distribution is not lower than $\text{var}[I(x \in C) - P(C) + \text{cov}(I(X \in C), g)Wg]$,

As applications of the empirical likelihood method, we leave following problems as exercises for readers.

Exercises on Empirical Likelihood Related Problems

Exercise 1 Suppose we are interested in testing the conditional density of Y given $X = x$

$$H_0 : f(y|x) = f(y|x, \beta).$$

Clearly under H_0 , the conditional score function $\partial \log f(y|x, \beta) / \partial \beta$ has zero mean. Let $h(x)$ be any vector function that depends on X only. Then

$$E \left[h(x) \frac{\partial \log f(y|x, \beta)}{\partial \beta} \right] = 0.$$

We may combine two sets of estimating functions

$$\left(\frac{\partial \log f(y|x, \beta)}{\partial \beta}, h(x) \frac{\partial \log f(y|x, \beta)}{\partial \beta} \right),$$

and use the empirical likelihood method to estimate β . What is the relationship between the maximum empirical likelihood estimator and the maximum likelihood estimator? Construct a test statistic for testing H_0 .

Exercise 2 Consider a parametric likelihood $f(x) = f(x, \theta)$. One may construct the empirical likelihood by using the score estimating equation

$$g(x, \theta) = \frac{\partial \log f(x, \theta)}{\partial \theta}.$$

On the other hand the parametric log-likelihood is

$$\ell(\theta) = \sum_{i=1}^n \log f(x_i, \theta).$$

Plot the two likelihoods vs. θ . Which one should we use? Why?

Exercise 3 Consider a simple bivariate score

$$S_i(y_{ij}, x_{ij}) = \frac{\partial \log f(y_{ij}|x_{ij}\beta)}{\partial \beta}, \quad i = 1, 2, \quad j = 1, 2, \dots, n,$$

discussed in Liang and Zeger (1986), with the joint density of (y_{1j}, y_{2j}) given (x_{1j}, x_{2j}) unspecified. Then the empirical likelihood method can be used to combine estimating functions $S_i, i = 1, 2$. Note the empirical likelihood approach does not need to choose a “working correlation matrix”. Moreover we may use extra estimating equations

$$h_i(y_{ij}, x_{ij}) = \frac{\partial^2 f(y_{ij}|x_{ij}\beta)/\partial \beta^2}{f(y_{ij}|x_{ij}\beta)}, \quad i = 1, 2.$$

Use empirical likelihood to combine estimating equations (S_1, S_2, h_1, h_2) . Compare this resultant estimator with the empirical likelihood estimator by combining S_1 and S_2 only. It was shown in Corollary 2 that extra estimating equations would not increase any information if the score equations are already included. However, in this bivariate case, only parametric models for the two margins are available. Are the extra estimating equations useful for estimating β ? Furthermore if $\beta = (\beta_1, \beta_2)$, develop the empirical likelihood ratio test for $\beta_1 = \beta_1^0$.

Exercise 4 Arnold and Strauss (1988) discussed conditional parametric models

$$f(y|x) = f_1(y|x, \beta), \quad f(x|y) = f_2(x|y, \beta).$$

In general the two conditional densities can uniquely determine the joint density. However the joint density may not have a closed form. Instead they proposed to use the composite likelihood

$$\ell_c = \sum_{i=1}^n \log f_1(y_i|x_i, \beta) + \sum_{i=1}^n \log f_2(x_i|y_i, \beta).$$

We may use the empirical likelihood method to combine $\partial \log f_1(y_i|x_i, \beta)/\partial \beta$ and $\partial \log f_2(x_i|y_i, \beta)/\partial \beta$. Compare the efficiency between the composite likelihood method and the empirical likelihood method.

Exercise 5 Conditional independence assumption.

Suppose the parameter β is defined through

$$E[\psi(X, Y, Z, \beta)] = 0.$$

In addition, assume that conditional on Z , Y and X is independent. Let $h_1(y, z, \theta)$ and $h_2(x, z, \gamma)$ be the conditional score of $Y|Z$ and $Z|X$, respectively. Then

$$E[h_1(Y, Z, \theta)h_2(X, Z, \gamma)] = 0, \quad E[h_1(Y, Z, \theta)] = 0, \quad E[h_2(X, Z, \gamma)] = 0,$$

Use the empirical likelihood method to combine the above estimating equations, then derive large sample results.

Exercise 6 A parametric characteristic function problem.

Chan et al. (2009) found an application of the empirical likelihood method for Levy processes based on its characteristic function. Let

$$\phi(t; \theta) = E[\exp(itX)]$$

be the characteristic function. In general it would be very difficult to find the density function by the Fourier inverse transformation. Chan et al. (2009) proposed maximizing the empirical likelihood $\prod_{i=1}^n p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0,$$

and

$$\sum_{i=1}^n p_i \cos(tx_i) = \phi^R(t; \theta), \quad \sum_{i=1}^n p_i \sin(tx_i) = \phi^I(t; \theta),$$

where ϕ^R and ϕ^I are real and image parts of ϕ , respectively. Since the above constraints hold true for any t , they considered an integrated version to estimate θ . Define

$$T(\theta) = \int_{-a}^a \ell(t; \theta) dG(t),$$

where G is a given smoothing distribution function. Their maximum empirical likelihood estimate is defined as

$$\hat{\theta} = \operatorname{argmax}_{\theta} T(\theta).$$

- (a) Derive large sample results for $\hat{\theta}$.
- (b) Discuss the optimal choice of G .

8.3 Miscellaneous Topics on Empirical Likelihood

1. Non-smooth Constraints

So far we have assumed that the underlying estimating functions are differentiable with respect to θ . However, in some cases, this may not be true, for example, the estimating problem for quantiles,

$$I(Y \leq \theta(q)) = q.$$

A general result on empirical likelihood method based on non-smooth estimating equations is given by Molanes Lopez et al. (2009). They derived large sample results based on the modern empirical process theory. Note that the expected value of an indicator function is smooth, even though this function itself is non-smooth. We consider two examples.

(a) The Empirical Likelihood Method for Quantile Function

We start from the construction of empirical likelihood for a quantile function. Denote the observed data as

$$X_1, \dots, X_n \sim dF(x).$$

Suppose we are interested in estimating the quantile of F . Then the estimating equation is defined as

$$E[I(X \leq \theta(\tau)) - \tau] = 0, \quad 0 < \tau < 1.$$

The profile empirical likelihood is

$$\ell = - \sum_{i=1}^n \log[1 + \lambda(I(x_i \leq \theta) - \tau)] - n \log n$$

subject to the constraint

$$\sum_{i=1}^n \frac{I(x_i \leq \theta) - \tau}{1 + \lambda(I(x_i \leq \theta) - \tau)} = 0.$$

Let

$$k = \sum_{i=1}^n I(x_i \leq \theta).$$

The constraint equation is

$$k \frac{1 - \tau}{1 + \lambda(1 - \tau)} - (n - k) \frac{\tau}{1 - \lambda\tau} = 0.$$

$$\lambda = \frac{k - n\tau}{n\tau(1 - \tau)}.$$

The profile log empirical likelihood is

$$\begin{aligned}\ell &= k \log \tau + (n - k) \log(1 - \tau) - k \log(k/n) - (n - k) \log((n - k)/n) - n \log n \\ &= k \log \tau + (n - k) \log(1 - \tau) - k \log k - (n - k) \log(n - k).\end{aligned}$$

Note that we do not need the observed data to have the same distribution. In fact the above approach is valid as long as $E[I(X \leq \theta) - \tau] = 0$. Kalbfleisch (1978) recommended using

$$\binom{n}{k} \tau^k (1 - \tau)^{n-k}$$

to make inference for the τ -th quantile.

(b) Empirical Likelihood for Copula Function Estimation

Consider a bivariate copula model with observed data

$$(X_{i1}, X_{i2}), i = 1, 2, \dots, n \sim H(x_1, x_2) = C(F_1(x_1), F_2(x_2)),$$

where F_1 and F_2 are two marginal distributions, and $C(\cdot, \cdot)$ is an unknown copula function. In probability theory and statistics, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables

We would like to make inference for the copula function

$$C(u_1, u_2) = H(F_1^{-1}(x_1), F_2^{-1}(x_2)).$$

Define

$$v_1 = F_1^{-1}(u_1), \quad v_2 = F_2^{-1}(u_2).$$

We have estimating equations

$$g_1(X_1, X_2, u, v) = I(X_1 \leq v_1, X_2 \leq v_2) - C(u_1, u_2),$$

and

$$g_2(X_1, u, v) = I(X_1 \leq v_1) - u_1, \quad g_3(X_2, u, v) = I(X_2 \leq v_2) - u_2.$$

Denote $g = (g_1, g_2, g_3)$. Based on the estimating functions g , we should be able to construct an empirical likelihood based confidence interval for the copula function.

2. Some Strategies for Numerical Solutions

Many softwares are available for estimation of the sample mean, for example, Owen's Stanford web page on the empirical likelihood and Yang and Small's (2013) R package etc. It becomes complex in the general estimating equations setup.

The numerical method suggested by Chen et al. (2002) is effective for solving the Lagrange multipliers. The key is that in each iteration, the Lagrange multipliers satisfy the condition

$$1 + \lambda^T g(x_i, \theta) > 0, \quad i = 1, 2, \dots, n.$$

If this is not the case, the step length must be reduced by half until this condition is true.

As an alternative, we also have found the following strategy works well for most practical problems. Suppose we need to solve an equation

$$Q(\theta) = 0$$

in the neighbourhood of θ_0 . We can choose a constant r , say, $r = 20$. First we solve

$$Q(\theta) - Q(\theta_0) = 0.$$

Clearly θ_0 is the solution. Next we solve

$$Q(\theta) - \frac{r-1}{r} Q(\theta_0) = 0$$

by using θ_0 as the initial value. Denote the solution as θ_1 . Then we need to solve

$$Q(\theta) - \frac{r-2}{r} Q(\theta_0) = 0$$

using θ_1 as initial value. Denote the solution as θ_2 . Continue this procedure until the estimating equation becomes

$$Q(\theta) - \frac{r-r}{r} Q(\theta_0) = 0, \quad \text{i.e. } Q(\theta) = 0$$

using θ_{r-1} as the initial value. Finally the desire solution is θ_r .

3. Adjusted Empirical Likelihood Method to Deal with No Solution Problems

Suppose we have an unbiased estimating function $g(x, \theta)$. By definition, the true parameter value θ_0 is the unique solution of $E\{g(X; \theta)\} = 0$. Hence, under some moment conditions on $g(X, \theta)$ (Owen 2001), the convex hull $\{g(x_i, \theta_0), i = 1, 2, \dots, n\}$ contains 0 as its inner point with probability 1 as $n \rightarrow \infty$. When θ is not close to θ_0 , or when n is small, there is a considerable chance that the constraint

equations have no solutions. This can be a serious limitation in some applications. Chaudhuri et al. (2007) discussed the problem of estimating covariance matrix of a random vector with many known zero entries. Let $Y_i, i = 1, 2, \dots, n$ be a sample of 4×1 vector random variables. Suppose based on prior information we know that $\sigma_{ij} = 0, (i, j) = (1, 2), (2, 4), (3, 4)$. The empirical likelihood is

$$\max \prod_{i=1}^n p_i$$

subject to the constraints

$$\sum_{i=1}^n (y_{ij} - \mu_j) p_i = 0, \quad \sum_{i=1}^n p_i (y_{ij} - \mu_j)(y_{ik} - \mu_k) = 0, (j, k) = (1, 2), (2, 4), (3, 4), p_i \geq 0, \quad \sum_{i=1}^n p_i = 1.$$

Chen et al. (2008) found that as large as 759 times out of 1000 simulations the empirical likelihood has no solution for a sample size $n = 10$ if the underlying distribution is generated from a normal distribution.

As a remedy, Chen et al. (2008) added one or two extra artificial data points to guarantee the solution. They proposed the following adjusted empirical likelihood method. Denote $g_i = g_i(\theta) = g(x_i; \theta)$ and $\bar{g}_n = \bar{g}(\theta) = n^{-1} \sum_{i=1}^n g_i$ for any given θ . For some positive constant a_n , define $g_{n+1} = g_{n+1}(\theta) = -a_n \bar{g}(\theta)$. The adjusted profile empirical log-likelihood ratio function is defined as

$$\sup \left[\sum_{i=1}^{n+1} \log\{(n+1)p_i\} : \quad p_i = 0, i = 1, \dots, n+1; \quad \sum_{i=1}^{n+1} p_i = 1, \quad \sum_{i=1}^{n+1} p_i g_i(\theta) = 0 \right].$$

Since the convex hull of $\{g_i, i = 1, 2, \dots, n, n+1\}$ for any given θ contains 0, the empirical likelihood is well defined without exceptions. If $a_n = O_p(n^{2/3})$, it is not difficult to show that the large sample results remain the same since the extra data point is “absorbed” by the remaining n data points. In applications Chen et al. (2008) recommended to choose $a_n = (\log n)/2$.

To improve the coverage level, Tsao and Wu (2013) used a different approach. They extended the empirical likelihood by partitioning its domain into a collection of its contours and mapping the contours through a continuous sequence of similarity transformations onto the full parameter space.

We conclude this section with three exercises.

Exercise 1 A problem arising in many different contexts is the comparison between two different treatment conditions. Consider $N = m + n$ independent measurements in two samples. The first sample consists of n measurements x_1, \dots, x_n recorded under one set of conditions, and the second sample consists of m measurements y_1, \dots, y_m recorded under a different set of conditions. We are interested in comparing $\mu_1 = E(X)$ and $\mu_2 = E(Y)$.

(a) Derive the empirical likelihood if both distributions F and G are nonparametric, where $X \sim F$ and $Y \sim G$.

(b) If we can assume a parametric model for $G(y) = G(y, \theta)$, then $\mu_2 = \mu_2(\theta) = \int y dG(y, \theta)$. Let $\Delta = \mu_1 - \mu_2(\theta)$. Then the joint likelihood (one is the empirical likelihood and the other one is the parametric likelihood) is

$$\prod_{i=1}^n dF(x_i) \prod_{j=1}^m g(y_j, \theta).$$

Only under the constraint $\sum_{i=1}^n dF(x_i) = 1$, $dF(x_i) \geq 0$, it has maximum value $n^{-n} \prod_{j=1}^m g(y_j, \hat{\theta})$, where $\hat{\theta}$ is the MLE based on the second sample. Let

$$R(F, \theta) = \frac{\prod_{i=1}^n dF(x_i) \prod_{j=1}^m g(y_j, \theta)}{n^{-n} \prod_{j=1}^m g(y_j, \hat{\theta})},$$

$$C_r = \left\{ \int x dF(x) - \mu_2(\theta) \mid F \leq F_n, R(F, \theta) \geq r \right\},$$

$$\mathcal{R}(\Delta) = \sup_{F, \theta} \{R(F, \theta) \mid \int x dF(x) - \mu_2(\theta) = \Delta, F \leq F_n\},$$

where $F \leq F_n$ denotes that F is absolutely continuous with respect to F_n , i.e., the support of F is contained in the support of the empirical distribution F_n . Show that under some regularity conditions

$$-2 \log \mathcal{R}(\Delta_0) \rightarrow \chi^2(1),$$

where Δ_0 is the true value of Δ .

Exercise 2 Quite frequently in statistical theory the natural way of building up a mathematical model of an experiment leads to the description of the experiment by a random variable X whose distribution function F depends on p parameters $\theta_{p \times 1}$, which are not mathematically independent but satisfy q functional relationships $h(\theta) = 0$, where $h(\theta) = (h_1(\theta), \dots, h_q(\theta))^T$ and $q < p$. If f follows a parametric model $f(x) = f(x, \theta)$, Aitchison and Silvey (1958) introduced Lagrange multipliers to maximize the parametric likelihood subject to the constraints. In particular they showed that the constrained maximum likelihood estimators have many desirable properties.

A natural extension is to replace the parametric model by estimating equations

$$E_F[\psi(x, \theta)] = 0, \quad h(\theta) = 0,$$

where ψ is a p -dimensional vector function. We can construct empirical likelihood

$$\prod_{i=1}^n dF(x_i) = \prod_{i=1}^n p_i$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \psi(x_i, \theta) = 0.$$

The profile empirical likelihood is

$$\ell = - \sum_{i=1}^n \log\{1 + \lambda^T \psi(x_i, \theta)\}$$

subject to the constraint

$$\sum_{i=1}^n \frac{\psi(x_i, \theta)}{1 + \lambda^T \psi(x_i, \theta)} = 0.$$

Then we need to introduce another Lagrange multiplier ν

$$H = \sum_{i=1}^n \log\{1 + \lambda^T \psi(x_i, \theta)\} + \nu^T h(\theta).$$

Derive large sample results for the constrained maximum empirical likelihood estimate.

Exercise 3 Consider estimation of parameters in nonlinear implicitity models (Britt and Luecke 1973). Suppose a system in which a p -vector of unknown parameters θ and a m -vector of observable variables with true values at the time of measurement of ξ , are related though a q -vector function $g(\xi, \theta) = 0$, where $p < q \leq m$. The measurements of ξ contain random experimental errors so that

$$X_i = \xi + \epsilon_i.$$

In conventional approach the error is assumed to have a normal distribution.

$$E(X - \xi) = 0, \quad g(\xi, \theta) = 0.$$

Apply the empirical likelihood method to this problem and then develop large sample results.

8.4 Hybrid Likelihoods and Utilization Auxiliary Information

Utilizing auxiliary information to make a sharper inference in survey sampling is a very popular method, for example, Cochran (1977) and Sarndal et al. (1991) had a comprehensive discussion on the regression type estimators. The classic example is to use the sample mean $\bar{X} = N^{-1} \sum_{i=1}^N x_i$ from a large survey to enhance the mean estimation of Y . Due to restrictions such as cost effectiveness, convenience etc., only a small proportion of (y_i, x_i) , $i = 1, 2, \dots, n$ ($n << N$) is available. Chen and Qin (1993), Chen et al. (2002) and Wu and Sitter (2001) used empirical likelihood method to incorporate such information in finite population. With the advance of technology, many summarized statistical results are available in public domains. For example, many aggregated demographic and socioeconomic status data are given in the US census reports. The Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute provides the population-based cancer survival statistics, such as covariate specific survival probabilities. Imbens and Lancaster (1994) combined Micro and Macro data in economic studies through GMM. Chaudhuri et al. (2008) showed that the inclusion of the population level information can reduce bias and increase efficiency of the parameter estimates in a generalized linear model setup.

Next we will discuss the results by Qin (2000). Details were worked out in his Ph.D. thesis (University of Waterloo, Qin 1992).

Assume a conditional density model

$$f(y|x) = f(y|x, \beta),$$

and the marginal distribution $G(x)$ of X is not specified. Moreover assume some auxiliary information is given by

$$E[\phi(X, \beta)] = 0.$$

For example if we know the mean of Y , say, μ , then we have estimating equation

$$E[Y - \mu] = 0.$$

We can take

$$\phi(X, \beta) = \int (Y - \mu) f(y|x, \beta) dx = \int y f(y|x, \beta) dy - \mu.$$

We are interested in making inference for β . Furthermore we assume that the observed data may have missing value for response Y . Let D be the missing response indicator, being 1 if Y is available, and 0 otherwise. We assume a missing at random model

$$P(D = 1|y, x) = \pi(x),$$

where $\pi(x)$ only depends on x . The likelihood is

$$L = \prod_{i=1}^n [\pi(x_i) f(y_i | x_i \beta) dG(x_i)]^{d_i} [(1 - \pi(x_i)) dG(x_i)]^{1-d_i}.$$

We can maximize this likelihood subject to the constraints

$$\sum_{i=1}^n p_i = 1; \quad p_i \geq 0; \quad \sum_{i=1}^n p_i \phi(x_i, \beta) = 0.$$

The profile hybrid log empirical likelihood is

$$\ell = \sum_{i=1}^n [d_i \log f(y_i | x_i \beta) - \log \{1 + \nu^T \phi(x_i, \beta)\}], \quad (8.4.5)$$

where ν is the Lagrange multiplier determined by

$$Q_{1n}(\beta, \nu) := \sum_{i=1}^n \frac{\phi(x_i, \beta)}{1 + \nu^T \phi(x_i, \beta)} = 0. \quad (8.4.6)$$

The score for β is

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n \left[d_i \frac{\partial \log f(y_i | x_i \beta)}{\partial \beta} - \frac{\nu^T \partial \phi(x_i, \beta) / \partial \beta}{1 + \nu^T \phi(x_i, \beta)} \right] = 0.$$

In the special case that missing data is completely at random, i.e., $\pi(x_i)$ is a constant, Qin (1992, 2000) showed the following theorem.

Theorem 8.4 *Let $\tilde{\beta}$ be the maximum hybrid log empirical likelihood estimator, i.e., maximizing (8.4.5). Under some regularity conditions, in distribution*

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}, \quad (8.4.7)$$

$$\Sigma_{11} = \{\gamma_0 J + E(\partial \phi \phi^T / \partial \beta^T) (E \phi \phi^T)^{-1} E(\partial \phi / \partial \beta)\}^{-1}$$

and

$$\Sigma_{22} = \{E(\partial \phi / \partial \beta)(\gamma_0 J)^{-1} E(\partial \phi / \partial \beta^T) + E(\phi \phi^T)\}^{-1}.$$

Proof Without loss of generality we assume $d_i = 1, i = 1, 2, \dots, m$, i.e., $(y_i, x_i), i = 1, 2, \dots, m$ are available, but only $x_j, j = m+1, \dots, n$ are available. Denote

$$Q_{2n}(\beta, \nu) = \sum_{i=1}^n \frac{1}{1 + \nu^T \phi(x_i, \beta)} \nu^T \frac{\partial \phi(x_i, \beta)}{\partial \beta}, \quad Q_{3n}(\beta) = \sum_{i=1}^m \frac{\partial \log f(y_i | x_i, \beta)}{\partial \beta}.$$

Let $\tilde{\beta}$ and $\tilde{\nu}$ be a solution of the following equations

$$Q_{1n}(\beta, \nu) = 0, \quad Q_{2n}(\beta, \nu) - Q_{3n}(\beta) = 0.$$

Expanding $Q_{in}(\tilde{\beta}, \tilde{\nu}), i = 1, 2, 3$ at the point $(\beta_0, 0)$, we have

$$Q_{1n}(\tilde{\beta}, \tilde{\nu}) = Q_{1n}(\beta_0, 0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \beta} (\tilde{\beta} - \beta_0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \nu} (\tilde{\nu} - 0) + o_p(n^{1/2}),$$

$$\begin{aligned} Q_{2n}(\tilde{\beta}, \tilde{\nu}) - Q_{3n}(\tilde{\beta}) &= Q_{2n}(\beta_0, 0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \beta} (\tilde{\beta} - \beta_0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \nu} (\tilde{\nu} - 0) \\ &\quad - Q_{3n}(\beta_0) - \frac{\partial Q_{3n}(\beta_0)}{\partial \beta} (\tilde{\beta} - \beta_0) + o_p(n^{1/2}). \end{aligned}$$

In matrix form we have

$$\begin{pmatrix} Q_{3n}(\beta_0) \\ -Q_{1n}(\beta_0, 0) \end{pmatrix} = \begin{pmatrix} \frac{\partial Q_{3n}}{\partial \beta} & \frac{\partial Q_{2n}}{\partial \nu^T} \\ \frac{\partial Q_{3n}}{\partial \beta} & \frac{\partial Q_{2n}}{\partial \nu^T} \end{pmatrix}_{\beta_0, 0} \begin{pmatrix} \tilde{\beta} - \beta_0 \\ \tilde{\nu} - 0 \end{pmatrix} + o_p(n^{1/2}).$$

Therefore

$$\sqrt{n} \begin{pmatrix} \tilde{\beta} - \beta_0 \\ \tilde{\nu} - 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial Q_{3n}}{\partial \beta} & \frac{\partial Q_{2n}}{\partial \nu^T} \\ \frac{\partial Q_{3n}}{\partial \beta} & \frac{\partial Q_{2n}}{\partial \nu^T} \end{pmatrix}_{\beta_0, 0}^{-1} \begin{pmatrix} Q_{3n}(\beta_0) \\ -Q_{1n}(\beta_0, 0) \end{pmatrix} + o_p(1) \rightarrow N(0, U), \quad (8.4.8)$$

where

$$\Sigma = SVS^T, \quad S = \begin{pmatrix} -\gamma_0 J & E(\frac{\partial \phi}{\partial \beta^T}) \\ E(\frac{\partial \phi}{\partial \beta}) & E(\phi \phi^T) \end{pmatrix}^{-1}, \quad V = \begin{pmatrix} \gamma_0 J & 0 \\ 0 & E(\phi \phi^T) \end{pmatrix},$$

and

$$J = E \left[\frac{\partial \log f(y|x, \beta)}{\partial \beta} \frac{\partial \log f(y|x, \beta)}{\partial \beta^T} \right].$$

Easily we can show that Σ is given in (8.4.7).

Because the conditional score and estimating function $\phi(x, \beta)$ are orthogonal to each other, the information provided by the hybrid likelihood is the summation of the information provided by the conditional score and the auxiliary information $E[\phi] = 0$.

Remark Imbens and Lancaster (1994) discussed the same problem by using a method of moments approach. In particular, they directly combined the conditional score estimating equation $\partial \log f(y|x\beta)/\partial\beta$ and $\phi(x, \beta)$. Even though the first order large sample results are the same, the hybrid empirical likelihood based approach is more appealing since it respects the parametric conditional likelihood and only the marginal likelihood is replaced by the empirical likelihood. Numerical comparison between the two methods was given in Qin (2000).

Next we consider the hybrid likelihood ratio inference for the response mean,

$$\mu = E(Y) = E[E(Y|X)] = E\left[\int yf(y|x)dy\right] =: E[\mu(X, \beta)].$$

The hybrid likelihood for the mean is defined as

$$R(G, \beta) = \frac{\prod_{i=1}^n dG(x_i) \prod_{j=1}^m f(y_j|x_j, \beta)}{n^{-n} \prod_{j=1}^m f(y_j|x_j, \hat{\beta})},$$

$$\mathcal{R}(\mu) = \sup_{G, \theta} \{R(G, \beta) | \int \mu(x, \beta) dG(x) = \mu, G \leq G_n\},$$

where $G \leq G_n$ means that G is absolutely continuous with respect to G_n , the empirical distribution function of X . The hybrid likelihood confidence interval is defined as

$$\mathcal{C}_{r,n} = \{\mu | -2 \log \mathcal{R}(\mu) \geq r\}.$$

Then $\mu \in \mathcal{C}_{r,n}$ if and only if $\mathcal{R}(\mu) \leq \exp(-r/2)$. In order to get $\mathcal{R}(\mu)$, we need to maximize the hybrid likelihood subject to constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \psi(x_i, \mu, \beta) = 0,$$

where $\psi(x, \mu, \beta) = \mu - \mu(x, \beta)$.

Theorem 8.5 Under some regularity conditions, the hybrid likelihood ratio statistic satisfies

$$-2 \log \mathcal{R}(\mu_0) \rightarrow \chi^2(1).$$

Proof Let $\tilde{\beta}$ be the maximum hybrid likelihood estimator under the constraint $\mu = \mu_0$. From (8.4.8), we have

$$\tilde{\beta} - \beta_0 = (I_{p \times p}, 0_{p \times q}) S \begin{pmatrix} Q_{3n}(\beta_0) \\ Q_{1n}(\beta_0, \mu_0, 0) \end{pmatrix} + o_p(n^{-1/2}).$$

Denote $\hat{\beta}$ as the maximum likelihood estimator without the mean constraint $\mu = \mu_0$. Then $\hat{\beta}$ satisfies the score equation $\sum_{i=1}^n \partial \log f(y_i | x_i, \beta) / \partial \beta = 0$. Using Taylor's expansion, we have

$$\begin{aligned}\hat{\beta} - \beta_0 &= (\gamma_0 J)^{-1} Q_{3n}(\beta_0) + o_p(n^{-1/2}) \\ &= (I, 0) \begin{pmatrix} (\gamma_0 J)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q_{3n}(\beta_0) \\ Q_{1n}(\beta_0, \mu_0, 0) \end{pmatrix} + o_p(n^{-1/2}).\end{aligned}$$

Easily we have

$$\begin{aligned}\tilde{\beta} - \hat{\beta} &= -(\gamma_0 J)^{-1} E \left(\frac{\partial \psi}{\partial \beta} \right)^T (0_{q \times p}, I_{q \times q}) S \begin{pmatrix} Q_{3n}(\beta_0) \\ Q_{1n}(\beta_0, \mu_0, 0) \end{pmatrix} + o_p(n^{-1/2}) \\ &= -(\gamma_0 J)^{-1} E \left(\frac{\partial \psi}{\partial \beta} \right)^T \tilde{\nu} + o_p(n^{-1/2}).\end{aligned}$$

From $Q_{1n}(\tilde{\beta}, \tilde{\nu}, \mu_0) = 0$, we have

$$\begin{aligned}\sum_{i=1}^m \log \{1 + \tilde{\nu}^T \psi(x_i, \tilde{\beta}, \mu_0)\} &= \sum_{i=1}^n \tilde{\nu}^T \psi(x_i, \tilde{\beta}, \mu_0) - 0.5 \sum_{i=1}^n \tilde{\nu}^T \psi \psi^T \tilde{\nu} + o_p(1) \\ &= 0.5n \tilde{\nu}^T (E \psi \psi^T) + o_p(1),\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^m \{\log f(y_i | x_i, \tilde{\beta}) - \log f(y_i | x_i, \hat{\beta})\} \\ &= \sum_{i=1}^m \frac{\partial \log f(y_i | x_i, \hat{\beta})}{\partial \beta} (\tilde{\beta} - \hat{\beta}) + 0.5(\tilde{\beta} - \hat{\beta})^T \sum_{i=1}^m \frac{\partial^2 \log f(y_i | x_i, \hat{\beta})}{\partial \beta \partial \beta^T} (\tilde{\beta} - \hat{\beta}) + o_p(1) \\ &= 0 - 0.5n(\tilde{\beta} - \hat{\beta})^T (\gamma_0 J)^{-1} (E \frac{\partial \psi}{\partial \beta^T}) \tilde{\nu} + o_p(1) \\ &= -0.5n \tilde{\nu}^T (E \frac{\partial \psi}{\partial \beta^T}) (\gamma_0 J)^{-1} (E \frac{\partial \psi}{\partial \beta^T}) \tilde{\nu} + o_p(1).\end{aligned}$$

Hence the hybrid likelihood ratio statistic satisfies

$$\begin{aligned}&-2 \log \mathcal{R}(\mu_0) \\ &= 2 \sum_{i=1}^m \log \{1 + \tilde{\nu}^T \psi(x_i, \tilde{\beta}, \mu_0)\} - 2 \sum_{i=1}^m \{\log f(y_i | x_i, \tilde{\beta}) - \log f(y_i | x_i, \hat{\beta})\} \\ &= n \tilde{\nu}^T \left\{ (E \frac{\partial \psi}{\partial \beta^T}) (\gamma_0 J)^{-1} (E \frac{\partial \psi}{\partial \beta^T}) \tilde{\nu} + E(\psi \psi^T) \right\} + o_p(1) \\ &\rightarrow \chi^2(1).\end{aligned}$$

This theorem is a generalization of Wilks' theorem in a semiparametric model setup.

Example (Qin 1992 Thesis) Consider a simple logistic model for a binary indicator variable Y ,

$$P(Y = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Assume that the marginal probability $q_0 = P(Y = 0)$ is known. Hence

$$\psi(x, \alpha, \beta) = q_0 - \frac{1}{1 + \exp(\alpha + \beta x)}$$

the log hybrid likelihood is

$$\begin{aligned} \ell = & -\sum_{i=1}^n \log\{1 + \nu\psi(x_i, \alpha, \beta)\} + \sum_{i=1}^m (\alpha + \beta x_i) I(y_i = 1) \\ & - \sum_{i=1}^m \log\{1 + \exp(\alpha + \beta x_i)\}, \end{aligned}$$

where ν is determined by the constraint equation (8.4.6). The parametric log-likelihood is

$$\ell_p = \sum_{i=1}^m (\alpha + \beta x_i) I(y_i = 1) - \sum_{i=1}^m \log\{1 + \exp(\alpha + \beta x_i)\}.$$

After some simplifications, we have

$$\nu = \frac{mq_0 - m_0 + \sum_{i=m+1}^n \psi(x_i, \alpha, \beta)}{nq_0(1 - q_0)}.$$

In the special case $n = m$, $\nu = (mq_0 - m_0)/\{mq_0(1 - q_0)\}$ ($m_0 = \sum_{i=1}^m I(y_i = 0)$) is independent of α, β . Easily we can show that

$$\ell(\alpha, \beta) = \ell_p(\alpha^*, \beta) + C,$$

where

$$\alpha = \alpha + \log\left(\frac{1 + \nu q_0}{1 + \nu q_0 - \nu}\right), \quad C = -m \log(1 + \nu q_0 - \nu) - m_0 \log\left(\frac{1 + \nu q_0}{1 + \nu q_0 - \nu}\right).$$

Note that C is a constant independent of α, β . Therefore the maximum semiparametric likelihood estimators of β with auxiliary information on $q_0 = P(Y = 0)$ and without this auxiliary information have the same asymptotic distribution. We will discuss this further in Chap. 11.

Example 2 Consider a binary indicator variable Y with a Probit model

$$P(Y = 1|x) = 1 - \Phi(\alpha + x\beta),$$

where Φ is the standard normal distribution. Assume $q_0 = P(Y = 0)$ be known. Hence

$$\psi(x, \alpha, \beta) = q_0 - \Phi(\alpha + x\beta).$$

The hybrid log-likelihood is

$$\begin{aligned} \ell(\alpha, \beta) = & - \sum_{i=1}^n \log\{1 + \nu\psi(x_i, \alpha, \beta)\} + \sum_{i=1}^m I(y_i = 0) \log \Phi(\alpha + x_i\beta) \\ & + \sum_{i=1}^m I(y_i = 1) \log\{1 - \Phi(\alpha + x_i\beta)\}. \end{aligned}$$

We can derive the asymptotic formula for the maximum hybrid likelihood estimators $\hat{\alpha}, \hat{\beta}$ as

$$\begin{aligned} J = & \begin{pmatrix} E\{\phi^2(\alpha + x\beta)\xi(x) & E\{x\phi^2(\alpha + x\beta)\xi(x)\} \\ E\{x\phi^2(\alpha + x\beta)\xi(x)\} & E\{x^2\phi^2(\alpha + x\beta)\xi(x)\} \end{pmatrix}, \quad \phi(x) = \frac{d\Phi(x)}{dx} \\ E\{\psi^2(x, \alpha, \beta)\} = & E\{\Phi^2(\alpha + x\beta)\} - q_0^2 \end{aligned}$$

and

$$(E(\partial\psi/\partial\alpha), E(\partial\psi/\partial\beta)) = -(E\phi(\alpha + x\beta), E(x\phi(\alpha + x\beta))),$$

where $\phi(\cdot)$ is the standard normal density function and $\xi^{-1}(x) = \Phi(\alpha + x\beta)\{1 - \Phi(\alpha + x\beta)\}$.

Even though the asymptotic variances are different with or without using the auxiliary information, numerical studies showed (Qin 1992) that the efficiency gain in estimating β is not noticeable by using auxiliary information.

Example 3 Let T be a lifetime variable and X be some associated covariates. We assume a parametric model

$$f(t|x) = f(t|x\beta).$$

The observed data are

$$\{Y_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i), X_i\}, \quad i = 1, 2, \dots, n,$$

where $C_i, i = 1, 2, \dots, n$ are right censoring variables. The likelihood is

$$L = \prod_{i=1}^n f^{\delta_i}(y_i|x_i\beta) \bar{F}^{1-\delta_i}(y_i|x_i\beta) \prod_{i=1}^n dF(x_i).$$

Suppose the five year survival information for different covariate group is available from previous studies,

$$F(t_0|a_i < X \leq a_{i+1}) = \phi_i, i = 1, 2, \dots, I, t_0 = 5.$$

This information can be summarized as (Imbens and Lancaster 1994) estimating equation

$$E[\psi_i(X, \beta)] = 0, \quad \psi(X, \beta) = \{F(t_0|X\beta) - \phi_i\}I(a_i < X \leq a_{i+1}), i = 1, 2, \dots, I.$$

Let

$$\psi(X, \beta) = (\psi_1, \dots, \psi_I)^T.$$

We can construct the empirical likelihood subject to the constraint

$$\sum_{i=1}^n p_i \psi(x_i\beta) = 0.$$

After profiling out the p_i 's, the joint log-likelihood is

$$\ell = \sum_{i=1}^n \delta_i \log f(y_i|x_i\beta) + (1 - \delta_i) \log \bar{F}(y_i|x_i\beta) - \sum_{i=1}^n \log\{1 + \lambda^T \psi(x_i, \beta)\},$$

where λ is the Lagrange multiplier determined by

$$\sum_{i=1}^n \frac{\psi(x_i, \beta)}{1 + \lambda^T \psi(x_i, \beta)} = 0.$$

Next we can estimate β by maximizing ℓ . This estimator should be more efficient than the maximum likelihood estimate of β without using the auxiliary information.

Exercise Derive the large sample results for the maximum joint likelihood estimator.

Example 4 Utilizing auxiliary information in truncation problems. Consider a truncation model

$$(Y, A)|Y > A \sim \frac{f(y, \beta)}{\bar{F}(a, \beta)} \frac{\bar{F}(a, \beta)dG(a)}{\int \bar{F}(a, \beta)dG(a)},$$

where the density of Y is a specified parametric model but the density $dG(a)$ of the truncation variable A is arbitrary. Suppose we have auxiliary information

$$E[A] = \mu,$$

i.e., the mean truncation time is known.

Note

$$A|Y > A \sim \frac{g(a)\bar{F}(a, \beta)}{\int \bar{F}(a, \beta)dG(a)}.$$

The auxiliary information can be summarized as

$$E\left[\frac{A - \mu}{\bar{F}(A, \beta)}\right] = 0.$$

We can construct the empirical likelihood based on truncation data a_i 's.

Exercise Construct the empirical likelihood by using the auxiliary information on the mean truncation time and derive its large sample properties. Qin (1992 thesis) gave the details.

An idea similar to the hybrid empirical likelihood method was also used by Robert (2013) to construct confidence intervals for the extremal index. They combined an empirical likelihood for the cluster size observations and a parametric likelihood for the inter-cluster time observations.

8.5 Combine Summarized Information: A More Flexible Method in Meta Analysis

Meta analysis is a systematic way to combine published information. This method has become very popular since little extra cost is needed. The main restriction in meta analysis is that all studies must include the same variables in the analyses. The only allowed difference is the sample sizes. We have to discard some studies if they contain variables different from others. Summarized information is available from published results, such as census reports, national health studies, etc. Due to confidentiality or other reasons, we typically cannot gain access to the original data except for the summarized reports. Suppose we are interested in conducting a new study that may contain some new variables of interest that are not available from the summarized information, for example, in genetic studies, some new bio-markers and genes are newly discovered. Below we discuss a more flexible method to combine published information and individual study data for enhanced inference. Chatterjee et al. (2016) discussed a related problem on the utilization of auxiliary information. As Han and Lawless (2016) pointed out, however, their methodology and theoretical results were already developed by Qin (2000) and Imbens and Lancaster (1994) in the absence of selection bias sampling case. Under a case control sampling setup, we will further illustrate in Chap. 14 that their theoretical results are also covered in Qin et al. (2015) using the density ratio relationship between control population

and general population. Next we will discuss the conventional i.i.d sample case. In Chap. 14 we will present Qin et al. (2015)'s results on the over-identified parameter problems based on case-control sampling data.

We consider two cases. (1) Sample size for the summarized information is much larger than the sample size in the new study. (2) Sample sizes from the two data sources are comparable. In case 1, we can treat the summarized information as known, i.e., the variation in the summarized data is negligible compared to the variation in the new study. In case 2, we have to take the variation in the summarized information into consideration since it is comparable to the variation in the new study.

Suppose the summarized results are based on the statistical analysis from response Y and covariate variables X (though the original data are not available), and in the new study, in addition to Y , X , an extra covariate Z is included. We are interested in fitting a parametric model

$$f(y|x, z) = f(y|x, z, \beta).$$

Denote historical data as $(y_1^*, x_1^*), \dots, (y_N^*, x_N^*)$ even they are not available. The published information can be summarized as either

(I) \bar{h} is known, where

$$\bar{h} = N^{-1} \sum_{i=1}^N h(y_i^*, x_i^*),$$

or

(II) γ^* is the solution of the estimating equation

$$\sum_{i=1}^N h(y_i^*, x_i^*, \gamma) = 0,$$

where $h = h(y, x, \gamma)$ is a given function up to an unknown parameter γ .

Let $(y_1, x_1, z_1), \dots, (y_n, x_n, z_n)$ be observed data in the new study. The basic assumption is that (y_i, x_i) , $i = 1, 2, \dots, n$ and (y_i^*, x_i^*) have the same distribution. To utilize the summarized information \bar{h} in (I), we can define estimating functions

$$g = (g_1, g_3), \quad g_1 = \frac{\partial \log f(y|x, z, \beta)}{\partial \beta}, \quad g_3(y, x) = h(y, x) - \bar{h},$$

or

$$g = (g_1, g_3), \quad g_1 = \frac{\partial \log f(y|x, z, \beta)}{\partial \beta}, \quad g_3(y, x) = h(y, x, \gamma^*)$$

in (II). We only consider the situation that $n/N \rightarrow 0$. In other words, the variation in the auxiliary information is negligible. When this is not the case, we leave this as an exercise to the readers.

The empirical likelihood approach amounts to maximizing $\sum_{i=1}^n \log p_i$ subject to the constraint

$$\sum_{i=1}^n p_i g(y_i, x_i) = 0, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i = 1.$$

Based on Theorem 8.2 the asymptotic variance from the empirical likelihood method with estimating equations g is given by

$$[E(\partial g / \partial \beta) E(gg^T)^{-1} E(\partial g / \partial \beta^T)]^{-1}.$$

Denote

$$A = E(gg^T) = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix}, \quad A_{22.1} = A_{11} - A_{12}^T A_{11}^{-1} A_{12}.$$

Equivalently the asymptotic variance can be written as

$$[E(\partial g_1 / \partial \beta) A_{11}^{-1} E(\partial g_1 / \partial \beta^T) + E(\partial g_1 / \partial \beta) A_{11}^{-1} A_{12} A_{22.1}^{-1} A_{21} A_{11}^{-1} E(\partial g_1 / \partial \beta)]^{-1},$$

or

$$[J + A_{12} A_{22.1}^{-1} A_{21}]^{-1},$$

where $A_{11} = J$ is the Fisher's information matrix.

In the approach above the estimating function $g_3 = h(y, x) - \bar{h}$ does not involve the parameter β . As a result this method may not be efficient. As an alternative approach, we define

$$\psi(x, z, \beta) = E[h(Y, x)|x, z] - \bar{h} = \int h(y, x) f(y|x, z, \beta) dy - \bar{h},$$

and

$$g_2(x, z, \beta) = \psi(x, z, \beta).$$

Then

$$E[g_2(X, Z, \beta)] = 0.$$

If we combine estimating function g_2 and the log-likelihood $\sum_{i=1}^n \log f(y_i|x_i, z_i, \beta)$ or g_2 and g_1 as in last section (also Qin 2000), then the asymptotic variance of β is given by

$$[J + E(\partial \psi / \partial \beta) (E\psi\psi^T)^{-1} E(\partial \psi / \partial \beta)]^{-1}.$$

It is not clear which of the two, combining g_1, g_3 or g_1, g_2 is better if we directly compare their asymptotic variance formulae. Alternatively, we may enquire whether we should combine all three constraints $g = (g_1, g_2, g_3)$ together.

Using the results in Sect. 8.2 (also in Qin and Lawless 1994) and

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I & -B_{11}^{-1}B_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & B_{22,1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -B_{21}B_{11}^{-1} & I \end{pmatrix},$$

we find that the asymptotic variance of $\hat{\beta}$ by combining above three estimating equations is

$$[J + E(\partial\psi/\partial\beta)E^{-1}(\psi\psi^T)E(\partial\psi/\partial\beta^T) + E(\partial g_{12}/\partial\beta)B_{11}^{-1}B_{12}B_{22,1}^{-1}B_{21}B_{11}^{-1}E(\partial g_{12}/\partial\beta)]^{-1},$$

where $g_{12} = c(g_1, g_2)$, and

$$E(gg^T) = \begin{pmatrix} J & 0 \\ 0 & E(\psi\psi^T) \\ a^T & E(\psi\psi^T) \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix},$$

$$a = E\left(h^T(y, x) \frac{\partial \log f(y|x, z, \beta)}{\partial \beta}\right), \quad B_{11} = \begin{pmatrix} J & 0 \\ 0 & E(\psi\psi^T) \end{pmatrix},$$

$$B_{12} = \begin{pmatrix} a \\ E(\psi\psi^T) \end{pmatrix}, \quad B_{22,1} = B_{11} - B_{12}^T B_{11}^{-1} B_{12}.$$

We can easily show that

$$E(\partial g/\partial\beta) = (-J, E(\partial\psi/\partial\beta), 0),$$

$$E(\partial g_{12}/\partial\beta) = (-J, a).$$

Immediately, we have

$$E(\partial g_{12}/\partial\beta)B_{11}^{-1}B_{12} = (-J, a) \begin{pmatrix} J^{-1} & 0 \\ 0 & E^{-1}(\psi\psi^T) \end{pmatrix} \begin{pmatrix} a \\ E(\psi\psi^T) \end{pmatrix} = 0.$$

In other words the asymptotic variance of $\hat{\beta}$ by combining three estimating functions g_1, g_2, g_3 is the same as that of combining two estimating functions g_1, g_2 . As a result, this shows that adding g_3 does not provide any extra information about β if g_1 and g_2 are already used. The approach by combining g_2 and parametric likelihood $\prod_{i=1}^n f(y_i|x_i, z_i, \beta)$ is better than that by combining g_1 and g_3 .

Chapter 9

Kullback–Leibler Likelihood and Entropy Family

Besides empirical likelihood, the Kullback–Leibler likelihood is another popular method to calibrate auxiliary information. The entropy family has also been used extensively in information theory. We mainly focus on discussions for continuous random variable cases. The discrete cases can be treated similarly. More comprehensive discussions can be found in the book “Information Theory and Statistics” by Kullback (1959).

9.1 Minimize Kullback–Leibler Divergence Subject to Moment Constraints

Suppose $f_0(x)$ is a probability density. We know that another probability density $f(x)$ has mean $E_f[T(x)] = \theta$. It is well known that moment constraints cannot determine a density uniquely. We are interested in finding a probability density $f(x)$ such that

$$\min_f I(f, f_0) = \min_f \int f(x) \log \frac{f(x)}{f_0(x)} dx$$

subject to the constraints

$$\int f(x) dx = 1, \quad \int T(x)f(x) dx = \theta.$$

In other words, we are seeking a probability density f that satisfies the moment constraint and is as close to f_0 as possible in the sense of the Kullback–Leibler divergence. If $\int T(x)f_0(x) dx = \theta$, then $f = f_0$ is the trivial solution. Otherwise $f \neq f_0$.

By introducing Lagrange multipliers λ and ν , equivalently we need to minimize

$$\int \left(f(x) \log \frac{f(x)}{f_0(x)} + \lambda T(x)f(x) + \nu f(x) \right) dx.$$

Denote $g(x) = f(x)/f_0(x)$, then we need to minimize

$$\int \{g(x) \log g(x) + \lambda T(x)g(x) + \nu g(x)\} dF_0(x).$$

Write

$$\phi(t) = t \log t + \lambda Tt + \nu t, \quad t_0 = \exp(-\lambda T - \nu - 1).$$

For $t, t_0 > 0$, expanding $\phi(t)$ at t_0 gives

$$\phi(t) = \phi(t_0) + (t - t_0)\phi'(t_0) + 0.5(t - t_0)^2\phi''(t^*),$$

where t^* lies between t and t_0 . Moreover

$$\phi(t_0) = -t_0, \quad \phi'(t_0) = 0, \quad \phi''(t^*) = 1/t^* > 0,$$

$$\begin{aligned} \int \phi(g(x)) dF_0(x) &= - \int \exp\{-\lambda T(x) - \nu - 1\} dF_0(x) \\ &\quad + 0.5 \int [g(x) - \exp\{-\lambda T(x) - \nu - 1\}]^2 / h(x) dF_0(x), \end{aligned}$$

where $h(x)$ lies between $g(x)$ and $\exp\{-\lambda T(x) - \nu - 1\}$.

$$\int \phi(g(x)) dF_0(x) \geq - \int \exp\{-\lambda T(x) - \nu - 1\} dF_0(x)$$

with equality if and only if

$$g(x) = \exp\{-\lambda T(x) - \nu - 1\},$$

or

$$f(x) = \frac{\exp(\lambda T(x))f_0(x)}{\int \exp(\lambda T(x))f_0(x)dx}, \quad (9.1.1)$$

where λ is determined by

$$\frac{\int T(x) \exp(\lambda T(x))f_0(x)dx}{\int \exp(\lambda T(x))f_0(x)dx} = \theta$$

or

$$\int \{T(x) - \theta\} \exp(\lambda T(x)) f_0(x) dx = 0.$$

In statistical literature $f(x)$ is called the entropy family solution.

To make this problem easier to understand, in the following, we present a simple alternative proof that the entropy family is the constrained optimal solution. The result is given in Kagan et al. (1973).

Let f be the probability density given in (9.1.1). Denote

$$f(x)/f_0(x) = \frac{\exp(\lambda T(x))}{\int \exp(\lambda T(x)) f_0(x) dx} =: \exp(\lambda T(x) - \log c).$$

Let h be another probability density satisfying the constraint $E_h\{T(X)\} = \theta$. Using the non-negative property of Kullback–Leibler divergence,

$$\begin{aligned} H &= - \int h(x) \log\{h(x)/f_0(x)\} dx \leq - \int h(x) \log\{f(x)/f_0(x)\} dx \\ &= - \int h(x)\{\lambda T(x) - \log c\} dx = - \int f(x)\{\lambda T(x) - \log c\} dx \\ &= - \int f(x) \log\{f(x)/f_0(x)\} dx \end{aligned}$$

since both $h(x)$ and $f(x)$ satisfy the moment constraint.

Maxent Density

The maxent density is typically obtained by maximizing Shannons entropy (the Kullback–Leibler divergence relative to a uniform measure),

$$\max_f [- \int f(x) \log f(x) dx]$$

subject to the moment constraints $\int x^i f(x) da = \theta_i, i = 0, 1, 2, \dots, k$. Use the same argument as before, the optimal solution is

$$f^*(x) = \exp\left(- \sum_{i=0}^k \lambda_i x^i\right), \quad (9.1.2)$$

where λ_i 's are the Lagrange multipliers determined by

$$\int x^i \exp\left(- \sum_{i=0}^k \lambda_i x^i\right) dx = \theta_i, \quad \theta_0 = 1.$$

Most of the well-known distributions may be characterized as maxent densities subject to certain moment constraints. For example, the normal distribution is a

maxent density with characterizing moments x and x^2 , the gamma distribution is characterized by x and $\log x$ for $x > 0$, and the beta distribution by $\log(x)$ and $\log(1 - x)$ for $0 < x < 1$. More comprehensive discussions can be found in Golan et al. (1996).

Applications

Inference under moment constraints has been discussed, among others, by Haberman (1984) and Sheehy (1988). Efron (1981) considered an exponential tilting bootstrap method.

Suppose $X_1, \dots, X_n \sim i.i.d. f_0(x)$, where it is known that $E_{f_0}[T(X)] = \theta$. Since in general $dF_0(x) = f_0(x)dx$ is unknown, a natural method is to replace dF_0 by dF_n , the empirical probability measure. Unfortunately, the constraint $\int T(x)dF_n(x) = n^{-1} \sum_{i=1}^n T(x_i) = \theta$ may not be satisfied. We need to find a discrete probability measure \hat{F} such that $\int T(x)d\hat{F}(x) = \theta$ and also is as close to dF_n as possible. Using the same argument as above, we know the solution is

$$dF(x_i) = \frac{\exp\{\lambda T(x_i)\}dF_n(x_i)}{\sum_{j=1}^n \exp(\lambda T(x_j))}, \quad i = 1, 2, \dots, n,$$

where the Lagrange multiplier is determined by

$$\sum_{i=1}^n \{T(x_i) - \theta\} \exp\{\lambda T(x_i)\} = 0.$$

Let X_1, \dots, X_n be the observed data. Denote X_1^*, \dots, X_n^* as the bootstrap sample. It seems reasonable to impose the sample mean constraint when we construct empirical likelihood or Kullback–Leibler likelihood based on bootstrap samples. More specifically we can maximize

$$\sum_{i=1}^n \log p_i^*, \quad p_i^* = dF(x_i^*), \quad i = 1, 2, \dots, n,$$

or minimize

$$-\sum_{i=1}^n p_i^* \log p_i^*$$

subject to the constraints

$$p_i^* \geq 0, \quad \sum_{i=1}^n p_i^* = 1, \quad \sum_{i=1}^n p_i^*(x_i^* - \bar{x}) = 0.$$

The above result for the inference on mean can be generalized immediately to the general estimating equation case. Denote the observed i.i.d. data as

$$X_1, \dots, X_n \sim f(x).$$

If we have prior information that $E[g(X, \theta)] = 0$, where g are $r \times 1$ estimating functions and θ are $p \times 1$ unknown parameters ($r \geq p$). We can use entropy family to incorporate the prior information.

Replacing $T(x)$ by $g(x, \theta)$, we have

$$d\hat{F}(x_i) = \frac{\exp\{\lambda^T g(x_i, \theta)\} dF_n(x_i)}{\sum_{j=1}^n \exp\{\lambda^T g(x_j, \theta)\}}, \quad i = 1, 2, \dots, n,$$

where λ satisfies

$$\int g(x, \theta) d\hat{F}(x) = 0.$$

or

$$\sum_{i=1}^n g(x_i, \theta) \exp\{\lambda^T g(x_i, \theta)\} = 0.$$

We call $d\hat{F}(x)$ an entropy family.

As an alternative, in discrete version, we may directly minimize the Kullback–Leibler divergence $-\sum_{i=1}^n p_i \log p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0 \quad \sum_{i=1}^n p_i g(x_i, \theta) = 0.$$

After introducing the Lagrange multiplier λ , we can easily show that

$$p_i = \frac{\exp\{\lambda^T g(x_i, \theta)\}}{\sum_{j=1}^n \exp\{\lambda^T g(x_j, \theta)\}}.$$

Schennach (2005, 2007) interpreted the function $-\sum_{i=1}^n p_i \log(np_i)$ with p_i replaced by \hat{p}_i 's as a valid likelihood for Bayesian inference. The profile Kullback–Leibler divergence is

$$\ell(\theta) = \sum_{i=1}^n \lambda^T g(x_i, \theta) - n \log \left[\sum_{j=1}^n \exp\{\lambda^T g(x_j, \theta)\} \right].$$

Inference for θ can be based on $\ell(\theta)$. Noting that λ is an implicit function of θ , we can derive the score function for θ as

$$\begin{aligned}\frac{\partial \ell}{\partial \theta} &= \sum_{i=1}^n \left[\frac{\partial \lambda^T}{\partial \theta} g(x_i, \theta) + \lambda^T \frac{\partial g(x_i, \theta)}{\partial \theta} \right] \\ &\quad - n \frac{\sum_{j=1}^n \exp\{\lambda^T g(x_j, \theta)\} \{\partial \lambda^T / \partial \theta g(x_i, \theta) + \lambda^T \partial g(x_j, \theta) / \partial \theta\}}{\sum_{j=1}^n \exp\{\lambda^T g(x_j, \theta)\}} = 0.\end{aligned}$$

Exercise Derive the large sample results for the minimum Kullback–Leibler divergence estimate.

9.2 Entropy Family in the Presence of Covariates

Let $f(y)$ be a specified baseline density and X be covariates. Suppose the conditional density $f(y|x)$ satisfies

$$E_{f(y|x)}[Y - \mu(x\beta)] = 0,$$

where $\mu(\cdot)$ is a specified function. For given X , we would like to minimize the Kullback–Leibler divergence such that $f(y|x)$ satisfies the mean constraint and is as close to $f(y)$ as possible. The solution is the entropy family given by

$$f(y|x) = \frac{\exp\{\tau(y - \mu(x\beta))\}f(y)}{\int \exp\{\tau(y - \mu(x\beta))\}f(y)dy} = \frac{\exp(\tau y)f(y)}{\int \exp(\tau y)f(y)dy}, \quad (9.2.3)$$

where $\tau = \tau(x, \beta)$ is determined by

$$\int \{y - \mu(x\beta)\} \exp(\tau y) f(y) dy = 0.$$

Example 1 Let $f(y)$ be the standard normal density. Then its moment generating function is

$$\int \exp(\tau y) f(y) dy = \exp(\tau^2/2).$$

Differentiating both sides with respect to τ , we have

$$\int y \exp(\tau y) f(y) dy = \tau \exp(\tau^2/2).$$

Since $E(Y|x) = \mu(x\beta)$, then we have the following equation for τ

$$\tau \exp(\tau^2/2) - \mu(x\beta) \exp(\tau^2/2) = 0.$$

Therefore $\tau = \mu(x\beta)$. The entropy family is given by

$$f(y|x) = \exp\{\mu(x\beta)y - \mu^2(x\beta)/2\}f(y).$$

In other words, the entropy family is a normal with mean $\mu(x\beta)$ and unit variance. In terms of estimating β , we can pretend the underlying distribution to be normal.

Example 2 Let $f(y)$ be the standard exponential density. It can be shown that the entropy family with moment constraint $E(Y|x) = \mu(x\beta)$ is

$$f(y|x) = \mu^{-1}(x\beta) \exp\{1 - y/\mu(x\beta)\}f(y) = \mu^{-1}(x\beta) \exp\{-y/\mu(x\beta)\}, \quad y > 0.$$

Exercise 1 To estimate β , we may use the normal likelihood in Example 1 or the exponential likelihood in Example 2. Find the maximum likelihood estimates based on the two different likelihoods and compare their asymptotic efficiencies.

Exercise 2 Let the baseline density be the standard normal. Find the entropy family with moment constraints $E(Y|x) = \mu(x\beta)$ and $\text{Var}(y|x) = \sigma^2(x\beta)$.

In the mean restricted model $E(Y|x) = \mu(x\beta)$, different choices of the baseline “carrier density” $f(y)$ may produce different entropy families. All of them produce consistent estimate of β . However, the true “carrier density” corresponds to the true likelihood, which gives the most efficient estimate. A good choice of $f(y)$ is very important in practical applications.

Copas and Eguchi’s Envelop Likelihood

Copas and Eguchi (2005) considered the so-called “envelop likelihood”. Suppose $X \sim f(x, \theta)$ is a working model, but the true model is

$$g(x) = \exp\{\epsilon u(x, \theta) - b(\theta)\}f(x, \theta),$$

where $E_g[u(x, \theta)] = 0$ and was assumed that $\epsilon = O_p(n^{-1/2})$. This can be thought of as departing from the true one locally in the direction $u(x, \theta)$. Essentially this is similar to the entropy family approach. Some robustness properties were discussed in their paper.

9.3 Some Miscellaneous Results

1. Power mixture family

Let $f_0(x)$ and $f_1(x)$ be two different density functions. Define

$$T(x) = \log[f_1(x)/f_0(x)].$$

We are seeking a density f that minimizes

$$I(f, f_0) = \int f(x) \log \frac{f(x)}{f_0(x)} dx$$

subject to the constraint

$$\theta = \int T(x)f(x)dx = \int f(x) \log \frac{f_1(x)}{f_0(x)} dx.$$

Kullback (1959) found the entropy solution as

$$\min_{\tau} I(f, f_0) = \theta\tau - \log M(\tau),$$

where

$$M(\tau) = \int f_0(x) \exp \left\{ \tau \log \frac{f_1(x)}{f_0(x)} \right\} dx,$$

and

$$\theta = \frac{d}{d\tau} \log M(\tau) = \frac{\int f_0(x) \log \{f_1(x)/f_0(x)\} \exp \{ \tau \log \frac{f_1(x)}{f_0(x)} \} dx}{M(\tau)}.$$

Finally

$$f(x) = f_0(x)M^{-1}(\tau) \exp \{ \tau \log (f_1(x)/f_0(x)) \} = \frac{f_1^\tau(x)f_0^{1-\tau}(x)}{\int f_1^\tau(x)f_0^{1-\tau}(x)dx}.$$

The value of τ is determined by θ . For different values of θ , it introduces a power mixture density of $f_1^\tau(\cdot)$ and $f_0^{1-\tau}(\cdot)$. Note that $\int f_1^\tau(x)f_0^{1-\tau}(x)dx \neq 1$ in general. This is in contrast to the conventional mixture model $\lambda f_0(x) + (1 - \lambda)f_1(x)$, where the normalizing constant is not needed.

The model comprehensively discussed by Atkinson (1970) and Cox (1961, 1962) for discriminating between two non-nested models is exactly based on the power mixture model. Interestingly we identify its connection with the entropy family here.

2. Shannon's mutual information

Suppose two random variables X and Y have joint and marginal densities $f(x, y), f_1(x), f_2(y)$, respectively. The Shannon's mutual information is defined as

$$I(X, Y) = E_f[\log \{f(X, Y)/f_1(X)f_2(Y)\}] = E[\log f(X|Y)/f_2(Y)].$$

By using iterated expectation,

$$\begin{aligned} I(X, Y) &= E \left[-E \left\{ \log \frac{f_2(Y)}{f(X|Y)} \right\} | Y = y \right] \\ &\geq E \left[-\log E \left\{ \frac{f_2(Y)}{f(X|Y)} | Y = y \right\} \right] \\ &= E[\log(1)] = 0, \end{aligned}$$

where the convexity inequality has been used.

Applying a two-term Taylor's expansion gives

$$\log \left\{ \frac{f(x, y)}{f_1(x)f_2(y)} - 1 + 1 \right\} = \frac{f(x, y)}{f_1(x)f_2(y)} - 1 + 0.5 \left\{ \frac{f(x, y)}{f_1(x)f_2(y)} - 1 \right\}^2 + \dots$$

$$\begin{aligned} & \int \left\{ \frac{f(x, y)}{f_1(x)f_2(y)} - 1 \right\} f_1(x)f_2(y) dx dy + 0.5 \int \left\{ \frac{f(x, y)}{f_1(x)f_2(y)} - 1 \right\}^2 f_1(x)f_2(y) dx dy \\ &= 0.5 \Phi_{XY}^2 \geq 0, \end{aligned}$$

where

$$\Phi_{XY}^2 = \int \left\{ \frac{f(x, y)}{f_1(x)f_2(y)} - 1 \right\}^2 f_1(x)f_2(y) dx dy = \int \frac{f^2(x, y)}{f_1(x)f_2(y)} dx dy - 1.$$

Φ_{XY}^2 equals 0 if and only if X and Y are independent.

Observe that in the definition of $I(X, Y)$, the expectation is with respect to the joint density $f(x, y)$, while in the definition of $\Phi_{X,Y}^2$, the expectation is with respect to the two marginal densities.

The conditional Shannon's mutual information is defined by

$$I(Y, X|Z) = E \left\{ \frac{f(Y, X|Z)}{f(Y|Z)f(X|Z)} | Z \right\}.$$

Note that $I(Y, X|Z) = 0$ if and only if Y and X are conditionally independent for given Z . Similarly we have

$$\Phi_{X,Y|Z}^2 = \int \frac{f^2(x, y|z)}{f_1(x|z)f_2(y|z)} dx dy - 1.$$

3. Partial association measures of conditional independence

Daudin (1980) discussed the following results.

Let X_1, X_2, X_3 be three random variables. X_2 and X_3 are called conditionally independent for given X_1 if

$$E[g(X_1, X_2)h(X_1, X_3)|X_1] = E[g(X_1, X_2)|X_1]E[h(X_1, X_3)|X_1]$$

for any square integrable functions $g(X_1, X_3)$ and $h(X_1, X_3)$.

Define

$$\mathcal{E}_1 = \{g(X_1, X_2), E(g|x_1) = 0, Eg^2 < \infty\}, \quad \mathcal{E}_2 = \{g(X_1, X_3), E(h|x_1) = 0, Eh^2 < \infty\}.$$

Daudin (1980) showed that X_2 and X_3 are conditionally independent for given X_1 if and only if for any $g \in \mathcal{E}_1$ and $h \in \mathcal{E}_2$,

$$E[g(X_1, X_2)h(X_1, X_3)] = 0.$$

A weaker condition for conditional independence is

$$E[\{r(X_2) - E(r(X_2)|X_1)\}\{s(X_3) - E(s(X_3)|X_1)\}] = 0,$$

where $r(\cdot)$ and $s(\cdot)$ are any given functions. This can be used to test conditional independence between X_2 and X_3 for given X_1 .

9.4 Entropy Family with Fixed Margins in Discrete Case

Ireland and Kullback (1968a,b) considered an estimation problem in contingency tables with fixed margins. Suppose the observations $n_{ij} > 0$ on the interior of an $I \times J$ contingency table with unknown cell probability p_{ij} . Suppose the two margins are known, i.e.,

$$p_{i+} = \sum_{j=1}^J p_{ij}, \quad i = 1, 2, \dots, I,$$

$$p_{+j} = \sum_{i=1}^I p_{ij}, \quad j = 1, 2, \dots, J.$$

Given a contingency table $\pi_{ij}, i = 1, 2, \dots, I; j = 1, 2, \dots, J$ with $\pi_{ij} > 0$ and $\sum_{i=1}^I \sum_{j=1}^J \pi_{ij} = 1$, we can consider all contingency tables $\{p_{ij}\}$ of the same dimension such that the marginal probability p_{i+} and p_{+j} are given and fixed. Define

$$I(p; \pi) = \sum_{ij} p_{ij} \log\{p_{ij}/\pi_{ij}\}. \quad (9.4.4)$$

Using the Lagrange multiplier method, one can show that the minimum is attained at

$$p_{ij}^* = a_i b_j \pi_{ij}, \quad i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J,$$

where a_i, b_j are determined by the marginal constraints. The minimum value of $I(p; \pi)$ is

$$\begin{aligned} I(p^*, \pi) &= \sum_{ij} p_{ij}^* (\log a_i + \log b_j + \log \pi_{ij} - \log \pi_{ij}) \\ &= \sum_i p_{i+}^* \log a_i + \sum_j p_{+j}^* \log b_j \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^I p_{i+} \log a_i + \sum_{j=1}^J p_{+j} \log b_j \\
&= \sum_{i=1}^I p_{i+} \log \{p_{i+} / \sum_j b_j \pi_{ij}\} + \sum_{j=1}^J p_{+j} \log \{p_{+j} / \sum_i a_i \pi_{ij}\}.
\end{aligned}$$

Exercise Show the Pythagorean identity

$$I(p, \pi) = I(p^*, \pi) + I(p, p^*) \quad (9.4.5)$$

for the optimal solution p^* , where both p^* and p have common specified margins.

Exercise Generalize the two-way contingency table to a three-way contingency table with fixed margins.

A Classical Example

Fisher considered estimation of the linkage from the progeny of self-fertilized heterozygotes for two factors in maize: Starchy versus Sugary and Green versus White base leaf. The seedling counts and the probability parameter in terms of θ are given in Tables 9.1 and 9.2 below, respectively.

In this example we take π as the vector empirical frequencies,

$$\hat{p} = 1997/3839, 906/3839, 904/3836, 32/3839.$$

Let \hat{p}^* be the optimal solution for (9.4.4) under the marginal constraints. We can show that

$$\begin{aligned}
2nI(\hat{p}^*, \hat{p}) &= 2(3839)[0.5089 \log \{0.5089/(1997/3839)\} + \dots + 0.0089 \log \{0.0089/(32/3839)\}] \\
&= 2.039,
\end{aligned}$$

which has an asymptotic chi-square with two degrees of freedom.

Table 9.1 Fisher's genetic data

Type	Green	White	Total
Starchy	1997	906	2903
Sugary	904	32	936
Total	2901	938	3839

Table 9.2 Joint probabilities

-	Green	White	Total
Starchy	0.25(2 + θ)	0.25(1 - θ)	0.75
Sugary	0.25(1 - θ)	0.25 θ	0.25
Total	0.75	0.25	1

Alternatively, we can minimize

$$M = 0.25(2 + \theta) \log\{0.25(2 + \theta)/\hat{p}_{11}\} + 0.25(1 - \theta) \log\{0.25(1 - \theta)/\hat{p}_{12}\} \\ + 0.25(1 - \theta) \log\{0.25(1 - \theta)/\hat{p}_{21}\} + 0.25\theta \log(0.25\theta/\hat{p}_{22}),$$

to arrive at

$$\frac{(2 + \theta)\theta}{n_{11}n_{22}} = \frac{(1 - \theta)^2}{n_{12}n_{21}}.$$

This exactly matches the maximum likelihood estimating equation.

Ireland and Kullback (1968a,b) suggested the following iterative procedure for finding the optimal solution to (9.4.4):

For $n = 1, 2, \dots$, define

$$p_{ij}^{(2n-1)} = \frac{p_{i+}}{p_{i+}^{(2n-2)}} p_{ij}^{(2n-2)}, \quad p_{ij}^{(2n)} = \frac{p_{+j}}{p_{+j}^{(2n-1)}} p_{ij}^{(2n-1)}, \quad i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J,$$

where $p_{ij}^0 = \pi_{ij}$.

An excellent related work for contingency tables with known margins when the target and sampled population differ can be found in Little and Wu (1991).

Application of Entropy Family in Lottery Problem

Stern and Cover (1989) found another application of entropy family in lottery problems.

In a Canadian lottery game, the “Lotto 6/49”, players select 6 numbers from the first 49 integers without replacement on \$1 ticket to participate. Six winning numbers and a bonus number are selected at random by the Lottery Commission. In 161 games through July 6, 1985, Stern and Cover (1989) carried out an uniformity test for the winning number and found no violation of the uniformity. However, it has been noticed that players in a random lottery game with parimutuel prizes win larger prizes if they choose numbers that are not popular with other players. Stern and Cover (1989) found that number 7 appeared more than 750, 000 times than its mean under the uniformity assumption. They rejected the uniformity assumption by using a chi-squared test. In this game, the marginal probability of a number being selected is available.

In general a player selects m numbers out of M ($M > m$) without replacement. There are $N = M!/\{m!(M-m)!\}$ different choices. This number becomes exponentially large when M and m get large. In Lotto 6/49, $M = 49$, $m = 6$. Let $\mathbf{T}_j, j = 1, 2, \dots, n$ be n independent tickets from unknown distribution $P(\mathbf{t})$, where \mathbf{T}_j is a random m -tuple chosen from the first M integers. All possible outcomes for a ticket T can be listed as $\mathbf{T} = \mathbf{t}_1, \dots, \mathbf{t}_N$; each \mathbf{t}_i contains m numbers. We can treat $P(\mathbf{t})$ as a multinomial distribution with N categories. Since the marginal probability $P(i \in \mathbf{T}) = \pi_i, i = 1, 2, \dots, M$ is known, we are interested in finding a probability $P(\mathbf{t})$ that satisfies the marginal moment constraints and is as close to the uniform probability as possible.

Csiszar (1984) and Van Campenhout and Cover (1981) established the following result:

If the unconditional distribution on tickets is uniform, i.e., $P(\mathbf{T} = \mathbf{t}_i) = m!(M - m)!/M!$, $i = 1, 2, \dots, N$, then as $n \rightarrow \infty$,

$$P\{\mathbf{T}_1 = \mathbf{t}|n^{-1} \sum_{j=1}^n I(i \in \mathbf{T}_j) = \pi_i, i = 1, 2, \dots, M\} \rightarrow P^*(\mathbf{t}),$$

where $P^*(\mathbf{t})$ maximizes the entropy $H(P) = -\sum_{\mathbf{t}} P(\mathbf{t}) \log P(\mathbf{t})$ over the probability mass function $P(\mathbf{t})$ under the marginal constraints

$$\sum_{\mathbf{t}} P(\mathbf{t}) I(i \in \mathbf{t}) = \pi_i, \quad i = 1, 2, \dots, M.$$

As shown before, the solution $P^*(\mathbf{t})$ is given by the entropy family through

$$P^*(\mathbf{t}) = \exp\{\lambda_0 + \sum_{j=1}^M \lambda_j I(j \in \mathbf{t})\} = c \prod_{j=1}^M \theta_j^{I(j \in \mathbf{t})}.$$

More discussions can be found in Stern and Cover (1989).

2. Entropy family with fixed marginal density in continuous case.

Let $\pi(x, y)$ be a bivariate density. We are interested in finding a bivariate density $f(x, y)$ that has fixed marginal densities $f_X(x) = g(x)$ and $f_Y(y) = h(y)$ and minimizes

$$I(f, \pi) = \int \int f(x, y) \log\{f(x, y)/\pi(x, y)\} dx dy.$$

Kullback (1968) defined

$$T(x) = \delta(x - t), \quad T(y) = \delta(y - t),$$

where δ is the Dirac delta-function. Let

$$f^*(x, y) = a(x)b(y)\pi(x, y),$$

where $a(x)$ and $b(y)$ are functions to be determined by

$$g(x) = a(x) \int b(y)\pi(x, y) dx dy, \quad h(y) = b(y) \int a(x)\pi(x, y) dx, \quad \int \int a(x)b(y)\pi(x, y) dx dy = 1.$$

The following iterative procedure works well in practice to obtain the two functions

$$b_0 = 1, \quad a_0 = g(x),$$

$$b_1(y) = \frac{h(y)}{\int \pi(x, y)a_0(x)dx}, \quad a_1(x) = \frac{g(x)}{\int \pi(x, y)b_1(y)dy},$$

and in general

$$b_{k+1}(y) = \frac{h(y)}{\int \pi(x, y)a_k(x)dx}, \quad a_{k+1}(x) = \frac{g(x)}{\int \pi(x, y)b_{k+1}(y)dy}.$$

Ruschendorf (1995) showed the proposed iterative algorithm converges.

Bickel et al. (1991) proposed efficient estimators of linear functionals of a probability measure with known margins.

Iterative Scaling

Next we briefly discuss the convergence aspects of the iterative algorithm above. The general theory was given by Csiszar and Tusnády (1984).

For any two probability measures P and Q , it is known

$$I(P, Q) = \int f_P(x) \log\{f_P(x)/f_Q(x)\}dx \geq 0,$$

where f_P and f_Q are the corresponding probability densities. If P satisfies some moment constraints, the maximum entropy solution subject to the same constraint is denoted as P^* . We can easily show the Pythagorean identity (Csiszar and Tusnády 1984)

$$I(P, Q) = I(P, P^*) + I(P^*, Q).$$

Denote Π_1 and Π_2 as the set of a (two dimensional) probability distribution with the fixed first marginal distribution P_{i+} or second marginal distribution P_{+j} , respectively. According to k is odd or even let $P^k \in \Pi_1$ or $P \in \Pi_2$ be the projection of P , i.e., the optimal solutions under marginal constraints. Then

$$I(P, P^{k-1}) = I(P, P^k) + I(P^k, P^{k-1}),$$

for each $P \in \Pi_1$ (k odd) or $P \in \Pi_2$ (k even). We need to show that P^k converges. It suffices to show for any subsequence $P^{n_i} \rightarrow P^0$, say, we have $P^0 \in \Pi_1, \Pi_2$ and for each $P \in \Pi_1, \Pi_2$,

$$I(P, Q) = I(P, P^0) + I(P^0, Q).$$

Recursively using the Pythagorean identity gives

$$I(P, Q) = I(P, P^0) + \sum_{k=1}^{\infty} I(P^k, P^{k-1}).$$

Then $I(P, Q) < \infty$ implies $I(P^k, P^{k-1}) \rightarrow 0$. Consequently $\lim P^{n_i+1} = \lim P_{n_i} = P^0$. Therefore $P^0 \in \Pi_1, \Pi_2$.

$$I(P^0, Q) = I(P^0, P^0) + \sum_{k=1}^{\infty} I(P^k, P^{k-1}) = \sum_{k=1}^{\infty} I(P^k, P^{k-1}).$$

Along the same line Csiszar and Tusnády (1984) demonstrated the convergence of the EM algorithm discussed in Chap. 4 (Dempster et al. 1977). It is equivalent to an alternating minimization of $I(P, Q)$ for $P \in \Pi_1$ and $Q \in \Pi_2$ with suitably constructed Π_1 and Π_2 . Interested readers may go through their works.

9.5 Generalized Empirical Likelihoods

There are many generalizations of the empirical likelihood method by modifying the objective function. Next we discuss a few examples.

1. Euclidean distance

It is known that the nonparametric MLE of F has jumps at each of the observed data point with size $1/n$. With the auxiliary information $E[g(X, \theta)] = 0$, at each data point we hope F has a jump size not too far away from $1/n$ but also satisfies the moment constraint. This motivates us to minimize the Euclidean distance

$$\sum_{i=1}^n (p_i - 1/n)^2$$

subject to the constraint

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i g(x_i, \theta) = 0.$$

By using the Lagrange multiplier method, we have

$$p_i = \frac{1}{n} + \lambda^T (\bar{g} - g_i), \quad \bar{g} = \frac{1}{n} \sum_{i=1}^n g_i \lambda = \frac{1}{n} S^{-1} \bar{g}, \quad S = \frac{1}{n} \sum_{i=1}^n (g_i - \bar{g})(g_i - \bar{g})^T,$$

where $g_i = g(x_i, \theta)$. Therefore the empirical Euclidean log-likelihood is

$$\ell_E = -\bar{g}^T S^{-1} g.$$

Clearly this is equivalent to the GMM with covariance replaced by the sample version.

2. Kullback–Leibler divergence and Hellinger distance

As discussed before the Kullback–Leibler divergence and the Hellinger distance are respectively defined as

$$KL = \sum_{i=1}^n p_i \log(np_i),$$

and

$$H = \sum_{i=1}^n (p_i^{1/2} - n^{-1/2})^2.$$

We can minimize them subject to the moment constraints. As an exercise, readers can go through the details.

3. The Cressie-Read power divergence statistic

This statistic is originally defined for a multinomial distribution. Let Q_i and E_i be the observed number and expected number in category i , $i = 1, 2, \dots, k$. The Cressie-Read power divergence statistic is defined as

$$CR(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^k O_i \left[\left(\frac{O_i}{E_i} \right)^\lambda - 1 \right],$$

where $-\infty < \lambda < \infty$. The degenerate cases $\lambda = -1$ or 0 are treated as the limiting values. For n observations from a continuous distribution, it naturally becomes

$$CR(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^n [(np_i)^{-\lambda} - 1].$$

Corcoran (1998) and Imbens et al. (1998) focused on the Cressie-Read discrepancy statistic, for a fixed λ , as a function of two vectors p and q of dimension n .

The method of empirical likelihood can be viewed as one of allocating probabilities to an n -cell contingency table so as to minimize a goodness-of-fit criterion. Baggerly (1998) showed that when the Cressie-Read power-divergence statistic is used as the criterion, confidence regions enjoying the same convergence rates as those found for the empirical likelihood can be obtained for the entire range of values of the Cressie-Read parameter λ , including $\lambda = -1$, the maximum entropy; $\lambda = 0$, the empirical likelihood; and $\lambda = 1$, the Pearson's χ^2 . It is noted that, in the power-divergence family, the empirical likelihood is the only member that is Bartlett correctable. However, simulation results suggest that, for the mean, the simple method of using a scaled F -distribution yields more accurate coverage levels for moderate sample sizes.

Newey and Smith (2001) compared higher-order asymptotic properties between empirical likelihood (EL), generalized empirical likelihood (GEL) and generalized method of moments (GMM). They found that the asymptotic bias of EL often does not grow with the number of moment restrictions, while those of GMM and other GEL estimators grow without bound. Also Imbens and Spady (2001) found that GEL estimators of over-identified moment models are unambiguously superior to two-step

GMM. This nice property of empirical likelihood has a simple explanation. When the data are discrete, having a finite support, EL estimator is equal to the maximum likelihood estimator. Furthermore the bias correction based on EL probabilities is identical to the discrete data bias correction for the MLE. Consequently, for discrete data, EL inherits the well known higher order efficiency of the MLE. Since discrete distributions can be used to approximate moments of a continuous distribution, the efficiency of EL for the discrete case leads to more efficiency in general.

9.6 Inference for Exponential Family with Specified Mean Function

We have found the maximum entropy family via minimizing the Kullback–Leibler divergence between $f(y|x)$ and $f(y)$ subject to some moment constraints in Sect. 9.2. This section discusses the detailed inference in an exponential family with a specified mean function. It is an application of the entropy family, though Zhao et al. (1992) called this a partially specified exponential distribution family.

Let Y be a $r \times 1$ vector response variable. Suppose the conditional mean of Y given X is pre-specified by $E(Y|X) = \mu(X\beta)$, where $\mu(\cdot)$ is a $r \times 1$ known vector function and β is a $p \times 1$ unknown parameter. We are interested in finding a density $f(y|x)$ that satisfies the given expectation and minimizes the Kullback–Leibler divergence from $f(y, \lambda)$, where $f(y, \lambda)$ is a specified baseline density with an unknown parameter λ . Then the solution is given by the entropy family

$$f(y|x) = \frac{\exp(y^T \theta) f(y, \lambda)}{\int \exp(y^T \theta) f(y, \lambda) dy} =: \Delta^{-1}(\theta, \lambda) \exp(y^T \theta) f(y, \lambda), \quad (9.6.6)$$

where θ is a $r \times 1$ unknown parameter determined by the conditional mean constraint and Δ is a normalizing constant. Using the chain rule in differentiation, we derive the score equations below.

Differentiating

$$1 = \Delta^{-1} \int \exp(y^T \theta) f(y, \lambda) dy$$

with respect to θ , we can derive $\mu = \partial \log \Delta / \partial \theta$. Similarly differentiating

$$\mu = \Delta^{-1} \int y \exp(y^T \theta) f(y, \lambda) dy,$$

we have $\partial \mu / \partial \theta = \text{Cov}(y)$. Differentiating

$$1 = \Delta^{-1} \int \exp(y^T \theta) f(y, \lambda) dy$$

with respect to λ , we have

$$E\left[\frac{\partial \log f(y, \lambda)}{\partial \lambda}\right] = E\left[\frac{\partial \log \Delta}{\partial \lambda}\right].$$

Finally differentiating

$$\mu = \Delta^{-1} \int y \exp(y_k^T \theta_k) f(y, \lambda) dy$$

with respect to λ , we have

$$\frac{\partial \mu}{\partial \lambda} = \text{Cov}(y, w), \quad w = \partial \log f / \partial \lambda.$$

The log-likelihood of the k -th observation is

$$\ell_k = \ell(\theta_k, \lambda) = y_k^T \theta_k + \log f(y_k, \lambda) - \log \Delta_k(\theta_k, \lambda).$$

Note that the parameters (θ_k, λ) are transformed to $\mu_k = \mu_k(\theta_k, \lambda)$, $\mu_k = \mu_k(x_k \beta)$.

$$\begin{pmatrix} \partial \ell_k / \partial \theta_k \\ \partial \ell_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} y_k - \partial \log \Delta_k / \partial \theta_k \\ \partial \log f(y_k) / \partial \lambda - \partial \log \Delta_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} y_k - \mu_k \\ w_k - \eta_k \end{pmatrix},$$

where

$$\eta_k = E(w_k) = \partial \log \Delta_k / \partial \lambda.$$

By using chain rule,

$$\begin{pmatrix} \partial \ell_k / \partial \theta_k \\ \partial \ell_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} \partial \ell_k / \partial \mu_k \partial \mu_k / \partial \theta_k + \partial \ell_k / \partial \lambda \partial \lambda / \partial \theta_k \\ \partial \ell_k / \partial \mu_k \partial \mu_k / \partial \lambda + \partial \ell_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} \Sigma_k & 0 \\ B_k & I \end{pmatrix} \begin{pmatrix} \partial \ell_k / \partial \mu_k \\ \partial \ell_k / \partial \lambda \end{pmatrix},$$

where $\Sigma_k = \text{Cov}(y_k) = \partial \mu_k / \partial \beta$ and $B_k = \partial \mu_k / \partial \lambda = \text{Cov}(w_k, y_k)$. Therefore

$$\begin{pmatrix} \partial \ell_k / \partial \mu_k \\ \partial \ell_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} \Sigma_k & 0 \\ B_k & I \end{pmatrix}^{-1} \begin{pmatrix} y_k - \mu_k \\ w_k - \eta_k \end{pmatrix},$$

$$\begin{aligned} \begin{pmatrix} \partial \ell_k / \partial \beta \\ \partial \ell_k / \partial \lambda \end{pmatrix} &= \begin{pmatrix} \partial \ell_k / \partial \mu_k \partial \mu_k / \partial \beta \\ \partial \ell_k / \partial \lambda \end{pmatrix} = \begin{pmatrix} \partial \mu_k / \partial \beta & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \partial \ell_k / \partial \mu_k \\ \partial \ell_k / \partial \lambda \end{pmatrix} \\ &= \begin{pmatrix} D_k & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_k & 0 \\ B_k & I \end{pmatrix}^{-1} \begin{pmatrix} y_k - \mu_k \\ w_k - \eta_k \end{pmatrix}, \end{aligned}$$

where $D_k = \partial \mu_k / \partial \beta$. Using the fact that

$$\begin{pmatrix} \Sigma_k & 0 \\ B_k & I \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_k^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -B_k \Sigma^{-1} & I \end{pmatrix},$$

the score equations become

$$\sum_{k=1}^K \begin{pmatrix} D_k^T \Sigma_k^{-1} (y_k - \mu_k) \\ -B_k \Sigma^{-1} (y_k - \mu_k) + w_k - \eta_k \end{pmatrix} = 0. \quad (9.6.7)$$

The corresponding Fisher information matrix is (for the k -th individual)

$$I_k = \begin{pmatrix} D_k^T \Sigma_k^{-1} D_k & 0 \\ 0 & \text{Var}(w_k) - B_k \Sigma^{-1} B_k \end{pmatrix}.$$

The asymptotic variance of $\tilde{\beta}$, the solution of the estimating equation

$$\sum_{k=1}^K \frac{\partial \mu_k}{\partial \beta} V_k^{-1} (y_k - \mu_k) = 0$$

is

$$\frac{\partial \mu_k}{\partial \beta} V_k^{-1} \Sigma_k V_k \frac{\partial \mu_k}{\partial \beta^T}, \quad (9.6.8)$$

where V_k is the “working covariance”. If $V_k = \Sigma_k$, then $\tilde{\beta}$ is coincident to the MLE based on the exponential family.

The mean constrained partial exponential family has automatically determined the conditional covariance matrix of Y and imposes a stronger assumption than the estimating function approach where the “working covariance matrix” may not be correctly specified. Of course if the true underlying conditional density happens to come from the partial exponential family, then it would be more efficient than the estimating equation based approach. Zhao et al. (1992) conducted some numerical comparisons between these two methods.

As an alternative approach, it is possible to leave the baseline density $f(y)$ arbitrary. Then model (9.6.6) becomes the so-called “density ratio” model with continuous covariates. Huang (2014a) discussed an approach by profiling out dF subject to the moment constraints. In general this is a computationally intensive method. We will revisit this problem in Chap. 21.

A closely related work on the pseudo maximum likelihood method can be found in Gourieroux et al. (1984). They are interested in examining the properties of the estimators obtained by maximizing a likelihood function associated with a “working” family of probability distributions, which does not necessarily contain the true one.

Chapter 10

General Theory on Biased Sampling Problems

As indicated in the beginning of this book, the biased sampling is a ubiquitous problem in many disciplines. Pioneering works on this problem can be found, among others, in Wicksell (1925, 1926), Fisher(1934), Rao (1965), Cox (1969), Patil and Rao (1978), Anderson (1979), Prentice and Pyke (1979), Vardi (1982a,b), Vardi (1985, 1989), Heckman (1979) and McFadden (1980). In this chapter, we discuss this problem comprehensively. In a series of papers, Vardi had made outstanding contributions on the nonparametric estimation of the underlying distribution, which are among the most interesting applications.

First we start from a discussion on the nonparametric maximum likelihood estimation of the baseline distribution function based on the simplest one sample biased sampling problem.

10.1 Maximum Likelihood Estimation for Length Biased Sampling Problems

In the study of positive random variables, Cox (1969) and Vardi (1982a) considered a length biased sampling problem with independent and identically distributed observations

$$y_1, \dots, y_n \sim dG(y) = \frac{ydF(y)}{\int ydF(y)}.$$

The goal is to estimate the underlying distribution function F . We may solve F by using

$$dF(y) = \frac{y^{-1}dG(y)}{\int_0^\infty y^{-1}dG(y)}.$$

Clearly the nonparametric MLE of G is the empirical distribution of the y_i 's, i.e.,

$$G_n(t) = \frac{1}{n} \sum_{i=1}^n I(y_i \leq t).$$

Using the invariance property of MLE, the nonparametric MLE of F is

$$\hat{F}(t) = \int_0^t \frac{y^{-1} dG_n(y)}{\int y^{-1} dG_n(y)}.$$

As an alternative, we may directly maximize the likelihood

$$\prod_{i=1}^n \frac{y_i dF(y_i)}{\int y dF(y)}$$

with respect to F subject to the constraint that F is a distribution function. Let $p_i = dF(y_i)$, $i = 1, 2, \dots, n$. Then we need to maximize

$$L = \prod_{i=1}^n \frac{y_i p_i}{\sum_{j=1}^n p_j y_j}$$

subject to the constraint

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0.$$

It is not hard to show that the nonparametric MLE is

$$d\hat{F}(y_i) = y_i^{-1} / \sum_{j=1}^n y_j^{-1}.$$

Next we use the concept of truncation to interpret the length biased sampling problems. It is a very useful method to understand the structure of biased sampling problems. This idea can be used in more complex setups where the maximum likelihood estimate cannot be easily calculated.

EM Algorithm in the Length Biased Sampling Problems

We can treat the length biased sampling as a missing data problem. The EM-algorithm can be applied to find the MLE. Even though in the one sample problem this method does not look very appealing, we may find this method to be very powerful later. Moreover, this approach may shed light on solving many difficult problems (Details are given in Chap. 25).

Based on the length biased sampling data t_1, \dots, t_n , the likelihood is

$$L = \prod_{i=1}^n \frac{t_i dF(t_i)}{\mu} = \prod_{i=1}^n \frac{t_i p_i}{\sum_{j=1}^n p_j t_j}.$$

Statistically we can construct a different model but with the same likelihood. Define

$$c = \sum_{i=1}^n t_i, \quad \mu = \sum_{i=1}^n p_i t_i / c.$$

Let T be a discrete random variable with a probability distribution given by

$$P(T = t_j) = p_j, \quad j = 1, 2, \dots, n.$$

Similarly define another discrete random variable Z which is independent of T , and its probability distribution is given by

$$P(Z = t_j) = \frac{t_j}{c}, \quad j = 1, 2, \dots, n, \quad c = \sum_{i=1}^n t_i.$$

Instead of observing T and Z , we only observe those pairs such that $T = Z$. We do not observe pairs with $Z \neq T$ and we have no idea how many of them are not observed. Note that

$$P(T = t_i, Z = t_i | T = Z) = \frac{P(T = t_i)P(Z = t_i)}{\sum_{i=1}^n P(T = Z = t_i)} = \frac{p_i t_i}{\sum_{j=1}^n t_j p_j}.$$

This is a truncation data likelihood and can also be treated as a missing data problem. We can use an EM algorithm to find the maximum likelihood estimator.

For a cohort of subjects, let

$$\mathbf{O} = \{(T_1, Z_1), \dots, (T_n, Z_n), \quad T_i = Z_i, i = 1, 2, \dots, n\}$$

be the observed data with $T_i = Z_i, i = 1, 2, \dots, n$, whereas the truncated subjects are denoted by

$$\mathbf{O}^* = \{(T_1^*, Z_1^*), \dots, (T_m^*, Z_m^*), \quad T_i^* \neq Z_i^*, i = 1, 2, \dots, m\}.$$

The random integer m then follows a negative binomial distribution with parameter $\pi = \sum_{i=1}^n t_i p_i / c$. The probability mass function of m is

$$\binom{m+n-1}{m} (1-\pi)^m \pi^n, \quad m = 0, 1, 2, \dots,$$

and

$$E(m|\mathbf{O}) = n(1 - \pi)/\pi.$$

If all data T_i, Z_i are observable regardless of whether $Z_i = T_i$, the complete data log likelihood is

$$\ell = \sum_{j=1}^h \left\{ \sum_{i=1}^n I(T_i = t_j) + \sum_{k=1}^m I(T_k^* = t_j) \right\} \log p_j,$$

where $t_1 < t_2 < \dots < t_h$ ($h \leq n$) are observed distinct data points, and m is the number of those $T_i \neq Z_i$ (which are truncated). Let $\xi_j = \sum_{i=1}^n I(T_i = t_j)$. Since $I(T_k^* = t_j)$'s are not observable, we need to impute them by conditioning on the observed data \mathbf{O} . Let

$$w_j = E \left[\left\{ \sum_{i=1}^n I(T_i = t_j) + \sum_{k=1}^m I(T_k^* = t_j) \right\} \mid \mathbf{O} \right] := \xi_j + np_j(1 - t_j/c)/\pi.$$

Easily

$$w_+ = \sum_{j=1}^h w_j = n + n \left(1 - \sum_{j=1}^h t_j p_j / c \right) / \pi = n/\pi.$$

For convenience we suppose that there are no ties, i.e., $\xi_j = 1$. The imputed log-likelihood is $\sum_{i=1}^n w_i \log p_i$. As a consequence, the MLE satisfies

$$p_j = w_j/w_+ = \frac{\xi_j + np_j(1 - t_j/c)/\pi}{n/\pi} = \pi/n + p_j(1 - t_j/c),$$

or

$$p_j = \frac{\pi c}{t_j n} = \frac{\sum_{j=1}^h t_j p_j}{t_j n}.$$

Clearly

$$p_j = \frac{t_j^{-1}}{\sum_{i=1}^h t_i^{-1}}$$

is the solution. Therefore we end up with the same MLE by using EM algorithm for the truncation problem.

A more interesting problem discussed by Vardi (1982a) is the two sample problem where

$$X_1, \dots, X_m \sim i.i.d. F(x), \quad Y_1, \dots, Y_n \sim i.i.d. \frac{ydF(y)}{\int ydF(y)}.$$

In this case, we have to work on the likelihood

$$L = \prod_{i=1}^m dF(x_i) \prod_{j=1}^n \frac{y_j dF(y_j)}{\int y dF(y)}$$

to find the nonparametric MLE.

Without loss of generality, we assume there are no ties in the observed data. Denote the pooled data as

$$(t_1, \dots, t_N) = (x_1, \dots, x_m; y_1, \dots, y_n), \quad N = m + n.$$

The log-likelihood can be written as

$$\ell = \sum_{i=1}^N \log p_i - n \log \mu, \quad \mu = \sum_{i=1}^N p_i t_i.$$

We need to maximize this likelihood subject to the constraints

$$\sum_{i=1}^N p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^N p_i t_i = \mu.$$

Moreover, if we are interested in constructing a confidence interval for $H_0 : \mu = \mu_0$, then the empirical likelihood ratio method can be applied here (Qin 1993).

Exercise 1 Give the details on the empirical likelihood based inference.

Exercise 2 Generalize the EM algorithm from a one sample example to two sample problems discussed above.

Exercise 3 A goodness of fit test.

In absence of the length bias sampling, we may form a binomial likelihood using $I(x_i \leq t)$,

$$\ell_B(t) = \sum_{i=1}^n I(x_i \leq t) \log F(t) + \sum_{i=1}^n I(x_i > t) \log \bar{F}(t).$$

To test $F(t) = F_0(t)$, we may use the likelihood ratio

$$R(t) = \sum_{i=1}^n I(x_i \leq t) \log \{F(t)/F_0(t)\} + \sum_{i=1}^n I(x_i > t) \log \{\bar{F}(t)/\bar{F}_0(t)\}.$$

The Berk–Jones test statistic is $\max_t R(t)$. Owen (1995) discussed the computation aspect of this test statistic. Jager and Wellner (2007) gave a comprehensive discussion of related theoretical results.

In the presence of length bias sampling, an analogous version can be developed as follows. Since the nonparametric MLE of F under length biased sampling is

$$\hat{F}(t) = \frac{\sum_{i=1}^n I(x_i \leq t)/x_i}{\sum_{i=1}^n 1/x_i}.$$

Naturally we may define the likelihood ratio test as

$$\begin{aligned} R(t) &= 2[\hat{F}(t) \log \hat{F}(t) + \{1 - \hat{F}(t)\} \log \{1 - \hat{F}(t)\} - \hat{F}(t) \log F_0(t) \\ &\quad + \{1 - \hat{F}(t)\} \log \{1 - F_0(t)\}] \\ &= 2 \left[\hat{F}(t) \log \frac{\hat{F}(t)}{F_0(t)} + \{1 - \hat{F}(t)\} \log \frac{1 - \hat{F}(t)}{1 - F_0(t)} \right]. \end{aligned}$$

Derive the large sample results.

10.2 Maximum Likelihood Estimation for Multiple Biased Sampling Problems

Next we present Vardi (1985) results based on multiple biased sampling problems.

Suppose we have multiple independent biased samples with different known weight functions. Specifically let

$$y_{i1}, \dots, y_{in_i} \sim i.i.d. dG_i(t) = \frac{w_i(t)dF(t)}{W_i(F)}, \quad W_i(F) = \int w_i(u)dF(u), \quad i = 1, 2, \dots, s,$$

where all weights $w_i(t)$'s are non-negative functions with the forms either completely known or known up to some finitely many parameters. In this section we only focus on the completely known weights case. Typically in multiple biased sampling problems, there is no closed form for the maximum likelihood estimation of F and an iterative procedure is needed to find the solution.

Let (t_1, \dots, t_h) be the pooled distinct data points, where $h \leq \sum_{i=1}^s n_i$. Define $\eta_{ij} = \sum_{k=1}^{n_i} I(y_{ik} = t_j)$ as the multiplicity at t_j for sample i , where $j = 1, 2, \dots, h; i = 1, 2, \dots, s$. We need to maximize

$$L = \prod_{j=1}^h \left\{ \prod_{i=1}^s \left(\frac{w_i(t_j)dF(t_j)}{W_i(F)} \right)^{\eta_{ij}} \right\}.$$

Let $p_j = dF(t_j)$, $j = 1, 2, \dots, h$ be the jumps. Clearly $L = 0$ if any one of $dF(t_j) = 0$, $j = 1, 2, \dots, h$. The support points of F should include all observed data points. On the other hand, if F has an extra mass p^* at t^* , which is not the observed data point, then the likelihood is

$$\begin{aligned} L^* &= \prod_{j=1}^h \left\{ \prod_{i=1}^s \left(\frac{w_i(t_j)p_j}{\sum_{k=1}^h \eta_{ik} w_i(t_k)p_k + w_i(t^*)p^*} \right)^{\eta_{ik}} \right\} \\ &\leq \prod_{j=1}^h \left\{ \prod_{i=1}^s \left(\frac{w_i(t_j)p_j}{\sum_{k=1}^h \eta_{ik} w_i(t_k)p_k} \right)^{\eta_{ik}} \right\} \\ &= \prod_{j=1}^h \left\{ \prod_{i=1}^s \left(\frac{w_i(t_j)p_j/(1-p^*)}{\sum_{k=1}^h \eta_{ik} w_i(t_k)p_k/(1-p^*)} \right)^{\eta_{ik}} \right\}, \end{aligned}$$

since $w_i(t^*) \geq 0$, $p^* \geq 0$. In other words, we have found a new probability distribution $P(T = t_i) = p_i/(1-p^*)$, $i = 1, 2, \dots, h$ such that it results in a larger likelihood. Therefore we have shown that the maximizer F of L has jumps only at the observed data points.

Nevertheless, in general, the nonparametric MLE may not exist or may not be unique. Below are two examples given by Vardi (1985).

Example 1 (Non-existence of the NPMLE)

Suppose we have a sample of size two from CDF truncated to $[4, 9]$, with observed values 6 and 8, and a sample size two from F , with observed values 1 and 3.

Denote $p_1 = P(X = 1)$, $p_2 = P(X = 3)$, $p_3 = P(X = 6)$, $p_4 = P(X = 8)$. Then $\sum_{i=1}^4 p_i = 1$, $p_i > 0$, $i = 1, 2, 3, 4$. The likelihood is

$$L = p_1 p_2 \frac{p_3 p_4}{(p_3 + p_4)^2} < 1/16,$$

where due to truncation the probabilities of observing 6 and 8 are, respectively, $P(X = 6)/P(4 \leq X \leq 9) = p_3/(p_3 + p_4)$ and $P(X = 8)/P(4 \leq X \leq 9) = p_4/(p_3 + p_4)$. Note

$$L(1/2 - \epsilon, 1/2 - \epsilon, \epsilon, \epsilon) = (1/2 - \epsilon)^2/4 \rightarrow 1/16$$

as $\epsilon \rightarrow 0$. This shows that indeed there does not exist a MLE.

However if the sample from F itself, y_2 , had included an observed value from the truncation interval $[4, 9]$ then $L(p)$ would have possessed a maximum. For example if $y_2 = (1, 5)$ (instead of $(1, 3)$); then the likelihood is

$$L(p) = p_1 p_2 \frac{p_3 p_4}{(p_2 + p_3 + p_4)^2}.$$

The maximum attains at $p = (1/2, 1/6, 1/6, 1/6)$.

Example 2 (Non-uniqueness of the NPMLE)

Consider a two sample case $s = 2$.

$$w_1(u) = I[u \leq 20], \quad w_2(u) = I[u \geq 10], \quad y_1 = (6, 8), \quad y_2 = (26, 28)$$

Denote

$$p_1 = P(X = 6), \quad p_2 = P(X = 8), \quad p_3 = P(X = 26), \quad p_4 = P(X = 28).$$

It can be shown that the likelihood

$$L(p) = \frac{p_1 p_2 p_3 p_4}{(p_1 + p_2)^2 (p_3 + p_4)^2}$$

is maximized by any p of the form

$$p = (\alpha/2, \alpha/2, (1 - \alpha)/2, (1 - \alpha)/2), \quad 0 < \alpha < 1.$$

In the following, we present the elegant result by Davidov and Iliopoulos (2009), which is a simplified version of Vardi's (1985) approach.

Without loss of generality, we assume there are no ties, i.e., $h = n = \sum_{i=1}^s n_i$. Define $w_{ij} = w_i(t_j)$. Replacing $\int w_i(u) dF(u)$ by $\sum_{j=1}^n p_j w_{ij}$, we have the following likelihood

$$\prod_{j=1}^n p_j \prod_{i=1}^s \frac{1}{(\sum_{j=1}^n p_j w_{ij})^{n_i}}.$$

We need to maximize it subject to

$$\sum_{j=1}^n p_j = 1, \quad p_j \geq 0, \quad j = 1, 2, \dots, n.$$

The log-likelihood is neither concave nor convex in p_i 's. If we replace $p_j, j = 1, 2, \dots, n$ by $p_j/c, j = 1, 2, \dots, n$, the likelihood does not change. Therefore we may drop the constraint $\sum_{j=1}^n p_j = 1$.

Note that the likelihood can be written as an exponential family with canonical parameter θ ,

$$C \exp \left\{ \sum_{j=1}^n \theta_j - \sum_{i=1}^s n_i \Omega_i(\theta) \right\},$$

where

$$\theta_j = \log p_j, \quad \Omega_i = \log \{W_i\}, \quad W_i = \sum_{j=1}^n w_{ij} \exp(\theta_j).$$

Lemma 10.1 *The likelihood is log concave in θ but not strict.*

Proof The log-likelihood is

$$\ell = \sum_{j=1}^n \theta_j - \sum_{i=1}^s n_i \Omega_i(\theta).$$

Denote

$$H_i = \frac{\partial^2 \Omega_i}{\partial \theta \partial \theta^T}$$

as the Hessian of Ω_i . Let

$$\gamma_i^T = (w_{i1} \exp(\theta_1), \dots, w_{in} \exp(\theta_n)), \quad i = 1, 2, \dots, s,$$

then

$$W_i = \sum_{j=1}^n \gamma_{ij}.$$

For any $x \in R^n$,

$$\begin{aligned} x^T H_i x &= x^T \frac{1}{W_i} (W_i \text{diag}(\gamma_i) - \gamma_i \gamma_i^T) x \\ &= \frac{1}{W_i^2} \left[\sum_{j=1}^n \gamma_{ij} \sum_{j=1}^n \gamma_{ij} x_j^2 - \left(\sum_{j=1}^n \gamma_{ij} x_j \right)^2 \right], \end{aligned}$$

where $\text{diag}(\gamma_i)$ is a $n \times n$ diagonal matrix with diagonal elements γ_{ii} . Using Cauchy-Schwartz's inequality we may show that it is non-negative for all $x \in R^n$. Therefore H_i is positive semi-definite and $\Omega_i(\theta)$ $1 \leq i \leq s$ are convex. Thus the log-likelihood is concave.

However if $x = (c, c, \dots, c)$, then $x^T H_i x = 0$ for all $1 \leq i \leq s$. This implies the log-likelihood is not strictly concave.

Existence and Uniqueness

To study the existence and uniqueness of the nonparametric MLE, we consider a simplified two-sample case below. Readers should be able to extend this to the multiple-sample case. Let

$$Y_{11}, \dots, Y_{1n_1} \sim \frac{w_1(y)dF(y)}{\int w_1(y)dF(y)},$$

$$Y_{21}, \dots, Y_{2n_2} \sim \frac{w_2(y)dF(y)}{\int w_2(y)dF(y)}.$$

The likelihood is

$$L = \prod_{i=1}^{n_1} \frac{w_1(y_{1i})dF(y_{1i})}{\int w_1(y)dF(y)} \prod_{i=1}^{n_2} \frac{w_2(y_{2i})dF(y_{2i})}{\int w_2(y)dF(y)}.$$

Let

$$p_i = dF(y_{1i}), i = 1, 2, \dots, n_1, \quad q_i = dF(y_{2i}), i = 1, 2, \dots, n_2, \quad n = n_1 + n_2.$$

Up to a constant the log-likelihood can be written as

$$\ell(p, q) = \ell_1(p, q) + \ell_2(p, q),$$

where

$$\ell_1(p, q) = \sum_{i=1}^{n_1} \log p_i - n_1 \log \left[\sum_{i=1}^{n_1} w_1(y_{1i})p_i + \sum_{i=1}^{n_2} w_1(y_{2i})q_i \right],$$

and

$$\ell_2(p, q) = \sum_{i=1}^{n_2} \log q_i - n_2 \log \left[\sum_{i=1}^{n_1} w_2(y_{1i})p_i + \sum_{i=1}^{n_2} w_2(y_{2i})q_i \right].$$

The constraints are

$$p_i \geq 0, i = 1, 2, \dots, n_1; \quad q_i \geq 0, i = 1, 2, \dots, n_2; \quad \sum_{i=1}^{n_1} p_i + \sum_{i=1}^{n_2} q_i = 1.$$

We will consider following three cases.

(1) If

$$\ell_1(p, q) = \ell_1(p), \quad \ell_2(p, q) = \ell_2(q),$$

that is $w_1(y_{i2}) = 0, i = 1, 2, \dots, n_2$ and $w_2(y_{i1}) = 0, i = 1, 2, \dots, n_1$. Therefore

$$\ell(p, q) = \ell_1(p) + \ell_2(q).$$

$$\max_{p,q} \ell(p, q) = \max_p \ell_1(p) + \max_q \ell_2(q).$$

Note that

$$\ell_1(cp) = \ell_1(p), \quad \ell_2(cq) = \ell_2(q).$$

Clearly the maximizer is not unique. For example

$$\hat{p}_i = \frac{\alpha w_1^{-1}(y_{1i})}{\sum_{j=1}^{n_1} w_1^{-1}(y_{1j})}, \quad i = 1, 2, \dots, n_1$$

and

$$\hat{q}_i = \frac{(1-\alpha)w_2^{-1}(y_{2i})}{\sum_{j=1}^{n_2} w_2^{-1}(y_{2j})}, \quad i = 1, 2, \dots, n_2$$

are the solutions for any $0 < \alpha < 1$.

$$(2) \quad \ell_1(p, q) \neq \ell_1(p), \quad \ell_2(p, q) = \ell_2(q).$$

Let q^* be a maximizer of $\ell_2(q)$ and it satisfies $\sum_{i=1}^{n_2} q_i^* = 1 - \alpha$ ($0 < \alpha < 1$). Note that

$$\begin{aligned} \ell_1(p, q) &= \sum_{i=1}^{n_1} \log p_i - n_1 \log \left[\sum_{i=1}^{n_1} w_1(y_{1i}) p_i + \sum_{i=1}^{n_2} w_1(y_{2i}) q_i \right] \\ &\leq \sum_{i=1}^{n_1} \log p_i - n_1 \log \left[\sum_{i=1}^{n_1} w_1(y_{1i}) p_i \right] = \ell_1(p). \end{aligned}$$

Therefore as $q_i^* \rightarrow 0$, $i = 1, 2, \dots, n_2$, or equivalently $\alpha \rightarrow 1$, $\ell(p, q)$ achieves the maximum. This is the boundary case. As a result the maximum does not exist!

$$(3) \quad \ell_1(p, q) \neq \ell_1(p), \quad \ell_2(p, q) \neq \ell_2(q).$$

Denote

$$\mathcal{P} = \{(p, q) : \sum_{i=1}^{n_1} p_i + \sum_{i=1}^{n_2} q_i = 1, p_i \geq 0, i = 1, 2, \dots, n_1, q_i \geq 0, i = 1, 2, \dots, n_2\}.$$

If any one of $p_i = 0$ or $q_j = 0$, then $\ell(p, q) = -\infty$. Since $\ell(p, q)$ is continuous and differentiable in \mathcal{P} and does not achieve the maximum at the boundary, the maximum must be inside \mathcal{P} .

Next we show uniqueness. If there are two maximizers

$$r^1 = (p^1, q^1), \quad r^2 = (p^2, q^2),$$

then

$$\ell(p^1, q^1) = \ell(r^1) \geq \ell(r), \quad \ell(p^2, q^2) = \ell(r^2) \geq \ell(r), \quad r \in \mathcal{P}.$$

Since $\ell(r)$ is concave

$$\ell(r) \leq \alpha\ell(r^1) + (1 - \alpha)\ell(r^2) \leq \ell(\alpha r^1 + (1 - \alpha)r^2).$$

Therefore $\ell\{\alpha r^1 + (1 - \alpha)r^2\}$ is a constant for any $0 \leq \alpha \leq 1$. Differentiating $\ell(\alpha r^1 + (1 - \alpha)r^2)$ with respect to α gives

$$\frac{\partial \ell}{\partial \alpha}(r^1 - r^2) = 0,$$

which implies $r^1 = r^2$.

Vardi (1985) and Gill et al. (1988) gave general necessary and sufficient conditions for the maximization problem to have a unique solution. Define a graph G on the s vertices $i = 1, 2, \dots, s$ by identifying vertex i with k ($i \neq k$) if and only if

$$\int I[w_i(y) > 0]I[w_k(y) > 0]dF(y) > 0.$$

The graph G is connected if every pair of vertices is connected by a path. If this is true, then the underlying distribution F is identifiable. In other words, in order for F to be identifiable, for each y in the support of F there is a positive probability of y being observed.

More generally, Gilbert et al. (1999) allowed $w_i(y) = w_i(y, \theta)$, $i = 1, 2, \dots, s$ ($s \geq 2$) to depend on an unknown parameter θ . Suppose there exist functions w_i and w_k , for $i, k \in \{1, 2, \dots, s\}$, $i \neq k$, such that the functions $w_i(y, \theta)w_k(y, \tilde{\theta})$ and $w_i(y, \tilde{\theta})w_k(y, \theta)$ are linearly independent in y for all $\theta, \tilde{\theta} \in \Theta$ (the parameter space) with $\theta \neq \tilde{\theta}$. Then the s -sample selection bias model is identifiable.

The large sample results based on multi-sample biased sampling data were discussed thoroughly by Gill et al. (1988) and Gilbert et al. (1999) using the empirical process theory. Readers may go through their papers for details.

Exercise Consider a two-sample biased sampling data

$$X_1, \dots, X_m \sim f(x), \quad Y_1, \dots, Y_n \sim g(y) = \frac{w(y)f(y)}{\int w(y)f(y)dy}, \quad w(y) \geq 0,$$

where the weight function $w(y)$ is known.

(1) Assume there exists a r -vector estimating equation $\psi(X, \beta)$ such that $E_F[\psi(X, \beta)] = 0$, where β is a $p \times 1$ unknown parameter, $p \leq r$. Generalize Qin and Lawless' (1994) results to this two-sample biased sampling problem.

(2) Suppose $f(y) = f(y, \theta)$ is a parametric model. Find the MLE for θ . On the other hand, one may find the nonparametric MLE for F by using biased sampling method. Denote the nonparametric MLE as

$$\hat{F}(t) = \sum_{i=1}^N \hat{p}_i I(t_i \leq t),$$

where $(t_1, \dots, t_N) = (x_1, \dots, x_m, y_1, \dots, y_n)$, $N = m+n$, and \hat{p}_i , $i = 1, 2, \dots, N$ are the jumps derived from the nonparametric MLE. Define

$$\ell(\theta) = \sum_{i=1}^N \hat{p}_i \log f(t_i, \theta).$$

Then θ can be estimated by maximizing $\ell(\theta)$. Compare the two different parametric MLEs.

10.3 Weight Function Depends on the Underlying Distribution Problems

So far we have considered biased sampling problems where the weight functions are completely known. Sometimes, however, the weight functions may depend on the underlying distribution as well or are monotonic functions with unspecified forms. In these cases the maximum likelihood estimates are difficult to find. Next we give three examples.

Example 1 Maximum likelihood estimation based on ranked set samples

In order to optimize precision and allow for generalization of results, while keeping the costs of associated field and laboratory work at a reasonable level, rank set sampling is a popular method in environmental studies. Ranked set sampling involves an initial ranking of n samples of size n followed by observing the first order statistic (smallest observation) from the first sample, the second order statistic (second smallest observation) from the second sample, and so on, until the n -th order statistic from the n -th sample yields a secondary sample of size n from the initial n^2 data points. More specifically let X_{i1}, \dots, X_{in} , $i = 1, 2, \dots, n$ be independent and identically distributed data from F . Instead of observing the n^2 random variables, $X_{(1)}, \dots, X_{(n)}$ are recorded, where $X_{(i)}$ is the i -th order statistic among X_{i1}, \dots, X_{in} . Detailed discussions of this problem can be found in the excellent monograph by Chen et al. (2004). More generally we assume the r_i -th order statistic out of k_i observations has density

$$X_i \sim r_i \binom{k_i}{r_i} F^{r_i-1}(x) \{1 - F(x)\}^{s_i} dF(x), \quad s_i = k_i - r_i, \quad i = 1, 2, \dots, n$$

In this case we may treat

$$w_i(x) = w_i(F(x)) = r_i \binom{k_i}{r_i} F^{r_i-1}(x) \{1 - F(x)\}^{s_i},$$

as the weight function. The likelihood is

$$L = \prod_{i=1}^n F^{r_i-1}(x_i) \{1 - F(x_i)\}^{s_i} dF(x_i) = \prod_{i=1}^n F^{r_{(i)}-1}(x_{(i)}) \{1 - F(x_{(i)})\}^{s_{(i)}} dF(x_{(i)}).$$

Kvam and Samaniego (1994) studied the nonparametric MLE for F .

Let $\phi_j = F(x_{(j)})$, $j = 1, 2, \dots, n$. The parameter space is

$$\Phi = \{\phi : 0 < \phi_1 < \phi_2 < \dots < \phi_n \leq 1\}$$

Therefore the log-likelihood is

$$\ell = \sum_{i=1}^n [(r_{(i)} - 1) \log \phi_i + s_{(i)} \log(1 - \phi_i) + \log(\phi_i - \phi_{i-1})],$$

where $\phi_0 = 0$.

First we will show that the likelihood is a concave function. Let

$$Q = \frac{\partial^2 \ell}{\partial \phi \partial \phi^T}$$

be the Hessian matrix. Direct calculation show that it is tri-diagonal, i.e., it has nonzero entries only along its diagonal and in elements adjacent to its diagonal. It can be shown (Exercise) that for any vector $z \in R^n$,

$$z^T Q z = - \sum_{i=1}^n z_i^2 \left(\frac{r_{(i)} - 1}{\phi_i^2} + \frac{s_{(i)}}{(1 - \phi_i)^2} \right) - \sum_{i=1}^n (z_i - z_{i-1})^2 \left(\frac{1}{\phi_i - \phi_{i-1}} \right)^2 \leq 0,$$

where z_0 is defined as 0. Note that $z^T Q z = 0$ if and only if $z = 0$. Therefore the log-likelihood is strictly concave. By the standard result in convex analyses, the MLE exists. Using the result by Csiszar and Tusnády (1984), moreover we can claim that the EM-algorithm converges.

The following three algorithms can be used to find the maximum likelihood estimate of F .

(1) Direct maximization.

In this case, we need to be careful with the monotonicity and convergence, especially for the large sample size.

(2) Maximization by coordinator recycling method.

We may fix $\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n$, and maximize the likelihood with respect to ϕ_i , $i = 1, 2, \dots, n$. The first and second derivatives are, respectively,

$$\frac{\partial \ell}{\partial \phi_i} = \frac{r_i - 1}{\phi_i} - \frac{s_i}{1 - \phi_i} + \frac{1}{\phi_i - \phi_{i-1}} - \frac{1}{\phi_{i+1} - \phi_i}, \quad r_i \geq 1,$$

and

$$\frac{\partial^2 \ell}{\partial \phi_i^2} = -\frac{r_i - 1}{\phi_i^2} - \frac{s_i}{(1 - \phi_i)^2} - \frac{1}{(\phi_i - \phi_{i-1})^2} - \frac{1}{(\phi_{i+1} - \phi_i)^2} < 0.$$

From $\partial \ell / \partial \phi_i = 0$ we can solve for ϕ_i if $\phi_{i-1} \leq \phi_i \leq \phi_{i+1}$. Then we can iteratively solve for $\phi_1, \phi_2, \dots, \phi_n$.

(3) The EM algorithm.

Kvam and Samaniego (1994) studied the EM algorithm. Unfortunately this algorithm converges slowly. Moreover Kvam and Samaniego (1994) and Huang (1997) obtained the following self-consistent estimating equation

$$\hat{F}_n(t) = n^{-1} \sum_{i=1}^I (x_i \leq t) + n^{-1} \sum_{i=1}^n \left[\left\{ \frac{r_i - 1}{\hat{F}_n(x_i)} - \frac{k - r_i}{1 - \hat{F}_n(x_i)} \right\} \{\hat{F}_n(\min(x_i, t)) - \hat{F}_n(x_i)\hat{F}_n(t)\} \right].$$

The consistency of $\hat{F}_n(t)$ and limiting distribution were derived in Huang (1997).

Next we consider two special cases.

(1) Out of m samples, only the largest one is available. This is called “nomination sampling”. We can easily show

$$P(X_{(m_i)} > t) = 1 - P(X_{(m_i)} < t) = 1 - F^{m_i}(t).$$

The corresponding density is

$$P(X_{m_i} = t) = m_i F^{m_i-1}(t) f(t),$$

which is in the form of the Lehmann's parametric family.

(2) Only the minimum out of m_i samples is available. The density is

$$P(X_{(1)} = t) = m_i \bar{F}^{m_i-1} f(t),$$

which becomes a special case of the Cox proportional hazards model.

In these two special cases the MLE has closed form. More discussions on the non-parametric MLE for the Cox proportional hazards model will be discussed in Chaps. 24 and 25.

Example 2 Transformation model

Consider a two-sample transformation problem, where the two underlying distribution functions are linked by

$$dF_1(x) = C(F_0(x), \theta) dF_0(x),$$

where $C(\cdot, \cdot)$ is a given function up to an unknown parameter θ , and $F_0(x)$ is not specified. The function $C(F_0(x), \theta)$ can be treated as a weight. In general, the maximum profile likelihood has no closed form. It can be shown that the MLE of the

baseline distribution has jumps at all observed failure time points. More details for different transformation models with covariates and right censored data can be found in Zeng and Lin (2007).

Example 3 A monotonic two sample density ratio model

Instead of specifying the weight functions in the biased sampling problems, a natural generalization is

$$f_1(x)/f_0(x) = w(x),$$

where $f_1(x)$ and $f_0(x)$ are two density functions, and $w(x)$ is a monotonic non-increasing function but its form is unknown. As discussed in Chap. 1, this is called the likelihood ratio ordering in statistical literature. Equivalently we can write $dF_1(x) = \bar{G}(x)dF_0(x)/\int \bar{G}(x)dF_0(x)$, where $\bar{G}(x)$ is a survival function but the form is not specified. Suppose we have two samples of observed data

$$X_1, \dots, X_m \sim dF_0(x), \quad Y_1, \dots, Y_n \sim dF_1(y).$$

Dykstra et al. (1995) used the pool adjacent violation algorithm to find the maximum likelihood estimate. We will discuss this in detail in Chap. 26.

Chapter 11

General Theory for Case-Control Studies

So far we have assumed that the weight functions in biased sampling problems are either completely known or depend on the underlying distribution. In many applications, the weight functions may depend on some unknown finite dimensional parameters. We are not going to discuss in detail the general forms of the weight functions. We refer readers to the work by Gilbert et al. (1999). Instead we focus on a specific case, i.e., the log density ratio is a linear function of the observed data. We call this as a density ratio model or an exponential tilting density ratio model. It originates from the logistic regression model using case and control data. Prentice and Pyke (1979)'s and Anderson (1979)'s classical results on the validity of using prospective logistic likelihood inference for the odds ratio parameter for retrospectively collected case and control data have a fundamental impact on genetic and cancer studies. In this Chapter we use different methods to interpret their results. Furthermore we discuss the results by Whittemore (1995) on family-based case-control studies. We also outline methodologies for case and control studies based on the liability threshold model or the Probit model in genetic studies.

11.1 Semiparametric Inference for Logistic Regression Analysis Based on Case-Control Data

The logistic regression model is one of the most commonly used methods to study the relationship between a binary response variable and continuous or discrete covariates. Let D be disease indicator, assuming 1 if an individual has the disease of interest, and 0 otherwise. Suppose X are $1 \times p$ covariates. The standard logistic regression model has the form of

$$P(D = 1|x) = \frac{\exp(\alpha^* + x\beta)}{1 + \exp(\alpha^* + x\beta)} \equiv \pi(x), \quad X \sim f(x), \quad (11.1.1)$$

where α^* is an intercept parameter and β is a $p \times 1$ vector parameter. The marginal density $f(x)$ is unspecified.

Instead of collecting data prospectively, the retrospective sampling (Prentice and Pyke 1979) or case-control sampling is one of the most popular methods in cancer epidemiological studies. This is mainly due to the fact that it is the most convenient, economic and effective method. Especially in the study of rare diseases, one has to collect large samples in order to get a reasonable number of cases by using prospective sampling, which may not be practical. Using the case-control sampling, a pre-specified number of cases (n_1) and controls (n_0) are collected retrospectively from case and control populations separately. Typically this can be accomplished by sampling cases from hospitals, and sampling controls from the general disease free population. In general, the disease prevalence $\pi = P(D = 1)$ is different from the disease proportion $n_1/(n_1 + n_0)$ in the samples.

Let x_1, \dots, x_{n_0} be a random sample from $F(x|D = 0)$ and, independent of the x_i 's, let z_1, \dots, z_{n_1} be a random sample from $F(x|D = 1)$, where $F(x|D = 0)$ and $F(z|D = 1)$ are, respectively, the covariate distribution functions for controls and cases. Let $n = n_0 + n_1$ and write

$$\{t_1, t_2, \dots, t_n\} = \{x_1, \dots, x_{n_0}; z_1, \dots, z_{n_1}\}$$

as the combined sample. We assume that $n_i/n \rightarrow \rho_i > 0$ as $n \rightarrow \infty$ for $i = 0, 1$. Moreover the corresponding density functions of the covariates are denoted as, respectively, $f(x|D = i) = dF(x|D = i)/dx$, $i = 0, 1$. The Bayes' rule gives

$$f(x|D = 1) = \frac{\pi(x)}{\pi} f(x), \quad f(x|D = 0) = \frac{1 - \pi(x)}{1 - \pi} f(x).$$

It is seen that

$$\frac{f(x|D = 1)}{f(x|D = 0)} = \frac{1 - \pi}{\pi} \frac{\pi(x)}{1 - \pi(x)}.$$

Let $g(x) = f(x|y = 0)$ and $h(x) = f(x|D = 1)$. The corresponding cumulative distribution functions are denoted as $G(x)$ and $H(x)$, respectively. Then

$$h(x) = f(x|D = 1) = \frac{1 - \pi}{\pi} \frac{\pi(x)}{1 - \pi(x)} g(x) = \exp(\alpha + x\beta)g(x),$$

where $\alpha = \alpha^* + \log\{(1 - \pi)/\pi\}$. As a result, we arrive at the following two-sample exponential tilting model or density ratio model in which (x_1, \dots, x_{n_0}) and (z_1, \dots, z_{n_1}) are independent and

x_1, \dots, x_{n_0} are independent with density $g(x)$,

z_1, \dots, z_{n_1} are independent with density $h(x) = \exp(\alpha + x\beta)g(x)$. (11.1.2)

Table 11.1 Density ratio models and commonly used exponential families

$X D = j$	$R_1(x)$	$Q_1(\theta_j)$	$R_2(x)$	$Q_2(\theta_j)$
Binomial(n, p_j)	x	$\log(p_j/(1-p_j))$	–	–
Poisson(λ_j)	x	$\log \lambda_j$	–	–
Normal(μ_j, σ_j^2)	x	μ_j/σ_j	x^2	$-0.5/\sigma_j^2$
Gamma(α_j, β_j)	x	$-\beta_j$	$\log x$	α_j
Beta(α_j, β_j)	$\log x$	α_j	$\log(1-x)$	β_j

Therefore, this is a biased sampling model with weight function $\exp(\alpha + x\beta)$ depending on the unknown parameters α and β . Since the form of $g(x)$ is not specified, statistical inferences based on the exponential tilting model would be more robust than those based on a full parametric model in which the form of $g(x)$ is assumed to be known. Moreover, this model is invariant under various forms of selection bias, including truncation. We discuss this point in more detail in Chap. 24. The exponential tilting model encompasses many common distributions, including exponential distributions with different rates and normal distributions with common variance but different means. Recall in Sect. 6.3 of Chap. 6 we calculated the information lower bound for β . This model will be discussed extensively in later chapters.

Before discussing the semiparametric MLE, we first present the results by Kay and Little (1987).

Suppose given D , the density of X is a member of the exponential family

$$f(x; \theta) = B(\theta)h(x) \exp \left\{ \sum_{i=1}^p R_i(x) Q_i(\theta) \right\}, \quad \theta^T = (\theta_1, \dots, \theta_p).$$

Denoting the conditional density of $X|D = j$ by $f(x, \theta_j) = f_j(x)$, then we have

$$\log\{f_1(x)/f_0(x)\} = \log\{B(\theta_1)/B(\theta_0)\} + \sum_{i=1}^p R_i(x)\{Q_i(\theta_1) - Q_i(\theta_0)\}.$$

For some familiar exponential families, we tabulate the forms of $R_i(x)$, $Q_i(x)$, $i = 1, 2$ in Table 11.1.

Maximum Semiparametric Likelihood Estimation Method I

Next we discuss maximum semiparametric likelihood estimation. We adopt Anderson's (1972, 1979) approach by employing the Lagrange multiplier method to maximize the semiparametric likelihood. Some other alternative methods to derive the maximum semiparametric likelihood estimates will be given later.

Using the exponential tilting model, we can write the likelihood as

$$\mathcal{L}(\alpha, \beta, G) = \prod_{i=1}^{n_0} dG(x_i) \prod_{j=1}^{n_1} w(z_j) dG(z_j) = \left\{ \prod_{i=1}^n p_i \right\} \left\{ \prod_{j=1}^{n_1} w(z_j) \right\}, \quad (11.1.3)$$

where $w(x) = \exp(\alpha + x\beta)$ and $p_i = dG(t_i)$, $i = 1, 2, \dots, n$, are (nonnegative) jumps with total unit mass.

The first step is, for fixed (α, β) , to maximize \mathcal{L} with respect to p_i , $i = 1, \dots, n$, subject to constraints $\sum p_i = 1$, $p_i \geq 0$, $\sum p_i \{w(t_i) - 1\} = 0$, where the last constraint reflects the fact that $\int w(x)dG(x) = 1$. As discussed in Chap. 10, the maximum value of \mathcal{L} is attained at $p_i = n_0^{-1}\{1 + \rho \exp(\alpha + t_i\beta)\}^{-1}$ by using a Lagrange multiplier method, where $\rho = n_1/n_0$. Therefore, ignoring constants, the log-likelihood function is

$$l(\alpha, \beta) = \sum_{j=1}^{n_1} (\alpha + z_j\beta) - \sum_{i=1}^n \log\{1 + \rho \exp(\alpha + t_i\beta)\}. \quad (11.1.4)$$

Next we maximize l over (α, β) . Let $(\tilde{\alpha}, \tilde{\beta})$ satisfy the following system of score equations:

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \alpha} &= n_1 - \sum_{i=1}^n \frac{\rho \exp(\alpha + t_i\beta)}{1 + \rho \exp(\alpha + t_i\beta)} = 0, \\ \frac{\partial l(\alpha, \beta)}{\partial \beta} &= \sum_{j=1}^{n_1} z_j - \sum_{i=1}^n \frac{t_i \rho \exp(\alpha + t_i\beta)}{1 + \rho \exp(\alpha + t_i\beta)} = 0. \end{aligned} \quad (11.1.5)$$

Then we have

$$\tilde{p}_i = [n_0\{1 + \rho \exp(\tilde{\alpha} + t_i\tilde{\beta})\}]^{-1}. \quad (11.1.6)$$

Note that the above score equations are identical to those of Prentice and Pyke (1979). By using this maximum semiparametric likelihood approach, we have confirmed the well known result that prospective likelihood may be used to make inference for the log odds ratio parameter β even if the sampling design is retrospective.

On the basis of the \tilde{p}_i 's, naturally we can estimate $G(t)$ by

$$\tilde{G}(t) = \sum_{i=1}^n \tilde{p}_i I(t_i \leq t) = \frac{1}{n_0} \sum_{i=1}^n \frac{I(t_i \leq t)}{1 + \rho \exp(\tilde{\alpha} + t_i\tilde{\beta})}. \quad (11.1.7)$$

If $\hat{G}(t)$ denotes the empirical distribution function based on control data x_1, \dots, x_{n_0} , the difference

$$\Delta(t) = n^{\frac{1}{2}} |\hat{G}(t) - \tilde{G}(t)|, \quad \Delta = \sup_{-\infty \leq t \leq \infty} \Delta(t) \quad (11.1.8)$$

may be used to measure the departure from the logistic regression model assumption (11.1.1). Note that, for two vectors $a = (a_1, \dots, a_p)$ and $b = (b_1, \dots, b_p)$, $a \leq b$ and $-\infty \leq a \leq \infty$ stand for, respectively, $a_i \leq b_i$ and $-\infty \leq a_i \leq \infty$ for $i = 1, \dots, p$.

Remark 1 Similarly, $H(t) = F(t|D = 1)$ can be estimated by $\hat{H}(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} I(z_j \leq t)$ based on the case data z_1, \dots, z_{n_1} and by $\tilde{H}(t) = \sum_{i=1}^n \tilde{p}_i \exp(\tilde{\alpha} + t_i \tilde{\beta}) I(t_i \leq t)$ based on the case-control data t_1, \dots, t_n under model (11.1.2). If

$$\Delta_1(t) = n^{\frac{1}{2}} |\tilde{H}(t) - \hat{H}(t)|, \quad \Delta_1 = \sup_{-\infty \leq t \leq \infty} \Delta_1(t),$$

then Δ_1 is an alternative test statistic. However, since $\Delta_1(t) = \frac{n_0}{n_1} \Delta(t)$ and $\Delta_1 = \frac{n_0}{n_1} \Delta$, there is a symmetry between the case and control designations for such a global test. Note that both Δ and Δ_1 reduce to the Kolmogorov–Smirnov two-sample statistic when $\alpha = \beta = 0$. Note also that distance measures for distribution functions other than Δ or Δ_1 can also be employed to test the validity of the logistic regression model or the exponential tilting model, for example, the Cramer–von Mises’ statistic, etc.

Remark 2 The test statistic Δ can also be applied to mixture sampling data in which a sample of $n = n_0 + n_1$ members is randomly selected from the whole population with both n_0 and n_1 being random. Let (y_i, x_i) , $i = 1, 2, \dots, n$, be a random sample from the joint distribution of (Y, X) , then the likelihood has the form of

$$\mathcal{L} = \prod_{i=1}^n P(y_i|x_i) f(x_i) = \prod_{y_j=1} \{\pi f(x_j|y=1)\} \prod_{y_j=0} \{(1-\pi) f(x_j|y=0)\},$$

where $\pi = P(Y = 1)$.

Remark 3 In terms of estimating G and H , the model based semiparametric estimators $\tilde{G}(t)$ and $\tilde{H}(t)$ are more efficient than the nonparametric MLEs $\hat{G}(t)$ and $\hat{H}(t)$ based on the control and case data separately. However for the mean estimation of H , the model based semiparametric estimator $\int t d\tilde{H}(t)$ and nonparametric estimator $\int t d\hat{H}(t)$ based on the case data only are exactly the same.

In fact by the semiparametric score estimating Eqs. (11.1.5)–(11.1.7),

$$\sum_{j=1}^{n_1} z_j - n_0 \rho \int t d\hat{H}(t) = 0,$$

i.e.,

$$n_1 \int t d\hat{H}(t) - n_0 \rho \int t d\tilde{H}(t) = 0.$$

As a consequence $\int t d\hat{H}(t) = \int t d\tilde{H}(t)$. Similarly it can be shown that $\int t d\tilde{G}(t) = \int t d\hat{G}(t)$.

However in general $\int \phi(t)d\hat{H}(t) \neq \int \phi(t)d\tilde{H}(t)$ if $\phi(t) \neq t$.

Next we derive some asymptotic results. Let (α_0, β_0) be the true value of (α, β) under the logistic regression model and assume that $\rho = n_1/n_0$ remains fixed as $n \rightarrow \infty$. Then

$$-\frac{1}{n} \begin{pmatrix} \frac{\partial^2 l(\alpha_0, \beta_0)}{\partial \alpha^2} & \frac{\partial^2 l(\alpha_0, \beta_0)}{\partial \alpha \partial \beta^\top} \\ \frac{\partial^2 l(\alpha_0, \beta_0)}{\partial \beta \partial \alpha} & \frac{\partial^2 l(\alpha_0, \beta_0)}{\partial \beta^2} \end{pmatrix} \rightarrow \frac{\rho}{1+\rho} \begin{pmatrix} \int \frac{\exp(\alpha_0 + t\beta_0)}{1+\rho \exp(\alpha_0 + t\beta_0)} dG(t) & \int \frac{t \exp(\alpha_0 + t\beta_0)}{1+\rho \exp(\alpha_0 + t\beta_0)} dG(t) \\ \int \frac{t^\top \exp(\alpha_0 + t\beta_0)}{1+\rho \exp(\alpha_0 + t\beta_0)} dG(t) & \int \frac{t^\top t \exp(\alpha_0 + t\beta_0)}{1+\rho \exp(\alpha_0 + t\beta_0)} dG(t) \end{pmatrix} = S$$

in probability, where $n_1/n = \rho/(1+\rho)$. Write

$$A_0(t) = \int_{-\infty}^t \frac{\exp(\alpha_0 + \beta_0 y)}{1 + \rho \exp(\alpha_0 + \beta_0 y)} dG(y), \quad A_1(t) = \int_{-\infty}^t \frac{\exp(\alpha_0 + \beta_0 y)}{1 + \rho \exp(\alpha_0 + \beta_0 y)} y dG(y),$$

$$A_2(t) = \int_{-\infty}^t \frac{\exp(\alpha_0 + \beta_0 y)}{1 + \rho \exp(\alpha_0 + \beta_0 y)} y^\top dG(y), \quad A_0 = A_0(\infty), \quad A_1 = A_1(\infty), \quad A_2 = A_2(\infty),$$

$$A = \begin{pmatrix} A_0 & A_1 \\ A_1^\top & A_2 \end{pmatrix}, \quad S = \frac{\rho}{1+\rho} A, \quad \Sigma = \frac{1+\rho}{\rho} \left[A^{-1} - \begin{pmatrix} 1+\rho & 0 \\ 0 & 0 \end{pmatrix} \right].$$

Then it can be seen that

$$n^{-\frac{1}{2}} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} \rightarrow N(0, V) \text{ in distribution}, \quad V = \frac{\rho}{1+\rho} A - \rho \begin{pmatrix} A_0 \\ A_1^\top \end{pmatrix} (A_0, A_1).$$

Now we can easily show

Theorem 11.1 *Under some regularity conditions, then in distribution*

$$n^{\frac{1}{2}} \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} \rightarrow N(0, S^{-1} V S^{-1}) = N(0, \Sigma) \text{ in distribution}.$$

In the following we study the difference between the empirical distribution and the semiparametric distribution function estimator. For simplicity we only consider the case of $p = 1$, though all the results can be naturally generalized to the case of $p > 1$.

Theorem 11.2 *Under the logistic regression model and suitable regularity conditions,*

$$\tilde{G}(t) - \hat{G}(t) = H_1(t) - \hat{G}(t) - H_2(t) + o_p(n^{-1/2}), \quad (11.1.9)$$

where

$$H_1(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{I(t_i \leq t)}{1 + \rho \exp(\alpha_0 + \beta_0 t_i)}, \quad H_2(t) = \frac{\rho}{n} (A_0(t), A_1(t)) S^{-1} \begin{pmatrix} \partial l(\alpha_0, \beta_0)/\partial \alpha \\ \partial l(\alpha_0, \beta_0)/\partial \beta \end{pmatrix}.$$

As a result, as $n \rightarrow \infty$, $n^{\frac{1}{2}}(\tilde{G}(t) - \hat{G}(t))$ converges weakly to a Gaussian process $W(t)$ with mean 0 and covariance function

$$E\{W(s)W(t)\} = \rho(1 + \rho)(A_0(s), A_1(s)) \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right\}.$$

Proof Let

$$H_0(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{\rho \exp(\alpha_0 + t_i \beta_0)}{[1 + \rho \exp(\alpha_0 + t_i \beta_0)]^2} I(t_i \leq t),$$

$$R_{1n}(t) = [H_0(t) - A_0(t), H_1(t) - A_1(t)] \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix}.$$

It can be shown by using a first-order Taylor's expansion that

$$\tilde{G}(t) - \hat{G}(t) = H_1(t) - \hat{G}(t) - H_2(t) + R_{1n}(t) + R_{2n}(t),$$

where $\sup_{-\infty \leq t \leq \infty} |R_{2n}(t)| = o_p(n^{-1/2})$. In the proof, we obtain the asymptotic expression

$$\begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} = \frac{1}{n} S^{-1} \begin{pmatrix} \frac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(n^{-1/2}),$$

which implies that $\sup_{-\infty \leq t \leq \infty} |R_{1n}(t)| = o_p(n^{-1/2})$. Let $R_n(t) = R_{1n}(t) + R_{2n}(t)$. Note that $\sup_{-\infty \leq t \leq \infty} |R_n(t)| = o_p(n^{-1/2})$. For the proof of weak convergence of $n^{1/2}(\tilde{G} - \hat{G})$, according to (11.1.9), it suffices to show that

$$n^{\frac{1}{2}}(H_1(t) - \hat{G}(t) - H_2(t)) \rightarrow W(t) \text{ weakly.}$$

It is easy to see that $E[H_1(t) - \hat{G}(t) - H_2(t)] = 0$. Moreover, some tedious algebra show that

$$Cov(n^{\frac{1}{2}}[H_1(s) - \hat{G}(s)], n^{\frac{1}{2}}H_2(t)) = Cov(n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}H_2(t))$$

and

$$\begin{aligned} & Cov(n^{\frac{1}{2}}[H_1(s) - \hat{G}(s)] - n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}[H_1(t) - \hat{G}(t)] - n^{\frac{1}{2}}H_2(t)) \\ &= Cov(n^{\frac{1}{2}}[H_1(s) - \hat{G}(s)], n^{\frac{1}{2}}[H_1(t) - \hat{G}(t)]) - Cov(n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}H_2(t)) \\ &= \frac{n\rho}{n_0} A_0(s) - \frac{n^2 \rho^2}{n_0 n_1} A_0(s) A_0(t) - \rho(1 + \rho)(A_0(s), A_1(s)) A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} + \frac{n^2 \rho^2}{n_0 n_1} A_0(s) A_0(t) \\ &= \rho(1 + \rho) A_0(s) - \rho(1 + \rho)(A_0(s), A_1(s)) A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \\ &= \rho(1 + \rho)(A_0(s), A_1(s)) \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} - A^{-1} \begin{pmatrix} A_0(t) \\ A_1(t) \end{pmatrix} \right] = E W(s) W(t). \end{aligned}$$

This, together with the Central Limit Theorem for sample means and the Cramer-Wold device, imply that the finite-dimensional distribution of $n^{1/2}(H_1(t) - \hat{G}(t) - H_2(t))$ converges weakly to that of W . Thus, in order to prove weak convergence of $n^{1/2}(\tilde{G} - \hat{G})$, it is enough to show that the process $\{n^{1/2}(H_1(t) - \hat{G}(t) - H_2(t)), -\infty \leq t \leq \infty\}$ is tight (Van der Vaart and Wellner 1992). We leave this as an exercise.

A Bootstrap Method for Finding P Value

The simple bootstrap method by re-sampling case-control data separately does not work for model testing problems. Now we present a bootstrap method for finding critical values of Δ . Since α^* is not estimable in general on the basis of the case-control data t_1, \dots, t_n , if the logistic regression model is valid, we can generate bootstrap data from $d\tilde{G}(x)$ and $\exp(\tilde{\alpha} + x\tilde{\beta})d\tilde{G}(x)$, respectively. Specifically, suppose $x_1^*, x_2^*, \dots, x_{n_0}^*$ are drawn independently from $d\tilde{G}(x)$ and independent of the x_i^* , and suppose $z_1^*, z_2^*, \dots, z_{n_1}^*$ are drawn independently from $\exp(\tilde{\alpha} + x\tilde{\beta})d\tilde{G}(x)$. Note that some of the x_i^* could come from z_1, \dots, z_{n_1} and some of the z_j^* could be from x_1, \dots, x_{n_0} . Let t_1^*, \dots, t_n^* be the combined bootstrap sample and let $(\tilde{\alpha}^*, \tilde{\beta}^*)$ be the solution to the score equations with t_i replaced by t_i^* . Moreover, let $\hat{G}^*(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I(x_i^* \leq t)$ and

$$\tilde{p}_i^* = \frac{1}{n_0} \frac{1}{1 + \rho \exp(\tilde{\alpha}^* + t_i^* \tilde{\beta}^*)}, \quad \tilde{G}^*(t) = \frac{1}{n_0} \sum_{i=1}^n \frac{I(t_i^* \leq t)}{1 + \rho \exp(\tilde{\alpha}^* + t_i^* \tilde{\beta}^*)}. \quad (11.1.10)$$

Then a bootstrap version of the test statistic Δ is

$$\Delta^*(t) = n^{1/2} |\hat{G}^*(t) - \tilde{G}^*(t)|, \quad \Delta^* = \sup_{-\infty \leq t \leq \infty} \Delta^*(t). \quad (11.1.11)$$

We can approximate the critical values of Δ by those of Δ^* .

If the logistic regression model or its equivalent exponential tilting model is valid, the next problem is to test $\beta = 0$, that is, to test if X and Y are independent or to test $f(x|y=0) = f(x|y=1)$. The score-based test statistic is

$$S = \frac{n_1}{n} \sum_{i=1}^n t_i - \sum_{j=1}^{n_1} z_j,$$

which is equivalent to $(\bar{x} - \bar{z})$. Furthermore, the bootstrap percentile method (Efron and Tibshirani 1996) can be used to construct confidence intervals for the odds ratio parameter β . Alternatively, as in Qin and Lawless (1994), we can construct confidence intervals for β based on the empirical likelihood ratio statistics. Moreover, if we are interested in the underlying distribution function $G(x) = F(x|D=0)$, it is possible to construct confidence bands for G based on a bootstrap approach.

The Equivalency Between Prospectively and Retrospectively Estimated Baseline Distributions

In this section we show that it does not matter whether the available data are collected prospectively or retrospectively, the case and control distribution function estimates have the same formula. This is very useful in applications since we do not need to distinguish these two different sampling schemes.

Denote the control data as x_1, \dots, x_{n_0} and the case data as z_1, \dots, z_{n_1} , respectively. Let $(t_1, \dots, t_n) = (x_1, \dots, x_{n_0}; z_1, \dots, z_{n_1})$ be the pooled data, where $n = n_0 + n_1$. Define an artificial indicator $D_i = 0$, $i = 1, 2, \dots, n_0$ and $D_i = 1$, $i = n_0 + 1, \dots, n$. The logistic regression model is given by

$$P(D = 1|x) = \pi(x) = \frac{\exp(\alpha^* + x\beta)}{1 + \exp(\alpha^* + x\beta)}, \quad X \sim dF(x).$$

Pretending data are collected prospectively, then we can fit a prospective logistic regression model to find point estimators of α^* and β . Note that

$$H(t) = P(X \leq t | D = 1) = \frac{P(D = 1, X \leq t)}{P(D = 1)} = \frac{\int_{-\infty}^t \pi(x) dF(x)}{\int_{-\infty}^{\infty} \pi(x) dF(x)}$$

and

$$G(t) = P(X \leq t | D = 0) = \frac{\int_{-\infty}^t \{1 - \pi(x)\} dF(x)}{\int_{-\infty}^{\infty} \{1 - \pi(x)\} dF(x)}.$$

The case distribution H and the control distribution G can be estimated prospectively, by

$$\hat{H}(t) = \frac{\sum_{i=1}^n \hat{\pi}(t_i) I(t_i \leq t)}{\sum_{i=1}^n \hat{\pi}(t_i)}, \quad \hat{G}(t) = \frac{\sum_{i=1}^n \{1 - \hat{\pi}(t_i)\} I(t_i \leq t)}{\sum_{i=1}^n \{1 - \hat{\pi}(t_i)\}},$$

where

$$\hat{\pi}(t_i) = \frac{\exp(\hat{\alpha}^* + \hat{\beta}t_i)}{1 + \exp(\hat{\alpha}^* + \hat{\beta}t_i)}$$

and $\hat{\alpha}^*$ and $\hat{\beta}$ satisfy

$$n_1 - \sum_{i=1}^n \hat{\pi}(t_i) = 0, \quad n_1 \bar{z} - \sum_{i=1}^n \hat{\pi}(t_i) t_i = 0. \quad (11.1.12)$$

As we have shown, the profile log-likelihood based on retrospectively collected data is

$$\begin{aligned}\ell_R &= \sum_{i=1}^{n_1} (\alpha + z_j \beta) - \sum_{i=1}^n \log\{1 + \rho \exp(\alpha + t_i \beta)\} \\ &= \sum_{i=1}^{n_1} (\nu + z_j \beta) - \sum_{i=1}^n \log\{1 + \exp(\nu + t_i \beta)\} - n_1 \log \rho,\end{aligned}$$

where $\rho = n_1/n_0$, $\nu = \alpha + \log \rho = \alpha^* + \log\{(1 - \pi)/\pi\} + \log \rho$. Note that

$$p_i = \frac{1}{n_0} \frac{1}{1 + \rho \exp(\alpha + t_i \beta)} = \frac{1}{n_0} \frac{1}{1 + \exp(\nu + t_i \beta)}$$

The retrospective scores are

$$\frac{\partial \ell_R}{\partial \nu} = n_1 - \sum_{i=1}^n \frac{\exp(\nu + t_i \beta)}{1 + \exp(\nu + t_i \beta)} = 0$$

and

$$\frac{\partial \ell_R}{\partial \beta} = \sum_{i=1}^{n_1} z_i - \sum_{i=1}^n \frac{t_i \exp(\nu + t_i \beta)}{1 + \exp(\nu + t_i \beta)} = 0.$$

They are precisely the prospective score equations (11.1.12) except for different notations for α^* and ν . Moreover, we can write $p_i = \{1 - \hat{\pi}(t_i)\}/n_0$, $n_1 = \sum_{i=1}^n \hat{\pi}(t_i)$ and $n_0 = \sum_{i=1}^n \{1 - \hat{\pi}(t_i)\}$.

The control distribution function estimator based on the retrospective method is

$$\sum_{i=1}^n p_i I(t_i \leq t) = \frac{1}{n_0} \sum_{i=1}^n \{1 - \hat{\pi}(t_i)\} I(t_i \leq t) = \hat{G}(t),$$

which is the same as the one based on the prospective method. The same argument applies to the case distribution estimates.

11.2 Optimality of the Maximum Semiparametric Likelihood Estimating Equations

Now we show the optimality of the maximum semiparametric likelihood estimation of β . A slight generalization of the logistic regression model (11.1.1) is the multiplicative logistic regression model, given by

$$P(D = 1|x) = \frac{\exp(\alpha^* + \phi(x, \beta))}{1 + \exp(\alpha^* + \phi(x, \beta))},$$

where ϕ is a specified function. Using the retrospective representation,

$$f(x|D=1) = \exp\{\alpha + \phi(x, \beta)\} f(x|D=0), \quad \alpha = \alpha^* + \log\{(1-\pi)/\pi\}.$$

For any measurable function η ,

$$E_1[\eta(X, \beta)] = E_0[\eta(X, \beta) \exp(\alpha + \phi(x, \beta))],$$

where E_1 and E_0 are expectations with respect to $D = 1$ and $D = 0$, respectively.

Let $\rho_n = n_1/n_0$. We define a class of estimating functions

$$\mathcal{J} = \left\{ Q_\eta(\eta, \theta) \mid Q_\eta(\eta, \theta) = n^{-1} \left[\sum_{i=n_0+1}^n \eta(x_i) - \rho_n \sum_{i=1}^{n_0} \eta(x_i) \exp\{\alpha + \phi(x, \beta)\} \right] \right\}. \quad (11.2.13)$$

The question is, “What is the optimal choice of η ”? In the full parametric model it is shown in Chap. 5 that the score estimating function is optimal (Godambe 1960). Parallel to Godambe’s theory in the semiparametric density ratio model, we show that the semiparametric score estimating function falls in (11.2.13) and is optimal in this class. The optimality of the semiparametric MLE has been established by many authors, among others, Rabinowitz (1997) and Breslow et al. (2000). In Sect. 6.3, by using the projection method we found the optimal estimating function for β is orthogonal to the nuisance parameter space generated by the baseline density function. In this section we present an elementary proof (Qin 1998b) on the optimality of the semiparametric score equation based on case-control data.

Using the same exercise as in the exponential tilting model we can show that the profile log-likelihood for the general multiplicative model is

$$\ell = \sum_{i=n_0+1}^n \{\alpha + \phi(x_i, \beta)\} - \sum_{i=1}^n \log[1 + \rho_n \exp(\alpha + \phi(x_i, \beta))], \quad \rho_n = n_1/n_0.$$

Let $g(x, \theta) = \exp\{\alpha + \phi(x, \beta)\}$ be the density ratio between the cases and controls.

The maximum semiparametric likelihood score function is equivalent to taking

$$\eta(x) = \xi(x) = (\xi_1(x), \xi_2(x)) = \frac{1}{1 + \rho_n g(x, \theta)} \frac{\partial \log g(x, \theta)}{\partial \theta},$$

where

$$\xi_1(x) = \frac{1}{1 + \rho_n \exp\{\alpha + \phi(x, \beta)\}}, \quad \xi_2(x) = \frac{\partial \phi / \partial \beta}{1 + \rho_n \exp\{\alpha + \phi(x, \beta)\}}.$$

Let $(\tilde{\alpha}_\eta, \tilde{\beta}_\eta)$ be the solution of the equation

$$Q_\eta(\eta, \theta) = 0, \quad \theta^T = (\alpha, \beta^T),$$

where

$$Q_\eta(\eta, \theta) = n^{-1} \left[\sum_{i=n_0+1}^n \eta(x_i) - \rho_n \sum_{i=1}^{n_0} \eta(x_i) \exp\{\alpha + \phi(x, \beta)\} \right].$$

Theorem 11.3 *Under certain regularity conditions, in distribution*

$$\sqrt{n} \begin{pmatrix} \tilde{\alpha}_\eta - \alpha_0 \\ \tilde{\beta}_\eta - \beta_0 \end{pmatrix} \rightarrow N(0, A_\eta),$$

where

$$A_\eta = U_\eta^{-1} V_\eta (U_\eta^{-1})^T,$$

$$V_\eta = \rho_1 \text{var}_{F_1}\{\eta(x, \theta_0)\} + \rho \rho_1 \text{var}_{F_0}\{\eta(x, \theta_0)g(x, \theta_0)\},$$

$$U_\eta = -\rho_1 E_{F_0} \left\{ \eta(x, \theta_0) \frac{\partial g(x, \theta_0)}{\partial \theta} \right\},$$

$$\lim_{n \rightarrow \infty} \rho_n = \rho, \quad \lim_{n \rightarrow \infty} n_1/n = \rho_1,$$

In particular, if $\eta(x) = \xi(x, \theta)$, i.e., the maximum likelihood estimating function, then

$$A_\xi = -U_\xi^{-1} - \frac{1}{\rho_0 \rho_1} \begin{pmatrix} 1 & 0_1^T \\ 0_1 & 0_2 \end{pmatrix},$$

$$\rho_0 = 1 - \rho_1$$

where 0_1 is a $(p-1)$ -vector of zeros and 0_2 is a $(p-1) \times (p-1)$ matrix of zeros, and

$$U_\xi = -\rho_1 E_{F_0} \left\{ \left(\frac{1}{\partial \phi / \partial \beta} \frac{(\partial \phi / \partial \beta)^T}{\partial \phi / \partial \beta (\partial \phi / \partial \beta)^T} \right) \frac{g}{1 + \rho g} \right\}.$$

Theorem 11.4 *The maximum semiparametric likelihood estimating equation $Q_\xi(\xi, \theta) = 0$ is optimal in the class \mathcal{J} defined in (11.2.13), in the sense that $A_\eta - A_\xi$ is positive semidefinite for any measurable function $\eta(x)$ provided its second moment exists under F_0 and F_1 .*

Proof Write $\eta^T = (\eta_1^T, \eta_2^T)$, where η_1 and η_2 are scale and $(p-1)$ vector-valued functions, respectively. Since $g(x, \theta) = \exp\{\alpha + \phi(x, \beta)\}$, we have,

$$U_\eta = -\rho_1 E_{F_0} \left\{ \left(\begin{matrix} \eta_1 \\ \eta_2 \eta_2 \partial \phi / \partial \beta \end{matrix} \right) \frac{\eta_1 \partial \phi / \partial \beta}{\eta_2 \eta_2 \partial \phi / \partial \beta} \right\} g,$$

$$V_\eta = \rho_1 E_{F_0} \{ \eta \eta^T g(1 + \rho g) \} - \rho_0^{-1} \rho_1 (E_{F_0} \eta g)(E_{F_0} \eta g)^T.$$

Note that

$$0 \leq \text{var}[n^{1/2} \{ U_\eta^{-1} Q_\eta(\eta, \theta) - U_\xi^{-1} Q_n(\xi, \theta) \}] = A_\eta + A_\xi - 2B,$$

where

$$B = n U_\eta^{-1} \text{cov}(Q_n(\eta, \theta), Q_n(\xi, \theta)) U_\xi^{-1}.$$

After some algebra we have

$$B = A_\xi.$$

Therefore

$$0 \leq A_\eta + A_\xi - 2A_\xi = A_\eta - A_\xi.$$

Exercise 1 Is there any optimal result for the distribution function estimation?

11.3 Different Semiparametric MLE Methods for Case-Control Data

The Prentice and Pyke (1979)'s result on using prospective likelihood to make inference for the log odds ratio parameter for retrospectively collected case-control data has a fundamental impact on cancer case-control studies. Below we will use different methods to interpret this result. We believe the different derivations offer useful perspectives to understanding this important result, especially for those who are new in this research area. We begin from an intuitive argument.

Semiparametric MLE Method II

First we assume that data $(D_i, x_i), i = 1, 2, \dots, n$ are collected prospectively. There are two ways to decompose the likelihood function.

(1) Prospective likelihood decomposition

$$L_1 = \prod_{i=1}^n \left[\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right]^{D_i} \left[\frac{1}{1 + \exp(\alpha + \beta x_i)} \right]^{1-D_i} \prod_{i=1}^n dF(x_i),$$

where $F(x)$ is the marginal distribution of X . For fixed α and β , the maximum profile prospective likelihood with respect to F is $\prod_{i=1}^n (1/n) = (1/n)^n$.

(2) Retrospective likelihood decomposition

$$L_2 = \prod_{i=1}^n \{P(D = 1)\}^{D_i} \{1 - P(D = 1)\}^{1-D_i} \left[\prod_{i=1}^n \{f(x_i | D = 1)\}^{D_i} \{f(x_i | D = 0)\}^{1-D_i} \right].$$

Clearly for fixed $f(x|D = 1)$ and $f(x|D = 0)$, the maximum likelihood estimation of $P(D = 1)$ is n_1/n . As a result the profile likelihood with retrospective decomposition is

$$L_2 = \prod_{i=1}^n [n_1/n]^{D_i} [1 - n_1/n]^{1-D_i} \left\{ \prod_{i=1}^n [f(x_i|D = 1)]^{D_i} [f(x_i|D = 0)]^{1-D_i} \right\}.$$

Using the fact that $L_1 = L_2$, we can find

$$\begin{aligned} \prod_{i=1}^n [f(x_i|D = 1)]^{D_i} [f(x_i|D = 0)]^{1-D_i} &\propto \prod_{i=1}^n \left[\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right]^{D_i} \\ &\times \left[\frac{1}{1 + \exp(\alpha + \beta x_i)} \right]^{1-D_i}. \end{aligned}$$

The key is that

$$\begin{aligned} P(D = 1) &= \int \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)} dF(x) = \int \frac{\exp(\alpha + t)}{1 + \exp(\alpha + t)} dF(t/\beta) \\ &= \int \frac{\exp(\alpha + t)}{1 + \exp(\alpha + t)} dG(t). \end{aligned}$$

Since F is completely unknown, then the disease prevalence only depends on α and F ! This is the reason that the prospective likelihood can be used for estimating β .

Furthermore, suppose there are n_1 cases and n_0 controls in a retrospective study. We can choose α^* and F^* such that

$$P(D = 1) = n_1/(n_1 + n_0) = \int \frac{\exp(\alpha^* + t)}{1 + \exp(\alpha^* + t)} dF^*(t/\beta).$$

Then we can conduct a prospective study with a total sample size of $n = n_0 + n_1$. We end up with roughly n_1 cases and n_0 controls. Then all cases and controls are included in data collection stage. Finally based on the above arguments, we can again use the prospective likelihood to make inference on the log odds ratio parameter β .

Semiparametric MLE Method III

Instead of profiling out the control distribution function $dF(x|D = 0)$ in Sect. 11.1, here, we profile out the marginal distribution function $F(x)$ in the general population. Note that

$$P(D = 1|x) = \pi(\alpha, \beta, x), \quad \pi(x) = 1 - \frac{1}{1 + \exp(\alpha + \beta x)}$$

and

$$\pi = P(D = 1) = \int \pi(x) dF(x).$$

Based on the case data x_{11}, \dots, x_{n_11} and the control data x_{10}, \dots, x_{n_00} , the likelihood is

$$\begin{aligned} & \prod_{i=1}^{n_1} dF(x_{i1}|D=1) \prod_{i=1}^{n_0} dF(x_{i0}|D=0) \\ &= \prod_{i=1}^{n_1} \frac{\exp(\alpha + x_{i1}\beta)}{1 + \exp(\alpha + x_{i1}\beta)} \frac{dF(x_{i1})}{\pi} \prod_{i=1}^{n_0} \frac{1}{1 + \exp(\alpha + x_{i0}\beta)} \frac{dF(x_{i0})}{1 - \pi}. \end{aligned}$$

Denote $n = n_0 + n_1$, $(x_1, \dots, x_n) = (x_{11}, \dots, x_{n_11}; x_{10}, \dots, x_{n_00})$ and $d_i = 1$, $i = 1, 2, \dots, n_1$ for the cases and $d_i = 0$, $i = n_1 + 1, \dots, n$ for the controls. Write $p_i = dF(x_i)$, $i = 1, 2, \dots, n$. Then

$$\pi = \sum_{i=1}^n \pi(x_i) p_i$$

For fixed α, β, π , we can maximize the above likelihood with respect to the p_i s subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \{\pi(x_i) - \pi\} = 0,$$

to obtain the profile the log-likelihood

$$\begin{aligned} \ell &= \sum_{i=1}^n d_i(\alpha + \beta x_i) - \log\{1 + \exp(\alpha + \beta x_i)\} - n_1 \log \pi - n_0 \log(1 - \pi) \\ &\quad - \sum_{i=1}^n \log[1 + \lambda\{\pi(x_i, \alpha, \beta) - \pi\}], \end{aligned}$$

where the Lagrange multiplier λ is determined by

$$\sum_{i=1}^n \frac{\pi(x_i) - \pi}{1 + \lambda\{\pi(x_i) - \pi\}} = 0.$$

Taking derivative with respect to π , we have

$$\frac{\partial \ell}{\partial \pi} = -n_1/\pi + n_0/(1 - \pi) + \lambda \sum_{i=1}^n \frac{1}{1 + \lambda\{\pi(x_i, \alpha, \beta) - \pi\}} = 0,$$

or

$$-n\lambda = \frac{-n_1 + n\pi}{\pi(1 - \pi)}, \quad \pi\lambda = \frac{\rho_1 - \pi}{1 - \pi}, \quad \rho_1 = n_1/n, \quad \rho_0 = n_0/n,$$

After some simplifications, we obtain

$$\begin{aligned} 1 + \lambda\{\pi(x) - \pi\} &= 1 - \lambda\pi + \lambda\pi(x) \\ &= 1 - \lambda\pi + \lambda \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \\ &= [1 - \lambda\pi + (\lambda + 1 - \lambda\pi)\exp(\alpha + \beta x)]/[1 + \exp(\alpha + \beta x)]. \end{aligned}$$

From

$$\frac{\partial \ell}{\partial \alpha} = 0,$$

we have

$$\lambda = \frac{\rho_1 - \pi}{(1 - \pi)\pi}.$$

Replacing λ in the ℓ , we have

$$\ell = \sum_{i=1}^n d_i \{\alpha^* + x_i \beta\} - \log\{1 + \exp(\alpha^* + x_i \beta)\},$$

where

$$\alpha^* = \alpha + \log\{1 + \lambda(1 - \pi)\} - \log(1 - \lambda\pi) = \alpha + \log\{(1 - \pi)/\pi\} - \log\{(1 - \rho_1)/\rho_1\}.$$

This is the prospective log logistic likelihood except for a possibly different intercept.

Semiparametric MLE Method IV

Prentice and Pyke (1979) and Dykstra et al. (1995) used the following argument to find the semiparametric MLE.

In order to avoid using double index, we assume there are m controls and n cases. Denote $t_1, \dots, t_h, h \leq m + n$ as the pooled case and control samples. We assume that controls and cases are taken from (t_1, \dots, t_h) with multinomial distributions $\mathbf{p} = (p_1, \dots, p_h)$ and $\mathbf{q} = (q_1, \dots, q_h)$, respectively. The corresponding observed frequencies for controls and cases are denoted as $\mathbf{m} = (m_1, \dots, m_h)$ and $\mathbf{n} = (n_1, \dots, n_h)$, where $m = m_1 + \dots + m_h$ and $n = n_1 + \dots + n_h$.

Let

$$\theta_i = mp_i/(mp_i + nq_i), \quad \phi_i = mp_i + nq_i,$$

$$p_i = \theta_i \phi_i / m, \quad q_i = \phi_i (1 - \theta_i) / n.$$

As derived before the case and control densities are linked by the exponential tilting model, we have

$$\theta_i = \frac{1}{1 + \rho q_i / p_i} = \frac{1}{1 + \rho \exp(\alpha + \beta t_i)}, \quad \rho = n/m.$$

The retrospective likelihood can be written as

$$L = \prod_{i=1}^h p_i^{m_i} q_i^{n_i} = \left(\frac{1}{m}\right)^m \left(\frac{1}{n}\right)^n \prod_{i=1}^h \theta_i^{m_i} (1 - \theta_i)^{n_i} \prod_{i=1}^h \phi_i^{m_i + n_i},$$

Note that

$$\sum_{i=1}^h \phi_i = m + n, \quad \sum_{i=1}^h \theta_i \phi_i = \sum_{i=1}^h \frac{1}{1 + \rho \exp(\alpha + \beta t_i)} \phi_i = m.$$

Based on the binomial likelihood $\prod_{i=1}^h \theta_i^{m_i} (1 - \theta_i)^{n_i}$, the log version is

$$\ell_B = - \sum_{i=1}^h (m_i + n_i) \log\{1 + \rho \exp(\alpha + t_i \beta)\} + n_i (\log \rho + \alpha + t_i \beta).$$

If we directly maximize it without imposing the constraint $\sum_{i=1}^h \theta_i \phi_i = m$, then we have

$$\frac{\partial \ell_B}{\partial \alpha} = \sum_{i=1}^h n_i - (m_i + n_i) \frac{\rho \exp(\alpha + t_i \beta)}{1 + \rho \exp(\alpha + t_i \beta)} = 0,$$

or

$$\sum_{i=1}^h m_i = \sum_{i=1}^h \frac{m_i + n_i}{1 + \rho \exp(\alpha + t_i \beta)}.$$

However, if we directly maximize $\sum_{i=1}^h (m_i + n_i) \log \phi_i$ subject to constraint $\sum_{i=1}^h \phi_i = m + n$, then the MLE is $\hat{\phi}_i = m_i + n_i$. Note that

$$\sum_{i=1}^h \hat{\theta}_i \hat{\phi}_i = \sum_{i=1}^h \frac{1}{1 + \rho \exp(\hat{\alpha} + t_i \hat{\beta})} (m_i + n_i) = \sum_{i=1}^h m_i = m.$$

Therefore the constrained MLE (with $\sum_{i=1}^h \theta_i \phi_i = m$) and unconstrained MLE are the same. This completes the proof.

Semiparametric MLE Method V

Finally we use the missing data information principle and EM algorithm to derive the score estimating equations. Suppose

$$Z_1, \dots, Z_m \sim dF(t)$$

and

$$Y_1, \dots, Y_n \sim \frac{\exp(t\beta) dF(t)}{\int \exp(t\beta) dF(t)}.$$

Let $t_1 < \dots < t_h$ be the pooled data. Write

$$c(\beta) = \sum_{j=1}^h \exp(t_j \beta), \quad \pi = \sum_{j=1}^h p_j \exp(t_j \beta) / c(\beta).$$

As we did for length biased sampling problems using the EM algorithm in Sect. 10.1, we can create a new random variable Y with probability masses

$$P(Y = t_i) = p_i, \quad i = 1, 2, \dots, h.$$

Define another independent random variable A with masses

$$P(A = t_i) = \frac{\exp(t_i \beta)}{c}, \quad c = \sum_{i=1}^h \exp(t_i \beta) \quad i = 1, 2, \dots, h.$$

We observe Y if and only if $A = Y$. Note that

$$P(Y = t_i | A = Y) = \frac{P(Y = A = t_i)}{P(A = Y)} = \frac{\exp(t_i \beta) p_i}{\sum_{j=1}^h \exp(t_j \beta) p_j}, \quad i = 1, 2, \dots, h.$$

In the absence of truncation, the log-full likelihood is

$$\begin{aligned} \ell = & \sum_{j=1}^h \left[\left\{ \sum_{i=1}^m I(Z_i = t_j) \right\} \log p_j + \left\{ \sum_{i=1}^n I(Y_i = A_i = t_j) \{ \log p_j + t_j \beta - \log c(\beta) \} \right\} \right] \\ & + \sum_{j=1}^h \sum_{i=1}^{n^*} I(Y_i^* = t_j) I(A_i^* \neq t_j) \{ \log p_j + \log(1 - \exp(t_j \beta) / c(\beta)) \}, \end{aligned}$$

where n^* is the number of unobserved (Y^*, A^*) 's with $Y^* \neq A^*$. Denote

$$\xi_j = \sum_{i=1}^m I(Z_i = t_j), \quad \eta_j = \sum_{i=1}^n I(Y_i = t_j),$$

Also denote the observed data as O . The conditional expectation is

$$\begin{aligned} a_j &= E[\sum_{i=1}^{n^*} I(Y_i^* = t_j) I(A_i^* \neq t_j) | O] = n \frac{1 - \pi}{\pi} \frac{p_j (1 - \exp(t_j \beta) / c(\beta))}{1 - \pi} \\ &= \frac{n}{\pi} p_j (1 - \exp(y_j \beta) / c). \end{aligned}$$

Denote

$$w_j = \xi_j + v_j, \quad v_j = \eta_j + \frac{n}{\pi} p_j (1 - \exp(y_j \beta) / c),$$

where ξ_j is the observed frequency at t_j from sample Z_i 's, and v_j is the summation of observed and imputed frequencies from sample Y_i 's. Note that

$$w_+ = \sum_{j=1}^h w_j = m + n/\pi.$$

Using the same arguments as in length biased sampling problem with EM algorithm in Chap. 10, we have the expected log-likelihood

$$\begin{aligned}\ell &= \sum_{j=1}^h [w_j \log p_j + \eta_j t_j \beta] - n \log c(\beta) + \sum_{j=1}^h a_j \log[1 - \exp(t_j \beta)/c(\beta)] \\ &= \sum_{j=1}^h [w_j \log p_j + \eta_j t_j \beta] - \frac{n}{\pi} \log c(\beta) + \sum_{j=1}^n a_j \log\{c(\beta) - \exp(t_j \beta)\}.\end{aligned}$$

Clearly the maximum likelihood estimates satisfy

$$p_j = \frac{w_j}{w_+} = \frac{\xi_j + \eta_j + np_j(1 - \exp(y_j \beta)/c)/\pi}{m + n/\pi},$$

or

$$p_j = \frac{\xi_j + \eta_j}{m + n \exp(y_j \beta)/c/\pi} = \frac{1}{m} \frac{\xi_j + \eta_j}{1 + \rho \exp(\alpha + y_j \beta)},$$

where

$$\alpha = -\log(c\pi).$$

This exactly matches the previous results in method I. Moreover the score equation is

$$\frac{\partial \ell}{\partial \beta} = \sum_{j=1}^h \eta_j t_j - n/\pi \frac{\sum_{j=1}^h t_j \exp(t_j \beta)}{\sum_{j=1}^h \exp(t_j \beta)} + \sum_{j=1}^n a_j \frac{C'(\beta) - t_j \exp(t_j \beta)}{C(\beta) - \exp(t_j \beta)} = 0.$$

Noting

$$a_j = \frac{n}{\pi c} p_j \{c(\beta) - \exp(t_j \beta)\},$$

we have

$$\begin{aligned}\sum_{j=1}^n a_j \frac{C'(\beta) - t_j \exp(t_j \beta)}{C(\beta) - \exp(t_j \beta)} &= \frac{n}{\pi c} \sum_{j=1}^h p_j \{c'(\beta) - t_j \exp(t_j \beta)\} \\ &= \frac{n}{\pi c} c'(\beta) - \frac{n}{\pi c} \sum_{j=1}^h p_j t_j \exp(t_j \beta).\end{aligned}$$

Therefore the score equation for β is

$$\begin{aligned}
 \frac{\partial \ell}{\partial \beta} &= \sum_{j=1}^h \eta_j t_j - \frac{n}{\pi c} c'(\beta) + \frac{n}{\pi c} c'(\beta) - \frac{n}{\pi c} \sum_{j=1}^h p_j t_j \exp(t_j \beta) \\
 &= \sum_{j=1}^h \eta_j t_j - \frac{n}{\pi c} \sum_{j=1}^h p_j t_j \exp(t_j \beta) \\
 &= \sum_{j=1}^h \eta_j t_j - n \sum_{j=1}^h p_j t_j \exp(\alpha + t_j \beta) \\
 &= \sum_{j=1}^h \eta_j t_j - \sum_{j=1}^h \frac{(\xi_j + \eta_j) \rho t_j \exp(\alpha + t_j \beta)}{1 + \rho \exp(\alpha + t_j \beta)}.
 \end{aligned}$$

Again this is precisely the same estimating equation as derived in method I.

The EM algorithm derived score for the semiparametric model is rarely seen in the statistical literature, nevertheless it provides a valuable tool for us to find maximum score estimating equations.

Logistic Discrimination Versus Normal Discrimination

As observed before, in the standard logistic regression analysis the marginal distribution of covariate X is not specified. The full likelihood is

$$L = \left\{ \prod_{i=1}^n \frac{\exp(D_i x_i \beta)}{1 + \exp(D_i x_i \beta)} \right\} \prod_{i=1}^n dF(x_i),$$

where the first factor is the conditional likelihood of D_i given X_i . In comparison with the full normal discrimination analysis where $F(x|D=0)$ and $F(x|D=1)$ are normal distribution functions, the logistic regression analysis does not use the baseline normal density information. As a consequence it may lead to a loss of information. Efron (1975) showed that when the marginal information is used, the logistic discrimination is between 1/2 and 2/3 as effective as normal discrimination. Therefore, normal discrimination analysis is preferred if this information is available. However, logistic discrimination is more robust against departure from the normal distribution.

11.4 Family-Based Case-Control Studies

The case-control design was adapted to family-based case-control studies by Whittemore (1995). Similar to the conventional case-control design, in the family-based case-control design, again fixed numbers of individuals with disease or without disease are accrued. For each identified individual (called a proband), the information

of his/her environment covariates, family structure, and the disease status and covariates of his/her relatives is collected. As a consequence the numbers of disease and disease free for relatives are random. For simplicity we assume that each family has two members. Let $y = (y_1, y_2)$ and $z = (z_1, z_2)$ be disease status and covariates for the proband and relative, respectively. Suppose the probability of disease occurrence in the population satisfies a parametric model

$$P(y|z) = P(y_1, y_2|z_1, z_2, \theta).$$

A common choice of the probability model is the bivariate logistic regression model. We do not need to assume

$$P(Y_i = 1|z) = P(Y_i = 1|z_i), \quad i = 1, 2,$$

that is the regression parameters have the same interpretation in the full model as in marginal distributions obtained by summing over the response of one individual. Whittemore (1995) made this assumption and discussed its rationality, and argued how this assumption would be inappropriate in some situations.

The family case-control study involves two separate samples, one from $P(Y_2, Z = z|Y_1 = 1)$ and one from $P(Y_2, Z = z|Y_1 = 0)$. Note that the marginal model $P(Y_1 = 1|z)$ can be derived from the joint model through

$$P(Y_1 = 1|z, \theta) = P(Y_1 = 1, Y_2 = 0|z, \theta) + P(Y_1 = 1, Y_2 = 1|z, \theta)$$

Without loss of generality we may write

$$\frac{P(Y_1 = 1|z)}{P(Y_1 = 0|z)} = \exp\{\alpha^* + \phi(z, \beta)\},$$

where β is a $p \times 1$ subset parameter of θ and α^* is a scale parameter.

Denote $\pi = P(Y_1 = 1)$ and let $f(z)$ be the marginal density of z . We have

$$f(z|y_1 = 1) = P_\theta(y_1 = 1|z)f(z)/\pi, \quad f(z|y_1 = 0) = P_\theta(y_1 = 0|z)f(z)/(1 - \pi),$$

where

$$\begin{aligned} P_\theta(y_1 = 1|z) &= P_\theta(y_1 = 1, y_2 = 0|z) + P_\theta(y_1 = 1, y_2 = 1|z), \\ P_\theta(y_1 = 0|z) &= 1 - P_\theta(y_1 = 1|z). \end{aligned}$$

Hence

$$\begin{aligned} f(z|y_1 = 1) &= \frac{1 - \pi}{\pi} \frac{P_\theta(y_1 = 1|z)}{P_\theta(Y_1 = 0|z)} f(z|y_1 = 0) \\ &= \frac{1 - \pi}{\pi} \exp\{\alpha^* + \phi(z; \beta)\} f(z|y_1 = 0) \\ &= w(z; \alpha, \beta) f(z|y_1 = 0), \end{aligned}$$

where

$$w(z; \alpha, \beta) = \exp\{\alpha + \phi(z; \beta)\}, \quad \alpha = \alpha^* + \log\{(1 - \pi)/\pi\}.$$

Denote the observed case data as $(y_{1i} = 1, y_{2i}, z_i), i = 1, 2, \dots, n_1$ and control data as $(y_{1j} = 0, y_{2j}, z_j), j = n_1 + 1, \dots, n_1 + n_0 = n$, respectively. Note that

$$P(Y_2 = y_2, z|y_1 = 0) = f(z|y_1 = 0) P_\theta(y_2|y_1 = 0, z), \quad (11.4.14)$$

$$\begin{aligned} P(Y_2 = y_2, z|y_1 = 1) &= f(z|y_1 = 1) P_\theta(y_2|y_1 = 1, z) \\ &= w(z; \alpha, \beta) P_\theta(y_2|y_1 = 1, z) f(z|y_1 = 0). \end{aligned} \quad (11.4.15)$$

Based on the observed data, the likelihood function is

$$\begin{aligned} L &= \prod_{i=1}^{n_1} P(y_{2i}, z_i|y_{1i} = 1) \prod_{j=n_1+1}^n P(y_{2j}, z_j|y_{1j} = 0) \\ &= \prod_{i=1}^{n_1} \{f(z_i|y_{1i} = 1) P_\theta(y_{2i}|y_{1i} = 1, z_i)\} \prod_{j=n_1+1}^n \{f(z_j|y_{1j} = 0) P_\theta(y_{2j}|y_{1j} = 0, z_j)\} \\ &= \prod_{i=1}^{n_1} \{w(z_i; \alpha, \beta) P_\theta(y_{2i}|y_{1i} = 1, z_i) dG(z_i)\} \prod_{j=n_1+1}^n \{P_\theta(y_{2j}|y_{1j} = 0, z_j) dG(z_j)\}, \end{aligned}$$

where $n = n_1 + n_0$ and $dG(z) = f(z|y_1 = 0) dz$. Therefore the log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^{n_1} \log w(z_i; \alpha, \beta) + \log P_\theta(y_{2i}|y_{1i} = 1, z_i) + \sum_{j=n_1+1}^n \log P_\theta(y_{2j}|y_{1j} = 0, z_j) \\ &\quad + \sum_{i=1}^{n_1} \log p_i, \end{aligned}$$

where

$$p_i = dG(z_i) \geq 0, \quad i = 1, 2, \dots, n, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \{w(z_i; \alpha, \beta) - 1\} = 0.$$

For fixed $(\alpha, \alpha^*, \beta)$, maximizing ℓ with respect to p_i , we have

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda \{w(z_i; \alpha, \beta) - 1\}}, \quad (11.4.16)$$

where λ is the Lagrange multiplier which is determined by

$$\frac{1}{n} \sum_{i=1}^n \frac{w(z_i) - 1}{1 + \lambda \{w(z_i) - 1\}} = 0. \quad (11.4.17)$$

Plugging the values of p_i 's into the log-likelihood, we have a profile log-likelihood

$$\begin{aligned} \ell &= \sum_{i=1}^{n_1} \log w(z_i; \alpha, \beta) + \log P_\theta(y_{2i} | y_{1i} = 1, z_i) + \sum_{j=n_1+1}^n \log P_\theta(y_{2j} | y_{1j} = 0, z_j) \\ &\quad - \sum_{i=1}^n \log[1 + \lambda \{w(z_i; \alpha, \beta) - 1\}]. \end{aligned} \quad (11.4.18)$$

Let $(\tilde{\lambda}, \tilde{\alpha}, \tilde{\theta})$ maximize ℓ . Taking derivative with respect to α , we have

$$\frac{\partial \ell}{\partial \alpha} = n_1 - \sum_{i=1}^n \frac{\lambda w}{1 + \lambda(w - 1)} = 0.$$

Hence

$$\lambda = n_1/n, \quad p_i = \frac{1}{n_0} \frac{1}{1 + \rho \exp(\alpha + \phi(z_i, \beta))}.$$

After plugging $\lambda = n_1/n$ in ℓ , the log profile likelihood function is

$$\begin{aligned} \ell &= \sum_{i=1}^{n_1} \log P_\theta(y_{2i} | y_{1i} = 1, z_i) + \sum_{j=n_1+1}^n \log P_\theta(y_{2j} | y_{1j} = 0, z_j) \\ &\quad + \sum_{i=1}^{n_1} \{\alpha + \phi(z_i, \beta)\} - \sum_{i=1}^n \log[1 + \rho \exp(\alpha + \phi(z_i, \beta))]. \end{aligned} \quad (11.4.19)$$

Differentiating ℓ with respect to $(\alpha, \alpha^*, \beta)$, we have estimating equations

$$\frac{\partial \ell}{\partial \alpha} = n_1 - \sum_{i=1}^n \frac{\rho \exp(\alpha + \phi(z_i, \beta))}{1 + \rho \exp(\alpha + \phi(z_i, \beta))} = 0,$$

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha^*} &= \sum_{i=1}^{n_1} \frac{\partial \log P_\theta(y_{2i}|y_{1i}=1, z_i)}{\partial \alpha^*} + \sum_{j=n_1+1}^n \frac{\partial \log P_\theta(y_{2j}|y_{1j}=0, z_j)}{\partial \alpha^*} = 0, \\ \frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{n_1} \frac{\partial \log P_\theta(y_{2i}|y_{1i}=1, z_i)}{\partial \beta} + \sum_{j=n_1+1}^n \frac{\partial \log P_\theta(y_{2j}|y_{1j}=0, z_j)}{\partial \beta} \\ &\quad + \sum_{i=1}^{n_1} \frac{\partial \phi(z_i, \beta)}{\partial \beta} - \sum_{i=1}^n \frac{\rho \exp(\alpha + \phi(z_i, \beta))}{1 + \rho \exp(\alpha + \phi(z_i, \beta))} \frac{\partial \phi(z_i, \beta)}{\partial \beta} = 0. \end{aligned}$$

Whittemore (1995) showed that the case-control family data can be analyzed as if they were obtained from a prospective study, with the baseline disease probabilities of case and control probands differing from that of their relatives. Readers should be able to verify her results by working out the details.

Note that $P(y_2, z|y_1 = 1)$ and $P(y_2, z|y_1 = 0)$ are linked through

$$P(y_2, z|y_1 = 1) = \left\{ w(z; \alpha, \beta) \frac{P_\theta(y_2|y_1 = 1, z)}{P_\theta(y_2|y_1 = 0, z)} \right\} P(y_2, z|y_1 = 0).$$

Hence for any η ,

$$E\{\eta(y_2, z)|y_1 = 1\} = E \left\{ \eta(y_2, z) w(z; \alpha, \beta) \frac{P_\theta(y_2|y_1 = 1, z)}{P_\theta(y_2|y_1 = 0, z)} | y_1 = 0 \right\}.$$

We can construct a class of estimating functions

$$\mathcal{Q} = \left\{ Q_n(\eta) = \frac{1}{n_1} \sum_{i=1}^{n_1} \eta(y_{2i}, z_i) - \frac{1}{n_0} \sum_{j=n_1+1}^n \eta(y_{2j}, z_j) w(z_j; \alpha, \beta) \frac{P_\theta(y_{2j}|y_{1j}=1, z_j)}{P_\theta(y_{2j}|y_{1j}=0, z_j)} \right\}.$$

Note that $A(t) = P(y_2 = 1, z \leq t|y_1 = 0)$ can be estimated by

$$A_{1n}(t) = \sum_{i=1}^n \tilde{p}_i P_{\tilde{\theta}}(y_2 = 1|y_1 = 0, z_i) I(z_i \leq t).$$

where

$$\tilde{p}_i = \frac{1}{n_0} \frac{1}{1 + \rho \exp\{\tilde{\alpha} + \phi(z_i; \tilde{\beta})\}} \quad (11.4.20)$$

Without parametric model assumption, the empirical distribution estimation of $A(t)$ is

$$A_{2n}(t) = \frac{1}{n_0} \sum_{j=n_1+1}^n y_{2j} I(z_j \leq t).$$

Define a stochastic process

$$U_n(t) = \sqrt{n}\{A_{1n}(t) - A_{2n}(t)\}. \quad (11.4.21)$$

Then $\sup_t |U_n(t)|$ can be used to measure the departure from probability model assumption.

Exercise 1 Under suitable regularity conditions, show that

$$\sqrt{n} \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\alpha}^* - \alpha_0^* \\ \tilde{\beta} - \beta_0 \end{pmatrix} \rightarrow N(0, V),$$

and find an expression for V .

Exercise 2 Under suitable regularity conditions, show that the score estimating equations are optimal in the class \mathcal{Q}

Exercise 3 The likelihood ratio statistic for β is defined as

$$R(\beta) = 2\{\max_{\alpha, \alpha^*, \beta} \ell(\alpha, \alpha^*, \beta) - \max_{\alpha, \alpha^*} \ell(\alpha, \alpha^*, \beta)\}.$$

Then show that as $n \rightarrow \infty$, $R(\beta_0)$ converges to a χ^2 random variable.

Exercise 4 Consider a bivariate logistic model $P(y_1, y_2|z, \theta)$. The collection of covariate Z is based on the status of Y_1 and Y_2 ,

$$z_{ij1}, \dots, z_{ijk_{ij}} \sim P(Z|Y_1 = i, Y_2 = j), \quad i = 0, 1; \quad j = 0, 1,$$

where k_{ij} , $i, j = 0, 1$ are fixed numbers. Derive the details based on the bivariate case and control data.

Existence of Solution in Logistic Regression Model

Silvapulle (1981) gave necessary and sufficient conditions for the existence of the MLE of the linear parameter in binomial responses, including the Logistic and Probit models.

Define the relative interiors of the convex cones generated by case data x_1, \dots, x_{n_1} and control data x_{n_1+1}, \dots, x_n as S and F , respectively, where

$$S = \left\{ \sum_{i=1}^{n_1} k_i x_i | k_i > 0 \right\}, \quad F = \left\{ \sum_{i=n_1+1}^n k_i x_i | k_i > 0 \right\}.$$

The basic requirement for the existence of a finite solution is the intersection of S and F is not empty.

Albert and Anderson (1984) classified logistic regression data into three mutually exclusive and exhaustive categories: (1) Complete separation, (2) Quasi-separation and (3) Overlap. The MLE exists only if the observed data are in case (3), i.e., overlap between cases and controls. More detailed discussions can be found in Santner and Duffy (1986).

Exercise 1 It was shown (Anderson et al.) that there is no solution when the case and control data are completely separated. In that case, can we swap one case and control data to get a solution? Can Chen et al. (2002) method of adding one artificial data point for empirical likelihood be used here?

11.5 Genetic Liability Model or Probit Model

For binary response data, the logistic regression model is a standard method in epidemiology studies. The liability threshold model introduced by Pearson and Lee (1901) is very popular in genetic study due to its nice interpretation. The liability-threshold model assumes an underlying continuous variable called liability, Z , that has, or can be transformed to have a normal distribution with mean 0 and variance 1 in the general population. The disease is assumed to be present in all individuals whose liability is above a certain threshold t . Let $D = 1$ if $Z > t$ and $D = 0$ otherwise. The disease prevalent probability π is

$$\pi = P(Z > t) = 1 - \Phi(t),$$

or $t = \Phi^{-1}(1 - \pi)$.

If a population is stratified into subgroups according to the presence or absence of risk factors that may be genetic or environmental, denote by Z_i the liability in subgroup i . Suppose $Z_i \sim N(\mu_i, 1)$. Then the risk in that stratum is

$$\pi_i = P(Z_i > t) = 1 - \Phi(t - \mu_i)$$

In regression analysis, denote X as the covariate information and $\mu_i = \alpha^* + x_i\beta^*$.

$$P(D_i = 1|x_i) = P(Z_i > t|x_i) = P\{Z_i - E(Z_i|x_i) > t - E(Z_i|x_i)\} = 1 - \Phi(\alpha + x_i\beta),$$

where $\alpha = t - \alpha^*$ and $\beta = -\beta^*$.

If data are collected prospectively, then α and β may be estimated by maximizing the binomial likelihood

$$\prod_{i=1}^n \{1 - \Phi(\alpha + x_i\beta)\}^{D_i} \{\Phi(\alpha + x_i\beta)\}^{1-D_i},$$

either directly or by using the EM algorithm.

Now suppose the data collection is retrospective. Denote respectively the case data as

$$X_{11}, \dots, X_{1n_1} \sim f(x|D = 1),$$

and the control data as

$$X_{01}, \dots, X_{0n_0} \sim f(x|D = 0).$$

How do we make inference for α and β ? Can we use prospective Probit likelihood to make inference for β even if the data are collected retrospectively?

Using Bayes' formula, we have

$$f(x|D = 1) = \frac{\{1 - \Phi(\alpha + x\beta)\}f(x)}{P(D = 1)}, \quad f(x|D = 0) = \frac{\Phi(\alpha + x\beta)f(x)}{P(D = 0)},$$

where $f(x)$ is the marginal density of X . The density ratio is

$$f(x|D = 1)/f(x|D = 0) = \frac{P(D = 0)}{P(D = 1)} \frac{\{1 - \Phi(\alpha + x\beta)\}}{\Phi(\alpha + x\beta)},$$

or

$$f(x|D = 1)/f(x|D = 0) = \exp\{\eta + \psi(x, \alpha, \beta)\},$$

where

$$\eta = \log \frac{P(D = 0)}{P(D = 1)}, \quad \psi(x, \alpha, \beta) = \log\{1 - \Phi(\alpha + x\beta)\} - \log \Phi(\alpha + x\beta).$$

Similar to previous approach, the log profile likelihood is

$$\ell = \sum_{i=1}^{n_1} \{\eta + \phi(x_i, \beta)\} - \sum_{i=1}^n \log[1 + \rho_n \exp(\eta + \phi(x_i, \beta))],$$

where $\rho_n = n_1/n_0$, and $(x_1, \dots, x_n) = (x_{11}, \dots, x_{1n_1}; x_{01}, \dots, x_{0n_0})$ are pooled data.

Due to model identification problem, in general it would be difficult to simultaneously maximize this profile likelihood with respect to (η, α, β) . However it would be helpful if there is prior information that the disease prevalence $P(D = 1)$ falls in a certain range. We can then profile out β and then search for the maximum of $\ell(\eta, \hat{\beta}(\eta))$ in the interval (η_0, η_1) . The more precise the auxiliary information (η_0, η_1) , the more accurate will be the estimation of β . There are three approaches for utilization of auxiliary:

(1) Restrict $\eta \in (\eta_0, \eta_1)$.

(2) Suppose the information on the number of $D = 1$ and $D = 0$ based on an independent study is available, then we can consider an augmented likelihood

$$\ell_A = \sum_{i=n_0+1}^n \{\eta + \phi(x_i, \beta)\} - \sum_{i=1}^n \log[1 + \rho_n \exp(\eta + \phi(x_i, \beta))] + m_1 \log \pi + m_0 \log(1 - \pi),$$

where $\pi = P(D = 1) = 1/\{1 + \exp(\eta)\}$.

(3) There is prior information on $\eta \sim h(\eta)$. Then we can maximize the augmented log-likelihood

$$\ell_A = \sum_{i=n_0+1}^n \{\eta + \phi(x_i, \beta)\} - \sum_{i=1}^n \log[1 + \rho_n \exp(\eta + \phi(x_i, \beta))] + \log h(\eta),$$

where $h(\eta)$ is a prior distribution of η .

Exercise Give details for the three approaches given above.

11.6 Miscellaneous Problems

Below we discuss different applications of the density ratio model or exponential tilting model in solving some seemly unrelated statistical problems.

Applications for One Way Layout

As shown before the exponential tilting model is equivalent to the logistic regression model under retrospective sampling. Interestingly, Fokianos et al. (2001) applied this model in one way layout problems. In classical analysis of variance, the response variable Y and group variable $Z = 1, 2, \dots, m$ are linked by

$$Y|Z \sim N(\mu_k, \sigma^2).$$

Denote $f_k(y)$ as the density of Y conditional on $Z = k$, then

$$f_k(y) = \exp(\alpha_k + y\beta_k) f_1(y), \quad k = 2, \dots, m,$$

where $f_1(x)$ is the density of a normal variable with mean μ_1 and variance σ^2 . If we do not specify the form of $f_1(y)$ then we have a multi-sample semiparametric exponential tilting model. We are interested in testing $H_0 : \beta_2 = \beta_3 = \dots = \beta_m = 0$.

Exercise Derive the maximum likelihood estimation in the semiparametric one way layout model. Furthermore extend the one-way anova to the two-way anova.

Importance Sampling and Numerical Integrations

Suppose X_1, \dots, X_n are a i.i.d. sample from

$$X \sim h(x) = \frac{f(x, \beta)}{\int f(x, \beta) dx},$$

where $f(x, \beta)$ is a given non-negative function. For example

$$X \sim \text{beta}(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}.$$

Suppose the integral has no close form. The following method may be used to find the maximum likelihood estimate without performing the integration.

Generate data

$$y_1, \dots, y_m \text{ i.i.d } g(y),$$

where g is a known density and m can be chosen as large as possible. Observe

$$h(x)/g(x) = \exp\{\alpha + \psi(x, \beta)\},$$

where α is a normalizing constant and

$$\psi(x, \beta) = \log f(x, \beta) - \log g(x).$$

Methods discussed in the two-sample density ratio model can be applied to find the semiparametric likelihood estimation of β .

Exercise (1) Perform numerical calculations. (2) Compare this with the maximum likelihood estimate by working out the integration. (3) Discuss the choice of $g(y)$.

The Combination of Importance Sampling and Side Information

For some function of X , say, $\psi(X)$, if one may easily work out the expectation $E_g[\psi(X)]$, say, $E_g[\psi(X)] = a$. Then one may impose the additional constraint $E_g[\psi(X) - a] = 0$ in searching for the semiparametric MLE.

Exercise Compare the efficiency for the constrained and unconstrained semiparametric MLEs.

Two-Stage Sampling

Wild (1991) and Scott and Wild (1997) studied two-stage sampling based on a logistic regression model. Suppose that a random sample with size N is drawn from the joint distribution (D, X) . If full data on all N individuals are available, then a prospective likelihood can be used. However, if all that is known about the N individuals is that there are N_i individuals with $D = i$, $i = 0, 1$ (This can be generalized to more than two categories). From each response class $D = i$, $i = 0, 1$, a simple random sample of size n_i is taken ($1 < n_i \leq N_i$), and X s are measured for all sampled individuals. Denote the observed control and case data as x_{0j} , $j = 1, 2, \dots, n_0$ and x_{1j} , $j = 1, 2, \dots, n_1$, respectively. The likelihood is

$$L = \left[\prod_{i=0}^1 \left\{ \prod_{j=1}^{n_i} P(X = x_{ij} | D = i) \right\} \right] P^{N_i}(D = i).$$

Clearly $P(Y = i)$ can be estimated by $N_i/(N_1 + N_2)$. The first part can be treated as a retrospective likelihood. In this case the information of $N_i, i = 0, 1$ is not useful for estimating the odds ratio parameter β . It only helps to identify the intercept in the retrospective likelihood.

Choice Based Sampling Problem in Econometrics

Suppose the discrete response variable D takes J possible values $1, 2, \dots, J$. The basic model assumption is

$$P(D = j|x) = P(D = j|x, \beta) = \psi_j(x, \beta), j = 1, 2, \dots, J,$$

where the marginal distribution of $X \sim f(x)$ is unspecified. Let

$$q_i = P(D = i), j = 1, 2, \dots, J$$

be the marginal distribution of D . Under random sampling, the likelihood can be written as

$$L = \prod_{i=1}^n \left\{ \prod_{j=1}^J \psi_j^{I(D_i=j)}(x_i, \beta) \right\} dF(x_i) = \prod_{i=1}^n \left\{ \prod_{j=1}^J q_i^{I(D_i=j)} f(x_i|D = j, \beta) \right\}.$$

Instead of simple random sampling, we first sample D with another marginal distribution $P(D = j) = p_j, j = 1, 2, \dots, J$. Then given $D = j$, we sample X from the conditional density $f(x|D = j)$. Denote the observed data as $D_i, x_i, i = 1, 2, \dots, n$. The likelihood is

$$L = \prod_{i=1}^n \prod_{j=1}^J \{p_j f(x_i|D = j, \beta)\}^{I(D_i=j)}.$$

It is possible to maximize this likelihood. Details was given in Cosslett (1981). Readers may follow the arguments in Sects. 11.1 and 11.2 of this chapter to derive the semiparametric MLE.

Missing Data Problems

In his thesis, Qin (1992) considered a missing data problem, where both the binary response binary D and covariate X can be missing. The observed data can be denoted as

$$(D_i, x_i), i = 1, 2, \dots, n_1, (D_{n_1+i}, ?), i = 1, 2, \dots, n_2, (?, x_{n_1+n_2+i}), i = 1, 2, \dots, n_3$$

If a logistic model is assumed, it is possible to find the semiparametric likelihood estimation based on the above incomplete data.

Exercise Give details of the analysis in this model.

Combine Prospective and Retrospective Studies

Based on the results discussed so far, it would be not difficult to combine prospectively collected data and retrospectively collected data.

Suppose we have two data sets. Let

$$(D_i, X_i), i = 1, 2, \dots, m$$

be the prospective sampling data. Also denote the retrospective sampling data as

$$X_{01}, \dots, X_{0n_0} \sim f(x|D=0), \quad X_{11}, \dots, X_{1n_1} \sim f(x|D=1).$$

where n_0 and n_1 are fixed. Again the logistic regression model $P(D=0|x) = [1 + \exp(\alpha + x\beta)]^{-1}$ is equivalent to

$$f(x|D=1) = \exp(\alpha^* + x\beta)f(x|D=0), \quad \alpha^* = \alpha + \log\{(1-\pi)/\pi\}, \quad \pi = P(D=1).$$

Without loss of generality we assume $D_i = 1$ for the first m_1 observations in prospective sampling and $D_i = 0$ for the rest. Denote $m_0 = m - m_1$. Therefore the prospective likelihood is

$$L_P = \prod_{i=1}^m P(D_i) f(x_i|D_i) = \pi^{m_1} (1-\pi)^{m_0} \prod_{i=1}^{m_1} \exp(\alpha^* + x_i\beta) \prod_{i=1}^m dF(x_i|D=0).$$

The retrospective likelihood contribution is

$$L_R = \prod_{i=1}^{n_1} \exp(\alpha^* + x_{1i}\beta) \prod_{i=1}^{n_1} dF(x_{1i}|D=0) \prod_{i=1}^{n_0} dF(x_{0i}|D=0).$$

The overall likelihood is the product of L_P and L_R . Note that the disease prevalence π can also be estimated.

Exercise Derive the large sample results.

Exercise Consider a two sample density ratio model with moment constraints. Suppose

$$X_1, \dots, X_m \sim f_1(x), \quad Y_1, \dots, Y_n \sim f_2(x) = \phi(x)f_1(x),$$

where the form of $\phi(x)$ is known or known up to some finite parameters. In addition we have prior information such as

$$E_1[\psi(X, \theta)] = 0.$$

Denote the pooled data as $(t_1, \dots, t_N) = (y_1, \dots, y_n; x_1, \dots, x_m)$. Let $\hat{q}_i, i = 1, 2, \dots, N = m + n$ maximize

$$\sum_{i=1}^N \log q_i + \sum_{i=1}^n \log \phi(y_i)$$

subject to the constraints

$$q_i \geq 0, \quad \sum_{i=1}^N q_i = 1, \quad \sum_{i=1}^N q_i \phi(y_i) = 1, \quad dF(t_i) = q_i.$$

On the other hand, maximizing the Kullback–Leibler information subject to the moment constraint $\int \psi(x, \theta) dF_1(x) = 0$, we have the entropy family given by (Chap. 9) that

$$f_1^*(y) = \frac{\exp\{\tau\psi(y, \theta)\} f_1(y)}{\int \exp\{\tau\psi(y, \theta)\} f_1(y) dy},$$

Replacing $dF_1(t_i)$ by \hat{q}_i in the entropy family, we have the log entropy likelihood

$$\ell_{EN} = \sum_{i=1}^n \tau\psi(y_i, \theta) - n \log \left[\sum_{j=1}^N \hat{q}_j \exp\{\tau\psi(t_j, \theta)\} \right].$$

Derive large sample results.

As an alternative, we may directly maximize

$$\prod_{i=1}^m dF_1(x_i) \prod_{j=1}^n \phi(y_j) dF_1(y_j)$$

subject to the constraints

$$\sum_{i=1}^N \phi(t_i) p_i = 1, \quad \sum_{j=1}^N p_j = 1, \quad p_j \geq 0, \quad \sum_{i=1}^N p_i \psi(y_i, \theta) = 0,$$

Derive large sample results and then compare the two approaches.

Using Pooled Exposure Assessment in Case-Control Studies

In epidemiological studies, sometimes it is impossible to perform assay analysis on each biological specimen due to the fact that assays can be too expensive. Weinberg and Umbach (1999) proposed to pool equal volume aliquots from randomly grouped sets of cases and controls. It would be interesting to develop methods to analyze the pooled case-control data.

For ease of exposition, we assume two cases are pooled together and two controls are also pooled together. Again we assume case and control densities are linked by the exponential tilting model

$$f_1(x) = f_0(x) \exp(\alpha + x\beta),$$

where $f_1(x)$ and $f_0(x)$ are case and control densities, respectively. For the pooled case $T = X_1 + X_2$, the density is

$$\begin{aligned} \frac{P(T = t | D = 1)}{dt} &= \int f_1(t - x_1) f_1(x_1) dx_1 \\ &= \int f_0(t - x_1) \exp\{\alpha + (t - x_1)\beta\} f_0(x_1) \exp(\alpha + x_1\beta) dx_1 \\ &= \exp(2\alpha + t\beta) \int f_0(t - x_1) f_0(x_1) dx_1. \end{aligned}$$

Note that $\int f_0(t - x_1) f_0(x_1) dx_1$ is the density for the pooled controls. Let

$$g_1(t) = \int f_1(t - x_1) f_1(x_1) dx_1$$

and

$$g_0(t) = \int f_0(t - x_1) f_0(x_1) dx_1.$$

Then

$$g_1(t) = \exp(2\alpha + t\beta) g_0(t).$$

In other words, the exponential tilting model applies to the convolution densities of cases and controls. Then β can be estimated by the standard logistic regression method.

Chapter 12

Conditioning Approach for Discrete Outcome Problems

In this chapter we study conditional likelihood-based inference in discrete outcome problems. This method is very useful for sparse data where there exists a large number of nuisance parameters. Moreover it is used extensively in matched case-control studies where some baseline covariates or survival times are matched at the data collection stage.

12.1 Eliminate Nuisance Parameters in Logistic Regression Models

First we present Farewell's (1979) work for the logistic regression model with case-control sampling. In this example the intercept in the logistic regression is treated as a nuisance parameter and the odds ratio is the parameter of interest.

Without loss of generality, assume the first n_0 observations are controls and the remaining n_1 observations are cases. Let $D = 1$ denote a case and $D = 0$ otherwise. Furthermore, let X be a covariate and x_1, \dots, x_n denote its value on the $n = n_0 + n_1$ observations, where n_0 and n_1 are numbers of controls and cases, respectively. The likelihood conditioning on the n observations is

$$L = \prod_{i=1}^{n_0} f(x_i | D = 0) \prod_{j=n_0+1}^{n_0+n_1} f(x_j | D = 1),$$

where $f(x | D = 1)$ and $f(x | D = 0)$ are the densities of X in case and control groups, respectively. Given the observed values x_1, \dots, x_n and n but without knowledge of the case-control status of each observation, the conditional likelihood for the observed n_1 cases and n_0 controls is

$$L_C = \left[\prod_{i=1}^{n_0} f(x_i | D=0) \prod_{j=n_0+1}^n f(x_j | D=1) \right] / \left\{ \sum_{i_1, \dots, i_n} \prod_{j=1}^{n_0} f(x_{i_j} | D=0) \prod_{k=n_0+1}^n f(x_k | D=1) \right\},$$

where i_1, \dots, i_n range over $n!/(n_0!n_1!)$ permutations of x_1, \dots, x_n in the n_1 cases and n_0 controls. Under the logistic regression model assumption, we have the density ratio model

$$f(x | D=1) = f(x | D=0) \exp(\alpha^* + x\beta).$$

It can easily be observed that $f(x | D=0)$ and $\exp(\alpha^*)$ are cancelled out in the likelihood. The conditional likelihood becomes

$$\prod_{j=n_0+1}^n \exp(x_j \beta) / \left[\sum_{i_1, \dots, i_n} \sum_{k=n_0+1}^n \exp(x_{i_k} \beta) \right].$$

The summation is possible numerically only in situations where n_0 and n_1 are small or X is categorical. For any moderate values of n_0 and n_1 , the summation is not practically feasible. In stratified analyses, this method is preferred to the profile likelihood approach discussed in last chapter (Farewell 1979). First of all, it is an effective way to eliminate the intercept and baseline density from the analysis. Furthermore, in many stratified studies, as the number of strata increases, the limited sample size within each stratum may not be large enough to perform a valid profile likelihood analysis. In fact, in survival analysis, the Cox partial likelihood method is exactly a conditional likelihood approach by appropriately defining “cases” and “controls” at each failure time point. We will discuss this in Chaps. 24 and 25.

Next we consider the situation with two covariates X and Z in the logistic regression model

$$P(Y=1|x, z) = \frac{\exp(\alpha + x\beta + z\gamma)}{1 + \exp(\alpha + x\beta + z\gamma)}.$$

Suppose we are interested in making inference for β by treating α and γ as nuisance parameters.

Based on the observed data, the likelihood is

$$\begin{aligned} L &= P(Y_1 = y_1, \dots, Y_n = y_n | x_1, \dots, x_n; z_1, \dots, z_n) \\ &= \frac{\exp(\sum_{i=1}^n y_i \alpha + \sum_{i=1}^n x_i y_i \beta + \sum_{i=1}^n z_i y_i \gamma)}{\prod_{i=1}^n \{1 + \exp(\alpha + x_i \beta + z_i \gamma)\}}. \end{aligned}$$

Let

$$T_0 = \sum_{i=1}^n Y_i, \quad T_1 = \sum_{i=1}^n x_i Y_i, \quad T_2 = \sum_{i=1}^n z_i Y_i.$$

Then the joint distribution of (T_0, T_1, T_2) is

$$P(T_0 = t_0, T_1 = t_1, T_2 = t_2) = \sum_{y_1, \dots, y_n} \frac{\exp(t_0\alpha + t_1\beta + t_2\gamma)}{\prod_{i=1}^n \{1 + \exp(\alpha + x_i\beta + z_i\gamma)\}},$$

where the summation is with respect to all y_i 's such that

$$\sum_{i=1}^n y_i = t_0, \quad \sum_{i=1}^n x_i y_i = t_1, \quad \sum_{i=1}^n z_i y_i = t_2. \quad (12.1.1)$$

By conditioning on $T_0 = t_0, T_2 = t_2$,

$$P(T_1 = t_1 | T_0 = t_0, T_2 = t_2) = \frac{P(T_0 = t_0, T_1 = t_1, T_2 = t_2)}{P(T_0 = t_0, T_2 = t_2)} = \frac{\sum_{y_1, \dots, y_n} \exp(t_1\beta)}{\sum_{y_1, \dots, y_n} \exp(\sum_{i=1}^n x_i y_i \beta)},$$

where the summation in the numerator is subject to the constraints in (12.1.1) while the summation in the denominator is subject to the first and second constraints in (12.1.1). Note that the conditional density belongs to the exponential family. If we are interested in testing $\beta = \beta_0$ versus $\beta > \beta_0$, then we can use the Neyman–Pearson lemma to find the rejection region. The drawback of this approach is the computational burden if the sample size is large. The elimination of nuisance parameter method used in prospective studies can also be adapted in retrospective studies. Some related works can be found in Mehta and Patel (1995). Analysis can be carried out using the StatXact software.

12.2 Matched Case Control Study

In his Fisher lecture, Breslow (1996) pointed out that the development of case-control methodology over the last half century is one of the most important contributions in public health and biomedicine. It is well known that the case-control study is one of the most convenient and economic method to study risk factors in medical research. The case and control study however may suffer from many different kinds of biases. Rothman and Greenland (1998) classified biases into selection bias, information bias and confounding. To minimize bias, Wacholder et al. (1992) discussed some strategies on the choices of controls in case-control study, particularly the matched case-control method. Matching addresses the issue of confounding in the design stage of a study as opposed to the analysis phase. It is a means of providing a more efficient stratified analysis, rather than directly avoiding confounding.

The following reasons were given as an advantage of matching in Wacholder et al. (1992).

- (1) To make control for confounding more efficient when the sample size is small. Without matching, control for confounding in the analysis will result in many strata with sparse data. By balancing the distribution across strata, estimation of the odds ratio will be more stable – with smaller standard errors, and leading to narrower confidence intervals.
- (2) Even if the sample size is not small, if there are many confounders with many categories, data can be sparse in any given stratum. In that case, an alternative is to use multivariate analysis.
- (3) If obtaining information from subjects is expensive, e.g., running expensive lab tests on blood samples, matching will ensure control of confounding and will not lead to loss of information. This is especially true if the cost of matching is small compared to the cost of expanding the study size.
- (4) Sometimes, control of confounding is only possible by matching, e.g., controlling for sibship.

Even though matching is a good method, “over matching” is a potential concern (Rothman and Greenland 1998). “Avoidance of overmatching” is a term originally referred to the loss of validity in a case-control study stemming from a control group that is so closely matched to the case group that the exposure distributions differed very little. Once a factor is matched, it cannot be included in the analysis since it is deemed no longer significantly associated with the outcome. If a matched factor is correlated highly with another factor, then essentially both factors have been matched.

Suppose a risk factor with I levels is matched between the controls and cases. Denote the remaining risk factors as X . The exponential tilting model is specified by

$$f_i(x|D = 1) = f_i(x|D = 0) \exp(\alpha_i + x\beta), i = 1, 2, \dots, I,$$

where $D = 1$ if it is a case, and 0 otherwise. We can use the conditional approach to eliminate the nuisance parameters $f_i(x|D = 0)$ and α_i from the i -th stratum. Then we can combine all I strata to draw inferences on the log odds ratio parameter β .

More generally in a stratified case-control study, denote x_{i1}, \dots, x_{in_i} as the risk factors of n_i cases and $x_{i,n_i+1}, \dots, x_{N_i}$ as the risk factor of $m_i = N_i - n_i$ controls in the i -th stratum, $i = 1, 2, \dots, I$. The logistic regression model in the i -th stratum is

$$\text{logit}P(D = 1|x, \text{ in stratum } i) = \alpha_i + \beta x.$$

To lessen computational burden, we may construct a pair-wise conditional likelihood from the following consideration. For the (j, l) pair from the i -th stratum ($j = 1, \dots, n_i; l = n_i + 1, \dots, N_i$), the conditional probability that x_{ij} is a case given x_{il} and x_{il} are observed is,

$$\begin{aligned}
& P(X_{ij} \text{ is from case} | x_{ij}, x_{il}, \text{ one from case and one from control}) \\
& = \frac{P(X_{ij} = \text{case}, X_{il} = \text{control})}{P(X_{ij} = \text{case}, X_{il} = \text{control}) + P(X_{ij} = \text{control}, X_{il} = \text{case})} \\
& = \frac{\exp(\alpha_i + \beta x_{ij}) f_i(x_{ij}) f_i(x_{il})}{\exp(\alpha_i + \beta x_{ij}) f_i(x_{ij}) f_i(x_{il}) + \exp(\alpha_i + \beta x_{il}) f_i(x_{ij}) f_i(x_{il})} \\
& = \frac{\exp(\beta x_{ij})}{\exp(x_{ij}\beta) + \exp(x_{il}\beta)}.
\end{aligned}$$

Naturally we can combine all pairwise conditional likelihood as a pseudo likelihood:

$$L_P = \prod_{i=1}^I \prod_{j=1}^{n_i} \prod_{l=n_i+1}^{N_i} \frac{\exp(\beta x_{ij})}{\exp(x_{ij}\beta) + \exp(x_{il}\beta)}.$$

Two types of large sample results can be considered (Liang 1987):

1. n_i and m_i are both fixed and $I \rightarrow \infty$.
2. I is fixed, but n_i and m_i go to ∞ .

Yanagawa and Fujii (1995) introduced a projection method to generalize the Mantel–Haenszel estimator for estimating the common odds ratio in $K 2 \times J$ ($J > 2$) tables.

General 1: M Matched Case-Control Study

Prentice and Breslow (1978) and Breslow (1996) discussed the 1: M matched case-control study in detail. In a 1: M design, every case is matched with M controls. Based on the same argument as before, the conditional likelihood for a 1: M matched case-control study is

$$\prod_{i=1}^I \prod_{j=1}^{n_i} \prod_{k=1}^M \frac{\exp(\beta x_{ijk})}{\sum_{k=1}^M \exp(x_{ij}\beta)},$$

where $i = 1, 2, \dots, I$ denote the number of strata, $j = 1, \dots, n_i$ denote the cases in i -th stratum, and $k = 1, 2, \dots, M$ are the corresponding matched controls. The question is, “How many controls should be matched to each case?” Numerical studies have shown that the value of M needs not be more than 4, as there is little efficiency gain in using more than 4 controls per case.

There are many softwares available for implementing a 1: M case-control study, e.g., Splus and SAS. Interestingly, one may use the “coxph” program in R used for survival analysis to calculate β estimation. In order to do so, one can treat cases as “failures” and controls as “no failures” in each stratum. Moreover one can assign each case with a failure time, say, 1, and each control with a “censoring” time, say, 2. In this setup, all controls will contribute in the risk sets in the stratified survival analysis. More details on survival analysis can be found in Chaps. 24–26.

12.3 Matched Case and Control Sampling for Survival Analysis

So far we have made simplification in a case-control study when survival time does not play a role. In other words, we have assumed that the cases and controls in a case control study are selected randomly over the same fixed time interval. In a survival analysis, however, it is often desirable to take survival time into consideration. If “death” is treated as a case, then we need to select one or more “controls” with survival times at least as long as the case. In other words, the survival time plays the role of matching. The fundamental mathematical theory for the matched survival analysis is based on the well known Cox proportional hazards model

$$\lambda(t|x) = \lambda(t) \exp(x\beta),$$

where $\lambda(t|x)$ is the conditional hazard for a given covariate x and $\lambda(t)$ is the baseline hazard, that is left arbitrary. We will use $f(y|x)$ and $\bar{F}(y|x)$ to denote the corresponding conditional density and survival functions, respectively. Thomas (1977) and Prentice and Breslow (1978) used the following conditional likelihood approach to eliminate $\lambda(t)$.

Suppose the observed death time for individual A is $T_a = t$. Select an individual B with a survival time or follow-up time $T_b > t$. Then the conditional likelihood is:

$$\begin{aligned} \mathcal{L} &= P[X_a = x_1, X_b = x_2 | T_a = t, T_b > t, (X_a, X_b) = \{(x_1, x_2) \text{ or } (x_2, x_1)\}] \\ &= \frac{P(X_a = x_1, X_b = x_2, T_a = t, T_b > t)}{P(X_a = x_1, X_b = x_2, T_a = t, T_b > t) + P(X_a = x_2, X_b = x_1, T_a = t, T_b > t)} \\ &= \frac{P(X_a = x_1)P(X_b = x_2)f(t|x_1)\bar{F}(t|x_2)}{P(X_a = x_1)P(X_b = x_2)f(t|x_1)\bar{F}(t|x_2) + P(X_a = x_2)P(X_b = x_1)f(t|x_2)\bar{F}(t|x_1)} \\ &= \frac{\lambda(t|x_1)}{\lambda(t|x_1) + \lambda(t|x_2)}. \end{aligned}$$

Under the proportional hazards model, the likelihood becomes

$$\mathcal{L} = \frac{\exp(x_1\beta)}{\exp(x_1\beta) + \exp(x_2\beta)},$$

where the baseline hazard $\lambda(t)$ has been eliminated from the analysis. Similarly in the $1:M$ matched case-control design, the conditional likelihood is

$$L(t) = \exp(x_i\beta) / \sum_{j \in R_i(t)} \exp(x_j\beta),$$

where $R_i(t)$ is the risk set, i.e., all those survivors at time $t-$. If $t_i, i = 1, 2, \dots, n$ are the observed death times, in general the overall likelihood is the product of

$L(t_i)$, $i = 1, 2, \dots, n$, i.e.,

$$L = \prod_{i=1}^n \frac{\exp(x_i \beta)}{\sum_{j \in R_i(t_i)} \exp(x_j \beta)}.$$

The score statistic can be calculated by taking derivative with respect to β . Thomas (1981) suggested a reparametrization of β which may give a better normal approximation.

Unfortunately, the above argument does not work under an additive hazards model (Breslow and Day 1986)

$$\lambda(t|x) = \lambda(t) + x\beta,$$

or a transformation model. The problem lies in that the baseline $\lambda(t)$ can no longer be eliminated from the conditional likelihood. Further discussion on the conditional likelihood approach will appear in Chaps. 24 and 25.

Chapter 13

Discrete Data Models

The logistic regression model has been widely used in statistical literature for analyzing categorical data. In this chapter we present many other useful discrete data models. If the data collection process is retrospective, then we end up with different biased sampling problems.

13.1 Regression and Ordered Categorical Variables

Consider an ordinal outcome Y with J possible categories: $Y = 1, \dots, J$. For a given covariate X , the proportional odds model for ordinal data discussed in McCullagh (1980) is

$$\frac{P(Y \leq j|X=x)}{1 - P(Y \leq j|X=x)} = \exp(\alpha_j + \beta x), \quad \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_J,$$

where α_j is the intercept and β is the log-odds parameter. Alternatively, instead of cumulative probability, this model can be written as

$$P(Y = j|x) = \frac{\exp(\alpha_j + x\beta)}{1 + \exp(\alpha_j + x\beta)} - \frac{\exp(\alpha_{j-1} + x\beta)}{1 + \exp(\alpha_{j-1} + x\beta)},$$

from which the maximum likelihood estimate can be readily obtained. If we denote a binary variable Y_j^* as 1 if $Y \leq j$ and 0, otherwise, then this model can be written as a familiar logistic regression model for Y_j^* ,

$$P(Y_j^* = 1|x) = P(Y \leq j|x) = \frac{\exp(\alpha_j + \beta x)}{1 + \exp(\alpha_j + \beta x)}.$$

This is a standard logistic regression and $Y_j^* = 1$ can be treated as case, and 0 as control, respectively.

Suppose covariates x_1, \dots, x_m are associated with m controls ($Y_j^* = 0$) and z_1, \dots, z_n are associated with the cases ($Y_j^* = 1$). Then the basic idea of the conditioning approach is, given the m controls and n cases and covariate information $(x_1, \dots, x_m; z_1, \dots, z_n)$, the probability of observing x_1, \dots, x_m as covariates from controls and z_1, \dots, z_n as covariates from cases is

$$\frac{\exp(\alpha_j + \sum_{i=1}^n z_i \beta)}{\sum \exp(\alpha_j + \sum_{i=1}^n z_i^* \beta)} = \frac{\exp(\sum_{i=1}^n z_i \beta)}{\sum \exp(\sum_{i=1}^n z_i^* \beta)},$$

where the summation in denominator is with respect to all possible $\binom{n+m}{n}$ combinations of $(x_1, \dots, x_m, z_1, \dots, z_n)$. In practice, m, n can both be large and this raises an important question on the computational burden of this method.

One way to ease the computational burden is to adopt the pairwise conditional likelihood approach. Given one covariate Z from “cases” ($Y \leq j$) and one covariate X from “controls” ($Y > j$), the observed data have a likelihood

$$\frac{\exp(\alpha_j + z\beta)}{\exp(\alpha_j + z\beta) + \exp(\alpha_j + x\beta)} = \frac{\exp(z\beta)}{\exp(z\beta) + \exp(x\beta)}.$$

We observe that α_j has also been cancelled out. The overall pairwise likelihood is

$$\prod_{i=1}^m \prod_{k=1}^n \frac{\exp(z_k \beta)}{\exp(z_k \beta) + \exp(x_i \beta)},$$

and the pseudo likelihood is the product of those pairwise likelihoods for each category j ,

$$\prod_{j=1}^J \prod_{i=1}^{m_j} \prod_{k=1}^{n_j} \frac{\exp(z_{jk} \beta)}{\exp(z_{jk} \beta) + \exp(x_{ij} \beta)},$$

where $x_{ij}, i = 1, 2, \dots, m_j$ and $z_{kj}, k = 1, 2, \dots, n_j$ are controls and cases corresponding to $Y > j$ and $Y \leq j$, respectively. This method can be generalized to the case where the log odds ratio parameter β_j ’s need not be the same. In that case, the pseudo likelihood becomes

$$\prod_{j=1}^J \prod_{i=1}^{m_j} \prod_{k=1}^{n_j} \frac{\exp(z_{jk} \beta_j)}{\exp(z_{jk} \beta_j) + \exp(x_{ij} \beta_j)}.$$

We may be interested in testing $H_0 : \beta_1 = \dots = \beta_J$ versus $H_A : \beta_1 \leq \dots \leq \beta_J$ for the trend of increasing log-odds ratios.

Unconditional Approach

Alternatively the following pseudo likelihood approach can be used. For category j , form a binary outcome through $Y \leq j$ versus $Y > j$. Then a log-pseudo likelihood can be written as

$$\ell_j = \sum_{i=1}^{n_j} (\alpha_j + \beta_j z_{ij}) - [\sum_{i=1}^{m_j} \log\{1+\exp(\alpha_j + \beta_j x_{ij})\} - \sum_{k=1}^{n_j} \log\{1+\exp(\alpha_j + \beta_j x_{kj})\}].$$

Combining all J categories, the overall log-pseudo-likelihood is

$$\ell_O = \sum_{j=1}^J \ell_j(\alpha_j, \beta_j).$$

Under this method, point estimation of the parameters can be obtained using a standard logistic regression program. However, due to the dependence in constructing this pseudo likelihood, the asymptotic variances must be derived separately.

Anderson (1984) proposed a regression model for an ordered categorical variable Y , where $Y = 1, 2, \dots, J$. Let $f_j(x)$ be the density for covariate X in the j -th category. He assumed that

$$f_j(x)/f_J(x) = \exp(\alpha_j + \phi_j x^T), \quad j = 1, 2, \dots, J-1,$$

where

$$1 = \phi_1 \geq \phi_2 \geq \dots \geq \phi_{J-1} \geq \phi_J = 0.$$

It would be interesting to develop a likelihood based test statistic for this stochastic ordering. Anderson (1984), Anderson and Philips (1981) gave some details. Instead of the prospective sampling, if data are collected by conditioning on the status of Y , then we end up with a multiple biased sampling problem, where the underlying densities for the covariates are linked by

$$f(x|Y=j)/f(x|Y=1) \propto \exp(\phi_j x), \quad j = 2, \dots, J,$$

This problem can be solved using the semiparametric MLE methods discussed in Chap. 11.

More generally we may consider the grouped continuous regression model

$$P(Y \leq y_s|x) = F(\theta_s - x^T \beta),$$

where F is a given distribution function, the parameters $\theta_s, s = 1, 2, \dots, k$ and β are unknown. Moreover

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{k-1}, \quad \theta_0 = -\infty, \quad \theta_k = \infty.$$

Instead of the logistic link function, another popular choice of F is the Probit model, i.e., $F(t) = \Phi(t)$, where $\Phi(t)$ is the standard normal distribution function.

13.2 Using Continuous Distributions to Construct Discrete Choice Models

Next we discuss how to use continuous distributions to construct discrete choice models. This method is very popular in the econometric literature.

1. Probit and logistic regression models

Consider a linear regression model

$$Y_i = x_i\beta + \epsilon_i,$$

and suppose instead of observing Y_i , it is only known that $Y_i > 0$ or $Y_i \leq 0$. Then

$$P(Y_i > 0|x_i) = P(\epsilon_i > -x_i\beta) = \bar{F}(-x_i\beta)$$

where \bar{F} is the survival function of ϵ_i . If $F = 1 - \bar{F}$ is the normal cumulative distribution, then the model becomes a Probit model. On the other hand, if F is the cumulative function of the logistic distribution, then the model becomes a logistic regression model.

2. Discrete skewed normal

Recently the skewed normal density has been used widely. Detailed discussions can be found in Azzalini's (2013) monograph. A skewed normal distribution can be considered as a truncated version of a random variable with normal distribution. Let X and Y be two independent normal random variables with common mean but different variances,

$$X \sim N(\mu, \sigma_1^2), \quad Y \sim N(\mu, \sigma_2^2).$$

In this model, X is observed only if $X < Y$. Hence the probability of observing $X = x$ is

$$P(X = x|X < Y) = \frac{P(X = x, Y > x)}{P(Y > X)} = 0.5 \frac{1}{\sigma_1} \phi\left(\frac{x - \mu}{\sigma_1}\right) \Phi\left(\frac{x - \mu}{\sigma_2}\right).$$

Note that in the skewed normal density σ_2 is allowed to take a negative value. This model can also be treated as a biased sampling problem, such that

$$P(X = x|X > Y) = 0.5 \frac{1}{\sigma_1} \phi\left(\frac{x - \mu}{\sigma_1}\right) \Phi\left(-\frac{x - \mu}{\sigma_2}\right).$$

For a given covariate Z and time t , the discrete version is

$$P(X \leq t|z) = \int_{-\infty}^t 0.5 \frac{1}{\sigma_1} \phi\left(\frac{x - \mu(z\beta)}{\sigma_1}\right) \Phi\left(\frac{x - \mu(z\beta)}{\sigma_2}\right) dx,$$

where $\mu(z\beta)$ is some specified function.

More discussions on the skewed link model for a dichotomous quantal response can be found in Chen et al. (1999).

3. Polytomous normal model

Aitchison and Bennett (1970) proposed a method to construct polytomous models based on a normal linear model.

We begin with a binary data problem. Consider two linear models

$$Y_i = \alpha_i + x^T \beta_i + \sigma \epsilon_i, \quad i = 1, 2,$$

where ϵ_1, ϵ_2 follow two independent standard normal distributions. Suppose there are two choices $D = 1$ or $D = 0$ depending on whether $Y_1 > Y_2$. The probability is

$$P(Y_1 > Y_2) = P\left(\frac{\epsilon_1 - \epsilon_2}{\sigma\sqrt{2}} > \frac{\alpha_2 - \alpha_1 + x^T(\beta_1 - \beta_2)}{\sigma\sqrt{2}}\right) = \Phi(\alpha + x^T\beta),$$

where $\alpha = (\alpha_1 - \alpha_2)/(\sigma\sqrt{2})$ and $\beta = (\beta_1 - \beta_2)/(\sigma\sqrt{2})$. Therefore we end up with a Probit model

$$P(D = 1|x) = \Phi(\alpha + x^T\beta).$$

This method can be generalized to the multi-category case. Let

$$Y_i = \alpha_i + x_i^T \beta + \sigma \epsilon_i, \quad i = 1, 2, \dots, I,$$

where $\epsilon_i, i = 1, 2, \dots, n$ are i.i.d. normal random variables. There are I choices, which are made according to $D = i$ if $Y_i = \max(Y_1, \dots, Y_I)$. We can easily write down the probability

$$\begin{aligned} P\{Y_i = \max(Y_1, \dots, Y_I)\} &= E\left\{\prod_{j \neq i} \Phi\left(\frac{y_j - \alpha_j - x^T \beta_j}{\sigma}\right)\right\} \\ &= E\left\{\prod_{j \neq i} \Phi\left(\frac{\epsilon_j + (\alpha_i - \alpha_j) + x^T(\beta_i - \beta_j)}{\sigma}\right)\right\} \\ &= \int_{-\infty}^{\infty} \left\{\prod_{j \neq i} \Phi\left(\epsilon + A_i - A_j + x^T(B_i - B_j)\right)\right\} \phi(\epsilon) d\epsilon := p_i(x), \end{aligned}$$

where

$$A_i = \frac{\alpha_i - \alpha_I}{\sigma}, \quad B_i = \frac{\beta_i - \beta_I}{\sigma}, \quad i = 1, 2, \dots, I-1.$$

Aitchison and Bennett (1970) showed the following:

Theorem 13.1 *For a model consisting of two multinomial trials, each with I possible outcomes. Suppose the probabilities associated with the I outcomes in the first trial are $p_1(x_1), \dots, p_I(x_1)$ and those for the second trial are $p_1(x_2), \dots, p_I(x_2)$, where $x_1 \neq x_2$. Then the parameters A_1, \dots, A_{I-1} and B_1, \dots, B_{I-1} are identifiable.*

Many existing statistical softwares can be used to calculate the integration in the $p_i(x)$'s. The EM-algorithm can be used to estimate $A_i, B_i, i = 1, 2, \dots, I-1$.

A natural generalization of this method is to consider the case that $\epsilon_1, \dots, \epsilon_I$ have a joint multivariate normal distribution or t -distribution. Of course one has to be careful on the model identifiability.

4. Logistic polytomous models

McFadden (1980) and Amemiya (1985) showed how to derive the multinomial logit model from utility maximization similar to the one discussed above but with the residual distribution replaced by the extreme value distribution. For the trinomial choice case, the method works as follows. Suppose an individual i whose utility associated with three different alternatives is given by

$$U_{ij} = \mu_{ij} + \epsilon_{ij}, \quad j = 0, 1, 2,$$

where μ_{ij} is a non-stochastic function of explanatory variables and some unknown parameters, and ϵ_{ij} 's follow independent extreme distributions, given by,

$$\epsilon_{ij} \sim \exp[-\epsilon \exp(-\epsilon)].$$

Define a random variable Y_i though

$$Y_i = j, \quad \text{if, } U_{ij} = \max(U_{i0}, U_{i1}, U_{i2}), \quad j = 0, 1, 2.$$

Easily we can find

$$\begin{aligned} P(Y_i = 2) &= P(U_{i2} > U_{i1}, U_{i2} > U_{i0}) \\ &= \frac{\exp(\mu_{i2})}{\exp(\mu_{i0}) + \exp(\mu_{i1}) + \exp(\mu_{i2})}. \end{aligned}$$

Along the same line, by different specifications of error distributions we can construct different parametric models for discrete data.

5. Discrete proportional hazards model

For an individual with covariate $X = x$, the continuous time proportional hazards model (Cox 1972) is

$$\lambda(t|x) = \lambda(t) \exp(x\beta),$$

where $\lambda(t)$ is the hazard function when $x = 0$ and β is an unknown vector parameter. In the survival function form it can be written as

$$-\log \bar{F}(t|x) = \Lambda(t) \exp(x^T \beta) = \exp(\theta(t) + x\beta),$$

where $\log \Lambda(t) = \theta(t)$ is monotonic increasing.

Let Y be an ordinal response variable that takes values in the domain $1, 2, \dots, k$. The discrete proportional hazards model (McCullagh 1980) for Y is

$$P(Y > j|x) = \exp\{-\exp(\theta_j + x\beta)\},$$

$$\log[-\log P(Y > j|x)] = \theta_j + x\beta.$$

In other words after a complementary log-log transformation, the survival function is a linear function of x .

Instead of fixing θ_j , Farewell (1982) assumed $\gamma = \exp(\theta_j)$ to have a gamma distribution with density given by

$$f(\gamma) = c\Gamma^{-1}(c\gamma^*) (c\gamma)^{c\gamma^*-1} \exp(-c\gamma), \quad \gamma > 0,$$

where its mean and variance are $\gamma^* = \exp(\theta^*)$ and γ^*/c , respectively. This gives a more general model

$$P(Y > j|x) = \left(\frac{c}{c + \exp(\delta_j + x\beta)} \right)^{c \exp(\theta^*)}, \quad j = 1, 2, \dots, k-1,$$

where $\delta_j = \theta_j - \theta_1$. As $c \rightarrow \infty$, this model becomes the proportional hazards model.

6. Proportional odds ratio model

Analogous to the proportional hazards model, the continuous time proportional odds ratio model given by Bennett (1983) is

$$\frac{F(t|x)}{1 - F(t|x)} = \exp(\alpha(t) + x\beta),$$

where $F(t|x)$ and $\bar{F}(t|x)$ are, respectively, the conditional cumulative distribution function and survival function, and $\alpha(t)$ is an unknown increasing function. The marginal density $g(x)$ of X is not specified.

The conditional likelihood argument can also be used to eliminate $\alpha(t)$ from the analysis. This follows by first constructing an artificial case-control sample as follows:

Cases:

$$X|T > t \sim \frac{\bar{F}(t|x)g(x)}{\bar{F}(t)} = h_1(t, x\beta).$$

Controls:

$$X|T \leq t \sim \frac{\{F(t|x)\}g(x)}{1 - F(t)} := h_0(t, x\beta).$$

Note that $h_1(t, x\beta)/h_0(t, x\beta) = \{1 - F(t)\}F^{-1}(t) \exp(\alpha(t) + x\beta) =: \exp(\alpha^*(t) + x\beta)$ is a density ratio model. Given the i -th individual being a case and j -th individual being a control, the observed data have a conditional likelihood

$$L_{i,j} = \frac{h_1(t, x_i\beta)h_0(t, x_j\beta)}{h_1(t, x_i\beta)h_0(t, x_j\beta) + h_0(t, x_i\beta)h_1(t, x_j\beta)} = \frac{\exp(x_i\beta)}{\exp(x_i\beta) + \exp(x_j\beta)}$$

and the overall likelihood is

$$L(t) = \prod_{T_i > t, T_j \leq t} \frac{1}{1 + \exp\{(x_j - x_i)\beta\}}.$$

Finally if we consider all possible t_i 's, the conditional likelihood is

$$L = \prod_{k=1}^n \prod_{T_i > t_k, T_j \leq t_k} \frac{1}{1 + \exp\{(x_j - x_i)\beta\}}.$$

This conditional approach has successfully eliminated the nuisance function $\alpha(t)$. Unfortunately, it is not clear how to modify this approach if there is right censoring for the survival time T . McCullagh (1984) discussed a related nuisance parameter elimination problem when the outcome variable is discrete.

7. Tobit model

The Tobit model was proposed by Tobin (1958) to describe the relationship between a non-negative dependent variable Y and an independent (vector) variable X . It has become very popular in econometric studies. The basic assumption is

$$Y_i^* = x_i\beta + \epsilon_i.$$

However, Y_i^* is observed if and only if $Y_i^* > 0$, i.e., $Y_i^* = \max(0, x_i\beta + \epsilon_i)$. Denote the observed value as Y_i , where $Y_i = Y_i^*$ if $Y_i^* > 0$ and $Y_i = 0$ if $Y_i^* \leq 0$. The most popular choice for the distribution of ϵ_i is the normal distribution. The conditional likelihood function is

$$\begin{aligned}
L &= \prod_{i=1}^n [f(y_i|x_i)]^{I(y_i>0)} [F(0|x_i)]^{I(y_i=0)} \\
&= \left[\prod_{i=1}^n \left\{ \frac{f(y_i|x_i)}{\bar{F}(0|x_i)} \right\}^{I(Y_i>0)} \right] \left[\prod_{i=1}^n \{\bar{F}(0|x_i)\}^{I(y_i>0)} \{F(0|x_i)\}^{I(y_i=0)} \right].
\end{aligned}$$

This can be considered as a Probit model (the second factor) augmented by a truncated normal (the first factor). Parametric maximum likelihood inference may be performed if a parametric model for F is imposed. There are some variations on the Tobit model. The best reference on this is Chap. 10 of Amemiya (1985). A closely related model is the Heckman selection bias model (Chap. 22).

A recent *Biometrika* paper by Peyhardi et al. (2015) has surveyed the latest development on decomposition of the link function into an inverse continuous cumulative distribution function and a ratio of probabilities in regression models for categorical responses.

Chapter 14

Gene and Environment Independence and Secondary Outcome Analysis in Case-Control Study

The testing of gene and environment interaction for a susceptibility disease such as cancer has become a very popular topic in genetic epidemiological studies. This is mainly due to the fact that relatively common polymorphisms in a wide spectrum of genes may be modified by environmental exposures. A well motivated example given in Begg and Zhang (1994) is the study of the possible role of smoking in causing bladder cancers that is characterized by p53 mutations. The association between smoking and bladder cancer is most pronounced among individuals with the NAT2 slow acetylation phenotype. Under the assumption of independence between gene and environment in the control population, Begg and Zhang (1994), and Umbach and Weinberg (1997) showed that efficient estimates of interaction for categorical environment and binary genotype variables can be done via the logistic regression model in a case-only analysis. Furthermore, they showed that, in general, the case-only analysis is more powerful than the case-control logistic regression analysis even though there are some limitations in the case-only analysis.

The case-only analysis depends heavily on the assumption of independence between gene and environment either in the control population or in the general population. In practical applications, however, it may not be certain that this assumption holds true. By using simulations, Albert et al. (2001) demonstrated that inferences on the multiplicative interaction with case-only design can be highly distorted when there are departures from this assumption. To avoid the possible bias in the lack of independence between gene and environment in the genome-wide association study, Murcay et al. (2009) proposed a two-stage approach. In the first stage, a logistic regression analysis is performed by using both case and control data, where the response variable is gene and the explanatory variables are environmental variables. The analysis continues to a second stage if the p -value derived from the likelihood ratio test for the coefficient of environment variable in the first stage is smaller than a specified level, say, α_1 . In the second stage, a likelihood ratio test for the interaction between gene and environment is performed by using a standard logistic regression

model with case-control data. A caveat for this approach is the difficulty in controlling the overall type-I error.

14.1 Score Test for Gene and Environment Independence

We discuss a score based test for the interaction between two covariates in case-control studies. The score test is a flexible test and is suitable for all types of covariates, i.e., discrete versus discrete, discrete versus continuous and continuous versus continuous. Therefore the results apply not only to the genetic case-control study but also to general case-control problems. In addition, when the gene and environment are independent of each other in the control population, we give a theoretical justification on why the case only analysis is more powerful than the conventional logistic regression analysis based on case-control data.

Let $D = 1$ or 0 be an indicator variable for disease or healthy individuals, respectively. Let Y and X be two covariates. The logistic regression model is given by

$$P(D = 1|Y = y, X = x) = \frac{\exp\{\alpha^* + y\beta + \gamma x + \xi\phi(x, y)\}}{1 + \exp\{\alpha^* + y\beta + \gamma x + \xi\phi(x, y)\}}, \quad (14.1.1)$$

where $\phi(x, y)$ is a given known function of x and y . The most common choice is $\phi(x, y) = xy$. In this model ξ characterizes the interaction between Y and X , we are interested in testing $H_0 : \xi = 0$.

Instead of prospectively collecting (D, Y, X) , (Y, X) can be collected by first conditioning on the status of D . This is the so-called retrospective sampling or case-control sampling. Let

$$(Y_{i1}, X_{i1}), i = 1, 2, \dots, n_1$$

be the covariate data for n_1 individuals with $D_i = 1$, where n_1 is a pre-specified number. Similarly let

$$(Y_{i0}, X_{i0}), i = 1, 2, \dots, n_0$$

be the covariate data from n_0 individuals with $D_i = 0$. Again, n_0 is a fixed number. Using Bayes formula, we have

$$f(y, x|D = 1) = \frac{P(D = 1|y, x)f(y, x)}{P(D = 1)}, \quad f(y, x|D = 0) = \frac{P(D = 0|y, x)f(y, x)}{P(D = 0)},$$

where $f(y, x)$ is the unspecified marginal density of (Y, X) in the general population. As discussed in Chap. 11, this model is equivalent to

$$f(y, x|D = 1) = \exp(\alpha + y\beta + \gamma x + \xi\phi(x, y))f(y, x|D = 0), \quad (14.1.2)$$

where $\alpha = \alpha^* + \log\{P(D = 0)/P(D = 1)\}$. In other words the density of (Y, X) in the case population is an exponential tilted version of the density of (Y, X) in the control population. Following Begg and Zhang (1994) and Umbach and Weinberg (1997), we assume in the control population the covariates Y and X are independent,

$$f(y, x|D = 0) = f_0(y, x) = f_0(y)g_0(x),$$

where $f_0(y)$ and $g_0(x)$ are marginal densities of Y and X in controls. With this assumption, it is clear that testing $H_0 : \xi = 0$ in the logistic regression model is equivalent to testing the independence between Y and X in the case population. Naturally Pearson's correlation coefficient test based on case-only data can be used.

If both case and control data are available, another approach to test $H_0 : \xi = 0$ is to conduct a logistic regression analysis without employing the assumption of independence between Y and X in the control population. In this approach, however, a large power loss has been found when compared with Pearson correlation coefficient test based on case data only.

The practical challenge is to combine case and control data and the auxiliary information on the independence between gene and environment in the control population to form a powerful test statistic. This motivates us to develop a semiparametric score-based test statistic. We will show that this test statistic is in general different from Pearson's correlation coefficient test statistic based on case-only data. However they are identical if $\phi(x, y)$ in (14.1.2) is given by $\phi(x, y) = xy$.

Under the assumption that gene and environment are independent in the control population, the density for cases can be written as

$$f_1(y, x) = f(y, x|D = 1) = \exp\{\alpha + y\beta + \gamma x + \xi\phi(y, x)\}f_0(y)g_0(x),$$

where α is a normalizing constant. Note that $f_1(x, y)$ can be written as

$$f_1(y, x) = f(y, x|D = 1) = \frac{\exp(y\beta + \gamma x + \xi\phi(x, y))f_0(y)g_0(x)}{\int \int \exp(y\beta + \gamma x + \xi\phi(x, y))f_0(y)g_0(x)dydx}.$$

Based on case data (y_{i1}, x_{i1}) , $i = 1, 2, \dots, n_1$, the log-likelihood can be written as

$$\begin{aligned} \ell_1 &= \sum_{i=1}^{n_1} \left[\{y_{i1}\beta + \gamma x_{i1} + \xi\phi(x_{i1}, y_{i1})\} \right. \\ &\quad \left. - \log \left\{ \int \int \exp\{y\beta + \gamma x + \xi\phi(x, y)\}f_0(y)g_0(x)dydx \right\} \right]. \end{aligned}$$

Differentiating ℓ_1 with respect to ξ , we have

$$\frac{\partial \ell_1}{\partial \xi} = \sum_{i=1}^{n_1} \left[\phi(x_{i1}, y_{i1}) - \frac{\int \int \phi(x, y) \exp\{y\beta + \gamma x + \xi\phi(x, y)\}f_0(y)g_0(x)dydx}{\int \int \exp\{y\beta + \gamma x + \xi\phi(x, y)\}f_0(y)g_0(x)dydx} \right].$$

The score statistic is

$$S = \frac{\partial \ell_1}{\partial \xi} |_{\xi=0} = \sum_{i=1}^{n_1} \left[\phi(x_{i1}, y_{i1}) - \frac{\int \int \phi(x, y) \exp(y\beta + \gamma x) f_0(y) g_0(x) dy dx}{\int \int \exp(y\beta + \gamma x) f_0(y) g_0(x) dy dx} \right].$$

Equivalently

$$S = n_1^{-1} \sum_{i=1}^{n_1} [\phi(y_{i1}, x_{i1}) - E_1\{\phi(Y, X)\}],$$

where E_1 denotes expectation under the case population.

Note that under $H_0 : \xi = 0$,

$$f_1(y, x) = \exp(\alpha + y\beta + x\gamma) f_0(y) g_0(x),$$

which shows that Y and X are independent in the case population with densities

$$f_1(y) = \exp(\alpha_1 + y\beta) f_0(y)$$

and

$$g_1(x) = \exp(\alpha_0 + x\gamma) g_0(x),$$

respectively, where $\alpha = \alpha_0 + \alpha_1$ and

$$1 = \int \exp(\alpha_1 + y\beta) f_0(y) dy, \quad 1 = \int \exp(\alpha_0 + x\gamma) g_0(x) dx.$$

Under $H_0 : \xi = 0$, case and control analyses based on only Y and only X can be performed separately. Using the results in Sect. 11.1, $E_1\{\phi(Y, X)\}$ can be estimated by

$$n_0^{-2} \sum_{i=1}^n \sum_{j=1}^n \frac{\phi(y_i, x_j) \exp(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}x_j)}{\{1 + \rho \exp(\hat{\alpha}_1 + \hat{\beta}y_i)\}\{1 + \rho \exp(\hat{\alpha}_0 + \hat{\gamma}x_j)\}},$$

where $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}, \hat{\gamma})$ are the maximum logistic likelihood estimate under H_0 , and $\hat{\alpha} = \hat{\alpha}_0 + \hat{\alpha}_1$.

Therefore the score test is

$$S_n = n_1^{-1} \sum_{i=1}^{n_1} \phi(y_{i1}, x_{i1}) - n_0^{-2} \sum_{i=1}^n \sum_{j=1}^n \frac{\phi(y_i, x_j) \exp(\hat{\alpha} + \hat{\beta}y_i + \hat{\gamma}x_j)}{\{1 + \rho \exp(\hat{\alpha}_1 + \hat{\beta}y_i)\}\{1 + \rho \exp(\hat{\alpha}_0 + \hat{\gamma}x_j)\}}.$$

In general the power by using the case and control data will be higher when compared to using the case data alone.

Remark 1 A common choice of the interaction term is $\phi(x, y) = xy$. In this case, the score test can be written as

$$S_n = \frac{1}{n_1} \sum_{i=1}^{n_1} [x_{i1}y_{i1} - E_1(X)E_1(Y)],$$

where E_1 denotes expectation under the case population. If $E_1(X)$ and $E_1(Y)$ are estimated based on the case-only data, then we have Pearson's correlation coefficient statistic. In this way, clearly the control data are not used at all. Alternatively, following the analysis above, under $H_0 : \xi = 0$, two logistic regression analyses can be performed separately. As shown in Sect. 11.1, (α_0, γ) and (α_1, β) may be estimated by

$$\begin{aligned} \sum_{i=1}^n \binom{1}{x_i} \left(d_i - \frac{\exp(\alpha_0 + \gamma x_i)}{1 + \rho \exp(\alpha_0 + \gamma x_i)} \right) &= 0, \\ \sum_{i=1}^n \binom{1}{y_i} \left(d_i - \frac{\exp(\alpha_1 + \beta y_i)}{1 + \rho \exp(\alpha_1 + \beta y_i)} \right) &= 0, \end{aligned}$$

where $d_i = 1, i = 1, 2, \dots, n_1$ and $d_i = 0, i = n_1, \dots, n$, $n = n_1 + n_0$, $(x_1, \dots, x_n) = (x_{11}, \dots, x_{n_11}, x_{10}, \dots, x_{n_00})$ and $(y_1, \dots, y_n) = (y_{11}, \dots, y_{n_11}, y_{10}, \dots, y_{n_00})$.

However as discussed in Remark 3 in Chap. 11, the estimate of $E_1(X)$ by using the case-control data and that using the case data alone are identical. The same identity applies to the estimation of $E_1(Y)$. As a result, the score test is precisely equivalent to Pearson's correlation coefficient test

$$S_n = n_1^{-1} \sum_{i=1}^{n_1} x_{i1}y_{i1} - \left[n_1^{-1} \sum_{i=1}^{n_1} x_{i1} \right] \left[n_1^{-1} \sum_{i=1}^{n_1} y_{i1} \right].$$

In other words, there is no improvement in using case-control data over Pearson's correlation coefficient test based on case-only data.

If the interaction term is specified as a known function of Y and X , e.g., $\phi(y, x) \neq yx$, then using the case-control data will lead to an increase in power than by using the cases alone. In that situation, however, the choice of an interaction term other than the conventional one must be justifiable.

Remark 2 We can also use the smooth goodness of fit test concept to understand this problem. We are interested in testing whether the case population has a density

$$f_1(y, x) = \exp(\alpha + \beta y + \gamma x) f_0(y) f_0(x).$$

We can embed $f_1(y, x)$ in a larger parametric family

$$f_1(y, x) = \exp\{\alpha + \beta y + \gamma x + \xi \phi(x, y)\} f_0(y) f_0(x),$$

where $\phi(x, y)$ is a specified function. In general it reflects the possible direction that the underlying model might depart from the H_0 . This leads to the test $\xi = 0$.

14.2 Inference Under the Assumption of Independence Between Gene and Environment

So far we have only discussed testing problems. Chatterjee and Carroll (2005) discussed inference aspects under the assumption of independence between gene and environment in the general population. More specifically, they let

$$P(D = 1|x, y) = \frac{\exp(\alpha^* + y\beta + x\gamma + \xi xy)}{1 + \exp(\alpha^* + y\beta + x\gamma + \xi xy)} =: \pi(x, y, \omega), \quad \omega = (\alpha^*, \beta, \gamma, \xi). \quad (14.2.3)$$

Moreover they assumed

$$P(X = x, Y = y) = f(x)g(y).$$

Consequently,

$$P(X = x, Y = y|D = 1) = \frac{\pi(x, y, \omega)f(x, y)}{P(D = 1)} = \frac{\pi(x, y, \omega)f(x)g(y)}{\int \int \pi(x, y, \omega)f(x)g(y)dxdy},$$

$$\begin{aligned} P(X = x, Y = y|D = 0) &= \frac{\{1 - \pi(x, y, \omega)\}f(x, y)}{P(D = 0)} \\ &= \frac{\{1 - \pi(x, y, \omega)\}f(x)g(y)}{\int \int \{1 - \pi(x, y, \omega)\}f(x)g(y)dxdy}. \end{aligned}$$

When both X and Y are continuous, the maximum semiparametric estimates of the underlying parameters is so complicated. Fortunately in gene-environment studies, gene is a discrete variable with values $Y = 0, 1, 2$. Next we outline the maximum semiparametric likelihood method.

Note that

$$X|D = 1 \sim \frac{w(x, \omega)f(x)}{\int w(x)f(x)dx}, \quad w(x) = \pi(x, 0, \omega)g(0) + \pi(x, 1, \omega)g(1) + \pi(x, 2, \omega)g(2),$$

$$X|D = 0 \sim \frac{\{1 - w(x, \omega)\}f(x)}{\int \{1 - w(x, \omega)\}f(x)dx},$$

where $g(2) = 1 - g(0) - g(1)$. The joint likelihood can be written as

$$L = \prod_{i=1}^{n_1} \frac{\pi(x_{1i}, y_{1i}, \omega)g(y_{1i})}{w(x_{1i}, \omega)} \frac{w(x_{1i})dF(x_{1i})}{\int w(x)dF(x)}$$

$$\prod_{i=1}^{n_0} \frac{\{1 - \pi(x_{0i}, y_{0i}, \beta)\}g(y_{0i})}{\{1 - w(x_{0i}, \omega)\}} \frac{\{1 - w(x_{0i})\}dF(x_{0i})}{\int \{1 - w(x)\}dF(x)}.$$

Denote $\theta = \int w(x)dF(x)$. After profiling out $dF(x)$, the log-likelihood becomes

$$\ell = \sum_{i=1}^{n_1} \log \pi(x_{1i}, y_{1i}, \omega) + \log g(y_{1i}) + \sum_{i=1}^{n_0} \log \{1 - \pi(x_{0i}, y_{0i}, \omega)\} + \log g(y_{0i})$$

$$- n_1 \log \theta - n_0 \log(1 - \theta) - \sum_{i=1}^{n_1} \log[1 + \lambda(w(x_{1i}) - \theta)]$$

$$- \sum_{i=1}^{n_0} \log[1 + \lambda(w(x_{0i}) - \theta)],$$

where the Lagrange multiplier λ is determined by

$$\sum_{i=1}^{n_1} \frac{w(x_{1i}) - \theta}{1 + \lambda(w(x_{1i}) - \theta)} + \sum_{i=1}^{n_0} \frac{w(x_{0i}) - \theta}{1 + \lambda(w(x_{0i}) - \theta)} = 0.$$

For a known $\theta = \theta_0$, this log-likelihood can be maximized with respect to β and p_0, p_1 .

Exercise 1 Derive the limiting distribution for the maximum semiparametric likelihood estimate.

Exercise 2 Using the same logistic regression model, but assuming gene and environment are independent in the control group, derive the maximum likelihood estimate and its limiting results.

Some Comments on the Assumption of Independence Between Gene and Environment

We have observed two different assumptions on the gene and environment independence. The first one is made in the control population, and the second one is made in the general population. Note that if the disease prevalence is low, then approximately, the two assumptions are equivalent. In fact in the general population

$$P(Y = y, X = x) = P(D = 1)P(Y = y, X = x|D = 1)$$

$$+ P(D = 0)P(Y = y, X = x|D = 0)$$

$$\approx P(Y = y, X = x|D = 0),$$

since $P(D = 1) = 0$ approximately. On the other hand, if the disease prevalence is high, the maximum likelihood estimates based on the independence assumption

in the general population might have convergence problems unless the true disease prevalence is known.

Furthermore, if the retrospective sampling is almost identical to the prospective sampling, i.e., the true disease prevalence matches the sampling fraction $n_1/(n_1+n_0)$ of cases, then Chatterjee and Carroll (2005) method has no improvement over the standard logistic regression estimation provided that the density $f(x, y)$ does not carry any information for the log odds ratio parameters. We can demonstrate this as follows.

Suppose there are n individuals and the sampling design is prospective, and there are n_1 cases and n_0 controls. The full likelihood based on all the data can be decomposed either prospectively or retrospectively as

$$\begin{aligned} & \prod_{i=1}^n \left[\frac{\exp\{d_i(\alpha^* + \beta y_i + \gamma x_i + \xi x_i y_i)\}}{1 + \exp(\alpha^* + \beta y_i + \gamma x_i + \xi x_i y_i)} f(y_i, x_i) \right] \\ &= \left\{ \prod_{i=1}^{n_1} f(y_{1i}, x_{1i} | D = 1) \right\} \left\{ \prod_{j=1}^{n_0} f(x_{0j}, y_{0j} | D = 0) \right\} \\ & \quad \times \{P^{n_1}(D = 1) P^{n_0}(D = 0)\}. \end{aligned}$$

Usually the density function $f(y, x)$ is unrelated to the parameters $(\alpha^*, \beta, \gamma, \xi)$ and the corresponding likelihood can be factored out of the prospective likelihood. Therefore, the full likelihood

$$\prod_{i=1}^n \left[\frac{\exp\{d_i(\alpha^* + \beta y_i + \gamma x_i + \xi x_i y_i)\}}{1 + \exp(\alpha^* + \beta y_i + \gamma x_i + \xi x_i y_i)} f(y_i, x_i) \right]$$

is exactly equivalent to the prospective conditional likelihood even if $f(y, x)$ is completely known. As a result, the retrospective likelihood

$$\left\{ \prod_{i=1}^{n_1} f(y_{1i}, x_{1i} | D = 1) \right\} \left\{ \prod_{j=1}^{n_0} f(y_{0j}, x_{0j} | D = 0) \right\}$$

is usually less informative than the prospective likelihood, and we cannot expect any improvement over the prospective likelihood by using the retrospective likelihood and auxiliary information on $f(y, x)$.

On the other hand any auxiliary information on $f(x, y | D = 0)$ will carry information for $(\alpha, \beta, \gamma, \xi)$. This can be easily seen through

$$\begin{aligned} f(y, x | D = 1) &= \exp(\alpha + y\beta + \gamma x + \xi xy) f(y, x | D = 0) \\ &= \frac{\exp(y\beta + \gamma x + \xi xy) f(y, x | D = 0)}{\int \int \exp(y\beta + \gamma x + \xi xy) f(y, x | D = 0) dx dy}. \end{aligned}$$

In the extreme case, if $f(x, y|D = 0)$ is completely known, then this becomes a full parametric model. More discussion on this can be found in Qin et al. (2016).

14.3 Secondary Outcome Analysis

In recent genetic epidemiological studies, a popular approach is to carry out a secondary analysis following a case-control study. Case-control genome-wide association studies (GWAS) provide a large amount of genetic information that can be used to conduct secondary phenotypes analyses. Lin and Zeng (2009) proposed a likelihood based method that combines the cases and controls efficiently to analyze secondary phenotypes. In their approach a parametric conditional density is assumed for the genetic phenotype given environment or other covariate information in the general population. As it is well known that the disease prevalence probability is not estimable based on case-control data, some identifiable problems may occur in their approach. They concluded that all standard methods based on controls only, case only, and the combination of cases and controls yield unbiased estimates only if the disease is rare. Li and Gail (2012) pointed out that the information on disease prevalence is crucial in the maximum likelihood estimates discussed in Lin and Zeng (2009). Without this piece of information, the resulting inference is unreliable, especially in the case where there is gene and environment interaction.

In this section, we discuss a unified approach for estimating and testing the interaction between two covariates in case-control studies as well as the secondary phenotype analysis in genome-wide association studies. Unlike Lin and Zeng (2009) approach, where a parametric model is assumed for the covariate Y (such as gene) given the other covariate X (such as environment) in the general population, in this section, a parametric assumption for $Y|X$ is made only in the control population in addition to the disease model (14.2.3) assumption. As a result, the method discussed here works whether the disease prevalence is low or high, and known or unknown.

We model the relationship between Y and X in the control group as

$$f_0(y|x) = f_0(y|x, \eta).$$

In general X and Y may not be independent in the controls. Under (14.2.3), the case and control densities are linked by

$$\begin{aligned} f(x, y|D = 1) &= \exp(\alpha + y\beta + x\gamma + \xi xy) f(x, y|D = 0), \\ \alpha &= \alpha^* + \log\{P(D = 1)/P(D = 0)\}. \end{aligned}$$

Then the marginal distribution $f_1(x)$ for cases is

$$\begin{aligned} f_1(x) &= \exp(\alpha + \gamma x) \int \exp(\beta y + \xi xy) f_0(y|x, \eta) dy f_0(x) \\ &= \exp(\alpha + \gamma x) \mu_1(x, \beta, \xi, \eta) f_0(x), \end{aligned}$$

where

$$\mu_1(x, \beta, \xi, \eta) = \int \exp(\beta y + \xi xy) f_0(y|x, \eta) dy.$$

For convenience let

$$\phi(x, \beta, \gamma, \xi, \eta) = \gamma x + \log \mu_1(x, \beta, \xi, \eta),$$

then

$$f_1(x) = \exp\{\alpha + \phi(x, \beta, \gamma, \xi, \eta)\} f_0(x).$$

The conditional density of $f_1(y|x)$ is given by

$$f_1(y|x) = \exp(\beta y + \xi xy) f_0(y|x, \eta) / \mu_1(x, \beta, \xi, \beta).$$

Therefore the log-likelihood is

$$\ell = \ell_c + \ell_M,$$

where

$$\begin{aligned} \ell_c &= \sum_{i=1}^{n_1} [\beta y_{i1} + \xi y_{i1} x_{i1} + \log f_0(y_{i1}|x_{i1}, \eta) - \log \mu_1(x_{i1}, \beta, \xi, \eta)] \quad (14.3.4) \\ &\quad + \sum_{i=1}^{n_0} \log f_0(y_{i0}|x_{i0}, \eta) \end{aligned}$$

and

$$\ell_M = \sum_{i=1}^{n_1} \{\alpha + \gamma x_{i1} + \log \mu_1(x_{i1}, \beta, \xi, \eta)\} + \sum_{i=1}^n \log dF_0(x_i).$$

After profiling out $dF_0(x_i)$, $i = 1, 2, \dots, n$ (Chap. 11),

$$\begin{aligned} \ell_M &= \sum_{i=1}^{n_1} \{\alpha + \gamma x_{i1} + \log \mu_1(x_{i1}, \beta, \xi, \eta)\} \quad (14.3.5) \\ &\quad - \sum_{i=1}^n \log [1 + \rho \exp\{\alpha + x_i \gamma + \log \mu_1(x_i, \beta, \xi, \eta)\}], \end{aligned}$$

where $\rho = n_1/n_0$. Define

$$\omega = (\alpha, \beta, \gamma, \eta, \xi),$$

and the maximum hybrid likelihood estimate of ω as $\hat{\omega}$.

The following results are given in Qin et al. (2016).

Theorem 14.1 *Under some regularity conditions specified in Qin et al. (2016), in distribution*

$$\sqrt{n}(\hat{\omega} - \omega_0) \rightarrow N(0, \Sigma),$$

where Σ is defined in the proof below.

Proof Denote $\omega = (\alpha, \beta, \gamma, \eta, \xi)$, $\omega_1 = (\beta, \gamma, \eta, \xi)$. Let $\hat{\omega} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\eta}, \hat{\xi})$.

Note that

$$\ell = \ell_c + \ell_M,$$

where

$$\ell_c = \sum_{i=1}^{n_1} \log f_1(y_{i1}|x_{i1}, \omega_1) + \sum_{i=1}^{n_0} \log f_0(y_{i0}|x_{i0}, \eta)$$

and

$$\ell_M = \sum_{i=1}^{n_1} \{\alpha + \phi(x_{i1}, \omega_1)\} - \sum_{i=1}^n \log[1 + \rho \exp\{\alpha + \phi(x_i, \omega_1)\}].$$

Differentiating ℓ with respect to ω , we have

$$\frac{\partial \ell}{\partial \omega} = \frac{\partial \ell_c}{\partial \omega} + \frac{\partial \ell_M}{\partial \omega},$$

where

$$\frac{\partial \ell_c}{\partial \omega} = \sum_{i=1}^{n_1} \frac{\partial \log f_1(y_{i1}|x_{i1}, \omega_1)}{\partial \omega} + \sum_{i=1}^{n_0} \frac{\partial \log f_0(y_{i0}|x_{i0}, \eta)}{\partial \omega}.$$

Let

$$g = \frac{\partial \ell}{\partial \omega} = \frac{\partial \ell_c}{\partial \omega} + \frac{\partial \ell_M}{\partial \omega}.$$

The two terms are orthogonal to each other since $\partial \ell_c / \partial \omega$ is the conditional score and $\partial \ell_M / \partial \omega$ only depends on the marginal data. It can be easily shown that, in distribution,

$$n^{-1/2} g \rightarrow N(0, V_c + V_M),$$

where

$$\begin{aligned} V_c &= \rho_1 E \left(\frac{\partial \log f_1(y|x, \omega)}{\partial \omega} \frac{\partial \log f_1(y|x, \omega)}{\partial \omega^T} \right) \\ &\quad + \rho_0 E \left(\frac{\partial \log f_0(y|x, \eta)}{\partial \omega} \frac{\partial \log f_0(y|x, \eta)}{\partial \omega^T} \right) \end{aligned}$$

$$V_M = \frac{\rho}{1+\rho} A - \rho \begin{pmatrix} A_0 \\ A_1^T \end{pmatrix} (A_0, A_1), \quad A = \begin{pmatrix} A_0 & A_1 \\ A_1^T & A_2 \end{pmatrix},$$

where

$$\begin{aligned} A_0 &= \int \frac{\exp\{\alpha + \phi(x, \omega_1)\}}{1 + \rho \exp\{\alpha + \phi(x, \omega_1)\}} dF_0(x), \\ A_1 &= \int \frac{\exp\{\alpha + \phi(x, \omega_1)\}}{1 + \rho \exp\{\alpha + \phi(x, \omega_1)\}} \frac{\partial \phi(x, \omega_1)}{\partial \omega_1} dF_0(x), \\ A_2 &= \int \frac{\exp\{\alpha + \phi(x, \omega_1)\}}{1 + \rho \exp\{\alpha + \phi(x, \omega_1)\}} \frac{\partial \phi(x, \omega_1)}{\partial \omega_1} \frac{\partial \phi(x, \omega_1)}{\partial \omega_1^T} dF_0(x). \end{aligned}$$

Furthermore, it is straightforward to show, in probability,

$$n^{-1} \frac{\partial^2 \ell_M}{\partial \omega \partial \omega^T} \rightarrow \frac{\rho}{1+\rho} A, \quad n^{-1} \frac{\partial^2 \ell_c}{\partial \omega \partial \omega^T} \rightarrow -V_c.$$

By expanding $\partial \ell(\hat{\omega})$ at ω_0 ,

$$n^{-1/2}(\hat{\omega} - \omega_0) = \left(\frac{1}{n} \frac{\partial^2 \ell(\omega_0)}{\partial \omega \partial \omega^T} \right)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell(\omega_0)}{\partial \omega} + o_p(1).$$

Finally we can show that, in distribution,

$$n^{-1/2}(\hat{\omega} - \omega_0) \rightarrow N(0, \Sigma),$$

where

$$\Sigma = (-V_c + \rho A / (1 + \rho))^{-1} (V_c + V_M) (-V_c + \rho A / (1 + \rho))^{-1}.$$

The likelihood ratio test for $H_0 : \xi = 0$ is

$$R = 2[\max_{(\alpha, \beta, \gamma, \xi, \eta)} \ell(\omega) - \max_{(\alpha, \beta, \gamma, \xi=0, \eta)} \ell(\omega)].$$

Theorem 14.2 *Under some regularity conditions,*

$$R = 2[\max_{(\alpha, \beta, \gamma, \xi, \eta)} \ell - \max_{(\alpha, \beta, \gamma, \xi=0, \eta)} \ell] \rightarrow \chi^2(1).$$

Proof

Denote $\omega = (\alpha, \beta, \gamma, \eta, \xi)$. Let $\hat{\omega} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\eta}, \hat{\xi})$ and $\tilde{\omega} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\eta}, 0)$ be the maximum semiparametric likelihood estimate and the constrained maximum semiparametric likelihood estimate, respectively. Expand $\ell(\tilde{\omega})$ at $\hat{\omega}$, we have

$$\ell(\tilde{\omega}) - \ell(\hat{\omega}) = 0.5(\tilde{\omega} - \hat{\omega})^T \frac{\partial^2 \ell}{\partial \omega \partial \omega^T} (\tilde{\omega} - \hat{\omega}) + o_p(1),$$

and

$$\frac{\partial^2 \ell}{\partial \omega \partial \omega^T} \rightarrow U = -V_c + \rho A / (1 + \rho) =: \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}.$$

$$\sqrt{n}(\hat{\omega} - \omega_0) = -U^{-1} \frac{\partial \ell(\omega_0)}{\partial \omega} + o_p(1) := -U^{-1}g + o_p(1),$$

$$\begin{pmatrix} \sqrt{n}(\tilde{\omega} - \omega_0) \\ 0 \end{pmatrix} = - \begin{pmatrix} U_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} g + o_p(1),$$

$$\sqrt{n}(\hat{\omega} - \tilde{\omega}) = U^{-1} \left[I - U \begin{pmatrix} U_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right] g := U^{-1}Bg,$$

$$R = g^T B^T U^{-1} B g.$$

Let $W = B^T U^{-1} B$. In order to show this theorem we only need to prove the conditions in Ogasawara–Takahashi’s Theorem (Rao 1973, p. 188) hold true, i.e.,

$$UWUWUW = WUW, \quad \text{trace}(UW) = p.$$

This can be done by matrix algebra.

Exercise 1 Suppose that

$$f_0(y|x) = f_0(y|\eta_1 + \eta_2 x),$$

where f_0 is a specified density. Define

$$R_1 = 2 \left[\max_{(\alpha, \beta, \gamma, \xi, \eta_1, \eta_2)} \ell(\omega) - \max_{(\alpha, \beta, \gamma, \xi=0, \eta_1, \eta_2=0)} \ell(\omega) \right].$$

for testing $H_0 : \xi = 0, \eta_2 = 0$, simultaneously, i.e., X and Y are independent in both the case and control groups. Show that under some regularity conditions, in distribution $R_1 \rightarrow \chi^2(2)$ under the null $H_0 : \xi = \eta_2 = 0$.

Next we propose a practical method to generate biased sampling data for $f(x, y|D=1)$ from $f(x, y|D=0)$.

Use Unequal Sampling

The basic idea comes from the unequal sampling method in survey sampling. Since

$$f_1(y, x) = \exp\{\alpha + \phi(y, x, \beta)\} f_0(y, x),$$

we can generate $(x_i, y_i), i = 1, 2, \dots, N$ from $f_0(y, x)$ first. Then we generate D_i from

$$P(D_i = 1|x_i, y_i, i = 1, 2, \dots, N) = \frac{\exp\{\phi(y_i, x_i, \beta)\}}{\sum_{i=1}^N \exp\{\phi(y_i, x_i, \beta)\}},$$

Note

$$f(y_i, x_i | D_i = 1) = \frac{P(D_i = 1|y_i, x_i) f_0(y_i, x_i)}{\int P(D_i = 1|x_i, y_i) f_0(x_i, y_i) dx_i dy_i}.$$

Therefore if N is large enough, then approximately

$$f(y_i, x_i | D_i = 1) = f_1(y, x) = \frac{\exp\{\phi(y, x, \beta)\} f_0(y, x)}{\int \exp\{\phi(y, x, \beta)\} f_0(y, x) dy dx}.$$

Numerical results can be found in Qin et al. (2016).

14.4 Use Covariate Specific Disease Prevalent Information

As discussed in previous Chapters, summarized statistics from previous studies can sometimes be utilized to enhance the estimation efficiency in a current study. This is especially important in the big data era where many types of information can be found through internet. More specifically, suppose the disease prevalence is known at various levels of a known risk factor X . We already showed how to use this type of information in Chap. 8 in prospective studies. In this section we combine this type of information in a case-control biased sampling setup. For a given risk factor X in a range $(a, b]$, the disease prevalence is

$$P(D = 1 | a < X \leq b) = \phi(a, b),$$

where $\phi(a, b)$ is known. Using Bayes' formula we have

$$\frac{\pi \int_a^b dF_1(x)}{P(a < X \leq b)} = \phi(a, b), \quad \pi = P(D = 1).$$

Similarly,

$$\frac{(1 - \pi) \int_a^b dF_0(x)}{P(a < X \leq b)} = 1 - \phi(a, b).$$

Taking the ratio of the above equations we have

$$\frac{\int_a^b dF_1(x)}{\int_a^b dF_0(x)} = \frac{1 - \pi}{\pi} \frac{\phi(a, b)}{1 - \phi(a, b)},$$

or

$$\int_a^b dF_1(x) = \frac{1-\pi}{\pi} \frac{\phi(a, b)}{1-\phi(a, b)} \int_a^b dF_0(x),$$

or

$$E_1[I(a < X \leq b)] = \frac{1-\pi}{\pi} \frac{\phi(a, b)}{1-\phi(a, b)} E_0[I(a < X \leq b)],$$

where E_0 and E_1 are, respectively, expectations with respect to the controls and cases. We assume that given covariates X and Y , the underlying disease model is given by the conventional logistic regression

$$P(D = 1|x, y) = \frac{\exp(\alpha^* + x\beta + y\gamma + yx\xi)}{1 + \exp(\alpha^* + x\beta + y\gamma + yx\xi)}. \quad (14.4.6)$$

As shown in Chap. 11, this is equivalent to the exponential tilting model

$$f_1(x, y) = f(x, y|D = 1) = \exp(\alpha + x\beta + y\gamma + yx\xi) f_0(x, y),$$

where $f_0(x, y) = f(x, y|D = 0)$ and $\alpha = \alpha^* - \log\{\pi/(1-\pi)\}$. As a consequence,

$$\begin{aligned} & E_0[I(a < X \leq b) \exp[\log\{(1-\pi)/\pi\} + \alpha + \beta X + \gamma Y + \xi XY]] \\ &= \frac{1-\pi}{\pi} \frac{\phi(a, b)}{1-\phi(a, b)} E_0[I(a < X \leq b)], \end{aligned}$$

or

$$E_0 \left[I(a < X \leq b) \exp\{\alpha + \beta X + \gamma Y + \xi XY\} - \frac{\phi(a, b)}{1-\phi(a, b)} I(a < X \leq b) \right] = 0. \quad (14.4.7)$$

Denote

$$g_0(X, Y) = \exp[\log\{(1-\pi)/\pi\} + \alpha + \beta X_i + \gamma Y_i + \xi X_i Y_i] - 1 := w(x, y) - 1,$$

and the summarized auxiliary information equations as

$$\begin{aligned} g_i(X, Y) &= I(a_{i-1} < X \leq a_i) \exp\{\alpha + \beta X + \gamma Y + \xi XY\} \\ &\quad - \frac{\phi(a_{i-1}, a_i)}{1-\phi(a_{i-1}, a_i)} I(a_{i-1} < X \leq a_i), \end{aligned}$$

$i = 1, 2, \dots, I$. Let

$$g(X, Y) = (g_0(X, Y), g_1(X, Y), \dots, g_I(X, Y)).$$

Then $E_0[g(X, Y)] = 0$. The log-likelihood is

$$\ell = \sum_{i=1}^n d_i [\log\{(1-\pi)/\pi\} + \alpha + \beta x_i + \gamma y_i + \xi x_i y_i] + \sum_{i=1}^n \log p_i,$$

where $p_i = dF_0(x_i)$, $i = 1, 2, \dots, n$, and the constraints are

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1,$$

and

$$\sum_{i=1}^n p_i g(x_i, y_i) = 0.$$

The profile log-likelihood is

$$\ell = \sum_{i=1}^n D_i [\eta + \alpha + \beta x_i + \gamma y_i + \xi x_i y_i] - \sum_{i=1}^n \log[1 + \lambda^T g(x_i, y_i)],$$

where the Lagrange multiplier λ is determined by the constraint equation

$$\sum_{i=1}^n \frac{g(x_i, y_i)}{1 + \lambda^T g(x_i, y_i)} = 0$$

and $\eta = \log\{(1-\pi)/\pi\}$. Finally, the underlying parameters can be obtained by maximizing ℓ .

If the overall disease prevalence probability $\pi = P(D = 1)$ is known, then η is known. On the other hand if it is unknown but $I \geq 1$, then π is identifiable. If $I > 1$, then we have an over-identified parameter problem. This can be treated as a generalization of the empirical likelihood method for estimating functions (Qin and Lawless 1994) to the biased sampling problems. Qin et al. (2015) considered the case that η is unknown and $I \geq 1$.

Let

$$\omega = (\eta, \alpha, \beta, \gamma, \xi, \lambda).$$

Since the first estimating function g_0 corrects biased sampling in a case-control study, the remaining estimating functions g_1, \dots, g_I are used for improving efficiency; the limiting value of λ is $\lambda_0 = n_1/n$, and $\lambda_1 = \dots = \lambda_I = 0$.

Qin et al. (2015)'s results can be summarized as follows.

If $\rho = n_1/n_0$ remains constant as $n \rightarrow \infty$ and $\rho \in (0, 1)$. Then under suitable regularity conditions $\sqrt{n}(\hat{\omega} - \omega_0)$ is asymptotically normally distributed with mean 0 and covariate matrix Σ specified in (A4) of the Appendix of their paper. Moreover, estimation of the logistic regression parameters (β, γ, ξ) improves as the number I of estimating functions gets larger. This means that, theoretically, a richer set of

auxiliary information will lead to better estimates. In practice, however, this must balanced with the numerical difficulty of solving a larger number of equations.

It is interesting to note that, auxiliary information is primarily informative for estimating β and ξ , but not for estimating γ . This can be observed through the following equations

$$\begin{aligned} & \int I(a < x < b) \exp(\alpha + \beta x + \gamma y + \xi xy) dF_0(x, y) \\ &= \int I(a < x < b) \exp(\alpha + \beta x + s + \xi xs/\gamma) dF_0(x, s/\gamma). \end{aligned}$$

Since the underlying distribution $F_0(x, y)$ is not specified, we can treat $F_0(x, s/\gamma)$ as a new underlying distribution $F_0^*(x, s)$. After profiling out F_0^* , the auxiliary information equation does not involve γ if $\xi = 0$. Hence, even if $\xi \neq 0$, the information for γ is minimal since γ and ξ cannot be separated.

Generalizations

In Qin et al. (2015)'s simulation studies, it looks like the maximum reduction of variance occurs for the coefficient of X . If the auxiliary information

$$P(D = 1 | b_{j-1} < Y \leq b_j) = \psi_j, \quad j = 1, 2, \dots, J$$

is also available, naturally we can combine them through estimating equations

$$\begin{aligned} g_i(X, Y) &= I(a_{i-1} < X \leq a_i) \exp\{\alpha + \beta X + \gamma Y + \xi XY\} \\ &\quad - \frac{\phi(a_{i-1}, a_i)}{1 - \phi(a_{i-1}, a_i)} I(a_{i-1} < X \leq a_i), \end{aligned}$$

$$\begin{aligned} h_j(X, Y) &= I(b_{j-1} < Y \leq b_j) \exp\{\alpha + \beta X + \gamma Y + \xi XY\} \\ &\quad - \frac{\psi(b_{j-1}, b_j)}{1 - \psi(b_{j-1}, b_j)} I(b_{j-1} < Y \leq b_j). \end{aligned}$$

It would be more informative if the auxiliary information $P(D = 1 | a < X < b, c < Y < d)$ is available.

More on the Use of Auxiliary Information in Case-Control Study

With a logistic regression model, the case-control densities are linked by the exponential tilting

$$f_1(x, y | D = 1) = f_0(x, y | D = 0) \exp(\alpha + x\beta + y\gamma + \xi xy).$$

Suppose in the general population we know

$$E(X) = \mu_1, \quad E(Y) = \mu_2, \quad E(XY) = \mu_3,$$

and $\pi = P(D = 1)$ is known or can be estimated using other data. Under the logistic regression model (14.4.6), the density $f(x, y)$ in the general population and the density $f(x, y|D = 0)$ in the control population are linked by

$$\begin{aligned} f(x, y) &= P(D = 1)f(x, y|D = 1) + P(D = 0)f(x, y|D = 0) \\ &= \{\pi \exp(\alpha + X\beta + Y\gamma + \xi XY) + (1 - \pi)\}f(x, y|D = 0). \end{aligned}$$

As a consequence

$$\begin{aligned} E(X) &= \pi E(X|D = 1) + (1 - \pi)E(X|D = 0) \\ &= E_0[X\{\pi \exp(\alpha + X\beta + Y\gamma + \xi XY) + (1 - \pi)\}] = \mu_1. \end{aligned}$$

The case-control log-likelihood is

$$\ell = \sum_{i=1}^n \log p_i + \sum_{i=1}^n D_i(\alpha + x_i\beta + y_i\gamma + x_i y_i \xi)$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0,$$

$$\sum_{i=1}^n p_i \exp(\alpha + x_i\beta + y_i\gamma + x_i y_i \xi) = 1$$

and

$$\sum_{i=1}^n p_i h(x_i, y_i) \{\pi \exp(\alpha + x_i\beta + y_i\gamma + x_i y_i \xi) + (1 - \pi)\} = 0,$$

where

$$h(x, y) = (x - \mu_1, y - \mu_2, xy - \mu_3)$$

with known μ_1, μ_2, μ_3 .

More generally, any information in the general population such as

$$E[\psi(Y, X)] = 0,$$

can be converted to the control population through

$$E_0[\{\pi \exp(\alpha + X\beta + Y\gamma + \xi XY) + (1 - \pi)\}\psi(Y, X)] = 0. \quad (14.4.8)$$

Therefore the results developed in Qin et al. (2015) can be applied too. Chatterjee et al. (2016)'s results for case-control data can be considered a special case of Qin et al. (2015).

14.5 Case-Control Study for Haplotype Data

Assuming a retrospective study design where a sample of n unrelated subjects, consisting of n_0 controls and n_1 cases are collected. Epstein and Satten (2003) and Satten and Epstein (2004) proposed maximum semiparametric estimation on haplotype effects in case-control studies using unphased genotype data. Their method can be implemented using the Chaplin case-control haplotype inference software:

<http://genetics.emory.edu/labs/epstein/software/chaplin/doc/chaplin.pdf>

First we consider the case where the haplotype is available. Given two phased genotypes, the probability of disease is given by

$$P\{D = 1|H = (h, h')\} = 1 - \frac{1}{1 + \exp[\alpha + \phi(h, h', \beta)]},$$

where $\phi(h, h', \beta)$ is a function of (h, h') up to an unknown parameter β . Assume the haplotype h in the control population is in Hardy-Weinberg Equilibrium (HWE) such that

$$\pi_{H_1=h, H_2=h'|D=0} = P(H_1 = h|D = 0)P(H_2 = h'|D = 0),$$

i.e., the two phases are independent of each other in the control population. We are interested in estimating β .

Using Bayes' formula, again we can find that the logistic regression model is equivalent to

$$P\{H = (h, h')|D = 1\} = P\{H = (h, h')|D = 0\}\exp[\alpha^* + \phi(h, h', \beta)],$$

where $\alpha^* = \alpha + \log\{P(D = 0)/P(D = 1)\}$ and the probability $P\{H = (h, h')|D = 0\}$ is unspecified. With the second assumption of HWE in the control population, we have

$$P\{H = (h, h')|D = 0\} = P(H_1 = h|D = 0)P(H_2 = h'|D = 0).$$

Furthermore we can assume $P(H_1 = h|D = 0) = P(H_2 = h|D = 0)$, i.e., the two phases have the same marginal probability distribution. Denote the observed data from the control group as

$$\{H_i^0 = (H_{1i}^0, H_{2i}^0) = (h_{1i}^0, h_{2i}^0), \quad i = 1, 2, \dots, n_0\},$$

and from the case group as

$$\{H_i^1 = (H_{1i}^1, H_{2i}^1) = (h_{1i}^1, h_{2i}^1), \quad i = 1, 2, \dots, n_1\}.$$

Let

$$(t_1, \dots, t_n) = (h_{1i}^0, h_{2i}^0, i = 1, 2, \dots, n_0; h_{1j}^1, h_{2j}^1, j = 1, 2, \dots, n_1), \quad n = 2n_0 + 2n_1$$

be the pooled data. Denote the corresponding probability as $p_i = P(H_1 = t_i) = P(H_2 = t_i)$, $i = 1, 2, \dots, n$. The constraints become

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0,$$

and

$$\sum_{i=1}^n \sum_{j=1}^n p_i p_j \exp\{\alpha^* + \phi(t_i, t_j, \beta)\} = 1.$$

Then the likelihood is

$$L = \prod_{i=1}^{n_0} P\{H_i^0 = (h_{1i}^0, h_{2i}^0) | D = 0\} \prod_{j=1}^{n_1} P\{H_j^1 = (h_{1j}^1, h_{2j}^1) | D = 1\},$$

which can be rewritten as

$$\begin{aligned} L(p, \beta) &= \prod_{i=1}^{n_0} P(H_{1i}^0 = h_{1i}^0 | D = 0) P(H_{2i}^0 = h_{2i}^0 | D = 0) \\ &\quad \prod_{j=1}^{n_1} P(H_{1j}^1 = h_{1j}^1 | D = 0) P(H_{2j}^1 = h_{2j}^1 | D = 0) \exp[\alpha^* + \phi(h_{1j}^1, h_{2j}^1, \beta)] \\ &= \left(\prod_{i=1}^n p_i \right) \left[\prod_{j=1}^{n_1} \exp\{\alpha^* + \phi(h_{1j}^1, h_{2j}^1, \beta)\} \right]. \end{aligned}$$

For fixed β , we need to maximize

$$\prod_{i=1}^n p_i$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n \sum_{j=1}^n p_i p_j \exp[\alpha^* + \phi(t_i, t_j, \beta)] = 1.$$

After profiling out $p_i = p_i(\beta)$, $i = 1, 2, \dots, n$, we can replace them in the likelihood $L(p(\beta), \beta)$. It can be shown that the maximum profile likelihood estimator $\hat{\beta}$ has an asymptotic normal distribution. Also confidence intervals for β or a subset of β can be constructed by using the profile likelihood ratio statistics.

Phase is Unknown

When phase is unknown, the probabilities for observed genotypes for control and case are, respectively,

$$\begin{aligned} P(G = g|D = 0) &= \sum_{(h, h') \in S(g)} P\{H = (h, h')|D = 0\} \\ &= \sum_{(h, h') \in S(g)} P\{H_1 = h|D = 0\}P\{H_2 = h'|D = 0\}, \end{aligned}$$

$$\begin{aligned} P(G = g|D = 1) &= \sum_{(h, h') \in S(g)} P\{H = (h, h')|D = 1\} \\ &= \sum_{(h, h') \in S(g)} P\{H_1 = h|D = 0\}P(H_2 = h'|D = 0) \exp[\alpha^* + \phi(x_{h, h'}, \beta)], \end{aligned}$$

where $S(g)$ denotes the set of haplotype pairs $\{H = (h, h')\}$ consistent with the observed genotype $G = g$. We have adopted the convention that $(h, h') \in S(g)$ directly implies that $(h, h') \in S(g)$ for $h \neq h'$. Again it is possible to profile out $P\{H_1 = h|D = 0\}$ and $P(H_2 = h'|D = 0)$ from the joint likelihood. For details, see Epstein and Satten (2003).

Chapter 15

Outcome Dependent Sampling and Maximum Rank Estimation

The case-control sampling for discrete outcomes has been generalized to continuous outcomes. The so-called outcome dependent sampling is a sampling scheme that depends on the outcome. In an outcome dependent sampling, the tails of a distribution are over-sampled to compensate for the low probability of making observation in the tails under a random sampling. For example, a covariate X will be sampled for those Y 's with $Y < c_0$ or $Y > c_1$, where c_0 and c_1 are two fixed numbers. Pioneering works on outcome dependent sampling have been done by, among others, Hausman and Wise (1981), Jewell (1985), Gill et al. (1988), Bickel and Ritov (1991) and Kalbfleisch and Lawless (1988). Imbens and Lancaster (1996), Liang and Qin (2000), Chen (2001) and Zhou et al. (2002) discussed semiparametric regression analyses based on outcome dependent sampled data. Godambe and Vijayan (1996) discussed estimating function based approaches. Recently Tao et al. (2015) have analyzed genetic sequence data using multivariate trait-dependent sampling.

There are two main approaches for outcome dependent sampling. The first approach assumes a linear regression model between a response Y and covariate(s) X , where the conditional distribution of Y given X and the marginal distribution of X are not specified. The second one postulates a parametric model $f(y|x) = f(y|x, \beta)$ with the marginal distribution of X left unspecified.

In many econometrical applications, it is reasonable to assume a monotonic relationship between a response variable and its index covariates. Manski (1975), Cosslett (1983), Han (1987), Cavanagh and Sherman (1998) and Abrevaya (1999) presented a class of rank estimators of scaled coefficients in semiparametric monotonic linear index models. In contrast to kernel estimation methods, rank based estimators do not require subjective bandwidth choices. Chen et al. (2014) showed that Han (1987) maximum rank correlation estimation is also valid for outcome dependent sampling data.

15.1 Linear Regression for Outcome Dependent Sampling

1. An iterative estimation method

Jewell (1985) considered an outcome dependent sampling problem under a linear regression model assumption

$$y = x\beta + \epsilon, \quad \epsilon \sim f(\epsilon).$$

Let

$$-\infty = K_0 < K_1 < \dots < K_{m-1} < K_m = +\infty$$

be a partition of $(-\infty, +\infty)$. Define

$$A_j = (K_{j-1}, K_j], \quad j = 1, 2, \dots, m.$$

Let $D_i = 1$ if the i -th individual is selected, and 0 otherwise. Given Y and X , the selection probability is a piecewise constant

$$P(D_i = 1|Y_i = y_i, X_i = x_i) = P(D_i = 1|y_i) = \pi(y_i) = p_j, \quad \text{if } y_i \in A_j.$$

This leads to

$$P(Y = y|D = 1, X = x) = \frac{P(Y = y, D = 1|x)}{P(D = 1|x)} = \frac{f(y|x)\pi(y)}{\int \pi(y)f(y|x)dy} = \frac{f(y|x) \sum_{j=1}^m p_j I(y \in A_j)}{\sum_{j=1}^m p_j P(Y \in A_j|x)}.$$

Let

$$\Delta(x) = \sum_{j=1}^m p_j P(Y \in A_j|x) = \sum_{j=1}^m p_j \int_{K_{j-1}-x\beta}^{K_j-x\beta} dF(\epsilon).$$

Then

$$\begin{aligned} E(Y|D = 1, X = x) &= \frac{\sum_{j=1}^m p_j \int y f(y|x) I(y \in A_j) dy}{\Delta(x)} \\ &= x\beta + \frac{\sum_{j=1}^m p_j \int (y - x\beta) f(y - x\beta) I(y \in A_j) dy}{\Delta(x)} \\ &= x\beta + \frac{\sum_{j=1}^m p_j \mu_j(x)}{\Delta(x)}, \quad \mu_j(x) = \int_{K_{j-1}-x\beta}^{K_j-x\beta} \epsilon dF(\epsilon). \end{aligned}$$

Define a pseudo random variable

$$Y^* = \Delta(x)Y/\pi(Y) = \Delta(x) \sum_{i=1}^m Y p_i^{-1} I(Y \in A_i), \quad i = 1, 2, \dots, m.$$

We can find

$$E[Y^*|D = 1, X = x] = \Delta(x) \frac{\sum_{j=1}^m p_j p_j^{-1} \int y f(y|x) I(y \in A_j) dy}{\Delta(x)} = \int y f(y|x) dy = x\beta.$$

Hence, β can be estimated by solving

$$\sum_{i=1}^n x_i (y_i^* - x_i \beta) = 0.$$

However, an iterative method is needed since the pseudo random variable Y^* involves the unknown parameter β and baseline density f . This bias corrected estimator is analogous to the Buckley and Buckley and James (1979) estimator for right censored data. We will discuss it in Chap. 24.

The joint distribution of (Y, X) given $D = 1$ is

$$(Y, X)|D = 1 \sim \frac{\pi(y)dF(x, y)}{\int \int \pi(y)dF(x, y)}.$$

Using Vardi's (1985) results, $F(x, y)$ can be estimated by

$$\hat{F}(x, y) = \frac{\sum_{i=1}^n \pi^{-1}(y_i) I(y_i \leq y, x_i \leq x)}{\sum_{i=1}^n \pi^{-1}(y_i)}.$$

The error distribution of $\epsilon = y - x\beta$ can be estimated by

$$\hat{F}(\epsilon) = \frac{\sum_{i=1}^n \pi^{-1}(y_i) I\{(y_i - x_i \beta) \leq \epsilon\}}{\sum_{i=1}^n \pi^{-1}(y_i)}.$$

Replacing F by \hat{F} in $\mu_j(x)$, we can iteratively estimate β and F .

In stratified sampling, n_i samples are selected from the conditional distribution $Y|Y \in A_i$. The selection probability is

$$\pi(y) = p_i = n_i/n, \quad y \in A_i, \quad i = 1, 2, \dots, m.$$

Note

$$E \left[x \frac{Y - x\beta}{\pi(Y)} \right] = 0.$$

This simple estimating equation can be used to estimate β . It would be interesting to compare this approach with the iterative estimation method.

2. An alternative iterative estimation method

Hausman and Wise (1981) considered a more general stratified sampling problem in linear regression models. Let

$$Y = X\beta + Z\gamma + \epsilon_1, \quad Z = X\xi + \epsilon_2.$$

Upon substitution, we have

$$Y = X(\beta + \gamma\xi) + \epsilon, \quad \epsilon = \epsilon_1 + \gamma\epsilon_2.$$

Without loss of generality we assume that ϵ_1 and ϵ_2 are independent of each other. Suppose the range of Z can be divided into m disjoint intervals S_j , $j = 1, 2, \dots, m$. Denote $D = 1$ if an individual is sampled, and 0, otherwise. Then the selection probability can be written as

$$P(D = 1|Y = y, X = x, Z = z) = P(D = 1|z) = w(z) = p_j, \quad \text{if } z \in S_j.$$

Suppose the observed data are

$$(x_{ij}, y_{ij}, z_{ij}), \quad z_{ij} \in S_i, \quad j = 1, 2, \dots, n_i; i = 1, 2, \dots, m.$$

Then conditioning on $D = 1$, the joint density of (Y, X, Z) is

$$f(y, x, z|D = 1) = \frac{P(D = 1|y, x, z)f(y, x, z)}{P(D = 1)} = \frac{w(z)f(y, x, z)}{W}, \quad W = \int w(z)f(z)dz.$$

At each observation (x_{ij}, y_{ij}, z_{ij}) , $z_{ij} \in S_i$, Vardi's nonparametric MLE mass is given by

$$p_i^{-1} / \left[\sum_{j=1}^m n_j / p_j \right].$$

Furthermore, given $D = 1$, the joint density of (Y, X) and the conditional density of $(Y|D = 1, X)$ are, respectively,

$$f(y, x|D = 1) = \frac{\int w(z)f(x, y, z)dz}{W},$$

$$f(y|D = 1, x) = \frac{P(Y = y, D = 1|x)}{P(D = 1|x)} = \frac{\int w(z)f(y, z|x)dz}{\int w(z)f(z|x)dz}.$$

Using these, we can write out the conditional expectation as

$$\begin{aligned} \int y f(y|D = 1, x) dy &= \frac{\int w(z)(x\beta + z\gamma)f(z|x)dz}{\int w(z)f(z|x)dz} \\ &= x\beta + \gamma \frac{\sum_{j=1}^m p_j \mu_j(x)}{\sum_{i=1}^m p_i G_i(x)}, \end{aligned}$$

where

$$\mu_i(x) = \int_{K_{i-1}}^{K_i} z f(z - x\xi) dz = \int_{K_{j-1}-x\xi}^{K_j-x\xi} (t + x\xi) f(t) dt,$$

$$\sum_{j=1}^m p_j \mu_j(x) = \sum_{j=1}^m p_j \int_{K_{j-1}-x\xi}^{K_j-x\xi} t f(t) dt + x\xi \sum_{j=1}^m p_j G_j,$$

$$G_j(x) = \int_{K_{j-1}}^{K_j} f(z - x\xi) dz = \int_{K_{j-1}-x\xi}^{K_j-x\xi} f(t) dt.$$

Estimation of β and γ is straightforward since

$$E[Y|x, z] = x\beta + z\gamma.$$

We can replace F in $\mu_i(x)$ and $G_j(x)$ by Vardi (1985) nonparametric MLE, and then iteratively estimate (β, γ, ξ) and F (Quesenberry and Jewell 1986). Note under the linear model assumption, the joint density of (Y, X, Z) given $D = 1$ is

$$f(y, x, z|D = 1) = \frac{w(z)f(y - x\beta - z\gamma)g(z - x\xi)h(x)}{W}, \quad W = \int w(z)g(z - x\xi)h(x)dx dz.$$

Vardi (1985) nonparametric MLE for the joint distribution of (Y, X, Z) may be rather inefficient. As a consequence the iterative estimation method discussed above may not be the most efficient one. Some future research is warranted.

3. Cosslett's approach (Journal of Econometrics 2013)

Finding the maximum likelihood estimate for outcome dependent sampling is a challenging task under a semiparametric linear model assumption. Note that

$$P(Y = y, X = x|Y > 0) = \frac{f(y - x\beta)g(x)}{\int F(-x\beta)dG(x)} = \frac{f(y - x\beta)}{\bar{F}(-x\beta)} \frac{\bar{F}(-x\beta)dG(x)}{\int \bar{F}(-x\beta)dG(x)}.$$

Similarly

$$P(Y = y, X = x|Y < 0) = \frac{f(y - x\beta)}{\bar{F}(-x\beta)} \frac{F(-x\beta)dG(x)}{\int F(-x\beta)dG(x)}.$$

Bickel and Ritov (1991) discussed biased sampling under a linear model setup when F and G are both unspecified. However they made the assumption that X has finite support points. This assumption may not be desirable in practical applications.

By using a nonparametric smoothing method, Cosslett (2013) found the semi-parametric maximum likelihood estimate of β . However, his method is not easy to be implemented.

Under the length biased sampling ($\pi(y) = y$), we already discussed the regression parameter estimation in the accelerated failure time regression model in Sect. 5.3. More discussions based on right censored data will be given in Sect. 25.6.

15.2 Semiparametric Approach

Next we study semiparametric approaches, where the conditional density of Y given X is given by a parametric model

$$f(y|x) = f(y|x\beta)$$

and the marginal density $g(x)$ of X is not specified. Let A_j , $j = 1, 2, \dots, m$ be subsets of the domain of Y . We can find

$$\begin{aligned} P(Y = y, X = x|Y \in A_j) &= \frac{P(Y = y, X = x, Y \in A_j)}{P(Y \in A_j)} \\ &= \frac{f(y|x\beta)dG(x)}{P(Y \in A_j)} \\ &= \frac{f(y|x\beta)}{P(Y \in A_j|x\beta)} \frac{P(Y \in A_j|x\beta)dG(x)}{\int P(Y \in A_j|x\beta)dG(x)}. \end{aligned}$$

Denote

$$(Y_{ij}, X_{ij}), i = 1, 2, \dots, n_j, \quad j = 1, 2, \dots, m$$

as a random sample from $P(Y = y, X = x|Y \in A_j)$. The likelihood is

$$\prod_{j=1}^m \prod_{i=1}^{n_j} \frac{f(y_{ij}|x_{ij}\beta)}{P(Y \in A_j|x_{ij}\beta)} \frac{P(Y \in A_j|x_{ij}\beta)dG(x_{ij})}{\int P(Y \in A_j|x\beta)dG(x)}.$$

The log-likelihood can be written as

$$\ell = \ell_c + \ell_M,$$

where

$$\ell_c = \sum_{j=1}^m \sum_{i=1}^{n_j} [\log f(y_{ij}|x_{ij}\beta) - \log P(Y \in A_j|x_{ij}\beta)]$$

is a full parametric log-likelihood, and

$$\ell_M = \sum_{j=1}^m \sum_{i=1}^{n_j} [\log P(Y \in A_j|x_{ij}\beta) + \log p_{ij} - n_j \log B_j], \quad B_k = \sum_{j=1}^m \sum_{i=1}^{n_j} p_{ij} P(A_k|x_{ij}\beta),$$

is a log semiparametric likelihood discussed in Chaps. 10 and 11 and $p_{ih} = dG(x_{ij})$, $i = 1, \dots, n_j$; $j = 1, 2, \dots, m$. For fixed β we can profile out p_{ij} subject to the constraints $\sum_{j=1}^m \sum_{i=1}^{n_j} p_{ij} = 1$, $p_{ij} \geq 0$. Then the second step is to maximize the profile likelihood with respect to β . Zhou et al. (2002) derived the

large sample results. Moreover they proved if $H_0 : \beta = \beta_0$, then the likelihood ratio statistic, defined by

$$2[\max_{\beta} \ell(\beta) - \ell(\beta_0)],$$

converges in distribution to a chi-squared distribution with degree of freedom p , the dimension of β .

An asymptotic equivalent estimator using GMM method was given by Imbens and Lancaster (1996).

A Choice Based Sampling for a Continuous Random Variable

Under the prospective sampling, (X, Y) are jointly sampled from the joint density $f(y|x)g(x)$. In a choice based sampling, Y is first sampled based on a density $h(y)$, where not necessarily $h(y) = \int f(y|x)g(x)dx$. Then given $Y = y$, X is sampled from the conditional density $f(x|y)$. The joint density based on this choice based sampling is

$$h(y)f(x|y) = h(y) \frac{f(y|x\beta)dG(x)}{\int f(y|x\beta)dG(x)}.$$

Practically this sampling design is easy to accomplish if Y is a discrete variable, for example, as in the case of case-control studies, where $Y = 0$ or 1 . However, it is not easy if Y is a continuous variable because there are very few individuals in which $Y = y$ exactly or approximately for whom X can be sampled.

Let (y_i, x_i) , $i = 1, 2, \dots, n$ be the observed data. The likelihood is

$$L = \left\{ \prod_{i=1}^n h(y_i) \right\} \left\{ \prod_{i=1}^n \frac{f(y_i|x_i\beta)dG(x_i)}{\int f(y_i|x\beta)dG(x)} \right\}.$$

Denote $p_i = dG(x_i)$, $i = 1, 2, \dots, n$, then we need to maximize

$$\prod_{i=1}^n \frac{f(y_i|x_i\beta)p_i}{\sum_{j=1}^n f(y_j|x_j\beta)p_j}$$

subject to the constraints

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1.$$

Details can be found in Chen (2001).

Connection with Truncation Problems

Truncation models (which will be discussed in Chap. 25) have been popular since the 1990's. There is a natural connection between truncation models and outcome dependent sampling. In a truncation model, an observation with outcome Y is available if and only if $Y > T$, where T is the truncation variable. Assume T and Y are

independent of each other and let X be a covariate and

$$Y|X \sim f(y|x\beta), \quad X \sim g(x), \quad T \sim h(t),$$

where $f(y|x\beta)$ is a specified parametric model and $g(x)$ and $h(t)$ are unspecified. The observed data are distributed as

$$(Y, T)|X, Y > T \sim \frac{f(y|x\beta)h(t)}{\int \bar{F}(t|x\beta)h(t)dt},$$

where $\bar{F}(t|x\beta)$ is the survival function corresponding to $f(y|x\beta)$. The joint conditional density of (Y, T) given X and $Y > T$ can be decomposed as

$$\frac{h(t)}{H(y)} \frac{H(y)f(y|x\beta)}{\int H(y)f(y|x)dy} = \frac{f(y|x\beta)}{\bar{F}(t|x\beta)} \frac{\bar{F}(t|x\beta)h(t)}{\int \bar{F}(t|x\beta)h(t)dt},$$

where H is the cumulative distribution of T . If T takes finitely many values, this model is equivalent to outcome dependent sampling.

15.3 Manski's Maximum Rank Score Method

In this section we discuss a binomial choice model and the maximum rank score method that is widely used in the econometric literature. So far we have assumed a parametric link function for binary response data. It is possible to relax this assumption to

$$P(Y = 1|x) = F(x\beta), \quad F(0) = 1/2,$$

where F is an unknown monotonic non-decreasing function. Under this formulation the intercept is absorbed by F . Furthermore, for a scalar constant c , $F(cx\beta) = F^*(x\beta)$, where $F^*(t) = F(ct)$. Without loss of generality we may assume there is no intercept, $F(0) = 1/2$ and $\beta^T\beta = 1$. It is also customary to assume the first component of β to be 1.

To estimate β , Manski (1975) defined the score function as

$$S_n(\beta) = \sum_{i=1}^n y_i I(x_i\beta \geq 0) + (1 - y_i) I(x_i\beta < 0) = \sum_{i=1}^n (2y_i - 1) I(x_i\beta \geq 0) + (1 - y_i).$$

Note that

$$E\left\{\sum_{i=1}^n (2y_i - 1) I(x_i\beta \geq 0)\right\} = \sum_{i=1}^n \{2F(x_i\beta_0) - 1\} I(x_i\beta \geq 0).$$

Since we have assumed that $F(0) = 1/2$, $x_i\beta \geq 0$ implies that $F(x_i\beta) \geq 1/2$. In fact the score is the number of correct predictions if y_i is simply assumed to be 1 whenever $x_i\beta \geq 0$ and 0 otherwise. Manski (1975) proposed estimating β by

$$\max_{\beta} S_n(\beta)$$

subject to the constraint $\beta^T \beta = 1$. Clearly $S_n(c\beta) = S_n(\beta)$ for any $c > 0$.

It was demonstrated in Sect. 6.6 of Chap. 6 that there is no root- n consistent estimator in this example. Kim and Pollard (1990) showed that the maximum score estimator has a cubic root- n consistency property, with a non-normal limiting distribution. We defer the discussion on the consistency of Manski (1975) maximum score estimation to the next section after introducing Han's maximum correlation coefficient approach. Han (1987) generalized Manski's method from discrete response model to continuous response transformation models.

15.4 Han's Maximum Rank Correlation Estimation

The maximum rank correlation method was proposed by Han (1987) under a prospective study setup. This method is based on Kendall's τ rank correlation coefficient between two continuous variables. Consider a simple linear regression model

$$Y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \dots, n.$$

If $x_i\beta > x_j\beta$, then it is very likely that $Y_i > Y_j$ since the random errors ϵ_i and ϵ_j would be dominated by $x_i\beta$ and $x_j\beta$. By noting the invariance of the rank statistic under monotonic transformations, Han (1987) considered a model specified by

$$Y = D(F(x\beta, \epsilon)),$$

where the composite transformation $D : R^1 \rightarrow R^1$ is non-degenerate monotonic and $F : R^2 \rightarrow R^1$ is strictly monotonic in each of its variates. This model does not require the separability of the systematic and the random components. Clearly the transformation linear model

$$h(y) = x\beta + \epsilon$$

is a special case of Han's model, where $h(y)$ is an unknown transformation function and the error distribution F_ϵ of ϵ is also unknown.

Under the transformation linear model, for $1 \leq i \neq j \leq n$,

$$P(Y_i < Y_j | x_i, x_j) = P(h(Y_i) < h(Y_j) | x_i, x_j) = P(\epsilon_i - \epsilon_j < (x_j - x_i)^T \beta) = F((x_j - x_i)\beta),$$

where F is the cumulative distribution function of $\epsilon_i - \epsilon_j$, which is symmetric around 0. If we treat $I(Y_i \leq Y_j)$ as an indicator function and $X_j - X_i$ as a covariate, then this becomes the binomial model discussed by Manski (1975).

Define

$$U_n(\beta) = \frac{2}{n^2 - n} \sum_{i < j} \{I(Y_i < Y_j)I(X_i\beta < X_j\beta) + I(Y_i > Y_j)I(X_i\beta > X_j\beta)\}.$$

Equivalently

$$U_n^*(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} I(Y_i < Y_j)I(X_i\beta < X_j\beta). \quad (15.4.1)$$

Han (1987) proposed to estimate β through

$$\max_{\beta} U_n(\beta)$$

subject to the constraint $\beta^T \beta = 1$.

By the Law of Large Numbers, it can be shown in probability

$$U_n(\beta) \rightarrow E[I(X_i\beta < X_j\beta)F\{(X_j - X_i)\beta_0\}] =: U(\beta, \beta_0).$$

We would like to show that under certain regularity conditions, $U(\beta, \beta_0)$ achieves the maximum if and only if $\beta = \beta_0$. It suffices if we can show that

$$U(\beta_0, \beta_0) > U(\beta, \beta_0),$$

for $\beta \neq \beta_0$. The above inequality is equivalent to

$$\begin{aligned} & E[I(X_i\beta > X_j\beta)I(X_i\beta_0 < X_j\beta_0)F\{(X_j - X_i)\beta_0\}] \\ & > E[I(X_i\beta < X_j\beta)I(X_i\beta_0 > X_j\beta_0)F\{(X_j - X_i)\beta_0\}]. \end{aligned}$$

If

$$E[I(X_i\beta > X_j\beta)I(X_i\beta_0 < X_j\beta_0)] > 0, \quad E[I(X_i\beta < X_j\beta)I(X_i\beta_0 > X_j\beta_0)] > 0$$

for any $\beta \neq \beta_0$, then $F\{(x_j - x_i)\beta_0\} > 1/2$ in the left hand side but $< 1/2$ in the right hand side. Therefore we require

$$P[(X_j - X_j)\beta] \neq P[(X_j - X_i)\beta_0]$$

for any $\beta \neq \beta_0$.

There are n and $n(n - 1)/2$ terms in the definitions of Manski's (1975) maximum score estimator and Han (1987) maximum correlation coefficient estimator, respectively. As a consequence the resulting estimators have different convergence rates. Sherman (1993) showed that Han's maximum rank correlation estimator is root- n consistent and with a limiting distribution of $n^{1/2}(\hat{\beta} - \beta)$. Furthermore instead of maximizing (15.4.1), Cavanagh and Sherman (1998) maximized

$$U_{CS}(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} (Y_i - Y_j) I(X_i \beta < X_j \beta). \quad (15.4.2)$$

with respect to β .

Exercise Show that $E[U_{CS}(\beta)]$ achieves the maximum if and only if at $\beta = \beta_0$, where $E(Y|x) = \mu(x\beta)$ is a monotonic function of $x\beta$ and X has a non degenerated distribution.

15.5 Triple and Quadruple Wise Rank Likelihood

Abrevaya (1999) extended Han's pairwise approach to triple-wise or quadruple-wise approach. By using the fact that

$$P(Y_i < Y_j | x_i, x_j) = F((x_j - x_i)\beta), \quad P(Y_k < Y_l | x_k, x_l) = F((x_l - x_k)\beta)$$

and $P(Y_i < Y_j | x_i, x_j) \leq P(Y_k < Y_l | x_k, x_l)$ if and only if $(x_j - x_i)\beta \leq (x_l - x_k)\beta$, Abrevaya (1999) proposed maximizing

$$\sum_{i,j,k,l} I\{(x_j - x_i)\beta \leq (x_l - x_k)\beta\} [I(y_i \leq y_j) - I(y_k \leq y_l)],$$

where the summation is over all distinct i, j, k, l . In general this estimator is more efficient than Han (1987) pairwise maximum rank approach. The drawback of this approach, however, is the computational burden. The order of summation terms is $O(n^4)$. Moreover, neither Han (1987) and Abrevaya (1999) addressed the problems on estimation of baseline distribution function $F_\epsilon(\cdot)$ and transformation function estimation $h(\cdot)$.

Instead of pairwise, triple-wise or quadruple-wise comparisons, Yu et al. (2017) recently have proposed a more effective estimation method by comparing distance among all pairs in the pairwise likelihood. They used the pool adjacent violation algorithm to account for all possible ranks of $(x_j - x_i)\beta$, $i \neq j$. We will discuss this in details in Chap. 26.

Furthermore Abrevaya (1999) proposed a method called leapfrog estimation for panel data analyses. Consider a panel data model

$$h(Y_{ij}) = \alpha_i + x_{ij}\beta + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n,$$

where ϵ_{ij} are independent and identically distributed, and α_i is a random variable which is independent of ϵ_{ij} . A simple method to eliminate α_i is to use

$$P(Y_{ij} < Y_{ik}|x) = F((x_{ik} - x_{ij})\beta), \quad k \neq j,$$

where F is the distribution function of $\epsilon_{ij} - \epsilon_{ik}$. This method may not be efficient since the pairwise comparisons are carried out only within each stratum. In order to extend the pairwise comparisons to different strata, Abrevaya (1999) used the fact that

$$P(Y_{ij_1} < Y_{kl_1}|x) = F(\alpha_k - \alpha_i + (x_{kl_1} - x_{ij_1})\beta),$$

$$P(Y_{ij_2} < Y_{kl_2}|x) = F(\alpha_k - \alpha_i + (x_{kl_2} - x_{ij_2})\beta).$$

Clearly $P(Y_{ij_1} < Y_{kl_1}|x) \leq P(Y_{ij_2} < Y_{kl_2}|x)$ if and only if $(x_{kl_1} - x_{ij_1})\beta \leq (x_{kl_2} - x_{ij_2})\beta$. Then the quadruple-wise rank objective function is defined as

$$Q(\beta) = \sum_{i \neq k} \sum_{j_1, l_1, j_2, l_2} \{I(Y_{ij_1} < Y_{kl_1}) - I(Y_{ij_2} < Y_{kl_2})\} I\{(x_{kl_1} - x_{ij_1})\beta < (x_{kl_2} - x_{ij_2})\beta\}.$$

The maximum quadruple-wise rank estimator is obtained by maximizing $Q(\beta)$ subject to the constraint $\|\beta\| = 1$, where $\|\cdot\|$ is the Euclidean norm for a vector.

15.6 Retrospective Sampling and Maximum Rank Estimation

Inspired by the fact that response-biased sampling would not change the positive correlation between the ranks of the response and the explanatory variables, Chen et al. (2014) considered maximum rank correlation estimation for transformation models with response-biased sampling. The nice feature of their method is that no additional assumption, such as the specification the form of selection biased sampling function, is needed.

Without loss of generality we assume the coefficient of the first component Z to be 1 in the transformation model,

$$h(Y) = Z + X\beta_0 + \epsilon, \quad \epsilon \sim g(\cdot).$$

Suppose the collected data are conditioned on the values of responses Y_i 's,

$$(Z_i, X_i | Y_i) \sim \frac{f(y|z, x)k(z, x)}{\int f(y|z, x)k(z, x)dzdx}, \quad i = 1, 2, \dots, n,$$

where $k(z, x)$ is the marginal density of (Z, X) . As an example, consider a selection bias sampling problem where the selection indicator D has a conditional probability given by

$$P(D = 1|z, x, y) = P(D = 1|y) = \pi(y).$$

Then the selected data have a density

$$(Z, X|D = 1, y) \sim \frac{\pi(y)f(y|z, x)k(z, x)}{\int \pi(y)f(y|z, x)k(z, x)dzdx} = \frac{f(y|z, x)k(z, x)}{\int f(y|x)k(z, x)dzdx},$$

which is independent of $\pi(y)$. On the other hand if conditioning on (Z, X) and $D = 1$, the density

$$(Y|D = 1, Z, X) \sim \frac{\pi(y)f(y|z, x)}{\int \pi(y)f(y|z, x)dy}$$

depends on $\pi(y)$, which is unknown in most cases.

Define

$$Q_n(\beta) = \frac{1}{n^2 - n} \sum_{i \neq j} I(Y_i < Y_j) I(z_i + x_i\beta < z_j + x_j\beta).$$

The maximum rank estimator is again defined as

$$\hat{\beta} = \operatorname{argmax}_{\beta} Q_n(\beta).$$

Under the mild requirement that the error density g is log-concave, the maximum rank estimator can be used for both prospective or retrospective data. This can be demonstrated using results by Chen et al. (2014). Define a random variable

$$\xi(\beta) = (\beta - \beta_0)(X_2 - X_1),$$

it has a symmetric distribution.

Denote

$$T_1 = Z_1 + X_1\beta_0 \sim f(t_1), \quad T_2 = Z_2 + X_2\beta_0 \sim f(t_2).$$

Noting that $h(Y)$ and (Z, X) are independent given $T = Z + X\beta_0$, and

$$h(Y)|T = Z + X\beta_0 = t \sim g(h(y) - t),$$

we may treat $\xi(\beta)$ as independent of $h(Y)$ for given T when evaluating the conditional expectation. Observe

$$\begin{aligned}
& P(Z_1 + X_1\beta < Z_2 + X_2\beta | Y_1 = y_1, Y_2 = y_2) \\
&= P\{T_1 - T_2 \leq (\beta_0 - \beta)(X_1 - X_2) | T_1 + \epsilon_1 = h(y_1), T_2 + \epsilon_2 = h(y_2)\} \\
&= \frac{P\{T_1 - T_2 \leq (\beta_0 - \beta)(X_1 - X_2), T_1 + \epsilon_1 = h(y_1), T_2 + \epsilon_2 = h(y_2)\}}{P\{T_1 + \epsilon_1 = h(y_1), T_2 + \epsilon_2 = h(y_2)\}} \\
&= \frac{E([P\{T_1 - T_2 \leq (\beta_0 - \beta)(X_1 - X_2), T_1 + \epsilon_1 = h(y_1), T_2 + \epsilon_2 = h(y_2) | T_1 = t_1, T_2 = t_2\}])}{P\{T_1 + \epsilon_1 = h(y_1), T_2 + \epsilon_2 = h(y_2)\}} \\
&= \frac{\int \int P\{t_1 - t_2 \leq (\beta_0 - \beta)(X_1 - X_2)\} f(t_1) f(t_2) g(h(y_1) - t_1) g(h(y_2) - t_2) dt_1 dt_2}{\int f(t_1) g(h(y_1) - t_1) dt_1 \int f(t_2) g(h(y_2) - t_2) dt_2}.
\end{aligned}$$

Following Chen et al. (2014), we can show the consistency of the maximum rank estimate for the response-biased sampling data.

Note that (exercise)

$$2P(\xi > s) = 1 - \text{sgn}(s)P(|\xi(\beta)| < |s|).$$

We only need to show that

$$II = - \int \int \text{sgn}(t_1 - t_2) P(|\xi(\beta)| < |t_1 - t_2|) f(t_1) f(t_2) g(h(y_1) - t_1) g(h(y_2) - t_2) dt_1 dt_2$$

is uniquely maximized at $\beta = \beta_0$. Denote

$$\Delta(|t_1 - t_2|) = P(|\xi(\beta)| < |t_1 - t_2|).$$

Note that

$$\begin{aligned}
II &= \int_{t_1 < t_2} \Delta(|t_1 - t_2|) f(t_1) f(t_2) g(h(y_1) - t_1) g(h(y_2) - t_2) dt_1 dt_2 \\
&\quad - \int_{t_1 > t_2} \Delta(|t_1 - t_2|) f(t_1) f(t_2) g(h(y_1) - t_1) g(h(y_2) - t_2) dt_1 dt_2 \\
&= \int_{t_1 < t_2} \Delta(|t_1 - t_2|) f(t_1) f(t_2) g(h(y_1) - t_1) g(h(y_2) - t_2) dt_1 dt_2 \\
&\quad - \int_{t_1 < t_2} \Delta(|t_1 - t_2|) f(t_1) f(t_2) g(h(y_1) - t_2) g(h(y_2) - t_1) dt_1 dt_2.
\end{aligned}$$

Therefore, it is sufficient to show that the quantity in the square brackets is positive for all $h(y_1) < h(y_2)$ and $t_1 < t_2$. Let $\eta = \log g$, we only need to show that

$$\eta(c_1 - t_1) + \eta(c_2 - t_2) > \eta(c_1 - t_2) + \eta(c_2 - t_1), \quad c_1 < c_2, t_1 < t_2,$$

or

$$\eta(c_1 - t_1) - \eta(c_1 - t_2) > \eta(c_2 - t_1) - \eta(c_2 - t_2),$$

or $\eta(t - t_1) - \eta(t - t_2)$ is a decreasing function.

$$\frac{\partial}{\partial t} \{ \eta(t - t_1) - \eta(t - t_2) \} = \int_{t-t_2}^{t-t_1} \eta''(s) ds < 0$$

since by assumption that g is log-concave.

Note that $\xi(\beta) = (\beta - \beta_0)(x_2 - x_1) = 0$ if and only if $\beta = \beta_0$. $P(|\xi(\beta)| < |t_1 - t_2|)$ has a maximum of 1.

The large sample property can be derived by studying the local behavior of $Q_n(\beta)$ in the ball $||\beta - \beta_0|| \leq n^{-1/3}$. We refer readers to Chen et al. (2014) for details.

A practical resampling method for variance estimation can be based on the following approach. Define

$$Q_n^*(\beta) = \frac{1}{n^2 - n} \sum_{i \neq j} e_i e_j I(Y_i < Y_j) I(z_i + x_i \beta < z_j + x_j \beta),$$

where $e_i, i = 1, 2, \dots, n$ are i.i.d $\text{exp}(1)$. Let

$$\hat{\beta}^* = \operatorname{argmax}_{\beta} Q_n^*(\beta).$$

Then $\sqrt{n}(\hat{\beta} - \beta_0)$ and $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ have the same asymptotic distribution.

Note that $e_i, i = 1, 2, \dots, n$ must be drawn from a distribution function with nonnegative value. Otherwise the orders of $I(Y_i < Y_j)$ and $I(z_i + x_i \beta < z_j + x_j \beta)$ will be changed by the sign of $e_i e_j$. As a consequence maximization of the objective function $Q_n^*(\beta)$ will lead to an inconsistent bootstrap estimator.

Chapter 16

Noncentral Hypergeometric Distribution and Poisson Binomial Distribution

So far we have discussed continuous and mixture of continuous and discrete covariate discrete choice models. In applications, especially in biomedical researches, a series of 2×2 tables have been used extensively. Next we present a fundamental result by Kou and Ying (1996) on the i.i.d. representation of a noncentral hypergeometric distribution as the summation of independent Bernoulli trials with possible different success probabilities. As a consequence, it will be very easy to conduct statistical inferences and to derive the Central Limit Theorem for a series of 2×2 tables.

A Poisson-Binomial distribution of order n is the distribution of a sum $Y = \sum_{i=1}^n X_i$, where X_1, \dots, X_n are independent Bernoulli random variables with success probability $p_i, i = 1, 2, \dots, n$. If all p_i 's are the same, then this model becomes the standard Binomial distribution. Based on Kou and Ying's (1996) result, the non-central hypergeometric distribution can be treated as a special case of the Poisson and Binomial distribution.

16.1 I.I.D. Representation of the Hypergeometric Distribution

Hypergeometric distribution is a basic distribution in the study of two by two contingency tables. It comes from the following application. Suppose an urn contains N_1 and N_2 white and red balls, respectively. Let $N = N_1 + N_2$ be the total number of balls in the urn. Denote X as the number of white ball if M balls are randomly selected without replacement from the urn. Then

$$P(X = x) = \binom{N_1}{x} \binom{N_2}{M-x} / \binom{N}{M}, \quad L \leq x \leq S,$$

where $L = \max(0, M - N_2)$ and $S = \min(N_1, M)$.

As an alternative, the hypergeometric distribution can also be generated as follows. Suppose two coins A and B with same success probability are tossed N_1 and N_2 times, respectively, and let $M_1(M_2)$ denote the total number of heads (tails) out of the $N = N_1 + N_2$ tosses and X the number of heads from the N_1 tosses of coin A . Then conditional on N_1, N_2, M_1, M_2 , X has a hypergeometric distribution as specified above.

On the other hand if the success probabilities π_A and π_B are different, then the conditional distribution is called a noncentral hypergeometric distribution with odds ratio parameter θ ,

$$P(X = x) = \binom{N_1}{x} \binom{N_2}{M-x} \theta^x / \sum_{u=L}^S \binom{N_1}{u} \binom{N_2}{M-u} \theta^u, \quad L \leq x \leq S,$$

where

$$\theta = \pi_A(1 - \pi_B)/[\pi_B(1 - \pi_A)], \quad L = \max(0, M - N_1), \quad S = \min(N_1, M).$$

Let

$$M(t, \theta) = E_\theta[\exp(X \log t)] = E(t^X)$$

be the moment generating function of X . Define

$$\phi(t) = \sum_{u=L}^S \binom{N_1}{u} \binom{N_2}{M-u} t^u.$$

It can easily be shown that

$$M(t, \theta) = E(t^X) = \sum_{x=L}^S \binom{N_1}{x} \binom{N_2}{M-x} \theta^x t^x / \phi(\theta) = \frac{\phi(t\theta)}{\phi(\theta)}.$$

Kou and Ying (1996) showed that all roots of the polynomial $\phi(t)$ are real and non-positive. Let $-\lambda_1, -\lambda_2, \dots, -\lambda_S$, $\lambda_i \geq 0$, $1 \leq i \leq S$ be the roots. Clearly there are exactly L of the roots equal to 0. Therefore for some constant C ,

$$\phi(t) = C \prod_{i=1}^S (t + \lambda_i).$$

Based on this, the moment generating function can be written as

$$M(t, \theta) = \frac{\phi(t\theta)}{\phi(\theta)} = \prod_{i=1}^S \frac{t\theta + \lambda_i}{\theta + \lambda_i} = \prod_{i=1}^S \frac{t + \theta^{-1}\lambda_i}{1 + \theta^{-1}\lambda_i}.$$

On the other hand if U_1, \dots, U_S are i.i.d. uniform random variables in $(0, 1)$, and let

$$Z = \sum_{i=1}^S I(U_i \leq (1 + \theta^{-1} \lambda_i)^{-1}).$$

Then, it can be shown that

$$E(t^Z) = \prod_{i=1}^S \frac{t + \theta^{-1} \lambda_i}{1 + \theta^{-1} \lambda_i}.$$

This is exactly the moment generating function of X defined before. Using the fact that the moment generating function determines a probability distribution uniquely, Kou and Ying (1996) found the following i.i.d. representation.

Theorem 16.1 *In distribution a noncentral hypergeometric random variable X can be represented as*

$$X = \sum_{i=1}^S I(U_i \leq (1 + \theta^{-1} \lambda_i)^{-1}), \quad (16.1.1)$$

where U_1, \dots, U_S are independent and identically distributed as uniform random variables in $(0, 1)$. Moreover in the decomposition, λ_i is independent of the odds ratio parameter θ . In other words they are the same for hypergeometric and noncentral hypergeometric distributions.

From this we can easily show that the mean and variance of X is

$$E_\theta(X) = \sum_{i=1}^S \frac{1}{(1 + \theta^{-1} \lambda_i)}, \quad \text{Var}_\theta(X) = \sum_{i=1}^S \frac{\theta^{-1} \lambda_i}{(1 + \theta^{-1} \lambda_i)^2}.$$

Note that

$$\log \phi(t) = \log C + \sum_{i=1}^S \log(t + \lambda_i).$$

Differentiating this equation with respect to t twice and letting $t = 1$, we can show that

$$E(X) = \sum_{i=1}^S \frac{1}{(1 + \lambda_i)} = M_1 N_1 / N, \quad \text{Var}(X) = \sum_{i=1}^S \frac{\lambda_i}{(1 + \lambda_i)^2} = N_1 N_2 M_1 M_2 / N^2 (N - 1)$$

in the central hypergeometric distribution $\theta = 1$ case.

Based on above i.i.d. representation, it is easy to show the following theorem:

Theorem 16.2 Let θ be the true odds ratio parameter. Then the following three statements are equivalent: (1). $(X - E(X))/\sqrt{\text{Var}_\theta(X)} \rightarrow N(0, 1)$ in distribution. (2). $\text{Var}_\theta(X) \rightarrow \infty$ and (3). $N_1 N_2 M_1 M_2 / N^3 \rightarrow \infty$.

A popular and simple estimator of θ is the empirical odds ratio

$$\hat{\theta}_e = \frac{X(X + N_2 - M_1)}{(N_1 - X)(M_1 - X)}.$$

In fact it is the maximum likelihood estimator for the odds ratio in the 2×2 table when one of its margins is not fixed. Thus, it is intuitively clear that $\hat{\theta}_e$ should be close to the maximum likelihood estimator $\hat{\theta}$. Indeed Kou and Ying (1996) showed this is the case. For this reason $\hat{\theta}_e$ is also called the asymptotic maximum likelihood estimator (Breslow and Day 1980, p. 130).

Next we study the maximum likelihood estimation of θ . Based on the i.i.d. representation, the log-likelihood is

$$\ell = X \log \theta - \log \phi(\theta) + \text{constant} = X \log \theta - \sum_{i=1}^S \log(\theta + \lambda_i) + C.$$

Differentiating with respect to θ gives

$$\frac{\partial \ell}{\partial \theta} = \frac{X}{\theta} - \sum_{i=1}^S \frac{1}{\theta + \lambda_i} = 0.$$

The maximum likelihood estimate of θ satisfies

$$X = \sum_{i=1}^S (1 + \hat{\theta}^{-1} \lambda_i)^{-1}.$$

A necessary and sufficient condition to guarantee the existence and uniqueness of $\hat{\theta}$ is

$$L < X < S.$$

Suppose $N_1 N_2 M_1 M_2 / N^3 \rightarrow \infty$, then

$$\sqrt{\text{Var}_\theta(X)} (\log \hat{\theta} - \log \theta) \rightarrow N(0, 1).$$

Due to the i.i.d. representation of X , it is very convenient to construct a confidence interval for θ using a resampling method. We can resample independent and identically distributed U_1^*, \dots, U_S^* from a uniform distribution on $(0, 1)$. Define

$$X^* = \sum_{i=1}^S I(U_i^* < (1 + \hat{\theta}^{-1} \lambda_i)^{-1}).$$

Then the bootstrap version $\hat{\theta}^*$ can be found by solving the equation

$$X^* = \sum_{i=1}^S (1 + \theta^{-1} \lambda_i)^{-1}$$

with respect to θ . Repeat this process B times (B is a large number), the confidence interval for θ can be constructed through the percentiles of $\hat{\theta}^{*b}$, $b = 1, 2, \dots, B$.

Next we consider the problem for a series of 2×2 tables. We assume that the odds ratio parameter θ is common for all 2×2 tables. Define the partial likelihood (conditional likelihood) for the k -th table as

$$f_k(\theta) \propto \theta^{X_k} / \phi_k(\theta).$$

The overall likelihood is

$$L = \prod_{k=1}^K f_k(\theta).$$

The corresponding log-likelihood is

$$\ell = \sum_{k=1}^K X_k \log \theta - \sum_{i=1}^{S_k} \log(\theta + \lambda_{ik}),$$

where λ_{ik} , $i = 1, 2, \dots, S_k$ can be found through the i.i.d. representation for the k -th table. The score estimating equation is

$$\sum_{k=1}^K X_k = \sum_{k=1}^K \sum_{i=1}^{S_k} \frac{1}{1 + \hat{\theta}^{-1} \lambda_{ik}}.$$

It can be shown that the likelihood identity holds true here, i.e.,

$$\begin{aligned} \theta^{-2} \left(\sum_{k=1}^K \text{Var}_{\theta}^{(k-1)}(X_k) \right) &= \sum_{k=1}^K E_{\theta}^{(k-1)} \left[\frac{\partial}{\partial \theta} \log f_k(\theta) \right]^2 \\ &= \sum_{k=1}^K E_{\theta}^{(k-1)} \left[-\frac{\partial^2}{\partial \theta^2} \log f_k(\theta) \right], \end{aligned}$$

where $E^{(k-1)}$ and $\text{Var}^{(k-1)}$ are expectation and variance conditioning on the first $(k-1)$ tables. Denote

$$W(\theta) = \sum_{k=1}^K X_k - \sum_{k=1}^K \sum_{i=1}^{S_k} \frac{1}{1 + \hat{\theta}^{-1} \lambda_{ik}}.$$

It can be shown that a necessary and sufficient condition for the existence of $\hat{\theta}$ is

$$\sum_{k=1}^K L_k < \sum_{k=1}^K X_k < \sum_{k=1}^K S_k.$$

Since the score function $W(\theta)$ is continuously monotone decreasing for $\theta > 0$, it follows that

$$P_{\theta_0}\{\hat{\theta} \geq \theta\} = P_{\theta_0}\{W(\theta) \geq 0\}.$$

Strawderman and Wells (1998) used saddle point approximations to the exact distribution of the conditional maximum likelihood estimator. The primary computational burden is in determining the roots of the polynomial $\phi_k(t)$, which needs to be done numerically but only once for each table. It is well known in the numerical analysis literature that finding the roots of any polynomial is equivalent to finding the eigenvalues of a certain upper Hessenberg matrix whose entries depend directly on the polynomial coefficients. When the roots or eigenvalues are real-valued, as is the case here, it is possible to find these eigenvalues in $O(N^2)$ operations.

We can also study the likelihood ratio statistic

$$R(\theta_0) = 2[\max_{\theta} \ell(\theta) - \ell(\theta_0)].$$

The null distribution can be simulated using the i.i.d. representation (16.1.1).

Next we discuss a regression analysis based on a series of 2×2 tables and covariate z_k , $k = 1, 2, \dots, K$. When the homogeneity assumption on the common odds ratio for a series of 2×2 tables is violated, Zelen (1971) and Breslow (1976) assumed a regression model

$$\log(\theta_k) = \alpha + z_k \beta, \quad k = 1, 2, \dots, K.$$

Using the i.i.d. representation (16.1.1), the score equations are

$$\sum_{k=1}^K \left(x_k - \sum_{i=1}^{S_k} \frac{1}{1 + \lambda_{ik}/\exp(\alpha + z_k \beta)} \right) \begin{pmatrix} 1 \\ z_k \end{pmatrix} = 0.$$

From this it is easy to test $\beta = 0$, i.e., to test the common odds ratio parameter.

16.2 Inferences for Poisson Binomial Distributions

Since the Poisson-Binomial distribution is the summation of a series of independent Bernoulli random variables with possibly different success probabilities, the non-central hypergeometric distribution discussed in the last section can be treated as a special case of the Poisson and Binomial distribution. Samaniego and Jones (1981) found applications of this distribution in reliability studies. Chen et al. (1994) found applications in finite survey sampling problems. Daskalakis et al. (2011) developed a highly efficient algorithm to approximate this distribution. The following example from them illustrates an application of this problem.

You are the manager of an independent weekly newspaper in a city of n people. Each week the i -th inhabitant of the city independently picks up a copy of your paper with probability p_i , $i = 1, 2, \dots, n$. Of course you do not know the values p_1, \dots, p_n ; each week you only see the total number of papers that have been picked up. For many reasons (advertising, production, revenue analysis, etc.) you would like to have a detailed “snapshot” of the probability distribution describing how many readers you have each week.

Next we present Samaniego and Jones’ (1981) results.

If k components are operating independently and their probabilities of operating successfully over a specified time period are p_i , $i = 1, \dots, k$, then the model provides the distribution of the number of components operating at the end of this period. An r out of k system is only as reliable as its r -th best component, and thus it is often of interest to estimate the ordered parameter vector. Let

$$Y_i = \sum_{j=1}^k x_{ij}, i = 1, 2, \dots, n,$$

where $X_{ij} \sim \text{Bernoulli}(1, p_j)$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, k$. Denote the frequency of the event $Y = i$ for $i = 0, 1, 2, \dots, k$ as n_i . Then the likelihood can be written as

$$L(n, p) = \left(\frac{n!}{\prod_{i=0}^k n_i!} \right) \left(\prod_{i=1}^k (1 - p_i) \right)^{n_0} \left(\sum_{i=1}^k p_i \prod_{j \neq i} (1 - p_j) \right)^{n_1} \dots \left(\prod_{i=1}^k p_i \right)^{n_k}.$$

Without loss of generality we assume $p_1 \leq p_2 \leq \dots \leq p_k$. Directly maximizing this likelihood would be difficult.

We may consider the case $(n_0, n_1, \dots, n_k) > 0$. Using the invariant principle of the MLE, we may solve

$$n_0/n = \hat{P}(Y = 0) = \prod_{i=1}^k (1 - p_i), \quad n_1/n = \hat{P}(Y = 1) = \sum_{i=1}^k p_i \prod_{j \neq i} (1 - p_j), \dots,$$

$$n_k/n = P(Y = k) = \prod_{i=1}^k p_i.$$

Dividing the last k equations by the first one, we have

$$\sum_{i=1}^k \frac{p_i}{1-p_i} = n_1/n_0, \quad \sum_{i < j} \frac{p_i}{1-p_i} \frac{p_j}{1-p_j} = n_2/n_0, \dots,$$

$$\sum_{i=1}^k \prod_{j \neq i} \frac{p_j}{1-p_j} = n_{k-1}/n_0, \quad \prod_{i=1}^k \frac{p_i}{1-p_i} = n_k/n_0.$$

Samaniego and Jones (1981) noticed that the left hand sides of the above equations are elementary symmetric functions of $p_i/(1-p_i)$. Moreover they found the above system has a unique solution (p_1, \dots, p_k) in the field of complex numbers. The solution may be obtained as

$$\hat{p}_i = \frac{\hat{\theta}_i}{1 + \hat{\theta}_i}, \quad i = 1, 2, \dots, k,$$

where $\hat{\theta}_i, i = 1, 2, \dots, k$ are the roots of the polynomial

$$p(x) = \sum_{i=0}^k (-1)^i n_i x^{k-i}.$$

Samaniego and Jones (1981) showed that as $n \rightarrow \infty$, in probability, $n_i > 0, i = 0, 1, \dots, k$ and $p(x)$ has k positive roots.

If $0 < p_{(1)} < p_{(2)} < \dots < p_{(k)} < 1$, then

$$\lim_{n \rightarrow \infty} P\left(\prod_{i=1}^k n_i > 0 \text{ and } p(x) \text{ has } k \text{ roots} \geq 0\right) = 1.$$

In fact

$$P\left(\prod_{i=1}^k n_i > 0\right) \geq 1 - \sum_{i=0}^k P(n_i = 0) = 1 - \sum_{i=0}^k (1 - P(Y = i))^n \rightarrow 1.$$

Consider a polynomial

$$f(x) = \begin{cases} n_0 p(x), & \text{if } n_0 > 0, \\ 1 & \text{if } n_0 = 0. \end{cases}$$

The coefficients of $f(x)$ are consistent estimators of the elementary symmetric functions of the ratios

$$\left\{ \theta_{(i)} = \frac{p_{(i)}}{1 - p_{(i)}}, \quad i = 1, 2, \dots, k \right\}.$$

Since $P(n_0 > 0) \rightarrow 1$, for each fixed x , $f(x)$ converges in probability to the polynomial with roots $\theta_{(1)}, \dots, \theta_{(k)}$. If $p_{(1)} < p_{(2)} < \dots < p_{(k)}$, these k roots are distinct.

Exercise Find the MLE in the case $k = 2$. List 8 configurations of n_0, n_1, n_2 for different MLEs.

Mazumdar and Jefferson (1983) found the duality theory of geometric programming can be used to find the maximum likelihood estimate that is more powerful than Samaniego and Jones's 1981 method. Lim et al. (2009) demonstrated that geometric programming is a very powerful tool for solving order restricted inference. We will go back to Poisson Binomial distributions in Chap. 20 in finite population problems.

Chapter 17

Inferences and Tests in Semiparametric Finite Mixture Models

Mixture models have been widely used in many disciplines, including econometric, psychosocial, genetic and medical researches and many others. A good collection of statistical books on mixture models includes, among others for example, Titterington et al. (1985), McLachlan and Peel (2004) (focusing on the finite mixture models) and Lindsay (1995) (mainly dealing with nonparametric mixture problems).

The existing literature on finite mixtures is based on full parametric modelling. As a consequence, it may not be flexible or robust. In this Chapter we first study a test problem in finite components semiparametric mixture models, followed then by examining statistical inference methods for estimating problems. Instead of modelling each component of a mixture model, we study a semiparametric mixture model, where only the density ratios between components are modelled.

17.1 Examples for Hypothesis Test in Semiparametric Mixture Models

For simplicity, we study a mixture model with two components. Consider two independent data sets

$$\begin{aligned} x_1, \dots, x_{n_0}, & \text{ i.i.d. with distribution } F(x), \\ y_1, \dots, y_{n_1}, & \text{ i.i.d. with distribution } H(y) = (1 - \lambda)F(y) + \lambda G(y). \end{aligned} \tag{17.1.1}$$

Denote the density functions by $f(x) = dF(x)/dx$, $g(y) = dG(y)/dy$ and $h(y) = dH(y)/dy$ respectively. Here, λ represents the proportion of the second sampled data set belonging to group two. A main hypothesis of interest is to test $H_0 : F(\cdot) = H(\cdot)$. The following six examples suggest that problems of this kind arise naturally with various applications.

Example 1 Test for partial differential gene expression in microarray studies

As a new technology, the microarray allows the monitoring of thousands of gene expression levels simultaneously in cancer study. The statistical analysis of microarray data focuses on the association between gene expressions and an outcome variable. To test the treatment effect in a two sample problem, a t -test or a Wilcoxon test can be performed for each gene. Typically, there are tens of thousands candidate genes available in microarray studies. As a result, there are thousands of p -values. This raises a challenging multiple comparison problem. Therefore it is desirable to find the most powerful test statistics.

Van Wieringen et al. (2008) studied a comparative microarray experiment involving a sample of n_0 normal tissues and n_1 cancerous tissues. Associated with each tissue is a column gene expression profile $X_i = (x_{i1}, \dots, x_{ip})^T$, where X_{ij} is a random variable representing the expression level of gene j , $j = 1, 2, \dots, p$, of tissue i , $i = 1, 2, \dots, n = n_0 + n_1$.

Let $f_j(x_{ij})$ be the density function of gene j in normal tissues, $i = 1, 2, \dots, n_0$. Also let

$$h_j(x_{ij}) = (1 - \lambda_j)f_j(x_{ij}) + \lambda_j g_j(x_{ij})$$

be the the density function of gene j in the cancerous tissues $i = 1, 2, \dots, n_1$, where $g_j(x_{ij})$ is a density modeling the expression level of gene j if it is differentially expressed and λ_j is the corresponding proportion. Denote F_j , G_j , H_j as the corresponding cumulative distributions of f_j , g_j , h_j , respectively. Define

$$\Theta(F_j, H_j) = 1 - \frac{\int F_j(x)H_j(x)dF_j(x)}{\int F_j^2(x)dF_j(x)}.$$

Van Wieringen et al. (2008) used

$$\hat{\Theta}(F_{n_0j}, H_{n_1j}) = 1 - \min \left\{ 1, \frac{\int F_{n_0j}(x)H_{n_1j}(x)dF_{n_0j}(x)}{\int F_{n_0j}^2(x)dF_{n_0j}(x)} \right\}$$

as a one sided test statistic, where F_{n_0j} , H_{n_1j} are empirical distribution functions for the j -th gene expression level based on normal tissues and cancerous tissues, respectively. The two-sided test statistic is defined as

$$\max\{\hat{\Theta}(F_{n_0j}, H_{n_1j}), \hat{\Theta}(1 - F_{n_0j}, 1 - H_{n_1j})\}.$$

A permutation method can be used to find p -values. This test is called a permutation differential expression test (PDE). Simulations show that PDE test is more powerful than the conventional t-test and Cramer-von Mises test in some situations.

Example 2 Infectious epidemiological studies

Vounatsou et al. (1998) analyzed a clinical malaria dataset involving children aged between 6 and 9 months. The data arose from repeated cross-sectional surveys

of parasitaemia and fever among 426 children up to a year old in a village in the Kilombero district in Tanzania. Briefly, clinical malaria can be diagnosed by the presence of parasites and fever. However, in endemic areas children can tolerate parasites without symptoms and may have fever due to other causes. Vounatsou et al. (1998) proposed a novel approach for estimating the mixing proportion of clinical malaria by formulating the problem as a mixture of distributions. The mixture consists of parasite densities in children with fever either due to malaria or due to other causes. Besides survey data from endemic areas, parasite levels in children from the community are available and are used as a training sample, i.e., a sample that comes from the component of the mixture corresponding to children without clinical malaria but who may have parasites. Vounatsou et al. (1998) presented a model for decomposing a two-component mixture distribution nonparametrically. The underlying assumptions are that a “training” sample is available from one of the components, and that one component of the mixture is stochastically smaller than the other one. They grouped the ordered observations from the training sample and the mixture into a number of categories and derived the likelihood as a product of two multinomial distributions. Bayesian approach was used to find the mixing proportion of clinical malaria. Qin and Leung (2005) discussed this problem in detail using semiparametric exponential tilting model approach.

Example 3 Clinical trial problems

In clinical trials, not all subjects receiving the new treatment necessarily respond. In this setting, $F(\cdot)$ represents the distribution of the outcome variable for the control group and λ is the proportion of “responders” in the treated group. Testing the efficacy of the new treatment allowing for non-responders in the treated group is equivalent to testing the null hypothesis of $H_0 : \lambda = 0$ (Good 1979). Locally most powerful rank tests were proposed by Johnson et al. (1987). In order to implement the optimal rank test, however, the underlying distributions are required to be fully specified. Boos and Brownie (1986) recommended the Wilcoxon test as a useful test statistic, especially for the case of $\lambda \geq 0.6$.

Example 4 Application in genetic association studies

It has become increasingly clear in the past decades that genetic factors play crucial etiological roles in many common diseases including cancer. Many genetic markers on a variety of chromosomes have been shown to have a close linkage to cancers. Testing for genetic linkage and homogeneity in the one sample mixture model setup has been discussed extensively in the literature. We refer readers to, among others, the papers by Lemdani and Pons (1995), Liang and Rathouz (1999) and the book by Sham (1998) and references therein. Devlin et al. (2000) found an interesting application of the two-sample mixture model in genetic association studies. Specifically, individuals who share a disease mutation from a common ancestor often share alleles at genetic markers adjacent to the mutation even if the common ancestor is remote. The alleles at these adjacent markers, known as the haplotype, can be visualized as a string of realizations of random variables, which may be dependent when individuals are related in some fashion. Ideally, for a sample of individuals all

having the same genetic disease, this dependence- measured as haplotype sharing will be greater in the vicinity of disease genes than in other regions of the genome. The amount of overlapping at markers far from the disease is treated as a random variable with unknown distribution F . Overlapping of markers surrounding disease genes are modeled as a mixture $(1 - \lambda)F(x) + \lambda G(x)$, where λ is the fraction of subjects with the disease mutation. In general it is hard to specify the distributions of F and G . To find a robust test statistic, a location shift model assumption $G(x) = F(x - \theta)$ for an unspecified F was imposed by Devlin et al. (2000).

Another application of the mixture model (17.1.1) in genetic quantitative trait loci (QTL) interval mapping was considered by Zou et al. (2002). The backcross design has become popular in animal studies and plant research. In the backcross design, the hybrid and the progenies in subsequent generations are repeatedly backcrossed to one of the parents. Backcrossing may be deliberately employed in animals to transfer a desirable trait in an animal of inferior genetic background to an animal of superior genetic background. As a result, the genotype of the backcross progeny becomes increasingly similar to that of the recurrent parent. In backcrossing studies, the collected data often follow complex mixtures of distribution functions where the mixing proportions are known (Zou et al. 2002).

Example 5 Case-control studies with contaminated controls

As demonstrated in Chap. 11, the case-control study is one of the most popular methods in epidemiology and many other related areas. However, case-control studies may suffer from contamination or misclassification bias. Lancaster and Imbens (1996) found applications of a case-control study with contaminated controls in econometric applications. Specifically, suppose two distinct samples are available. For example, the first one is a random sample of female labor force participants serving as controls. The second sample is working age women including both female labor force participants and no labor force participants working women. Here $1 - \lambda$ in model (17.1.1) represents the proportion of the second sample whose statuses are contaminated. Under a logistic regression model assumption, Lancaster and Imbens (1996) used generalized method of moments to estimate the underlying parameters.

Another application of this problem can be found in prostate cancer studies. The potential misclassification problems in a prostate cancer study were formulated by Begg (1994) and Godley and Schell (1999) as a contaminated case-control problem. Control groups in prostate case-control studies have misclassification rates of 20 to 40%, i.e., 20 to 40% of control series consist of (latent) prostate cancer cases. They pointed out that unsuspected prostate cancer cases may be misclassified into the control group, thereby obscuring the identification of prostate cancer risk factors in case-control studies. Unfortunately they did not propose any systematic statistical method to solve this problem.

Example 6 Applications in fishery

Hosmer Jr (1973) reported a study aiming to find the proportion of male and female halibut in each age class. The sex of halibut can be determined only by dissection of the fish. International Halibut Commission has two different sources of

data, namely, its own research cruises and commercial catches. Sex, age and length are available from fish taken on a research cruise. But only age and length can be obtained from commercial catches. For a fixed age group, let $f(x)$ and $g(x)$ denote the length density of female and male fish, respectively. Thus the length density of fish with unknown sex is $h(x) = (1 - \lambda)f(x) + \lambda g(x)$. There are actually three groups of observations:

$$x_1, \dots, x_{n_0}, \text{ i.i.d. with density } f(x),$$

$$y_1, \dots, y_{n_1}, \text{ i.i.d. with density } g(x),$$

and

$$z_1, \dots, z_{n_2}, \text{ i.i.d. with density } h(z) = (1 - \lambda)f(x) + \lambda g(x),$$

and Hosmer Jr (1973) assumed normality on both f and g allowing different mean values. General parametric estimation and Bayes method for λ were reviewed by Murray and Titterington (1978). Without any assumptions on f and g , Hall and Titterington (1984) developed some efficient estimators for λ .

Examples 1–6 involve data from mixture of two populations F and G and data from training samples either from F or G or both. Current approaches in finite mixture model mainly focus on a full parametric mixture model assumption. In this Chapter we adapt Anderson (1979) semiparametric approach and limit our attention to two-sample situations. Specifically, with the dataset (17.1.1) we assume

$$\log\{g(x)/f(x)\} = \alpha + \beta x \quad (17.1.2)$$

or

$$g(x) = \exp(\alpha + \beta x)f(x),$$

where $f(x)$ can be any density function. This model has been discussed by Lancaster and Imbens (1996). Note that testing the equality of $H(x) = F(x)$ is equivalent to testing $(\alpha, \beta) = (0, 0)$ or $\lambda = 0$.

17.2 A New Score Test Statistic

Based on the observed data derived from model (17.1.2), up to a constant term, the log-likelihood is

$$l(\lambda, \alpha, \beta) \propto \sum_{i=1}^{n_0} \log f(x_i) + \sum_{j=1}^{n_1} \log\{(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)\} + \sum_{j=1}^{n_1} \log f(y_j). \quad (17.2.3)$$

Differentiating with respect to λ , we have

$$\partial l(\lambda, \alpha, \beta) / \partial \lambda = \sum_{j=1}^{n_1} [-1 + \exp(\alpha + \beta y_j)] [(1 - \lambda) + \lambda \exp(\alpha + \beta y_j)]^{-1}.$$

Therefore, the score function for λ when evaluated at $\lambda = 0$ is

$$S(\alpha, \beta) = \partial l(0, \alpha, \beta) / \partial \lambda = \sum_{j=1}^{n_1} [-1 + \exp(\alpha + \beta y_j)]. \quad (17.2.4)$$

It is not difficult to prove that

$$E_H[S(\alpha, \beta)] = n_1 \lambda Var_F[\exp(\alpha + \beta y)].$$

As a result, the expectation of $S(\alpha, \beta)$ is always greater than or equal to zero. Any sensible testing procedure based on $S(\alpha, \beta)$ is necessarily one-sided. It is clear that standard asymptotic approaches to hypothesis testing do not apply to mixture models because the nuisance parameters (α, β) disappear from the likelihood under H_0 . In order to eliminate the nuisance parameters, Liang and Rathouz (1999) proposed a simple solution in the one sample full parametric mixture model case. Specifically, they estimated the nuisance parameters by maximizing $l(\lambda^*, \alpha, \beta)$ with respect to (α, β) where λ^* is a fixed value of λ with $0 < \lambda^* \leq 1$. In our semiparametric setup (17.1.2), $\lambda^* = 1$ provides the most convenient choice for the following reasons. With $\lambda^* = 1$, along with (1.2), this amounts to conducting a case-control study with y as the “risk factor.” As we have shown in Chap. 11 that by profiling out $f(\cdot)$ in the case of $\lambda = 1$, log-likelihood (17.2.3) becomes the logistic log-likelihood

$$l_p(\alpha, \beta) = \sum_{j=1}^{n_1} (\alpha + \beta y_j) - \sum_{i=1}^n \log\{1 + \rho \exp(\alpha + t_i \beta)\}, \quad (17.2.5)$$

where $(t_1, \dots, t_n) = (x_1, \dots, x_{n_0}, y_1, \dots, y_{n_1})$ and $\rho = n_1/n_0$. Consequently, (α, β) can be estimated by maximizing this logistic likelihood. In addition, it is straightforward to show that $(\tilde{\alpha}_{\lambda^*}, \tilde{\beta}_{\lambda^*})$, which maximizes $l(\lambda^*, \alpha, \beta)$ with fixed λ^* , is, under H_0 , consistent and asymptotically normal with the asymptotic covariance matrix inversely proportional to $(\lambda^*)^2$. With $0 < \lambda^* \leq 1$, $(\tilde{\alpha}_1, \tilde{\beta}_1)$ has the smallest covariance matrix among the class of $\{\tilde{\alpha}_{\lambda^*}, \tilde{\beta}_{\lambda^*}\}; 0 < \lambda^* \leq 1\}$.

We denote $(\tilde{\alpha}, \tilde{\beta})$ as the maximum logistic likelihood estimators of (α, β) , i.e. $(\tilde{\alpha}, \tilde{\beta}) = (\tilde{\alpha}_1, \tilde{\beta}_1)$ and assume that $n_i/n \rightarrow \rho_i > 0$, as $n \rightarrow \infty$, $i = 0, 1$. Evidently under H_0 , $(\tilde{\alpha}, \tilde{\beta}) \rightarrow (0, 0)$ in probability. Define

$$T_1 = S(\tilde{\alpha}, \tilde{\beta}) / (1 + \rho). \quad (17.2.6)$$

Theorem 17.1 When n goes to ∞ , under H_0 , T_1 converges to a chi-squared distribution with one degree of freedom, i.e., $\chi^2(1)$.

Note that this theorem is different from the result for the conventional score in which the limit distribution is normal.

Proof

Denote $\eta = (\alpha, \beta)'$ and $\tilde{\eta} = (\tilde{\alpha}, \tilde{\beta})'$. Under H_0 , $\eta = (0, 0)' = \mathbf{0}$. Expanding

$$S(\tilde{\eta}) = \sum_{j=1}^{n_1} \{\exp(\tilde{\alpha} + \tilde{\beta}y_j) - 1\}$$

at $\eta = \mathbf{0}$, we have

$$S(\tilde{\eta}) = S(\mathbf{0}) + \partial S(\mathbf{0})/\partial\eta(\tilde{\eta} - \mathbf{0}) + \frac{1}{2}(\tilde{\eta} - \mathbf{0})'[\partial^2 S(\mathbf{0})/\partial\eta\partial\eta'](\tilde{\eta} - \mathbf{0}) + o_p(1). \quad (17.2.7)$$

Note that

$$\begin{aligned} \partial S/\partial\alpha &= \sum_{j=1}^{n_1} \exp(\alpha + \beta y_j), \quad \partial S/\partial\beta = \sum_{j=1}^{n_1} y_j \exp(\alpha + \beta y_j), \\ \partial S/\partial\eta|_{\eta=(0,0)} &= n_1(1, \bar{y}). \end{aligned}$$

Similarly,

$$\frac{\partial^2 S}{\partial\eta\partial\eta'} = \begin{pmatrix} \sum_{j=1}^{n_1} \exp(\alpha + \beta y_j) & \sum_{j=1}^{n_1} y_j \exp(\alpha + \beta y_j) \\ \sum_{j=1}^{n_1} y_j \exp(\alpha + \beta y_j) & \sum_{j=1}^{n_1} y_j^2 \exp(\alpha + \beta y_j) \end{pmatrix}.$$

Hence

$$\frac{\partial^2 S(\mathbf{0})}{\partial\eta\partial\eta'} = n_1 \left(\frac{1}{\bar{y}} \frac{\bar{y}}{y^2} \right).$$

Also note that $\tilde{\eta} = (\tilde{\alpha}, \tilde{\beta})'$ maximizes

$$l_p(\alpha, \beta) = \sum_{j=1}^{n_1} (\alpha + \beta y_j) - \sum_{i=1}^n \log\{1 + \rho \exp(\alpha + \beta t_i)\}.$$

Furthermore, $\psi(\tilde{\alpha}, \tilde{\beta}) = 0$, where

$$\psi(\alpha, \beta) = \begin{pmatrix} \psi_1(\alpha, \beta) \\ \psi_2(\alpha, \beta) \end{pmatrix} = \begin{pmatrix} \partial l_p/\partial\alpha \\ \partial l_p/\partial\beta \end{pmatrix} = \begin{pmatrix} n_1 - \sum_{i=1}^n \frac{\rho \exp(\alpha + \beta t_i)}{1 + \rho \exp(\alpha + \beta t_i)} \\ \sum_{j=1}^{n_1} y_j - \sum_{i=1}^n \frac{\rho t_i \exp(\alpha + \beta t_i)}{1 + \rho \exp(\alpha + \beta t_i)} \end{pmatrix}.$$

It is not difficult to show that

$$\frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta'} = \left(\begin{array}{cc} \frac{\partial^2 \psi_1}{\partial \alpha^2} & \frac{\partial^2 \psi_1}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \psi_1}{\partial \beta \partial \alpha} & \frac{\partial^2 \psi_1}{\partial \beta^2} \end{array} \right)_{(\alpha, \beta)=(0,0)} = -\frac{\rho(1-\rho)n}{(1+\rho)^3} \left(\begin{array}{cc} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{array} \right),$$

$$\psi(\mathbf{0}) = \left(\begin{array}{c} \partial l_p / \partial \alpha \\ \partial l_p / \partial \beta \end{array} \right)_{\eta=(0,0)} = \left(\begin{array}{c} n_1 - \frac{\rho}{1+\rho} n \\ n_1 \bar{y} - \frac{\rho}{1+\rho} n \bar{t} \end{array} \right) = n_1 \left(\begin{array}{c} 0 \\ \bar{y} - \bar{t} \end{array} \right).$$

Also

$$\begin{aligned} \left(\begin{array}{cc} \partial^2 l_p / \partial \alpha \partial \alpha & \partial^2 l_p / \partial \alpha \partial \beta \\ \partial^2 l_p / \partial \beta \partial \alpha & \partial^2 l_p / \partial \beta \partial \beta \end{array} \right)_{\eta=(0,0)} &= \left(\begin{array}{cc} -\sum_{i=1}^n \frac{\rho \exp(\alpha + \beta t_i)}{[1+\rho \exp(\alpha + \beta t_i)]^2} & -\sum_{i=1}^n \frac{\rho t_i \exp(\alpha + \beta t_i)}{[1+\rho \exp(\alpha + \beta t_i)]^2} \\ -\sum_{i=1}^n \frac{\rho t_i \exp(\alpha + \beta t_i)}{[1+\rho \exp(\alpha + \beta t_i)]^2} & -\sum_{i=1}^n \frac{\rho t_i^2 \exp(\alpha + \beta t_i)}{[1+\rho \exp(\alpha + \beta t_i)]^2} \end{array} \right), \\ \left(\begin{array}{cc} \partial^2 l_p / \partial \alpha \partial \alpha & \partial^2 l_p / \partial \alpha \partial \beta \\ \partial^2 l_p / \partial \beta \partial \alpha & \partial^2 l_p / \partial \beta \partial \beta \end{array} \right)_{\eta=(0,0)} &= -\left(\begin{array}{cc} \frac{\rho}{[1+\rho]^2} n & \frac{\rho}{[1+\rho]^2} n \bar{t} \\ \frac{\rho}{[1+\rho]^2} n \bar{t} & \frac{\rho}{[1+\rho]^2} n \bar{t}^2 \end{array} \right) = -\frac{n_0 n_1}{n} \left(\begin{array}{cc} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{array} \right). \end{aligned}$$

By expanding $\partial l_p(\tilde{\eta})/\partial \eta = \psi(\tilde{\eta})$ up to the term $o_p(n^{-1/2})$, we have

$$\begin{aligned} \tilde{\eta} &= -(\partial^2 l_p(\mathbf{0})/\partial \eta \partial \eta')^{-1} \partial l_p(\mathbf{0})/\partial \eta + o_p(n^{-1/2}) \\ &= \frac{n}{n_0 n_1} \frac{1}{\Delta} \left(\begin{array}{cc} \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{array} \right) n_1 \left(\begin{array}{c} 0 \\ \bar{y} - \bar{t} \end{array} \right) + o_p(n^{-1/2}) \\ &= \frac{n}{n_0} \frac{\bar{y} - \bar{t}}{\Delta} \left(\begin{array}{c} -\bar{t} \\ 1 \end{array} \right) + o_p(n^{-1/2}), \end{aligned} \quad (17.2.8)$$

where

$$\Delta = \bar{t}^2 - \bar{t}^2.$$

Since $\partial S(\mathbf{0})/\partial \eta$ in the second term of (17.2.7) is of order $O_p(n)$, we have to find the exact expression of $\tilde{\eta}$ up to the order $O_p(n^{-1})$. Expanding ψ_1 and ψ_2 up to two terms,

$$\psi_i(\tilde{\eta}) = \psi_i(\mathbf{0}) + \frac{\partial \psi_i(\mathbf{0})}{\partial \eta} \tilde{\eta} + \frac{1}{2} \tilde{\eta}' \frac{\partial^2 \psi_i(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} + o_p(1), \quad i = 1, 2.$$

In vector form,

$$\mathbf{0} = \psi(\mathbf{0}) + \frac{\partial \psi(\mathbf{0})}{\partial \eta} \tilde{\eta} + \frac{1}{2} \left(\begin{array}{c} \tilde{\eta}' \frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \\ \tilde{\eta}' \frac{\partial \psi_2(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \end{array} \right) + o_p(1),$$

$$\tilde{\eta} = - \left(\frac{\partial \psi(\mathbf{0})}{\partial \eta} \right)^{-1} \psi(\mathbf{0}) - \frac{1}{2} \left(\frac{\partial \psi(\mathbf{0})}{\partial \eta} \right)^{-1} \left(\begin{array}{c} \tilde{\eta}' \frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \\ \tilde{\eta}' \frac{\partial \psi_2(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \end{array} \right) + o_p(n^{-1}).$$

Therefore the second term of (17.2.7) is

$$\begin{aligned}
& \frac{\partial S(\mathbf{0})}{\partial \eta} \tilde{\eta} \\
&= n_1(1, \bar{y}) \left[\frac{n}{n_0 n_1} \frac{1}{\Delta} \begin{pmatrix} \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix} n_1 \begin{pmatrix} 0 \\ \bar{y} - \bar{t} \end{pmatrix} + \frac{1}{2} \frac{n}{n_0 n_1} \frac{1}{\Delta} \begin{pmatrix} \bar{t}^2 & -\bar{t} \\ -\bar{t} & 1 \end{pmatrix} \begin{pmatrix} \tilde{\eta}' \frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \\ \tilde{\eta}' \frac{\partial \psi_2(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \end{pmatrix} \right] + o_p(1) \\
&= \frac{nn_1}{n_0} \frac{[\bar{y} - \bar{t}]^2}{\Delta} + \frac{1}{2} \frac{n}{n_0 \Delta} (\bar{t}^2 - \bar{y}\bar{t}, \bar{y} - \bar{t}) \begin{pmatrix} \tilde{\eta}' \frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \\ \tilde{\eta}' \frac{\partial \psi_2(\mathbf{0})}{\partial \eta \eta'} \tilde{\eta} \end{pmatrix} + o_p(1) \\
&= \frac{nn_1}{n_0} \frac{[\bar{y} - \bar{t}]^2}{\Delta} + \frac{1}{2} \frac{n}{n_0 \Delta} (\bar{t}^2 - \bar{y}\bar{t}) \tilde{\eta}' \frac{\partial \psi_1(\mathbf{0})}{\partial \eta \eta^T} \tilde{\eta} + o_p(1) \\
&= \frac{nn_1}{n_0} \frac{[\bar{y} - \bar{t}]^2}{\Delta} + \frac{1}{2} \frac{n}{n_0} \frac{n}{n_0} \frac{(\bar{y} - \bar{t})}{\Delta} (-\bar{t}, 1) \frac{-\rho(1-\rho)n}{(1+\rho)^3} \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{pmatrix} \frac{n}{n_0} \frac{\bar{y} - \bar{t}}{\Delta} \begin{pmatrix} -\bar{t} \\ 1 \end{pmatrix} + o_p(1) \\
&= \frac{nn_1}{n_0} \frac{[\bar{y} - \bar{t}]^2}{\Delta} - \frac{1}{2} \frac{n^3}{n_0^2} \frac{\rho(1-\rho)}{(1+\rho)^3} \frac{n(\bar{y} - \bar{t})^2}{\Delta} + o_p(1).
\end{aligned}$$

The third term of (17.2.7) is

$$\begin{aligned}
\frac{1}{2} \tilde{\eta}' [\partial^2 S(\mathbf{0}) / \partial \eta \partial \eta'] \tilde{\eta} &= \frac{1}{2} \frac{n}{n_0} \frac{\bar{y} - \bar{t}}{\Delta} (-\bar{t}, 1) n_1 \begin{pmatrix} 1 & \bar{y} \\ \bar{y} & \bar{y}^2 \end{pmatrix} \frac{n}{n_0} \frac{\bar{y} - \bar{t}}{\Delta} \begin{pmatrix} -\bar{t} \\ 1 \end{pmatrix} \\
&= \frac{1}{2} \frac{n^2 n_1}{n_0^2} \frac{[\bar{y} - \bar{t}]^2}{\mu_2 - \mu_1^2} + o_p(1),
\end{aligned}$$

where

$$\mu_2 = E_F(y^2), \quad \mu_1 = E_F(y), \quad \Delta \rightarrow \mu_2 - \mu_1^2.$$

Also note that

$$\bar{y} - \bar{t} = \bar{y} - (n_1 \bar{y} + n_0 \bar{x})/n = (\bar{y} - \bar{x}) n_0/n.$$

Finally,

$$S(\tilde{\alpha}, \tilde{\beta}) = (1 + \rho)(\bar{y} - \bar{x})^2 / [(\mu_2 - \mu_1^2)n/(n_1 n_0)] + o_p(1),$$

and

$$T_1 = S(\tilde{\alpha}, \tilde{\beta})/(1 + \rho) = (\bar{y} - \bar{x})^2 / [(\mu_2 - \mu_1^2)n/(n_1 n_0)] + o_p(1) \rightarrow \chi^2(1).$$

This completes the proof.

The calculation of T_1 involves no more than finding the MLE in a logistic model, which is available in most statistical software. For example, in Splus or R, $glm(d \sim t, family = binomial)$ may be used to find the point estimation of $(\tilde{\alpha}, \tilde{\beta})$, where d is an indicator variable of “cases” or “controls” and t is the pooled case and control data.

17.3 Inference in Three Samples Semiparametric Mixture Models

As shown in earlier Chapters, the logistic regression model with case and control data is equivalent to a two sample density ratio model, or the exponential tilting model. Here, we apply this model to mixture data. The existing methods in finite mixture models mainly focus on fully parametric approaches. The application of density ratio models provides an alternative robust inference method for mixture models.

We consider inference based on three groups of data.

$$X_1, \dots, X_{n_1} \sim i.i.d. F(x), \quad Y_1, \dots, Y_{n_2} \sim i.i.d. G(x), \quad Z_1, \dots, Z_{n_3} \sim i.i.d. \lambda F(z) + (1 - \lambda)G(z). \quad (17.3.9)$$

If we assume the X 's come from a “control” population, the Y 's from a “case” population, then the Z 's are from a mixture of “control” and “case” population. Anderson (1979) first applied the exponential tilting model

$$dG(x) = \exp(\beta_0 + \beta_1 x) dF(x),$$

to this mixture data problem. Qin (1999) derived large sample theories for the maximum semiparametric likelihood estimator. In particular, Qin (1999) constructed confidence interval for the disease prevalence parameter λ . In the special case that there are only two samples of either X 's and Z 's or Y 's and Z 's, Lancaster and Imbens (1996) and Qin and Leung (2005) discussed the so-called “case and control problem with contaminated cases or controls”. A related discussion on fitting binary regression models with case-augmented samples was given by Lee et al. (2006). Zhang (2002) discussed the EM algorithm to find the semiparametric MLE.

In general it is much easier to estimate λ using three groups of data specified above when compared with using two groups of data, either X 's and Z 's or Y 's and Z 's. If only the mixture data Z 's are available, then the sole specification of the density ratio between mixture components is not sufficient to identify the model parameters. In fact

$$Z \sim \frac{\{\lambda + (1 - \lambda) \exp(\beta_0 + \beta_1 x)\} dF(x)}{\int \{\lambda + (1 - \lambda) \exp(\beta_0 + \beta_1 x)\} dF(x)}.$$

Since $F(x)$ is left completely arbitrary, no information on $(\lambda, \beta_0, \beta_1)$ is contained in the sample of Z 's. To make inferences based on the Z 's alone, a full parametric mixture model assumption must be used. In next Chapter, we will demonstrate that it is possible to identify the mixture model semiparametrically if there are repeated measurements from each individual.

Denote

$$w_1(t) = w(t) = \exp(\beta_0 + \beta_1 t), \quad w_2(t) = \lambda + (1 - \lambda) \exp(\beta_0 + \beta_1 t).$$

Note that the original problem has nothing to do with the selection bias, however, it becomes a biased sampling problem by using the exponential tilting model

assumption. The second and third samples can be treated as biased sampling data from the first one with weight functions $w_1(t)$ and $w_2(t)$, respectively.

The log-semiparametric likelihood based on the data set (17.3.9) is

$$\ell = \sum_{i=1}^{n_1} \log dF(x_i) + \sum_{j=1}^{n_2} \{\log dF(y_j) + \log w(y_j)\} + \sum_{k=1}^{n_3} \{\log dF(z_k) + \log w_2(z_k)\},$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i \{w(t_i) - 1\} = 0, \quad p_i = dF(t_i) \geq 0,$$

where $n = n_1 + n_2 + n_3$ and $(t_1, \dots, t_n) = (x_1, \dots, x_{n_1}; y_1, \dots, y_{n_2}; z_1, \dots, z_{n_3})$ are pooled data. By going through standard Lagrange multiplier method, we can find

$$p_i = \frac{1}{n} \frac{1}{1 + \nu \{w(t_i) - 1\}}, \quad i = 1, 2, \dots, n,$$

where the Lagrange multiplier ν is determined by

$$\frac{1}{n} \sum_{i=1}^n \frac{w(t_i) - 1}{1 + \nu \{w(t_i) - 1\}} = 0.$$

We may change the variables (λ, β, ν) to (λ, β, α) , where

$$\alpha = \nu - n_2/n - n_3(1 - \lambda)/n.$$

Consequently

$$p_i = \frac{1}{n} \frac{\gamma^{-1}(t_i)}{1 + \alpha \{w(t_i) - 1\} \gamma^{-1}(t_i)}, \quad i = 1, 2, \dots, n.$$

where

$$\gamma(t_i) = n_1/n + n_3\lambda/n + w(t_i)\{n_2 + n_3(1 - \lambda)\}/n.$$

The log semiparametric likelihood can be decomposed as

$$\ell\{\lambda, \beta, \alpha(\lambda, \beta)\} = \ell_1\{\lambda, \beta, \alpha(\lambda, \beta)\} + \ell_2(\lambda, \beta),$$

where

$$\ell_1\{\lambda, \beta, \alpha(\lambda, \beta)\} = - \sum_{i=1}^n \log\{1 + \alpha g(t_i, \lambda, \beta)\},$$

$$\begin{aligned}\ell_2(\lambda, \beta) &= -\sum_{i=1}^n \log \gamma(t_i; \lambda, \beta) + \sum_{j=1}^{n_1} \log w(y_j, \beta) \\ &\quad + \sum_{k=1}^{n_2} \log w_2(z_k; \beta, \lambda), \\ g(t_i; \lambda, \beta) &= \frac{w(t_i) - 1}{\gamma(t_i)}.\end{aligned}$$

The constraint equation becomes

$$n^{-1} \sum_{i=1}^n \frac{g(t_i; \lambda, \beta)}{1 + \alpha g(t_i; \lambda, \beta)} = 0.$$

The advantage of the above variable transformation is $E [\sum_{i=1}^n g(t_i; \lambda, \beta)] = 0$. Therefore the constraint equation has the same form as those discussed in Chap. 8 on empirical likelihood based inferences for unbiased estimating equations.

Next maximize $\ell(\lambda, \beta, \alpha)$ with respect to (λ, β) .

The semiparametric likelihood ratio statistic for testing $H_0 : \lambda = \lambda_T$ is

$$R(\lambda) = 2[\max_{\lambda, \beta, F} \ell(\lambda, \beta, F) - \max_{\beta, F} \ell(\lambda, \beta, F)].$$

Theorem 17.2 Under the regularity conditions specified in Qin (1999),

$$R(\lambda_T) \rightarrow \chi^2(1). \quad (17.3.10)$$

Proof Denote the true value of (λ, β) by (λ_T, β_T) . The limiting value of α is 0. Using Taylor's expansion for the semiparametric score equations at $(\lambda_T, \beta_T, 0)$, we can solve

$$\begin{pmatrix} \hat{\lambda} - \lambda_T \\ \hat{\beta} - \beta_T \\ \hat{\alpha} - 0 \end{pmatrix} = -S_n^{-1} Q_n + o_p(n^{-1/2}),$$

where

$$S_n^{-1} = \begin{pmatrix} n^{-1} \frac{\partial^2 \ell}{\partial \lambda \partial \lambda} & n^{-1} \frac{\partial^2 \ell}{\partial \lambda \partial \beta^T} & n^{-1} \frac{\partial^2 \ell}{\partial \lambda \partial \alpha} \\ n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial \lambda} & n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} & n^{-1} \frac{\partial^2 \ell}{\partial \beta \partial \alpha} \\ n^{-1} \frac{\partial^2 \ell}{\partial \alpha \partial \lambda} & n^{-1} \frac{\partial^2 \ell}{\partial \alpha \partial \beta^T} & n^{-1} \frac{\partial^2 \ell}{\partial \alpha \partial \alpha} \end{pmatrix}_{(\lambda_T, \beta_T, 0)}^{-1}.$$

$$Q_n = \begin{pmatrix} n^{-1} \frac{\partial \ell(\lambda_T, \beta_T, 0)}{\partial \lambda} \\ n^{-1} \frac{\partial \ell(\lambda_T, \beta_T, 0)}{\partial \beta} \\ n^{-1} \frac{\partial \ell(\lambda_T, \beta_T, 0)}{\partial \alpha} \end{pmatrix}.$$

After some algebraic manipulations, we arrive at

$$S_n \rightarrow S = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12}^T & s_{22} & s_{23} \\ s_{13}^T & s_{23}^T & s_{33} \end{pmatrix} = \begin{pmatrix} e_{11} & e_{12} \\ e_{12}^T & s_{33} \end{pmatrix},$$

$$e_{11} = \begin{pmatrix} s_{11} & s_{12} \end{pmatrix}, \quad e_{12} = e_{21}^T = (s_{13}^T, s_{23}^T),$$

where

$$s_{11} = \rho_2^2 \eta_1 - \rho_2 \eta_2, \quad s_{12} = \rho_2 \phi_1 - \rho_2 \psi_1, \quad s_{13} = -\rho_2 \eta_1,$$

$$\gamma(t) = a + bw(t), \quad a = \rho_0 + \rho_2 \lambda_T, \quad b = \rho_1 + \rho_2(1 - \lambda_T),$$

$$\delta = \rho_0 b^2 + \rho_1 a^2 + \rho_2(\lambda_T - a)^2,$$

$$\eta_1 = \int \frac{\{1 - w(t)\}^2}{\gamma(t)} dF(t), \quad \eta_2 = \int \frac{\{1 - w(t)\}^2}{w_2(t)} dF(t),$$

$$\phi_1 = \int \frac{\partial w(t)}{\partial \beta} \frac{1}{\gamma(t)} dF(t), \quad \phi_2 = \int \frac{\partial^2 w(t)}{\partial \beta \partial \beta^T} \frac{1}{\gamma(t)} dF(t),$$

$$\psi_1 = \int \frac{\partial w(t)}{\partial \beta} \frac{1}{w_2(t)} dF(t), \quad \psi_2 = \int \frac{\partial^2 w(t)}{\partial \beta \partial \beta^T} \frac{1}{w_2(t)} dF(t).$$

Note that

$$Q_n = n^{-1} \left\{ \sum_{i=1}^{n_0} q_0(x_i) + \sum_{j=1}^{n_1} q_1(y_j) + \sum_{k=1}^{n_2} q_2(z_k) \right\},$$

where

$$q_0(x) = - \begin{pmatrix} \frac{\rho_2 \{1-w(x)\}}{\gamma(x)} \\ \frac{b \partial w(x)/\partial \beta}{\gamma(x)} \\ g(x, \lambda_T, \beta) \end{pmatrix}, \quad q_1(y) = - \begin{pmatrix} \frac{\rho_2 \{1-w(y)\}}{\gamma(y)} \\ \frac{b \partial w(y)/\partial \beta}{\gamma(y)} - \frac{\partial w(y)/\partial \beta}{w_2(y)} \\ g(y, \lambda_T, \beta) \end{pmatrix},$$

$$q_2(z) = - \begin{pmatrix} \frac{\rho_2 \{1-w(z)\}}{\gamma(z)} - \frac{1-w(z)}{w_2(z)} \\ \frac{b \partial w(z)/\partial \beta}{\gamma(z)} - \frac{(1-\lambda_T) \partial w(z)/\partial \beta}{w_2(z)} \end{pmatrix}.$$

By the Central Limit Theorem, we have

$$\sqrt{n} Q_n \rightarrow N(0, V),$$

where

$$V = \begin{pmatrix} -e_{11} - \delta e_{12} e_{12}^T & -\delta e_{12} s_{33} \\ -\delta e_{12}^T & s_{33} - \delta s_{33}^2 \end{pmatrix}.$$

Therefore

$$\sqrt{n} \begin{pmatrix} \tilde{\lambda} - \lambda_T \\ \tilde{\beta} - \beta \\ \tilde{\alpha} - 0 \end{pmatrix} \rightarrow N(0, U), \quad U = S^{-1} V S^{-1} = \begin{pmatrix} u_{11} & 0 \\ 0 & u_{22} \end{pmatrix},$$

where

$$u_{11} = - \begin{pmatrix} s_{11} - s_{13} s_{33}^{-1} s_{13}^T & s_{12} - s_{13} s_{33}^{-1} s_{23}^T \\ s_{12}^T - s_{23} s_{33}^{-1} s_{13}^T & s_{22} - s_{23} s_{33}^{-1} s_{23}^T \end{pmatrix}^{-1}, \quad u_{22} = s_{33}^{-1} e_{12}^T u_{11} e_{12} s_{33}^{-1} + s_{33}^{-1} - \delta.$$

$(\tilde{\lambda}, \tilde{\beta})$ and $\tilde{\alpha}$ are asymptotically independent since the off-diagonal elements of U are zeros.

To prove that $R(\lambda_T)$ converges to a chi-squared variable, we use an abbreviated notation $\tilde{\xi} = (\tilde{\beta} - \beta, \tilde{\alpha} - 0)$. Expanding the score equations,

$$\begin{aligned} \begin{pmatrix} \tilde{\lambda} - \lambda_T \\ \tilde{\xi} \end{pmatrix} &= - \begin{pmatrix} n^{-1} \frac{\partial^2 \ell}{\partial \lambda \lambda^T} & n^{-1} \frac{\partial^2 \ell}{\partial \lambda \xi^T} \\ n^{-1} \frac{\partial^2 \ell}{\partial \xi \lambda^T} & n^{-1} \frac{\partial^2 \ell}{\partial \xi \xi^T} \end{pmatrix} \begin{pmatrix} Q_{1n} \\ Q_{2n} \end{pmatrix} + o_p(n^{-1/2}) \\ &= - \begin{pmatrix} s_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \begin{pmatrix} Q_{1n} \\ Q_{2n} \end{pmatrix} + o_p(n^{-1/2}), \end{aligned}$$

where

$$c_{12} = c_{21}^T = (s_{12}, s_{13}), \quad c_{22} = \begin{pmatrix} s_{22} & s_{23} \\ s_{32} & s_{33} \end{pmatrix},$$

$$Q_{1n} = n^{-1} \frac{\partial \ell(\lambda_T, \beta, 0)}{\partial \lambda}, \quad Q_{2n} = n^{-1} \left(\frac{\partial \ell(\lambda_T, \beta, 0)}{\partial \beta}, \frac{\partial \ell(\lambda_T, \beta, 0)}{\partial \alpha} \right).$$

Suppose $\hat{\beta}$ maximizes $\ell(\lambda_T, \beta, \tilde{\alpha}(\lambda_T, \beta))$ with λ fixed at λ_T , then

$$\frac{\partial \ell(\lambda_T, \hat{\beta}, \hat{\alpha})}{\partial \beta} = 0, \quad \frac{\partial \ell(\lambda_T, \hat{\beta}, \hat{\alpha})}{\partial \alpha} = 0,$$

where $\hat{\alpha} = \hat{\alpha}(\lambda_T, \hat{\beta})$. Expanding $\partial \ell(\lambda_T, \hat{\beta}, \hat{\alpha})/\partial \alpha$ and $\partial \ell(\lambda_T, \hat{\beta}, \hat{\alpha})/\partial \beta$ at point $(\lambda_T, \beta_T, 0)$, we have

$$\hat{\xi} = \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\alpha} - 0 \end{pmatrix} = -c_{22}^{-1} Q_{2n} + o_p(n^{-1/2}).$$

Hence

$$\begin{aligned}
\begin{pmatrix} \lambda_T - \tilde{\lambda} \\ \hat{\xi} - \tilde{\xi} \end{pmatrix} &= - \begin{pmatrix} 0 & 0 \\ 0 & c_{22}^{-1} \end{pmatrix} \begin{pmatrix} Q_{1n} \\ Q_{2n} \end{pmatrix} + \begin{pmatrix} s_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}^{-1} \begin{pmatrix} Q_{1n} \\ Q_{2n} \end{pmatrix} + o_p(n^{-1/2}) \\
&= \left\{ \begin{pmatrix} 0 & 0 \\ 0 & c_{22}^{-1} \end{pmatrix} + \begin{pmatrix} s_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \right\} \begin{pmatrix} \tilde{\lambda} - \lambda_T \\ \hat{\xi} - \tilde{\xi} \end{pmatrix} + o_p(n^{-1/2}) \\
&= - \begin{pmatrix} I \\ c_{22}^{-1} c_{21} \end{pmatrix} (\tilde{\lambda} - \lambda_T) + o_p(n^{-1/2}).
\end{aligned}$$

Expanding $\ell(\lambda_T, \hat{\beta}, \hat{\alpha})$ at $(\tilde{\lambda}, \tilde{\beta}, \tilde{\alpha})$, we have

$$\begin{aligned}
\ell(\lambda_T, \hat{\beta}, \hat{\alpha}) - \ell(\tilde{\lambda}, \tilde{\beta}, \tilde{\alpha}) &= 0.5n(\lambda_T - \hat{\lambda}, \hat{\xi} - \tilde{\xi}) \begin{pmatrix} s_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \begin{pmatrix} \tilde{\lambda} - \lambda_T \\ \hat{\xi} - \tilde{\xi} \end{pmatrix} + o_p(1) \\
&= 0.5n(\tilde{\lambda} - \lambda_T)(s_{11} - c_{12}c_{22}^{-1}c_{21})(\tilde{\lambda} - \lambda_T) + o_p(1) \\
&\rightarrow -0.5\chi^2(1),
\end{aligned}$$

since the asymptotic variance of $\sqrt{n}(\tilde{\lambda} - \lambda_T)$ is $-(s_{11} - c_{12}c_{22}^{-1}c_{21})^{-1}$.

Exercise The underlying distribution F can be estimated by

$$\hat{F}(t) = \sum_{i=1}^n \hat{p}_i I(t_i \leq t).$$

Derive the limiting distribution for $\sqrt{n}\{\hat{F}(t) - F(t)\}$.

Numerical Results for Case-Control Study with Contaminated Cases or Controls

The simulation results based on three samples can be found in Qin (1999). Next we conduct some numerical studies to evaluate the bias and variance for the odds ratio parameter β in a case-control study with contaminated controls. In other words, there are only two groups of data, either case and contaminated data (Y_j 's and Z_k 's) or control and contaminated data (X_i 's and Z_k 's).

In general it would be difficult to estimate the underlying parameters if the two component distributions are close to each other. In fact, Lancaster and Imbens (1996) found that sometimes there is no solution by using the GMM method in their simulation studies. The empirical likelihood approach has the same problem, especially for small sample sizes.

Suppose the controls are generated from $N(\mu_0, 1)$ and the cases are generated from a mixture of $N(\mu_0, 1)$ and $N(\mu_1, 1)$, that is

$$\lambda N(\mu_0, 1) + (1 - \lambda)N(\mu_1, 1).$$

For different values of μ_0 , μ_1 and λ , we generated 1000 data sets with sample sizes $n_0 = 400$ and $n_1 = 200$. Means and variances based on the 1000 samples are reported in Tables 17.1, 17.2, 17.3 and 17.4.

Table 17.1 Case sample size = 200, control sample size = 400

$\lambda = 0$	mean1	var1	mean2	var2	var1/var2
α	2.01892	0.03098	2.07593	0.04273	73%
β	-2.02126	0.02923	-2.06710	0.03705	79%
λ	-	-	0.00453	0.00008	-

Table 17.2 Case sample size = 200, contaminated control sample size = 400

$\lambda = 0.1$	mean1	var1	mean2	var2	var1/var2
α	1.29783	0.01662	2.04787	0.13640	12%
β	-1.41344	0.01567	-2.06710	0.08308	19%
λ	-	-	0.09899	0.00135	-

Table 17.3 Case sample size = 200, contaminated control sample size = 400

$\lambda = 0.3$	mean1	var1	mean2	var2	var1/var2
α	0.61143	0.00501	2.11971	0.34649	1.5%
β	-0.90098	0.00743	-2.08420	0.16774	4.4%
λ	-	-	0.29795	0.00314	-

Table 17.4 Case sample size = 200, contaminated control sample size = 400

$\lambda = 0.5$	mean1	var1	mean2	var2	var1/var2
α	0.29926	0.00194	2.25311	0.91193	0.20%
β	-0.62269	0.00532	-2.16181	0.35559	1.5%
λ	-	-	0.49486	0.00510	-

When λ departs from 0, the log odds ratio parameter β estimate is attenuated by a large amount. If some external information is available for λ , then it would be much easier to estimate β . Qin and Leung (2005) applied this method to a malaria study.

17.4 Inference in Upgraded Mixture Models

Missing data is a ubiquitous problem in social and medical science research. We start from the simplest case of the so called “upgraded mixture model”. This model was introduced by Hasminskii and Ibragimov (1983). Bickel and Ritov (1993) and Van der Vaart and Wellner (1992) also studied this model. The upgraded mixture model is given by

$$(Y_1, \dots, Y_m) \sim i.i.d. \quad \int f(y|x)dG(x) = \int f(y|x, \theta)dG(x),$$

$$X_1, \dots, X_n \sim i.i.d. \quad dG(x),$$

where $f(y|x, \theta)$ is a specified parametric model and the marginal distribution function G is not specified. Without the X sample, this is the standard semiparametric mixture model discussed in Kiefer and Wolfowitz (1956), Laird (1978), Jewell (1982) and Lindsay (1995). Nonparametrically estimating G without a sample of X has a very slow convergence rate. Even though there exists some semiparametric methods for estimating both β and G , they are in general not easy at all. In the presence of a X sample, Qin (1998a) studied this problem by restricting the support of G for the X sample. In applications, Y may be a categorical variable. If it is not, we can artificially group it as

$$\xi_j = \sum_{i=1}^n I(a_{j-1} < Y_i \leq a_j), \quad j = 1, 2, \dots, k.$$

The likelihood is

$$L = L_M(w_1, \dots, w_k) \prod_{i=1}^n p_j,$$

where

$$L_M = \prod_{j=1}^k w_j^{\xi_j},$$

and

$$w_j = \int \psi_j(x, \theta) dG(x), \quad \psi_j(x, \theta) = F(a_j|x, \theta) - F(a_{j-1}|x, \theta),$$

$$p_i = dG(x_i), \quad i = 1, 2, \dots, n.$$

For discretized G at x_1, \dots, x_n , w_j becomes

$$w_j = \sum_{i=1}^n p_i \psi_j(x_i, \theta).$$

We will use a two-step approach.

Step 1. For fixed θ and w_j 's, maximize $\prod_{i=1}^n p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad j = 1, 2, \dots, n, \quad \sum_{i=1}^n p_i \{\psi_j(x_i, \theta) - w_j\} = 0, \quad j = 1, 2, \dots, k-1.$$

Profiling out the p_i 's, we have

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda^T g(x_i, \theta, w)}, \quad i = 1, 2, \dots, n,$$

where

$$g(x, \beta, w) = (g_1, \dots, g_{k-1}) = (\psi_1(x, \theta) - w_1, \dots, \psi(x, \theta) - w_{k-1}), \quad w = (w_1, \dots, w_{k-1}),$$

and the Lagrange multiplier λ is determined by

$$n^{-1} \sum_{i=1}^n \frac{g(x_i, \theta, w)}{1 + \lambda^T g(x_i, \theta, w)} = 0.$$

The log-profile likelihood is

$$\ell = - \sum_{i=1}^n \log\{1 + \lambda^T g(x_i, \theta, w)\} + \sum_{j=1}^{k-1} \xi_j \log w_j + \xi_k (1 - w_1 - \dots - w_{k-1}).$$

Step 2. Maximize ℓ with respect to (w, θ) , where w ranges over the full parameter set, that is the full unit simplex. Denote $\hat{\lambda}, \hat{w}, \hat{\theta}$ as the maximum log-likelihood estimates of ℓ . For the estimation of the distribution function $G(x)$, we can use

$$\hat{G}(x) = n^{-1} \sum_{i=1}^n \frac{1}{1 + \hat{\lambda}^T g(x_i, \hat{\theta}, \hat{w})} I(x_i \leq x).$$

Theorem 17.3 Let $\hat{\xi}^T = (\hat{w} - w_0, \hat{\theta} - \theta_0)$. Then in distribution

$$\sqrt{n} \begin{pmatrix} \hat{\lambda} \\ \hat{\xi} \end{pmatrix} \rightarrow N(0, \Sigma), \quad \Sigma = \begin{pmatrix} V & 0 \\ 0 & s_{22.1}^{-1} \end{pmatrix}.$$

Also the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is d^{-1} , where

$$d = E(\partial g / \partial \theta)^T \{(\rho_0 J)^{-1} + E(gg^T)\}^{-1} E(\partial g / \partial \theta).$$

and

$$\sqrt{n}[\hat{G}(x) - G(x)] \rightarrow N(0, \Gamma(x)), \quad \Gamma(x) = G(x)\{1 - G(x)\} - B^T(x)V B(x),$$

where $B(x) = E\{g(X, \theta_0, w_0)I(X \leq x)\}$, J is the Fisher information matrix based on the multinomial likelihood L_M .

Proof Let

$$Q_{n1}(\theta, w, \lambda) = \frac{1}{n} \sum_{i=1}^n \frac{g(x_i; \theta, w)}{1 + \lambda^T g(x_i, \theta, w)},$$

$$Q_{n3} = \sum_{i=1}^n \frac{1}{1 + \lambda^T g(x_i, \theta, w)} \lambda^T \frac{\partial g(x_i; \theta, w)}{\partial \theta},$$

$$Q_{n2}(\theta, w, \lambda) = -\lambda - h(w)/n.$$

We need to solve

$$(Q_{1n}, Q_{2n}, Q_{3n}) = 0.$$

Expanding $Q_{ni}(\tilde{\theta}, \tilde{w}, \tilde{\lambda}), i = 1, 2, 3$ at $(\theta_0, w_0, 0)$, and also noting $\partial g / \partial w = I_{(k-1) \times (k-1)}$, we have

$$\begin{aligned} \begin{pmatrix} \tilde{\lambda} - 0 \\ \tilde{w} - w_0 \\ \tilde{\theta} - \theta_0 \end{pmatrix} &= - \begin{pmatrix} -E(gg^T) & -I & W(\partial g / \partial \theta) \\ -I & \rho_0 J & 0 \\ E(\partial g / \partial \theta)^T & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} Q_{n1} \\ Q_{n2} \\ Q_{n3} \end{pmatrix}_{(w_0, \theta_0, 0)} + o_p(n^{-1/2}) \\ &= -S^{-1} Q_n + o_p(n^{-1/2}), \end{aligned}$$

where

$$J = E[h(w_0 h^T(w_0))] / m = -m^{-1} E \left\{ \frac{\partial^2 \ell_M(w_0)}{\partial w w^T} \right\},$$

is the Fisher information matrix based on ℓ_M . Partition S as

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}, \quad s_{22} = \begin{pmatrix} \rho_0 J & 0 \\ 0 & 0 \end{pmatrix},$$

where $s_{11} = -E[gg^T]$ and $s_{12} = s_{21}^T = (-I, E(\partial g / \partial \theta))$. Noting

$$\sqrt{n} Q_n|_{(\theta_0, w_0, 0)} \rightarrow N(0, U), \quad U = \begin{pmatrix} -s_{11} & 0 \\ 0 & s_{22} \end{pmatrix},$$

we can show that

$$\sqrt{n} \begin{pmatrix} \tilde{\lambda} \\ \tilde{\xi} \end{pmatrix} \rightarrow N(0, \Sigma), \quad \Sigma = S^{-1} U S^{-1}.$$

Also it can be shown that

$$\Sigma = \begin{pmatrix} V & 0 \\ 0 & s_{22,1}^{-1} \end{pmatrix},$$

where

$$V = -s_{11}^{-1} - s_{11}^{-1} s_{12} s_{22,1}^{-1} s_{11}^{-1},$$

and

$$\begin{aligned}s_{22.1} &= s_{22} - s_{21}s_{11}^{-1}s_{12} = \begin{pmatrix} \rho_0 J + (Egg^T)^{-1} & -(Egg^T)^{-1}E(\partial g/\partial\theta) \\ -E(\partial g/\partial\theta)^T(Egg^T)^{-1} & E(\partial g/\partial\theta)^T(Egg^T)^{-1}E(\partial g/\partial\theta) \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ b^T & c \end{pmatrix},\end{aligned}$$

$$a = \rho_0 J + E(gg^T)^{-1}, \quad aE(gg^T) = \rho_0 J(Egg^T) + I.$$

Hence

$$(\rho_0 J)^{-1} + E(gg^T) = (\rho_0 J)^{-1}[I + (\rho_0 J)E(gg^T)] = (\rho_0 J)^{-1}aE(gg^T) = -(\rho_0 J)^{-1}as_{11}.$$

Let

$$d = c - b^T a^{-1} b,$$

then

$$d = -b^T s_{11} b - b^T a^{-1} b = -b^T a^{-1} \{as_{11} + I\} b = -b^T a^{-1} \rho_0 J s_{11} b.$$

We can show that

$$\sqrt{n}(\tilde{\theta} - \theta_0) \rightarrow N(0, d^{-1}), \quad d = E\left(\frac{\partial g}{\partial\theta}\right)^T [(\rho_0 J)^{-1} + E(gg^T)]^{-1} E\left(\frac{\partial g}{\partial\theta}\right).$$

Note that

$$s_{22.1}^{-1} = \begin{pmatrix} I & -a^{-1}b \\ 0 & I \end{pmatrix} \begin{pmatrix} a^{-1} & 0 \\ 0 & d^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -b^T a^{-1} & I \end{pmatrix}$$

and

$$\begin{aligned}b^T a^{-1} s_{11}^{-1} + b^T &= b^T a^{-1} (ss_{11}^{-1} + a) = b^T a^{-1} \rho_0 J \\ s_{11}^{-1} a^{-1} b + b &= s_{11}^{-1} a^{-1} (I + as_{11}) b = s_{11}^{-1} a^{-1} \rho_0 J s_{11} b.\end{aligned}$$

Finally,

$$V = -s_{11}^{-1} a^{-1} \rho_0 J \{I + s_{11} b d^{-1} b^T a^{-1} \rho_0 J\}.$$

The the asymptotic variance of $\sqrt{n}\{\hat{G}(x) - G(x)\}$ can be derived easily. We leave it as an exercise.

Next we discuss the limiting distribution for the likelihood ratio statistic.

Theorem 17.4 Under $H_0 : f(y) = \int f(y|x, \theta) dG(x)$, the empirical likelihood ratio statistic satisfies

$$R = 2\{\max_w \ell_M(w) - \ell_M(\hat{w})\} + \sum_{i=1}^n \log\{1 + \hat{\lambda}^T g(x_i, \hat{\theta}, \hat{w})\} \rightarrow \chi^2(k - p - 1).$$

Proof From $\partial\ell_M(\hat{w})/\partial w = 0$, we have $-h(\hat{w})/n = 0$. Using Taylor's expansion, we have

$$\hat{w} - w_0 = \left(n^{-1} \frac{\partial h(w_0)}{w}\right)^{-1} (-n^{-1}h(w_0)) + o_p(n^{-1/2}) = -(\rho_0 J)^{-1} Q_{2n}(\theta_0, w_0, 0) + o_p(n^{-1/2}).$$

Furthermore,

$$\tilde{w} - w_0 = -(0_{(k-1) \times (k-1)}, I_{(k-1) \times (k-1)}, 0_{(k-1) \times p}) S^{-1} Q_n + o_p(n^{-1/2}).$$

Therefore it can be shown that

$$\hat{w} - \tilde{w} = -(\rho_0 J)^{-1} \tilde{\lambda} + o_p(n^{-1/2}).$$

Expanding $\ell_M(\tilde{w})$ at point \hat{w} and noting that $\partial\ell_M(\hat{w})/\partial w = 0$, we have

$$\begin{aligned} \ell_M(\tilde{w}) - \ell_M(\hat{w}) &= -0.5(\hat{w} - \tilde{w})^T \frac{\partial^2 \ell(\tilde{w})}{\partial w \partial w^T} (\hat{w} - \tilde{w}) + o_p(1) \\ &= 0.5(\hat{w} - \tilde{w})^T (\rho_0 J)^{-1} \tilde{\lambda} + o_p(1) \\ &= 0.5n \tilde{\lambda}^T (\rho_0 J)^{-1} \tilde{\lambda} + o_p(1). \end{aligned}$$

Also expanding the first term, we have

$$\sum_{j=1}^n \log\{1 + \tilde{\lambda}^T g(z_j, \tilde{\theta}, \tilde{w})\} = 0.5n \tilde{\lambda}^T (Egg^T) \tilde{\lambda} + o_p(1).$$

Finally the likelihood ratio statistic is

$$R = n \tilde{\lambda} \{(\rho_0 J)^{-1} + Egg^T\} \tilde{\lambda} + o_p(1) = n \tilde{\lambda} \{(\rho_0 J)^{-1} as_{11}\} \tilde{\lambda} + o_p(1).$$

To show R converges to a Chi-squared random variable, by the Ogasawara and Takahashi's theorem in Rao (1973), p. 188, we only need to show

$$VAVAV = VAV, \quad \text{trace}(AV) = k - 1 - p, \quad A = -(\rho_0 J)^{-1} as_{11}.$$

In fact

$$AV = I + s_{11} bd^{-1} b^T a^{-1} \rho_0 J.$$

We can show that

$$AVAV = AV, \quad VAVAV = VAV,$$

and

$$\text{trace}(AV) = \text{trace}(I_{(k-1) \times (k-1)}) + \text{trace}(d^{-1}b^T a^{-1} \rho_0 J s_{11} b) = k-1 - \text{trace}(I_{p \times p}) = k-1-p.$$

The above approach produces efficient estimation when Y is categorical. For the continuous case, if we use a sieve method allowing the lengths of intervals go to zero at a certain rate, it may extract the maximum information for estimating β .

Exercise

As an alternative method we may construct two empirical likelihoods based on y_i 's and x_j 's data, respectively. Denote $h(y) = \int f(y|x, \theta) dG(x)$. Let

$$L = \left(\prod_{i=1}^m p_i \right) \left(\prod_{j=1}^n q_j \right),$$

where $p_i = dF(y_i)$, $i = 1, 2, \dots, m$; $q_j = dG(x_j)$, $j = 1, 2, \dots, n$. For any measurable function $\psi(y, \beta)$,

$$E[\psi(y, \theta)] = \int \psi(y, \theta) dH(y) = \int \int \psi(y, \theta) f(y|x, \theta) dy dG(x) =: \int \phi(x, \theta) dG(x),$$

where $\phi(x, \theta) = \int \psi(y, \theta) f(y|x, \theta) dy$. Therefore we can impose constraints

$$\sum_{i=1}^m p_i = 1, \quad p_i \geq 0, \quad \sum_{j=1}^n q_j = 1, \quad q_j \geq 0,$$

and

$$\sum_{i=1}^m p_i \psi(y_i, \theta) = \sum_{j=1}^n q_j \phi(x_j, \theta).$$

Develop the two empirical likelihoods based approach. Discuss the possible choices of ψ for the efficient estimation of θ .

Vardi (1989) discussed a multiplicative censoring model where $X \sim dG(x)$ is a nonnegative random variable, and independently $Y = XU$, $U \sim U(0, 1)$. The observed data are

$$X_1, \dots, X_m \sim dG(x); \quad Y_1, \dots, Y_n \sim \int_y^\infty z^{-1} dG(z).$$

It can be treated as a special case of the upgraded mixture model. However, in this special case it is possible to find the nonparametric MLE for G by using all observed data points as the probability masses of G . We will discuss this in details in Chap. 25.

Strategies discussed above can also be used in more general missing data set up, for example, Lawless et al. (1999) discussed a missing covariate problem. Denote

the missing probability of a covariate X as

$$P(D = 1|Y = y, X = x, Z = z) = P(D = 1|y, z) = \pi(y, z, \theta),$$

where Y is the response variable and Z is another covariate. The underlying assumption is

$$f(y|x, z) = f(y|x, z, \beta),$$

where Y and Z are always observable. When both Y and Z are discrete such that $Y = 1, 2, \dots, I$, $Z = 1, 2, \dots, J$, the likelihood can be written as

$$\prod_{i=1}^n [f(y_i|x_i, z_i, \beta)dG(x_i|z_i)h(z_i)]^{d_i} \left[\int f(y_i|x, z_i, \beta)dG(x|z_i)h(z_i) \right]^{1-d_i}.$$

Denote

$$\begin{aligned} w_{ij}(\beta) &= P(Y = i|Z = j) \\ &= \int P(Y = i, X = x|Z = j)dx = \int P(Y = i|x, Z = j)P(X = x|Z = j)dx, \\ \xi_{ij} &= \sum_{k=1}^n I(Y_k = i, Z_k = j, D_k = 0), \end{aligned}$$

$n_j = \sum_{i=1}^n D_i I(Z_i = j)$. Let X_{jl} , $l = 1, 2, \dots, n_j$ be the observed X 's with $Z = j$ and $p_{lj} = dF(x_{jl}|Z = j)$.

The log-likelihood is

$$\ell = \sum_{i=1}^n D_i \log f(y_i|x_i, z_i, \beta) + \sum_{j=1}^k \sum_{l=1}^{n_j} \log p_{lj} + \sum_{j=1}^k \ell_{Mj},$$

where

$$\ell_{Mj} = \sum_{i=1}^{n_j} \xi_{ij} \log w_{ij},$$

$$\sum_{i=1}^{n_j} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \sum_{i=1}^{n_j} p_{ij} \{P(Y = i|x, Z = j, \beta) - w_{ij}\} = 0, \quad j = 1, 2, \dots, k.$$

Returning to the upgraded mixture model, Zhang and Rockette (2005) proposed profile likelihood approach by assigning mass at each of observed data points X_1, \dots, X_n , but no grouping is used. More specifically for the upgraded mixture model they used the EM algorithm to maximize

$$\prod_{i=1}^m \left[\sum_{j=1}^n p_j f(y_i | x_j, \beta) \right] \prod_{i=1}^n p_i$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0.$$

In the missing data case without additional covariate Z , it is equivalent to maximizing

$$\prod_{i=1}^n [f(y_i | x_i, \beta) p_i]^{D_i} \left[\sum_{j=1}^{n_1} p_j f(y_i | x_j, \beta) \right]^{1-D_i}.$$

It is straightforward to generalize the above method to the case of a categorical covariate Z .

Song et al. (2009) considered an outcome dependent sampling problem with two-stage sampling problem, where the likelihood is given by

$$[\pi(y) f(y|x, \beta) g(x)]^D [\{1 - \pi(y)\} \int f(y|x, \beta) g(x) dx]^{1-D}.$$

For a given X , let the probability of being observed be $\pi(y) = P(D = 1|y, x)$. Note that $\int f(y|x, \beta) dG(x)$ can be treated as a continuous version of a mixture model. Based on Lindsay (1995), Van der Vaart and Wellner (1992), the number of support points of G is finite (less than that of the observed data points). However, other than the observed X_i 's ($D_i = 1$), it is not easy to locate other support points. Zhang and Rockette (2005) and Song et al. (2009) showed that the restricted MLE with support points on only those observed X_i 's is also valid.

Chapter 18

Connections Among Marginal Likelihood, Conditional Likelihood and Empirical Likelihood

In this Chapter we present the results by Qin and Zhang (2005) and Li and Qin (2011) on the connection between marginal likelihood, conditional likelihood and empirical likelihood.

Marginal likelihood and conditional likelihood are two of the most popular methods to eliminate nuisance parameters in a parametric model. Let a random variable Y have a density $f_Y(y, \phi)$ depending on a vector parameter $\phi = (\theta, \eta)$. Consider the case where Y can be partitioned into the two components $Y = (Y_1, Y_2)$, possibly after a transformation. Based on $Y = (Y_1, Y_2)$, the full likelihood can be factorized into the product of a marginal likelihood and a conditional likelihood:

$$f_Y(y, \phi) = f_{Y_1}(y_1, \phi) f_{Y_2|Y_1}(y_2|y_1, \phi).$$

Let $\ell_F(\phi) = \ell_F(\phi; y) = \log f_Y(y, \phi)$ be the full log-likelihood, $\ell_C(\phi) = \ell_C(\phi; y) = \log f_{Y_2|Y_1}(y_2|y_1, \phi)$ be the conditional log-likelihood, and $\ell_M(\phi) = \ell_M(\phi; y_1) = \log f_{Y_1}(y_1, \phi)$ be the marginal log-likelihood. In terms of log-likelihood, the above factorization can be expressed as

$$\ell_F(\phi) = \ell_C(\phi) + \ell_M(\phi).$$

This type of factorization is widely used in parametric statistical inference, where we are only interested in θ with η treated as a nuisance parameter. In some cases, use of the joint maximum likelihood estimate of ϕ may lead to misleading results due to a high dimensional η . When only one of the two factors involves θ , it is possible to use just that factor to make statistical inference for θ . Some pioneering works can be found in Kalbfleisch and Sprott (1970) and Andersen (1970), among others. The above factorizations have been generalized to semiparametric models. Godambe (1976) discussed conditional likelihood and unconditional optimal estimating equations. In the development of methods based on appropriate factorization of the full likelihood, Cox (1972, 1975) introduced the concept of partial likelihood

to eliminate the baseline unknown hazard function in proportional hazards regression models (More details will be discussed in the survival analysis chapters 24 and 25). Moreover, Kalbfleisch and Prentice (1973) showed that the marginal rank likelihood is equivalent to the Cox partial likelihood. In a generalized linear model, Kalbfleisch (1978) used the conditional likelihood argument to eliminate unknown baseline functions.

The profile likelihood technique is another popular method to eliminate nuisance parameters or unknown baseline functions under semiparametric models. For the Cox proportional hazards regression model, Bailey (1984) proved that after profiling the baseline hazard function, the semiparametric likelihood is equivalent to the Cox partial likelihood. Murphy and Van der Vaart (2000) established many nice properties of profile likelihood under a general semiparametric model. The profile likelihood has a close connection with empirical likelihood. As a nonparametric method, the empirical likelihood was introduced by Owen (1988, 1990) for constructing confidence intervals for the mean and other parameters. In this Chapter we will demonstrate that a full empirical likelihood can be decomposed into a product of a conditional likelihood and a marginal empirical likelihood in some situations.

The best and simplest example is the unordered pairs in genetic studies and double blind clinical trials.

18.1 Unodered Pairs

As discussed by Hinkley (1973), unordered pairs of random variables occur when the ordering of variables is not observable. Let (X_{1i}, X_{2i}) , $i = 1, \dots, n$, be independent and identically distributed random pairs, where X_{1i} and X_{2i} are independent random variables with density functions g and f , respectively. For $i = 1, \dots, n$, let $Y_{1i} = \min(X_{1i}, X_{2i})$ and $Y_{2i} = \max(X_{1i}, X_{2i})$ be the observed minimum and maximum values of X_{1i} and X_{2i} . The identities of Y_{1i} and Y_{2i} are, however, not observable. Davies and Phillips (1988) gave an example in interim analysis of a double-blind clinical trial where this type of data might be collected. Another application of this problem can be found in genetic chromosome analysis (Lauder 1977).

Based on the observed data (y_{1i}, y_{2i}) , the likelihood is

$$L = \prod_{i=1}^n \{g(y_{1i})f(y_{2i}) + g(y_{2i})f(y_{1i})\}I(y_{1i} \leq y_{2i}).$$

Under normal assumptions on g and f , Hinkley (1973) considered maximum likelihood estimation for the underlying parameters. Following Anderson (1979), Qin and Zhang (2005) assumed that g and f are related by the density ratio model or exponential tilting model

$$f(x) = \exp(\alpha + \beta x)g(x), \quad (18.1.1)$$

where α and β are unknown scalar parameters. This semiparametric model is composed of a parametric part involving the finite dimensional parameter (α, β) and a nonparametric part involving the baseline density function $g(x)$, an unknown infinite dimensional parameter. For statistical inference on (α, β) , it is desirable to eliminate the baseline density function $g(x)$ (or distribution function $G(x)$).

Based on the observed data (y_{1i}, y_{2i}) the marginal log-likelihood function of (α, β, G) is given by

$$\ell_M(\alpha, \beta, G) = \sum_{i=1}^n \log\{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})\} + \sum_{i=1}^n \sum_{j=1}^2 \log p_{ji},$$

where $p_{ji} = dG(y_{ji})$ ($i = 1, \dots, n$, $j = 1, 2$) are the $2n$ nonnegative jump sizes of G at the observed y_{ji} with total mass unity. To eliminate G using profile likelihood or empirical likelihood, we maximize $\ell_M(\alpha, \beta, G)$ with respect to p_{ji} ($i = 1, \dots, n$, $j = 1, 2$) for fixed (α, β) , subject to the constraints

$$\sum_{i=1}^n \sum_{j=1}^2 p_{ji} = 1, \quad p_{ji} \geq 0, \quad \sum_{i=1}^n \sum_{j=1}^2 p_{ji} \{\exp(\alpha + \beta y_{ji}) - 1\} = 0.$$

Using the Lagrange multiplier argument, we can show that the semiparametric marginal profile log-likelihood function of (α, β) can be written as (**exercise**)

$$\ell_M(\alpha, \beta) = \sum_{i=1}^n \log\{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})\} - \sum_{i=1}^n \sum_{j=1}^2 \log\{1 + \exp(\alpha + \beta y_{ji})\}. \quad (18.1.2)$$

If (X_{1i}, X_{2i}) are observable instead of the unordered pairs (Y_{1i}, Y_{2i}) , then we can consider a full likelihood approach, which in itself motivates a conditional likelihood approach in the unordered pairs problem. Let $D_i = I(X_{1i} < X_{2i})$ for $i = 1, \dots, n$. Then the original data $\{(X_{1i}, X_{2i}), i = 1, \dots, n\}$ are equivalent to $\{(Y_{1i}, Y_{2i}, D_i), i = 1, \dots, n\}$. Based on (X_{1i}, X_{2i}) and the exponential tilting model, the full log-likelihood function of (α, β, G) is

$$\ell_F(\alpha, \beta, G) = \sum_{i=1}^n (\alpha + \beta x_{2i}) + \sum_{i=1}^n \sum_{j=1}^2 \log p_{ji},$$

where $p_{ji} = dG(y_{ji})$ ($i = 1, \dots, n$, $j = 1, 2$) are nonnegative jumps with total mass unity. In order to eliminate G , we need to maximize $\ell_F(\alpha, \beta, G)$ with respect to the p_{ji} s subject to the same constraints as above with X_{ji} in place of Y_{ji} . When $(X_{1i}, X_{2i}) = (Y_{1i}, Y_{2i})$, we have shown in Chap. 11 that the semiparametric full profile log-likelihood function of (α, β) is

$$\ell_F(\alpha, \beta) = \sum_{i=1}^n (\alpha + \beta y_{2i}) - \sum_{i=1}^n \sum_{j=1}^2 \log\{1 + \exp(\alpha + \beta y_{ji})\}. \quad (18.1.3)$$

Since $P(X_{1i}, X_{2i}) = P(Y_{1i}, Y_{2i})P(X_{1i}, X_{2i}|Y_{1i}, Y_{2i})$, in order for the factorization

$$\ell_F(\phi) = \ell_C(\phi) + \ell_M(\phi) \quad (18.1.4)$$

to hold in the unordered pairs problem, we need to show that $\ell_F(\alpha, \beta) - \ell_M(\alpha, \beta)$ is the semiparametric conditional log-likelihood function of (α, β) based on $P(X_{1i}, X_{2i}|Y_{1i}, Y_{2i})$ or $P(D_i|Y_{1i}, Y_{2i})$.

Using the exponential tilting model, the conditional distribution of D_i given the order statistics (Y_{1i}, Y_{2i}) is given by

$$\begin{aligned} P(D_i = 1|Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) &= \frac{P(X_{1i} = y_{1i}, X_{2i} = y_{2i})}{P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})} \\ &= \frac{g(y_{1i})f(y_{2i})}{g(y_{1i})f(y_{2i}) + g(y_{2i})f(y_{1i})} = \frac{\exp(\alpha + \beta y_{2i})}{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})}. \end{aligned}$$

Clearly $\exp(\alpha)$ is cancelled out. Therefore the log conditional likelihood is

$$\ell_C(\beta) = \sum_{i=1}^n \beta y_{2i} - \sum_{i=1}^n \log\{\exp(\beta y_{1i}) + \exp(\beta y_{2i})\}.$$

In addition,

$$\begin{aligned} P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) &= \frac{P(X_{1i} = y_{1i}, X_{2i} = y_{2i})}{P(X_{1i} = y_{1i}, X_{2i} = y_{2i}|Y_{1i} = y_{1i}, Y_{2i} = y_{2i})} \\ &= \frac{\exp(\alpha + \beta y_{2i})}{\{1 + \exp(\alpha + \beta y_{1i})\}\{1 + \exp(\alpha + \beta y_{2i})\}} / \frac{\exp(\alpha + \beta y_{2i})}{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})} \\ &= \frac{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})}{\{1 + \exp(\alpha + \beta y_{1i})\}\{1 + \exp(\alpha + \beta y_{2i})\}}. \end{aligned}$$

This provides an alternative way to derive the semiparametric marginal profile log-likelihood $\ell_M(\alpha, \beta)$ using the complete data (X_{1i}, X_{2i}) profile likelihood and the conditional likelihood $\ell_C(\beta)$.

Remark 1 The Fisher information based on the full log-likelihood $\ell_F(\alpha, \beta)$ is equal to the sum of the Fisher information based on the conditional log-likelihood $\ell_C(\beta)$ and the Fisher information based on the marginal log-likelihood $\ell_M(\alpha, \beta)$.

Remark 2 The marginal profile log-likelihood $\ell_M(\alpha, \beta)$ is invariant if we replace (α, β) by $(-\alpha, -\beta)$. Hence, we may restrict $\beta \geq 0$.

Next we derive the asymptotic results for the unordered pairs problem. As pointed out before, statistical inferences on (α, β) can be based on the semiparametric mar-

ginal profile log-likelihood ℓ_M . We can maximize $\ell_M(\alpha, \beta)$ over (α, β) . Let $(\hat{\alpha}, \hat{\beta})$ satisfy the following system of score equations:

$$\begin{aligned}\frac{\partial \ell_M(\alpha, \beta)}{\partial \alpha} &= n - \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\alpha + \beta y_{ji})}{1 + \exp(\alpha + \beta y_{ji})} = 0, \\ \frac{\partial \ell_M(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^n \frac{y_{1i} \exp(\alpha + \beta y_{1i}) + y_{2i} \exp(\alpha + \beta y_{2i})}{\exp(\alpha + \beta y_{1i}) + \exp(\alpha + \beta y_{2i})} - \sum_{i=1}^n \sum_{j=1}^2 \frac{y_{ji} \exp(\alpha + \beta y_{ji})}{1 + \exp(\alpha + \beta y_{ji})} = 0.\end{aligned}$$

Theorem 18.1 Let (α_0, β_0) be the true value of (α, β) under model (18.1.1). If $\beta_0 \neq 0$, i.e., if the two samples have different distributions, then under suitable regularity conditions, we can write

$$\begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} = \frac{1}{n} S^{-1} \begin{pmatrix} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(n^{-1/2}),$$

where

$$\begin{aligned}S &= \begin{pmatrix} S_0 & S_1 \\ S_1 & S_2 \end{pmatrix}, \quad S_k = S_k(\infty), \quad S_k(x) = \int_{-\infty}^x \frac{\exp(\alpha_0 + \beta_0 x)}{1 + \exp(\alpha_0 + \beta_0 x)} x^k g(x) dx, \quad k = 0, 1, \\ S_2 &= \int \frac{\exp(\alpha_0 + \beta_0 x)}{1 + \exp(\alpha_0 + \beta_0 x)} x^2 g(x) dx \\ &\quad - \frac{1}{2} \int \int \frac{(x-y)^2 \exp(\alpha_0 + \beta_0 x) \exp(\alpha_0 + \beta_0 y)}{\exp(\alpha_0 + \beta_0 x) + \exp(\alpha_0 + \beta_0 y)} g(x) g(y) dx dy.\end{aligned}$$

As a result,

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} \rightarrow N_2(0, \Sigma), \quad \Sigma = S^{-1} - \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \quad (18.1.5)$$

in distribution.

Remark 3 It can be proved that if $\beta = \alpha = 0$, then $S_0 = 0.5$, $S_1 = 0.5E(X_{1i})$, and $S_2 = 0.5\{E(X_{1i})\}^2$. Thus, the information matrix S is singular. Consequently, the classical development leading to the asymptotic distribution for the maximum likelihood estimation is not applicable in this case. This is similar to the parametric normal case, where Hinkley (1973) observed an irregular behavior when the two samples have the same distribution. This type of irregular phenomenon was observed in simulation studies carried out by Qin and Zhang (2005).

Proof Since the semiparametric marginal profile log-likelihood function $\ell_M(\alpha, \beta)$ is symmetric with respect to y_{1i} and y_{2i} , we can rewrite it as a function of the (x_{1i}, x_{2i}) :

$$\ell_M(\alpha, \beta) = \sum_{i=1}^n \log\{\exp(\alpha + \beta x_{1i}) + \exp(\alpha + \beta x_{2i})\} - \sum_{j=1}^2 \sum_{i=1}^n \log\{1 + \exp(\alpha + \beta x_{ji})\}.$$

The score estimating equations are

$$\begin{aligned}\frac{\partial \ell_M(\alpha, \beta)}{\partial \alpha} &= n - \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\alpha + \beta x_{ji})}{1 + \exp(\alpha + \beta x_{ji})} = 0, \\ \frac{\partial \ell_M(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^n \frac{x_{1i} \exp(\alpha + \beta x_{1i}) + x_{2i} \exp(\alpha + \beta x_{2i})}{\exp(\alpha + \beta x_{1i}) + \exp(\alpha + \beta x_{2i})} \\ &\quad - \sum_{i=1}^n \sum_{j=1}^2 \frac{x_{ji} \exp(\alpha + \beta x_{ji})}{1 + \exp(\alpha + \beta x_{ji})} = 0.\end{aligned}$$

It can be shown under suitable regularity conditions that $(\hat{\alpha}, \hat{\beta})$ is consistent.

It is also easy to show, that, under the exponential tilting model

$$E \left\{ \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} \right\} = 0, \quad E \left\{ \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \right\} = 0.$$

Furthermore,

$$\begin{aligned}\frac{\partial^2 \ell_M(\alpha, \beta)}{\partial \alpha^2} &= - \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\alpha + \beta x_{ji})}{\{1 + \exp(\alpha + \beta x_{ji})\}^2}, \quad \frac{\partial^2 \ell_M(\alpha, \beta)}{\partial \alpha \partial \beta} \\ &= - \sum_{i=1}^n \sum_{j=1}^2 \frac{x_{ji} \exp(\alpha + \beta x_{ji})}{\{1 + \exp(\alpha + \beta x_{ji})\}^2},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell_M(\alpha, \beta)}{\partial \beta^2} &= \sum_{i=1}^n \frac{(x_{1i} - x_{2i})^2 \exp(\alpha + \beta x_{1i}) \exp(\alpha + \beta x_{2i})}{\{\exp(\alpha + \beta x_{1i}) + \exp(\alpha + \beta x_{2i})\}^2} \\ &\quad - \sum_{i=1}^n \sum_{j=1}^2 \frac{x_{ji}^2 \exp(\alpha + \beta x_{ji})}{\{1 + \exp(\alpha + \beta x_{ji})\}^2}.\end{aligned}$$

Applying the Weak Law of Large Numbers gives

$$S_n = -\frac{1}{n} \begin{pmatrix} \frac{\partial \ell_M^2(\alpha_0, \beta_0)}{\partial \alpha^2} & \frac{\partial \ell_M^2(\alpha_0, \beta_0)}{\partial \alpha \partial \beta} \\ \frac{\partial \ell_M^2(\alpha_0, \beta_0)}{\partial \beta \partial \alpha} & \frac{\partial \ell_M^2(\alpha_0, \beta_0)}{\partial \beta^2} \end{pmatrix} \rightarrow S =: \begin{pmatrix} S_0 & S_1 \\ S_1 & S_2 \end{pmatrix},$$

in probability. Moreover, it follows from the Central Limit Theorem that

$$\frac{1}{\sqrt{n}} \begin{pmatrix} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} \rightarrow N(0, B),$$

in distribution, where we can show after some tedious algebra that $B = S - 2(S_0, S_1)^T(S_0, S_1)$. Expanding $\frac{\partial \ell_M(\hat{\alpha}, \hat{\beta})}{\partial \alpha}$ and $\frac{\partial \ell_M(\hat{\alpha}, \hat{\beta})}{\partial \beta}$ at (α_0, β_0) gives

$$\begin{aligned} 0 &= \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} + \frac{\partial^2 \ell_M(\alpha_0, \beta_0)}{\partial \alpha^2}(\hat{\alpha} - \alpha_0) + \frac{\partial^2 \ell_M(\alpha_0, \beta_0)}{\partial \alpha \partial \beta}(\hat{\beta} - \beta_0) + o_p(|\hat{\alpha} - \alpha_0| + |\hat{\beta} - \beta_0|), \\ 0 &= \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} + \frac{\partial^2 \ell_M(\alpha_0, \beta_0)}{\partial \alpha \partial \beta}(\hat{\alpha} - \alpha_0) + \frac{\partial^2 \ell_M(\alpha_0, \beta_0)}{\partial \beta^2}(\hat{\beta} - \beta_0) + o_p(|\hat{\alpha} - \alpha_0| + |\hat{\beta} - \beta_0|), \end{aligned}$$

$$S_n \begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} = \begin{pmatrix} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} \\ \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(1).$$

Finally, using Slutsky's theorem, we can show

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \end{pmatrix} \rightarrow N_2(0, \Sigma), \quad \Sigma = S^{-1} B S^{-1} = S^{-1} - \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix},$$

in distribution,

Next we consider the likelihood ratio test. The following theorem gives the asymptotic distribution of the likelihood ratio statistic for testing $H_0 : \beta = \beta_0$.

Theorem 18.2 *Let $\beta_0 \neq 0$. Under the conditions of Theorem 18.1, the likelihood ratio statistic*

$$R(\beta_0) = 2\{\max_{\alpha, \beta} \ell_M(\alpha, \beta) - \max_{\alpha} \ell_M(\alpha, \beta_0)\} = 2\{\ell_M(\hat{\alpha}, \hat{\beta}) - \ell_M(\tilde{\alpha}, \beta_0)\} \rightarrow \chi_1^2 \quad (18.1.6)$$

in distribution under the null hypothesis $H_0 : \beta = \beta_0$, where $\tilde{\alpha}$ is a solution to the following score equation:

$$\frac{\partial \ell_M(\alpha, \beta_0)}{\partial \alpha} = n - \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\alpha + \beta_0 y_{ji})}{1 + \exp(\alpha + \beta_0 y_{ji})} = 0.$$

Proof Similar to the proof of Theorem 18.1, we can write

$$\tilde{\alpha} - \alpha_0 = \frac{1}{n} S_0^{-1} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} + o_p(n^{-1/2}).$$

Let $d = |S| = S_0 S_2 - S_1^2$. Using a second-order Taylor's expansion,

$$\begin{aligned} R(\beta_0) &= 2\{\max_{\alpha, \beta_0} \ell_M(\alpha, \beta_0) - \max_{\alpha} \ell_M(\alpha, \beta_0)\} = 2\{\ell_M(\hat{\alpha}, \hat{\beta}_0) - \ell_M(\tilde{\alpha}, \beta_0)\} \\ &= \frac{1}{n} \frac{1}{d S_0} \left\{ S_1 \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} - S_0 \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \right\}^2 + o_p(1) = Z_n^2 + o_p(1), \end{aligned}$$

where

$$Z_n = \frac{1}{\sqrt{n}} \left\{ \frac{S_1}{\sqrt{dS_0}} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \alpha} - \sqrt{\frac{S_0}{d}} \frac{\partial \ell_M(\alpha_0, \beta_0)}{\partial \beta} \right\}.$$

It can be shown after some algebra that $E(Z_n) = 0$ and

$$\begin{aligned} \text{var}(Z_n) &= \frac{1}{n} \left\{ \frac{S_1^2}{dS_0} \text{var}(L_1) + \frac{S_0^2}{dS_0} \text{var}(L_2) - 2 \frac{S_0 S_1}{dS_0} \text{cov}(L_1, L_2) \right\} \\ &= \frac{1}{n} \left\{ \frac{S_1^2}{dS_0} n(S_0 - 2S_0^2) + \frac{S_0^2}{dS_0} n(S_2 - 2S_1^2) - 2 \frac{S_0 S_1}{dS_0} n(S_1 - 2S_0 S_1) \right\} \\ &= \frac{1}{dS_0} \{S_1^2(S_0 - 2S_0^2) + S_0^2(S_2 - 2S_1^2) - 2S_0 S_1(S_1 - 2S_0 S_1)\} \\ &= \frac{1}{dS_0} \{S_1^2 S_0 + S_0^2 S_2 - 2S_0 S_1^2\} = 1. \end{aligned}$$

By employing the Central Limit Theorem and Slutsky's Theorem, we can show that $Z_n \rightarrow N(0, 1)$ in distribution, and hence $R(\beta_0) = Z_n^2 + o_p(1) \rightarrow \chi_1^2$.

We can estimate the cumulative distribution functions G and F , and their respective means $\mu_1 = \int x dG(x)$ and $\mu_2 = \int x dF(x)$, by

$$\begin{aligned} \hat{G}(y) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 \frac{I(Y_{ji} \leq y)}{1 + \exp(\hat{\alpha} + \hat{\beta}Y_{ji})}, & \hat{F}(y) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\hat{\alpha} + \hat{\beta}Y_{ji}) I(Y_{ji} \leq y)}{1 + \exp(\hat{\alpha} + \hat{\beta}Y_{ji})}, \\ \hat{\mu}_1 &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 \frac{Y_{ji}}{1 + \exp(\hat{\alpha} + \hat{\beta}Y_{ji})}, & \hat{\mu}_2 &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 \frac{Y_{ji} \exp(\hat{\alpha} + \hat{\beta}Y_{ji})}{1 + \exp(\hat{\alpha} + \hat{\beta}Y_{ji})}. \end{aligned}$$

Exercise Derive the large sample properties.

18.2 Two-Component Mixture Model with Multiple Samples

In this section we discuss a two-component mixture model with multiple samples along with their applications in genetic quantitative trait loci analysis and estimation of tuberculous infection prevalence.

The standard method for quantitative trait loci is interval mapping (Lander and Botstein 1989). Since markers are observed at known locations, the genotypes between the locations are missing. In backcross studies, this leads to a two sample mixture model at putative loci. The component densities, f and g , are associated with two possible genotypes. The mixing probabilities are determined by the recombination fractions between a locus and the flanking markers. Zou, Fine and Yandell (2002) used the exponential tilting model in a semiparametric mixture model as

described below. The same problem was also discussed by Nagelkerke, Borgdorff and Kim (2001) in estimation of tuberculous prevalence based on a survey dataset carried out in South Korea, in which data from several populations with different proportions of tuberculosis infected cases were collected.

For $i = 1, \dots, I$ with $I > 1$, suppose

X_{i1}, \dots, X_{in_i} are independent with density $\lambda_i g(x) + (1 - \lambda_i) f(x)$,

where X_{ij} is the j th observation from the i th mixture and $\lambda_i \neq \lambda_j$ for $1 \leq i \neq j \leq I$. Assume that $\{(X_{i1}, \dots, X_{in_i}), i = 1, \dots, I\}$ are all independent. Let $n = \sum_{i=1}^I n_i$. Suppose D_{i1}, \dots, D_{in_i} are the indicator variables with $D_{ij} = 1$ for X_{ij} sampled from f . Let $P(D_{ij} = 1) = 1 - \lambda_i$. If we observe both $(X_{i1}, \dots, X_{in_i})$ and $(D_{i1}, \dots, D_{in_i})$, then under the exponential tilting model, the semiparametric full likelihood function of (α, β, G) is

$$\begin{aligned} L_F(\alpha, \beta, G) &= \prod_{i=1}^I \prod_{j=1}^{n_i} \{(1 - \lambda_i) f(x_{ij})\}^{D_{ij}} \{\lambda_i g(x_{ij})\}^{1-D_{ij}} \\ &= \prod_{i=1}^I \prod_{j=1}^{n_i} \{(1 - \lambda_i) \exp(\alpha + x_{ij}\beta)\}^{D_{ij}} \lambda_i^{1-D_{ij}} dG(x_{ij}). \end{aligned}$$

By profiling out G , the semiparametric full profile likelihood function of (α, β) is

$$L_F(\alpha, \beta) = \prod_{i=1}^I \prod_{j=1}^{n_i} \frac{\exp\{D_{ij}(\alpha + x_{ij}\beta)\}}{1 + \rho \exp(\alpha + x_{ij}\beta)} (1 - \lambda_i)^{D_{ij}} \lambda_i^{1-D_{ij}},$$

where $\rho = (\sum_{ij} D_{ij}) / \{\sum_{ij} (1 - D_{ij})\}^{-1}$.

Under the exponential tilting model and conditional on $(X_{i1}, \dots, X_{in_i}) = (x_{i1}, \dots, x_{in_i})$, the semiparametric conditional likelihood function of (α, β) based on $(D_{i1}, \dots, D_{in_i})$ is given by

$$L_C(\alpha, \beta) = \prod_{i=1}^I \prod_{j=1}^{n_i} \frac{\rho_i^{D_{ij}} \exp\{D_{ij}(\alpha + x_{ij}\beta)\}}{1 + \rho_i \exp(\alpha + x_{ij}\beta)},$$

where $\rho_i = (1 - \lambda_i) / \lambda_i$. Note that $L_C(\alpha, \beta)$ naturally eliminates the baseline density function $g(x)$, but, unlike the case in unordered pairs problems, it does not eliminate α . The ratio

$$\frac{L_F(\alpha, \beta)}{L_C(\alpha, \beta)} = \prod_{i=1}^I \lambda_i^{n_i} \left\{ \prod_{i=1}^I \prod_{j=1}^{n_i} \frac{1 + \rho_i \exp(\alpha + x_{ij}\beta)}{1 + \rho \exp(\alpha + x_{ij}\beta)} \right\},$$

is not dependent on the baseline density function $g(x)$ and may be employed to make statistical inferences on the underlying parameters if we replace the D_{ij} by their expectations $E(D_{ij})$ in ρ . Furthermore, it is of interest to know if $L_F(\alpha, \beta)/L_C(\alpha, \beta)$ is the semiparametric marginal profile likelihood function of (α, β) under the exponential tilting model based on the observed data $(X_{i1}, \dots, X_{in_i})$. This in general, is not true. Nevertheless, we have the following lemma.

Lemma 18.3 *Let $\ell_M(\alpha, \beta)$ represent the semiparametric marginal profile log-likelihood function of (α, β) based on the observed data $(X_{i1}, \dots, X_{in_i})$. Then*

$$\ell_M(\alpha, \beta) = \sum_{i=1}^I \sum_{j=1}^{n_i} \log\{\lambda_i + (1 - \lambda_i) \exp(\alpha + x_{ij}\beta)\} - \sum_{i=1}^I \sum_{j=1}^{n_i} \log[1 + \tau\{\exp(\alpha + x_{ij}\beta) - 1\}], \quad (18.2.7)$$

where τ is the Lagrange multiplier given by

$$\tau = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^{n_i} \frac{(1 - \lambda_i) \exp(\alpha + \beta x_{ij})}{\lambda_i + (1 - \lambda_i) \exp(\alpha + \beta x_{ij})}.$$

If we replace τ by $\tilde{\tau} = n^{-1} \sum_{i=1}^I \sum_{j=1}^{n_i} I(D_{ij} = 1)$ in the profile marginal empirical likelihood and define

$$\tilde{\ell}_M(\alpha, \beta) = \sum_{i=1}^I \sum_{j=1}^{n_i} \log\{\lambda_i + (1 - \lambda_i) \exp(\alpha + x_{ij}\beta)\} - \sum_{i=1}^I \sum_{j=1}^{n_i} \log[1 + \tilde{\tau}\{\exp(\alpha + x_{ij}\beta) - 1\}],$$

then $\log L_F(\alpha, \beta) = \log L_C(\alpha, \beta) + \tilde{\ell}_M(\alpha, \beta)$ so that the log joint profile empirical likelihood can be decomposed as the summation of a log conditional likelihood and a log marginal profile empirical likelihood.

Proof Based on the observed data $(X_{i1}, \dots, X_{in_i})$ and the exponential tilting model, the semiparametric marginal likelihood function of (α, β, G) is

$$\ell_M(\alpha, \beta, G) = \prod_{i=1}^I \prod_{j=1}^{n_i} \{\lambda_i + (1 - \lambda_i) \exp(\alpha + x_{ij}\beta)\} dG(x_{ij}).$$

After profiling out G in $\ell_M(\alpha, \beta, G)$, we arrive at the following semiparametric marginal profile log-likelihood function of (α, β)

$$\ell_M(\alpha, \beta) = \sum_{i=1}^I \sum_{j=1}^{n_i} \log\{\lambda_i + (1 - \lambda_i) \exp(\alpha + x_{ij}\beta)\} - \sum_{i=1}^I \sum_{j=1}^{n_i} \log[1 + \tau\{\exp(\alpha + x_{ij}\beta) - 1\}],$$

where τ is the Lagrange multiplier determined by the constraint equation

$$\sum_{i=1}^I \sum_{j=1}^{n_i} \frac{\exp(\alpha + x_{ij}\beta) - 1}{1 + \tau\{\exp(\alpha + x_{ij}\beta) - 1\}} = 0.$$

Using this equation and the fact that $\partial \ell_M(\alpha, \beta)/\partial \alpha = 0$, we have

$$n\tau - \sum_{i=1}^I (1 - \lambda_i) \sum_{j=1}^{n_i} \frac{\exp(\alpha + \beta x_{ij})}{\lambda_i + (1 - \lambda_i)\exp(\alpha + \beta x_{ij})} = 0.$$

Unlike the Lagrange multipliers in unordered pairs problems, the Lagrange multiplier τ here is not a constant. By taking expectation with respect to x_{ij} on the left-hand side of the above equation, we obtain

$$\tau = \frac{1}{n} \sum_{i=1}^I n_i (1 - \lambda_i) = E \left\{ \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^{n_i} I(D_{ij} = 1) \right\}.$$

We observe that $\log L_F(\alpha, \beta) = \log L_C(\alpha, \beta) + \tilde{\ell}_M(\alpha, \beta)$ if $\tilde{\tau} = n^{-1} \sum_{i=1}^I \sum_{j=1}^{n_i} I(D_{ij} = 1)$ replaces τ in $\ell_M(\alpha, \beta)$.

18.3 Genetic Linkage Mixture Models

Genetic linkage analysis has been discussed extensively in the statistics and genetics literature, see, e.g., Sham (1998). Most estimation methods used parametric assumptions. However, model misspecification is one of the main concerns for geneticists and biostatisticians (Sham 1998). In the following we show that the density ratio model based on semiparametric approach is possible in linkage analysis.

Consider a human pedigree of size n and let X_i ($i = 1, \dots, n$) denote the phenotype at one or multiple loci of the i -th pedigree member. The likelihood, being the probability of the observations, is then $P(x_1, \dots, x_n)$. A key assumption in linkage analysis is that an individual's phenotype depends only on his/her genotypes.

Individuals whose parents are not included in the pedigree are called founders. Suppose there are m founders and $n - m$ non-founders in a pedigree ($n > 1$). Denote the penetrance (probability density for a given genotype), population genotype frequency, and the transmission probability of genotype D_i for given genotypes of parents ($D_{i,F}, D_{i,M}$) by $f(X_i|D_i)$, $P(D_i)$, and $P(D_i|D_{i,F}, D_{i,M})$, respectively, then the observed phenotype data have the marginal likelihood (Sham 1998, Eq. 3.43)

$$L_M = \sum_{D_1, \dots, D_n} \prod_{i=1}^n f(x_i|D_i) \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M}).$$

For simplicity, assume the genotype can only take one of two values, 0 or 1. Even in the parametric likelihood setup, the computation of this likelihood is very complicated. There are n summations, each indexed by the possible ordered genotypes of a pedigree member. In the literature, different algorithms have been proposed, see, e.g., Elston and Stewart (1971), Lange and Elston (1975), and Lander and Green (1987). Instead of studying the algorithm, we explore the semiparametric likelihood estimation under the exponential tilting model, i.e.

$$f(x|D=1) = \exp(\alpha + x\beta)f(x|D=0).$$

In the complete data case where both phenotype and genotype are observed for each individual, the full likelihood is given by

$$L_F = \prod_{i=1}^n f(x_i|D_i) \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M}).$$

Similar to previous approaches, the semiparametric full profile likelihood function of (α, β) can be shown to be

$$L_F(\alpha, \beta) = \left[\prod_{i=1}^n \frac{\exp\{D_i(\alpha + x_i\beta)\}}{1 + \rho \exp(\alpha + x_i\beta)} \right] \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M}),$$

where $\rho = \sum_{i=1}^n D_i / \sum_{i=1}^n (1 - D_i)$.

Conditioning on the observed phenotype, the semiparametric conditional likelihood function of (α, β) based on the genotype is given by

$$L_C(\alpha, \beta) = \frac{\exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M})}{\sum_{D_1, \dots, D_n} \exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M})}. \quad (18.3.8)$$

Again, the unknown baseline density function g is cancelled out, but α is not. The semiparametric conditional likelihood $L_C(\alpha, \beta)$ allows an alternative way for statistical inferences on (α, β) when both phenotype and genotype are observed on each individual. The following lemma shows that $L_F(\alpha, \beta)/L_C(\alpha, \beta)$ is generally not identical to the semiparametric marginal profile likelihood function of (α, β) based on the observed phenotype data (x_1, \dots, x_n) .

Lemma 18.4 *Let $\ell_M(\alpha, \beta)$ denote the semiparametric marginal profile log-likelihood function of (α, β) under the exponential tilting model based on the observed phenotype data (x_1, \dots, x_n) . Then*

$$\begin{aligned} \ell_M(\alpha, \beta) &= \log \left\{ \sum_{D_1, \dots, D_n} \left[\prod_{i=1}^n \exp\{D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i|D_{i,F}, D_{i,M}) \right] \right\} \\ &\quad - \sum_{i=1}^n \log[1 + \tau\{\exp(\alpha + x_i\beta) - 1\}], \end{aligned}$$

where τ is the Lagrange multiplier given by

$$\tau = \frac{\sum_{D_1, \dots, D_n} \sum_{j=1}^n D_j \exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M})}{n \sum_{D_1, \dots, D_n} \exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M})}.$$

If we replace τ by $\tilde{\tau} = n^{-1} \sum_{i=1}^n D_i$ and define

$$\begin{aligned} \tilde{\ell}_M(\alpha, \beta) = & \log \left\{ \sum_{D_1, \dots, D_n} \left[\prod_{i=1}^n \exp\{D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}) \right] \right\} \\ & - \sum_{i=1}^n \log[1 + \tilde{\tau}\{\exp(\alpha + x_i\beta) - 1\}], \end{aligned}$$

then $\log L_F(\alpha, \beta) = \log L_C(\alpha, \beta) + \tilde{\ell}_M(\alpha, \beta)$ so that the joint profile empirical log-likelihood can be factorized as the summation of the marginal profile empirical log-likelihood and the log-conditional likelihood holds.

Proof Based on the observed phenotype data (x_1, \dots, x_n) and under the exponential tilting model, the semiparametric marginal likelihood function of (α, β, G) is given by

$$L_M(\alpha, \beta, G) = \sum_{D_1, \dots, D_n} \left[\prod_{i=1}^n \exp\{D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}) \right] \prod_{i=1}^n p_i,$$

where $p_i = dG(x_i)$ ($i = 1, \dots, n$) are nonnegative jumps with total mass unity. Using the same technique as in case-control studies, we can profile out G in $L_M(\alpha, \beta, G)$ with a Lagrange multiplier argument. The resulting semiparametric marginal profile log-likelihood function of (α, β) is

$$\begin{aligned} \ell_M(\alpha, \beta) = & \log \left\{ \sum_{D_1, \dots, D_n} \left[\prod_{i=1}^n \exp\{D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}) \right] \right\} \\ & - \sum_{i=1}^n \log[1 + \tau\{\exp(\alpha + x_i\beta) - 1\}], \end{aligned}$$

where τ is the Lagrange multiplier satisfying the following constraint equation:

$$n\tau - \frac{\sum_{D_1, \dots, D_n} \sum_{j=1}^n D_j \exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M})}{\sum_{D_1, \dots, D_n} \exp\{\sum_{i=1}^n D_i(\alpha + x_i\beta)\} \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M})} = 0.$$

Again, unlike the Lagrange multipliers in the unordered pairs problem, τ is not a constant.

If we take expectations on both sides of above equation, we obtain

$$n\tau - \sum_{D_1, \dots, D_n} \sum_{j=1}^n D_j \prod_{i=1}^n \int \exp\{D_i(\alpha + x_i\beta)\} dG(x_i) \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}) = 0.$$

Using the fact that $\int \exp(\alpha + x\beta) dG(x) = 1$ yields

$$n\tau - \sum_{D_1, \dots, D_n} \sum_{j=1}^n D_j \prod_{i=1}^m P(D_i) \prod_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}) = 0,$$

or

$$n\tau - \sum_{j=1}^m E(D_j) - \sum_{i=m+1}^n E(D_i | D_{i,F}, D_{i,M}) = 0,$$

or

$$\tau = \frac{n-m}{n} P(D_1) + \frac{1}{n} \sum_{i=m+1}^n P(D_i | D_{i,F}, D_{i,M}).$$

In other words, $\tau = E^*\{n^{-1} \sum_{i=1}^n D_i\}$, where E^* denotes expectation conditional on $n-m$ individual parents' genotypes. We can then show $\log L_F(\alpha, \beta) = \log L_C(\alpha, \beta) + \tilde{\ell}_M(\alpha, \beta)$ when τ is replaced by $\tilde{\tau} = n^{-1} \sum_{i=1}^n D_i$ in $\ell_M(\alpha, \beta)$.

18.4 Shannon's Mutual Information for Nuisance Parameter Elimination

In this section we introduce a new nuisance parameter elimination method by using Shannon's mutual information.

Let us go back to the unordered pairs problem. Recall the Shannon's mutual information discussed in Chap. 9 is the expectation of the log-joint density divided by its marginal densities. The joint density based on unordered pairs is

$$h(y_1, y_2) = g(y_1)f(y_2) + g(y_2)f(y_1).$$

Under the density ratio assumption

$$f(t) = \exp(\alpha + \beta t)g(t)$$

it can be written as

$$\{\exp(\alpha + \beta y_1) + \exp(\alpha + \beta y_2)\}g(y_1)g(y_2).$$

Since the density $g(t)$ is not specified we cannot directly make inferences based on the joint likelihood. On the other hand, we may treat $g(t)$ as a infinite dimensional

nuisance parameter. Marginally Y_1 and Y_2 have the same density $h_1(y) = h_2(y) = 0.5g(y) + 0.5f(y) = 0.5\{\exp(\alpha + \beta y) + 1\}g(y)$. In statistical literature, an effective method to eliminate a nuisance parameter is the conditional likelihood. For example, we may consider the conditional density of $Y_1|Y_2$. We can observe that only $g(y_1)$ is cancelled out but not $g(y_2)$.

Dividing the joint density by its two marginal densities, we can eliminate the nuisance density function $g(t)$ completely,

$$\begin{aligned}\frac{h(y_1, y_2)}{h_1(y_1)h_2(y_2)} &= \frac{\{\exp(\alpha + \beta y_1) + \exp(\alpha + \beta y_2)\}g(y_1)g(y_2)}{0.5\{\exp(\alpha + \beta y_1) + 1\}g(y_1)0.5\{\exp(\alpha + \beta y_2) + 1\}g(y_2)} \\ &= \frac{\{\exp(\alpha + \beta y_1) + \exp(\alpha + \beta y_2)\}}{0.5\{\exp(\alpha + \beta y_1) + 1\}0.5\{\exp(\alpha + \beta y_2) + 1\}}.\end{aligned}$$

This is equivalent to the marginal profile log-likelihood $l_M(\alpha, \beta)$ in (18.1.2).

The unbiasedness of the score estimating functions derived from

$$\frac{\partial \log h(y_1, y_2)}{\partial \alpha} - \frac{\partial \log h_1(y_1)}{\partial \alpha} - \frac{\partial \log h_2(y_2)}{\partial \alpha}$$

and

$$\frac{\partial \log h(y_1, y_2)}{\partial \beta} - \frac{\partial \log h_1(y_1)}{\partial \beta} - \frac{\partial \log h_2(y_2)}{\partial \beta}$$

is guaranteed since the first term is the score based on the full likelihood and the second and third terms are scores based on two marginal likelihoods.

Let β_0 be the true value of β . If $\beta_0 \neq 0$, in Sect. 18.1 we showed that the MLE of β based on ℓ_M has $n^{-1/2}$ convergence rate and asymptotic normality, and the likelihood ratio test (LRT) for testing $H_0 : \beta = \beta_0 \neq 0$ has a χ^2_1 limiting distribution. If $\beta_0 = 0$, Qin and Zhang (2005) noticed that the Fisher information matrix becomes degenerate. They also noted that the classical asymptotic distribution of the MLE and the LRT for testing $H_0 : \beta = \beta_0$ are not applicable.

Note that $\beta_0 = 0$ corresponds to the case that the pair (y_{i1}, y_{i2}) comes from the same distribution. Testing $H_0 : \beta = 0$ is one of the important problems in many applications. Next, we investigate the asymptotic properties of the MLE of β and derive the limiting distribution of the LRT for testing $H_0 : \beta = 0$.

Let $(\hat{\alpha}, \hat{\beta})$ be the MLE of (α, β) , i.e.,

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} \sup l_M(\alpha, \beta).$$

The LRT statistic for testing $H_0 : \beta = 0$ is defined as

$$R_n = 2\{l_M(\hat{\alpha}, \hat{\beta}) - \sup_{\alpha} l_M(\alpha, 0)\}.$$

We note that when $\beta = 0$, the MLE of α is 0. Therefore

$$R_n = 2\{l_M(\hat{\alpha}, \hat{\beta}) - l_M(0, 0)\}.$$

The next theorem (Li and Qin 2011) presents the main result.

Theorem 18.5 Suppose $E(x_{1i}^5) < \infty$. Under the null hypothesis of $\beta_0 = 0$, as $n \rightarrow \infty$

1. $\hat{\beta} = O_p(n^{-1/4})$;
2. $R_n \rightarrow 0.5\chi_0^2 + 0.5\chi_1^2$ in distribution.

The proof proceeds in the following two steps. In the first step, we show that $(\hat{\alpha}, \hat{\beta})$ are consistent for $(0, 0)$ under the null hypothesis. In the second step, a fourth-order Taylor's expansion is used to approximate the LRT statistic, from which the limiting distribution of the LRT is derived.

Step 1. Since the semiparametric marginal profile log-likelihood function $l_M(\alpha, \beta)$ is symmetric with respect to y_{1i} and y_{2i} , it can be rewritten as a function of (x_{1i}, x_{2i}) . Let

$$r_n(\alpha, \beta) = l_M(\alpha, \beta) - l_M(0, 0) = \sum_{i=1}^n \log\{h(\alpha, \beta; x_{1i}, x_{2i})\},$$

where

$$h(\alpha, \beta; x_{1i}, x_{2i}) = \frac{2 \exp(\alpha + \beta x_{1i}) + 2 \exp(\alpha + \beta x_{2i})}{\{1 + \exp(\alpha + \beta x_{1i})\}\{1 + \exp(\alpha + \beta x_{2i})\}}.$$

We first show that under the null hypothesis

$$E[\log\{h(\alpha, \beta; x_{1i}, x_{2i})\}] \leq 0, \quad (18.4.9)$$

with the equality holding if and only if $(\alpha, \beta) = (0, 0)$. Then the consistency of $(\hat{\alpha}, \hat{\beta})$ follow from Wald (1948)'s argument.

By Jensen's inequality,

$$E[\log\{h(\alpha, \beta; x_{1i}, x_{2i})\}] \leq \log[E\{h(\alpha, \beta; x_{1i}, x_{2i})\}],$$

where the inequality becomes equality if and only if $h(\alpha, \beta; x_{1i}, x_{2i})$ is a constant. Under the null hypothesis, x_{1i} and x_{2i} are independent and identically distributed. Therefore

$$\begin{aligned} E\{h(\alpha, \beta; x_{1i}, x_{2i})\} &= E\left[\frac{4 \exp(\alpha + \beta x_{1i})}{\{1 + \exp(\alpha + \beta x_{1i})\}\{1 + \exp(\alpha + \beta x_{2i})\}}\right] \\ &= E\left\{\frac{4 \exp(\alpha + \beta x_{1i})}{1 + \exp(\alpha + \beta x_{1i})}\right\} \times E\left\{\frac{1}{1 + \exp(\alpha + \beta x_{2i})}\right\} \\ &= 4E\left\{\frac{\exp(\alpha + \beta x_{1i})}{1 + \exp(\alpha + \beta x_{1i})}\right\} \times \left\{1 - E\left(\frac{\exp(\alpha + \beta x_{1i})}{1 + \exp(\alpha + \beta x_{1i})}\right)\right\} \\ &\leq 1. \end{aligned}$$

We have used the inequality $4x(1-x) \leq 1$ in the third step above. Combining the two inequalities above, then under the null hypothesis,

$$\log[E\{h(\alpha, \beta; x_{1i}, x_{2i})\}] \leq \log(1) = 0,$$

with equality if and only if $h(\alpha, \beta; x_{1i}, x_{2i}) = 1$. Note that

$$\begin{aligned} h(\alpha, \beta; x_{1i}, x_{2i}) = 1 &\iff \frac{\{\exp(\alpha + \beta x_{1i}) - 1\}\{\exp(\alpha + \beta x_{2i}) - 1\}}{\{1 + \exp(\alpha + \beta x_{1i})\}\{1 + \exp(\alpha + \beta x_{2i})\}} = 0 \\ &\iff \exp(\alpha + \beta x_{1i}) = 1 \text{ or } \exp(\alpha + \beta x_{2i}) = 1. \end{aligned}$$

Under the null hypothesis, $h(\alpha, \beta; x_{1i}, x_{2i}) = 1$ if and only if $(\alpha, \beta) = (0, 0)$.

Step 2. Without loss of generality, we assume that under the null hypothesis $E(x_{1i}) = E(x_{2i}) = 0$. Note that $R_n = 2r_n(\hat{\alpha}, \hat{\beta})$. Using a fourth-order Taylor's expansion for $r_n(\hat{\alpha}, \hat{\beta})$ around $(0, 0)$, we obtain

$$\begin{aligned} R_n &= 2r_n(\hat{\alpha}, \hat{\beta}) \\ &= -1/2\hat{\alpha}\hat{\beta} \sum_{i=1}^n (x_{1i} + x_{2i}) - n/2\hat{\alpha}^2 - \hat{\beta}^2/2 \sum_{i=1}^n x_{1i}x_{2i} \\ &\quad + 1/48n\hat{\alpha}^4 + 1/48\hat{\beta}^4 \sum_{i=1}^n (2x_{1i}^3x_{2i} - 3x_{1i}^2x_{2i}^2 + x_{1i}x_{2i}^3) \\ &\quad + 1/24\hat{\alpha}^3\hat{\beta} \sum_{i=1}^n (x_{1i} + x_{2i}) + 1/16\hat{\alpha}^2\hat{\beta}^2 \sum_{i=1}^n (x_{1i}^2 + x_{2i}^2) + 1/24\hat{\alpha}\hat{\beta}^3 \sum_{i=1}^n (x_{1i}^3 + x_{2i}^3) + \epsilon_n, \end{aligned}$$

with the remainder term

$$\epsilon_n = O_p(n)(\hat{\alpha}^5 + \hat{\alpha}^4\hat{\beta} + \hat{\alpha}^3\hat{\beta}^2 + \hat{\alpha}^2\hat{\beta}^3 + \hat{\alpha}\hat{\beta}^4 + \hat{\beta}^5) = o_p(n)(\hat{\alpha}^2 + \hat{\beta}^4).$$

In the above expansion, the first-order and third-order terms do not appear since the first and third derivatives or partial derivatives of $r_n(\alpha, \beta)$ with respect to α and β at $(0, 0)$ are all 0. Absorbing the higher order terms of $\hat{\beta}^4$ or $\hat{\alpha}^2$ into the remainder term, and noting that

$$\begin{aligned} |\hat{\alpha}\hat{\beta} \sum_{i=1}^n (x_{1i} + x_{2i})| &= o_p(n^{1/2}|\hat{\alpha}|) \leq o_p(1) + o_p(n\hat{\alpha}^2), \\ \hat{\beta}^4 \sum_{i=1}^n (2x_{1i}^3x_{2i} + 2x_{1i}x_{2i}^3) &= O_p(n^{1/2}\hat{\beta}^4) = o_p(n\hat{\beta}^4), \\ |\hat{\alpha}\hat{\beta}^3 \sum_{i=1}^n (x_{1i}^3 + x_{2i}^3)| &= o_p(n|\hat{\alpha}|\hat{\beta}^2) \leq o_p(n\hat{\alpha}^2 + n\hat{\beta}^4), \end{aligned}$$

we have

$$\begin{aligned} R_n &= -n/2\hat{\alpha}^2 - \hat{\beta}^2/2 \sum_{i=1}^n x_{1i}x_{2i} - 1/16\hat{\beta}^4 \sum_{i=1}^n x_{1i}^2x_{2i}^2 + o_p(n\hat{\alpha}^2 + n\hat{\beta}^4) + o_p(1) \\ &= -n/2\hat{\alpha}^2\{1 + o_p(1)\} + 2\hat{\beta}^2 \sum_{i=1}^n \frac{x_{1i}x_{2i}}{-4} - \hat{\beta}^4 \sum_{i=1}^n \left(\frac{x_{1i}x_{2i}}{-4}\right)^2 \{1 + o_p(1)\} + o_p(1). \end{aligned}$$

Therefore

$$\hat{\alpha} = o_p(n^{-1/2}), \quad \hat{\beta}^2 = \frac{\left\{\sum_{i=1}^n \frac{x_{1i}x_{2i}}{-4}\right\}^+}{\sqrt{\sum_{i=1}^n \left(\frac{x_{1i}x_{2i}}{-4}\right)^2}} + o_p(n^{-1/2}),$$

and

$$R_n = \frac{\left[\left\{\sum_{i=1}^n \frac{x_{1i}x_{2i}}{-4}\right\}^+\right]^2}{\sum_{i=1}^n \left(\frac{x_{1i}x_{2i}}{-4}\right)^2} + o_p(1).$$

Hence $\hat{\beta} = O_p(n^{-1/4})$ and the limiting distribution of R_n is $0.5\chi_0^2 + 0.5\chi_1^2$.

Goodness-of-Fit for the Exponential Tilting Model

Methods and theories developed in the last section rely on the exponential tilting model assumption. Next we consider a procedure to test this assumption.

Define a bivariate empirical distribution as

$$\hat{H}(s, t) = n^{-1} \sum_{i=1}^n \{I(y_{1i} \leq s, y_{2i} \leq t) + I(y_{2i} \leq s, y_{1i} \leq t)\}.$$

By symmetry, it can be written as

$$\hat{H}(s, t) = n^{-1} \sum_{i=1}^n \{I(x_{1i} \leq s, x_{2i} \leq t) + I(x_{2i} \leq s, x_{1i} \leq t)\}.$$

Therefore $E\{\hat{H}(s, t)\} = F(s)G(t) + F(t)G(s)$. Under the density ratio model, the semiparametric estimators of G and F are given by

$$\hat{G}(t) = n^{-1} \sum_{i=1}^n \sum_{j=1}^2 \frac{I(y_{ji} \leq t)}{1 + \exp(\alpha + \beta y_{ji})}, \quad \hat{F}(t) = n^{-1} \sum_{i=1}^n \sum_{j=1}^2 \frac{\exp(\hat{\alpha} + \hat{\beta} y_{ji}) I(y_{ji} \leq t)}{1 + \exp(\alpha + \beta y_{ji})}.$$

Naturally the discrepancy between the estimators can be used as a test statistic

$$\Delta_n = \sup_{-\infty < s, t < \infty} \sqrt{n} |\hat{H}(s, t) - \hat{F}(s)\hat{G}(t) - \hat{F}(t)\hat{G}(s)|.$$

The p -value can be derived by the following permutation method.

When the sample size is not large, the approximation of the limiting distribution by the finite sample distribution of the LRT may not be good. The following permutation method is then suggested to approximate the p -value of the LRT.

1. For each b from 1 to B

- (a) Generate a permutation sample as follows.
 - i. Let $y_{11}^{(b)} = \min\{y_{ji}, i = 1, \dots, n, j = 1, 2\}$ and y_{21}^b be a random observation from the remaining $2n - 1$ observations (not including $y_{11}^{(b)}$), each with the same probability $1/(2n - 1)$ of being sampled.
 - ii. Let $y_{12}^{(b)}$ be the minimum of the remaining $2n - 2$ observations (not including $y_{11}^{(b)}$ and y_{21}^b) and $y_{22}^{(b)}$ be a random observation from the remaining $2n - 3$ observations (not including $y_{11}^{(b)}$, y_{21}^b and $y_{12}^{(b)}$).
 - iii. Repeat the above process until we get $(y_{1n}^{(b)}, y_{2n}^{(b)})$.
- (b) For the permutation sample $\{(y_{1i}^{(b)}, y_{2i}^{(b)}), i = 1, \dots, n\}$, we calculate a LRT statistic $R_n^{(b)}$.

2. Approximate the p -value of R_n by

$$\frac{\#\{b : R_n^{(b)} \geq R_n\}}{B}.$$

A Real Data Example

This example considers the normalized measurements of C-band area on the number 9 chromosome pairs from a family of size three. There are 40, 18 and 31 unordered pairs of normalized measurements for the father, mother and offspring, respectively. The data are from Table 1 of Lauder (1977). For the purpose of illustration, we analyzed the 40 unordered pairs for the father, which is tabulated in the Table 18.1.

We first conduct a goodness-of-fit test for the exponential tilting model assumption on the original data and on the log-transformed data. The test statistics are $\Delta_n = 0.809$ and $\Delta_n = 0.798$, respectively. The bootstrap method with $B = 10000$ results in p -values of 0.08 for the original data and 0.1 for the log-transformed data. Therefore the exponential tilting model assumption is more reasonable for the log-transformed data. Figure 18.1 presents the plot of $\hat{H}(s, t)$ and $\hat{F}(s)\hat{G}(t) + \hat{F}(t)\hat{G}(s)$. These two estimates are very close to each other. Hence, we work on the log-transformed data.

We now apply the ELRT to test the null hypothesis that the unordered pairs for the fathers come from the same distribution. The ELRT for the transformed data is found to be 4.82. Calibrated by the limiting distribution, the p -value is 0.014. The bootstrap method with $B = 10000$ gives the p -value 0.025. Under the normality assumption for the log-transformed measurements, the PLRT statistic is found to

Table 18.1 Normalized measurements of C-band area (the units are $\mu\text{m}^2 \times 10^{-2}$) in number 9 chromosome pair for the father

Cell	Larger	Smaller
1	84.1	62.1
2	99.5	79.9
3	94.8	68.3
4	100.2	77.2
5	93.7	55.3
6	77.3	60.5
7	82.4	76.2
8	80.9	78.8
9	88.8	74.7
10	84.2	81.0
11	73.5	64.0
12	76.9	60.9
13	84.6	66.8
14	97.8	71.3
15	78.1	78.1
16	80.4	65.7
17	60.2	49.5
18	103.8	73.1
19	82.3	68.9
20	105.6	72.3
21	83.8	48.6
22	87.5	77.1
23	73.9	60.0
24	101.1	73.9
25	74.1	64.2
26	95.5	83.2
27	81.7	63.7
28	104.3	68.8
29	81.1	76.4
30	89.7	74.6
31	89.7	63.1
32	101.7	79.9
33	83.1	65.2
34	96.4	66.5
35	83.0	66.4
36	67.1	59.3
37	86.2	62.3
38	85.5	63.6
39	94.1	71.7
40	86.7	64.6

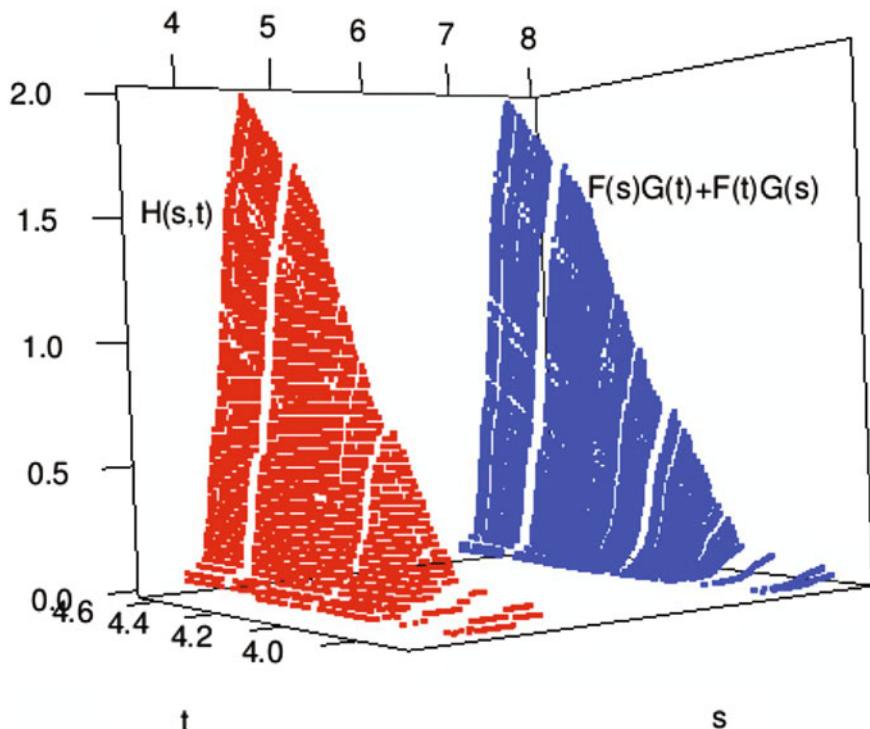


Fig. 18.1 Empirical distribution and semiparametric distribution functions based on unordered chromosome pairs

be 5.20. Calibrated by the limiting distribution, the p -value is 0.011. The bootstrap method with $B = 10000$ gives a p -value 0.023. Therefore the ELRT and PLRT based on the normal model almost have the same power to reject the null hypothesis. We have strong evidence against that the pairs come from the same distribution.

Chapter 19

Causal Inference and Missing Data Problems

Well established physical theories are developed under rigorous mathematical reasoning and tightly controlled laboratory experiment tests. In randomized clinical trials for the comparison between treatment and control, patients are randomly assigned to one of the two groups. As a consequence, baseline profiles are balanced between the two arms, i.e., they have the same distribution. In observational medical studies or epidemiological researches, however, insights from biology and intuition may suggest possible treatment effects while the underlying experiments may not have a rigorous design, which lead to unbalanced baseline patient characteristics between groups. Similarly in evidence based economic studies, there is no control over intervention programs. As a result participation may not be completely at random. Some individuals may be more likely to participate than others. The fundamental problem of causal inference is that we can only observe one of the two potential outcomes for a particular subject. It is impossible to conduct a paired t-test for the assessment of treatment effects. On the other hand, the unpaired two sample t-test or Wilcoxon test may produce biased results for treatment effects since they fail to adjust for baseline covariates.

Missing data are ubiquitous in social and medical studies. Interestingly enough, causal inference problems may be considered as special cases of missing data, where the baseline measurements are available for all individuals, while the treatment outcomes are available only for those assigned to the treatment group. The treatment outcomes for those assigned to controls are missing. On the contrary, the control outcomes are missing for all those assigned to the treatment group. In causal inference the main interest is the marginal average treatment effect or median treatment effect, while in missing data problems, in addition to the overall mean, any parameter defined by estimating equations would be of interest.

Over the last few decades, missing data problems and causal inferences have become two of the most active areas in statistical researches. The main applications, among others, include survey sampling, social and medical science and economic studies. Since the seminal work by Rosenbaum and Rubin (1983) on the adjustment of

unbalanced baseline covariates, the propensity score matching method and inverse probability weighted method have become very popular to estimate the marginal mean effect and to compare the marginal mean treatment effect in observational studies. Propensity score is the conditional probability of assigning an individual to treatment given his/her covariates.

In studying the relationship between a binary response variable and a continuous explanatory variable, the most popular models include logistic, probit, complementary log-log models. By fitting a parametric propensity score function, information in a vector of covariates can be transferred to a scale propensity score. Two individuals with similar propensity scores can be considered comparable in their baseline characteristics. On the other hand, the inverse probability weighted method originated from Horvitz and Thompson (1952), in which each observed complete outcome is inversely weighted by its propensity score, is widely used. Many variations of this method have been proposed in the literature, e.g., the design and model unbiased method discussed by Rao et al. (1990) in survey sampling and the augmented inverse probability weighted method proposed by Robins et al. (1994) in general superpopulations. Many excellent books about missing data and causal inference can be found, among other, for example Little and Rubin (2002), Kim and Shao (2013) and Imbens and Rubin (2015).

19.1 Definition of Three Types of Missing Data and Basic Concepts

In statistical literature missing data can be classified into one of three categories.

(1) Missing data do not depend on any observable or unobservable quantities. This is equivalent to randomly flipping a coin to decide whether data are missing or not. The probability of heads or tails is constant over all individuals. This type of data is called missing completely at random (MCAR). The complete data only inference method is valid, though it may not be the most efficient one.

(2) Missing data only depend on observable quantities. This is equivalent to randomly flipping a coin to decide whether data are missing or not, however the probability of heads or tails depends on observable quantities and may vary among individuals. This type of data is called missing at random (MAR). The complete data only inference method is not valid in general. It may produce biased results.

(3) When neither MCAR nor MAR hold, we say the data are missing not at random, abbreviated MNAR, or non-ignorably missing. This is equivalent to randomly flipping a coin to decide whether data are missing or not, however the probability of heads or tails depends on the underlying variable of interest (observable or unobservable) and may vary among individuals. Again the complete data only inference method is not valid. It produces biased results.

In this chapter we focus on missing at random data. Chapter 22 will discuss missing not at random data problems. Missing not at random data problems are far more difficult to deal with than missing at random data problems.

In a randomized experiment each individual is assigned to either treatment or control randomly, independent of the individual's characteristics. Therefore the baseline covariate X in the treatment and control arms should have the same distribution. Let $D = 1$ or 0 be the indicator of treatment or control, respectively, then $X|D = 1 \sim X|D = 0$ or $X \perp D$. As a result, the direct comparison between the treatment and control arms is meaningful and would produce unbiased results. In non-randomized experiments, however, direct comparisons between the treated and controls may be misleading because each individual, whether exposed to treatment or control, can differ systematically. In order to make the two groups comparable, Rosenbaum and Rubin (1983) introduced the concept of balancing score. The balancing score, $b(X)$, is a function of the observed covariates X such that the conditional distribution of X given $b(X)$ is the same for treated ($D = 1$) and untreated ($D = 0$), i.e.,

$$X \perp D|b(X).$$

In other words X and D are independent each other given the balancing score. A trivial balancing score is $b(x) = x$. However we can see later that there exist many balancing scores.

Denote Y_{0i} as the potential or latent outcome of individual i who is not treated and define Y_{1i} similarly if this individual is treated. The treatment effect for i is $Y_{1i} - Y_{0i}$. Unfortunately for each individual i , only one of Y_{1i} or Y_{0i} is observed. Denote the observed outcome for the i -th individual as

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}, \quad i = 1, 2, \dots, n,$$

where the treatment indicator D_i and baseline covariates X_i are available for everyone. The propensity score

$$\pi(x) = P(D = 1|x)$$

is the probability function of choosing the treatment for a given covariate X .

For simplicity we use Y_0 and Y_1 to denote generic outcomes for control and treatment, respectively. In a randomized experiment, $\pi(x) = \pi$ is independent of X and is known. This implies the treatment outcome Y_1 and control outcome Y_0 and covariate X are independent of D , or

$$Y_0, Y_1, X \perp D.$$

In observational studies, this assumption is unlikely to be true. Instead, two basic weaker assumptions are imposed in causal inference.

Assumption 1 In an un-confounded observational study, the basic assumption is that the treatment assignment D and outcomes (Y_0, Y_1) are conditionally independent given X , i.e.,

$$Y_0, Y_1 \perp D | X. \quad (19.1.1)$$

In this case X contains all variables that affect both treatment assignment and outcomes. If two individuals, $i \neq j$, have the same covariate $X_i = X_j$, but the first individual is assigned treatment ($D_i = 1$), and the second one is assigned control ($D_j = 0$), then the outcomes (Y_{0i}, Y_{1i}) (where Y_{0i} is not observable) and outcomes (Y_{0j}, Y_{1j}) (where Y_{1j} is not observable) should have the same distribution. Therefore the outcomes for the treated or untreated individuals are comparable as long as they have the same baseline covariates. In practical applications, unfortunately X is usually a high dimensional covariate vector. It would be very difficult to find two individuals who have exactly the same covariates.

Assumption 2 The i -th individual outcome does not affect the j -th individual outcome if $j \neq i$. This implies that individual results are independent of each other.

Since it is impossible to estimate the joint distribution of Y_0 and Y_1 . The best we can hope for is to recover the marginal distributions of Y_0 and Y_1 or functionals of them. In other words, only parameters defined by the marginal distributions are estimable. In particular, we will consider

- (1) Average treatment effect (ATE), defined as $\Delta_{ATE} = E(Y_1 - Y_0)$.
- (2) Average treatment effect on the treated, defined as

$$\Delta_{ATET} = E(Y_1 - Y_0 | D = 1).$$

Note that the treatment outcome Y_1 is available if $D = 1$, however, the control outcome Y_0 is not. Even though Y_0 is available from $D = 0$ group, we cannot directly use it to estimate Δ_{ATET} since $Y_0|D = 1$ and $Y_0|D = 0$ have different distributions. To assess the average treatment effect on treated, we have to find a way to evaluate $E(Y_0|D = 1)$ by borrowing information from the $Y_0|D = 0$ group.

A conventional method to analyze the relationship between an outcome variable Y , explanatory variables X and treatment variable D is the linear regression model

$$Y_i = \alpha + X_i\beta + D_i\gamma + \epsilon_i.$$

Since (D, X) and ϵ may be correlated,

$$E[X_i(Y_i - \alpha - X_i\beta - D_i\gamma_i)] \neq 0, \quad E[D_i(Y_i - \alpha - X_i\beta - D_i\gamma_i)] \neq 0.$$

Furthermore γ needs not to be the same for every individual in the population. Both of which may lead to biased inferences. This concludes that the simple regression method is undesirable.

Denote the observed data as

$$(Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}, D_i, X_i), i = 1, 2, \dots, n,$$

where again Y_{1i} is available if $D_i = 1$, and Y_{0i} is available if $D_i = 0$. Note that

$$Y_1, X|D = 1 \sim \frac{P(D = 1|Y_1, X)f_1(y_1|x)g(x)}{P(D = 1)} = \frac{\pi(x)f_1(y_1|x)g(x)}{P(D = 1)}$$

and

$$Y_0, X|D = 0 \sim \frac{P(D = 0|Y_0, X)f_0(y_0|x)g(x)}{P(D = 0)} = \frac{\{1 - \pi(x)\}f_0(y_0|x)g(x)}{P(D = 0)},$$

where $f_1(y_1|x)$ and $f_0(y_0|x)$ are, respectively, the conditional densities of treatment outcome Y_1 and control outcome Y_0 given the covariate X , and $g(x)$ is the marginal density of X .

Let $f_j(y)$ and $h_j(x|y)$, $j = 0, 1$ be the marginal and conditional densities for the control and treatment groups, respectively. Then the joint density of (Y_j, X) has two different decompositions

$$f_j(y_j|x)g(x) = f_j(y_j)h_j(x|y_j), \quad j = 0, 1.$$

The marginal density of Y_1 given $D = 1$ is

$$f_1(y_1|D = 1) = \frac{\int \pi(x)f_1(y_1|x)g(x)dx}{\int \int \pi(x)f_1(y_1|x)g(x)dx dy_1} = \frac{w_1(y_1)f_1(y_1)}{\int w_1(y_1)f_1(y_1)dy_1},$$

where $w_1(y_1) = \int \pi(x)h_1(x|y_1)dx$. Note that $f_1(y_2|D = 1)$ is a biased version of $f_1(y_1)$ if $\pi(x)$ is not a constant. Similarly

$$f_0(y_0|D = 0) = \frac{\int \{1 - \pi(x)\}f_0(y_0|x)g(x)dx}{\int \int \{1 - \pi(x)\}f_0(y_0|x)g(x)dx dy_0} = \frac{w_0(y_0)f_0(y_0)}{\int w_0(y_0)f_0(y_0)dy_0},$$

where

$$w_0(y_0) = \int \{1 - \pi(x)\}h_0(x|y_0)dx.$$

The equality of $f_1(y|D = 1) = f_0(y|D = 0)$ does not imply $f_1(y) = f_0(y)$ unless $\pi(x) = \pi$ is independent of x .

Moreover the conditional marginal densities

$$g(x|D = 1) = \frac{\pi(x)g(x)}{P(D = 1)}, \quad g(x|D = 0) = \frac{\{1 - \pi(x)\}g(x)}{P(D = 0)}$$

are different. The conditional density of Y_j given X and $D = j$ ($j = 0, 1$) is

$$f_j(y_j|x, D = j) = \frac{f_j(y_j, D = j|x)}{P(D = j|x)} = \frac{f_j(y_j|x)P(D = j|x, y_j)}{P(D = j|x)} = f_j(y_j|x)$$

since $P(D = j|x, y_j) = P(D = j|x)$ is independent of y_j .

In summary, even though marginally $f_j(y_j) \neq f_j(y_j|D = j)$, $j = 0, 1$, as long as conditioning on covariate x , the density for potential treatment outcome Y_1 is the same as the density for the observed treatment outcome group ($D = 1$), i.e., $f_1(y_1|x) = f_1(y_1|x, D = 1)$. The same is true for the control group.

19.2 Some Existing Methods in Casual Inferences

In this section we present some existing methods in causal inferences.

1. Parametric or nonparametric regression approaches

Since $F_i(y|x) = F_i(y|x, D = i)$, $i = 0, 1$, the key idea of both parametric and nonparametric methods is to estimate $F_1(y|x)$ and $F_0(y|x)$ based on data

$$[Y_{1i}|X_{1i}, D_i = 1], \quad [Y_{0i}|X_{0i}, D_i = 0],$$

respectively. Then $F_1(y)$ and $F_0(y)$ can be estimated through

$$n^{-1} \sum_{i=1}^n \hat{F}_1(y|x_i), \quad n^{-1} \sum_{i=1}^n \hat{F}_0(y|x_i),$$

where $\hat{F}_i(y|x)$, $i = 0, 1$ is either a parametric or nonparametric estimate of the conditional distribution. In this case, all baseline covariates are used.

If parametric models are postulated for $F_1(y|x) = F_1(y|x, \beta_1)$ and $F_0(y|x) = F_0(y|x, \beta_0)$, respectively, then it is straightforward to estimate β_1 and β_0 by using the maximum likelihood method. On the other hand, if regression models are given by $E(Y_1|x) = \mu_1(x\beta_1)$ and $E(Y_0|x) = \mu_0(x\beta_0)$, then either the least squares method or generalized estimating equation method can be used to estimate β . The average treatment effect can be estimated by

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n \{\mu_1(x_i\hat{\beta}_1) - \mu_0(x_i\hat{\beta}_0)\},$$

where the unbiasedness can be observed by applying iterated expectations $E(Y_i) = E[E(Y_i|X)] = E[E(Y_i|X, D = i)]$, $i = 0, 1$. In fact this is the most efficient estimator. For this estimator to be unbiased, the model assumption on $F_i(y|x, \beta_i)$, $i = 0, 1$ or $\mu_i(x\beta_i)$, $i = 0, 1$ must be correct. Cheng (1994) relaxed this assumption by using a kernel method to estimate $\mu_i(x) = E(Y_i|x)$. Then the average treatment effect can be estimated by

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)],$$

where $\hat{\mu}_1$ and $\hat{\mu}_0$ are kernel based estimators. Cheng's (1994) method can be shown to be the most efficient under no modelling assumptions. However his method may not work well if the dimension of X is high.

2. Propensity score matching method

Both parametric and nonparametric approaches for assessing the treatment effect may have their own weakness if the dimension of covariate X is high, which is this case most of times in applications. The former suffers from model misspecification, while the latter suffers from curse of dimensionality. To circumvent these problems, Rosenbaum and Rubin (1983) introduced the concept of propensity score.

Using the basic **Assumption 1**, we have shown that for given X , Y_1 is independent of D , i.e., $f_1(y_1|x, D = 1) = f_1(y_1|x)$. This property may not be transferable to other functions of X , say, $\psi(X)$. In other words, in general

$$f_1(y_1|\psi(x), D = 1) \neq f_1(y_1|\psi(x)).$$

Rosenbaum and Rubin (1983) proved two properties.

- (1) Balancing score property, i.e., $f_1(y_1|D = 1, \pi(x)) = f_1(y_1|\pi(x))$.
- (2) Sufficiency of the propensity score for un-confoundedness.

In the following we only consider Y_1 , the outcome for treated. A similar approach can be applied to Y_0 , the outcome for untreated.

Note that

$$E[I(D = 1)|\pi(x)] = E[E\{I(D = 1|x)\}|\pi(x)] = E[\pi(x)|\pi(x)] = \pi(x),$$

$$f_1(y_1|D = 1, \pi(x)) = \frac{P(D = 1|y_1, \pi(x))f_1(y_1|\pi(x))}{P(D = 1|\pi(x))}.$$

$$\begin{aligned} P\{D = 1|y_1, \pi(x)\} &= E\{P(D = 1|y_1, x, \pi(x))|y_1, \pi(x)\} = E[P(D = 1|x)|y_1, \pi(x)] \\ &= E[\pi(x)|y_1, \pi(x)] = \pi(x). \end{aligned}$$

Therefore

$$f_1(y_1|D = 1, \pi(x)) = \frac{\pi(x)f_1(y_1|\pi(x))}{\pi(x)} = f_1(y_1|\pi(x)).$$

This shows that Y_1 is independent of D given the propensity score $\pi(x)$. The same argument also applies to Y_0 , i.e., Y_0 is independent of D given $\pi(x)$.

Suppose we can assume a parametric model for $\pi(x)$, say,

$$P(D = 1|x) = \pi(x\beta) = \frac{\exp(x\beta)}{1 + \exp(x\beta)},$$

where β is an unknown parameter. We can maximize the binomial likelihood

$$L = \prod_{i=1}^n \{\pi(x_i\beta)\}^{d_i} \{1 - \pi(x_i\beta)\}^{1-d_i}$$

with respect to β . Denote the MLE as $\hat{\beta}$.

We can partition the propensity score interval $[0, 1]$ into

$$0 = a_0 < a_1 < \dots < a_{K-1} < a_K = 1,$$

where a_0, \dots, a_K are pre-specified real numbers. Using iterated expectation,

$$E(Y_0) = E[E(Y_0|\pi(X))] = E[E\{Y_0|\pi(x), D = 0\}],$$

we can define the matched mean estimators in k -interval as, respectively

$$\mu_{k1} = \frac{\sum_{i=1}^n d_i I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k) y_i}{\sum_{i=1}^n d_i I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k)}, \quad (19.2.2)$$

$$\mu_{k0} = \frac{\sum_{i=1}^n (1 - d_i) I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k) y_i}{\sum_{i=1}^n (1 - d_i) I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k)}. \quad (19.2.3)$$

The average treatment effect can be estimated by

$$ATE = \sum_{k=1}^K \{n^{-1} \sum_{i=1}^n I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k)\}(\mu_{k1} - \mu_{k0}). \quad (19.2.4)$$

Similarly the average treatment effect for treated can be estimated by

$$ATET = \frac{\sum_{i=1}^n d_i y_i}{\sum_{i=1}^n d_i} - \frac{\sum_{k=1}^K \{n^{-1} \sum_{i=1}^n I(a_{k-1} < \pi(x_i\hat{\beta}) \leq a_k)\} \mu_{k0}}{\sum_{i=1}^n d_i / n}. \quad (19.2.5)$$

Even though propensity score matching is one of the most popular methods in causal inference, it may not be the most efficient one. In the case where the bias is of sufficiently low order to be dominated by the variance, the matching estimators are not efficient given a fixed number of matches. To achieve efficient estimation, the number of matches needs to increase with the sample size. Suppose K is the total number of matches and n is the sample size, then similar to the sieve estimation

method, fully efficient estimation is achievable if $K \rightarrow \infty$ and $K/n \rightarrow 0$. In fact, a matching estimator is essentially a regression estimator, with the imputed missing potential outcomes in place of conditional expectations. However, the efficiency gain of such estimators is somewhat artificial. The optimal cut off points, the optimal number of matches and how to use data-dependent ways of choosing this number, remain open problems. For more details on propensity score matching method we refer readers to Abadie and Imbens (2011).

3. Inverse probability weighted and augmented inverse probability weighted methods

Due to the biased sampling feature,

$$Y_1, X|D = 1 \sim \frac{\pi(x\beta)dF_1(x, y)}{\int \int \pi(x\beta)dF_1(x, y)}.$$

One may use the Horvitz and Thompson inverse probability weighted method to estimate $\mu_1 = E(Y_1)$

$$\hat{\mu}_{HT} = \frac{1}{n} \sum_{i=1}^n \frac{Y_{1i} D_i}{\pi(x_i \beta)}. \quad (19.2.6)$$

A related estimator is called Hajek's estimator

$$\hat{\mu}_{HA} = \frac{\sum_{i=1}^n D_i Y_{1i} / \pi(x_i \beta)}{\sum_{i=1}^n D_i / \pi(x_i \beta)}. \quad (19.2.7)$$

This estimator can be thought as a nonparametric MLE in a biased sampling setup. In fact given $D_i = 1$, the nonparametric MLE of $F_1(x, y)$ is

$$\hat{F}_1(x, y) = \frac{\sum_{i=1}^n D_i I(X_i \leq x, Y_{1i} \leq y) / \pi(x_i \beta)}{\sum_{i=1}^n D_i / \pi(x_i \beta)}.$$

As a result, the mean of F_1 can be estimated by $\int y d\hat{F}_1(x, y)$, where β can be replaced by the maximum likelihood estimator based on (D_i, x_i) , $i = 1, 2, \dots, n$ and the propensity score model $\pi(x\beta)$.

Note that the X_i 's corresponding to $D_i = 0$ are not used in the inverse probability weighted estimator. This may lead to a loss of efficiency. Moreover this estimator is very unstable, especially in the case that $\pi(x)$ is close to 0. In a series of papers, Robins and his colleagues developed the augmented inverse probability weighted estimate by adding an extra term with mean 0 (if the propensity score is correctly specified) in the Horvitz and Thompson's estimate (Robins et al. 1994).

$$\hat{\mu}_{AI} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{D_i Y_{1i}}{\pi(x_i)} - \frac{D_i - \pi(x_i)}{\pi(x_i)} h(x_i) \right\}, \quad (19.2.8)$$

where $h(x)$ is a user specified function. Scharfstein et al. (1999) pointed out that this estimator has the so called doubly robust property, i.e., it is consistent when either the propensity score is correctly specified or the regression function $h(x) = E(Y_1|x)$ is correctly specified. However in general it is not a consistent estimator if both models are wrong.

In fact if $\pi(x)$ is correctly specified, then the first term has mean $E(Y_1)$ and the second term has mean 0. On the other hand, $\hat{\mu}_{AI}$ can be written as

$$\hat{\mu}_{AI} = \frac{1}{n} \sum_{i=1}^n \frac{D_i \{Y_{1i} - h(x_i)\}}{\pi(x_i)} + \frac{1}{n} \sum_{i=1}^n h(x_i).$$

Note that the conditional independence between Y_{1i} and D_i conditioning on x_i , the first term has mean 0 and the second term converges to $E(h(X)) = E(Y_1)$ if $h(x) = E(Y_1|x)$ is correctly specified.

This type of estimator was also discussed in the survey sampling literature by Rao et al. (1990). They called it design unbiased and model unbiased estimator. Even though the doubly robust estimator has very attractive asymptotic properties, Kang and Schafer (2007) demonstrated in some situations when both the propensity score and the “working regression model” $h(x)$ are slightly misspecified, the finite sample performance of the doubly robust estimator can perform poorly. This observation can be rationalized as follows:

(1) When the propensity score $\pi(x)$ is too close to 0, as long as $D = 1$ for one observation, then both Horvitz and Thompson estimator and augmented inverse probability weighted estimator become very unstable. Even though this is an event with a very small probability, but in simulation studies, this will happen eventually. Consequently, the mean square error of the population mean estimator is large. The erratic behaviors of these estimators are more likely to occur in larger sample sizes and larger simulation repetitions.

(2) It is observed that the inverse weighted and augmented inverse weighted estimates with parametrically fitted propensity score are better than their counterparts with the true propensity score, respectively. This is a counterintuitive phenomenon (Henmi and Eguchi 2004). We will formally demonstrate this in Sect. 19.5 by using a projection method. Intuitively, if the propensity score $\pi(x_i)$ is too close to 0 and $D_i = 1$ for some observation i , then the parametrically fitted propensity score $\pi(x_i \hat{\beta})$, in general, will be larger than the true $\pi(x_i)$. This will make the denominator stay away from 0. In Sect. 26.4 of Chap. 26 we will further show the robustness of inverse weighted and augmented inverse weighted estimates is improved greatly if the underlying propensity score is estimated nonparametrically or semiparametrically by using a single index model (Qin et al. 2017a).

(3) In applications, it is very unlikely to have a completely correct “working regression model” $h(x)$ (otherwise the sample mean $n^{-1} \sum_{i=1}^n h(x_i)$ is the best estimate, see Sect. 6.2), the augmented estimator by simply subtracting a term with mean zero from the inverse probability weighted estimator may not be optimal. It is desirable to consider a class of estimates

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{D_i Y_{1i}}{\pi(x_i)} - c \frac{D_i - \pi(x_i)}{\pi(x_i)} h(x_i) \right\},$$

where c is a constant. We may choose the optimal c such that the resulting estimator has the minimum variance.

Exercise Find the optimal c .

Next we will discuss the parameter estimation problem in the “working regression model” $E(Y_1|x) = h(x, \gamma)$, where the form of h is known but γ is an unknown parameter. To estimate γ we may use the method of least squares or weighted least squares. More specifically we minimize

$$n^{-1} \sum_{i=1}^n D_i \{y_{1i} - h(x_i, \gamma)\}^2,$$

or

$$n^{-1} \sum_{i=1}^n \frac{D_i}{\pi_i} \{y_{1i} - h(x_i, \gamma)\}^2,$$

with respect to γ . Unfortunately, neither method leads to the optimal solution for the augmented inverse probability weighted estimation of the mean.

A simple and effective method to estimate γ in the “working regression model” $E(Y_1|x) = h(x, \gamma)$ was discussed by Rubin and van der Laan (2008) and Cao et al. (2009) by using a new method of weighted least squares. For any fixed γ , we may find the asymptotic variance of the augmented estimator

$$\mu_A = n^{-1} \sum_{i=1}^n \frac{D_i y_{1i}}{\pi(x_i)} - \frac{D_i - \pi(x_i)}{\pi(x_i)} h(x_i, \gamma).$$

For the moment, we assume the propensity score is completely known. We can show that the asymptotic variance of μ_A is

$$\sigma^2(\gamma) = E \left[\frac{1 - \pi(X)}{\pi(X)} \{Y_1 - h(X, \gamma)\}^2 \right] + \text{Var}(Y_1).$$

Now we seek the optimal γ such that it has the minimum value. Since the second term is independent of γ , we need to minimize the first term with respect to γ . Essentially we need to fit a weighted least squares

$$\min_{\gamma} \sum_{i=1}^n D_i \frac{1 - \pi(x_i)}{\pi^2(x_i)} \{y_{1i} - h(x_i, \gamma)\}^2,$$

where $\{1 - \pi(x_i)\}/\pi^2(x_i)$ ’s are the optimal weights. Denote the solution as $\hat{\gamma}$.

Cao et al. (2009) proposed to minimize

$$\sum_{i=1}^n D_i \frac{1 - \pi(x_i)}{\pi^2(x_i)} \{y_{1i} - h(x_i, \gamma) - c^T \pi_\beta(x_i)/(1 - \pi(x_i))\}^2$$

with respect to γ and c , where $\pi_\beta(x) = \partial\pi(x, \beta)/\partial\beta$. The main motivation is the propensity score function may also carry information on the conditional mean of Y_1 , i.e., $E(Y_1|x)$. Finally we may replace the estimated $\hat{\gamma}$ in the augmented inverse weighted estimator μ_A .

When the parameter β in the propensity score is unknown, we can replace it by a maximum likelihood estimator based on the propensity score model. In general, the asymptotic variance of $\mu_A(\hat{\beta}, \gamma)$ is complicated. Fortunately the optimal estimator of $\mu_A(\hat{\beta}, \hat{\gamma})$ is almost the same as $\mu_A(\beta_0, \hat{\gamma})$ in numerical studies. Simulation results show that this method in general has a good finite sample performance even the variation of $\hat{\beta}$ is not taken into considered in the deriving the optimal estimating equation for γ .

Variance Estimation

Direct estimation of variance by using the asymptotic variance formula is not desirable since it involves terms such as $d_i/\pi^2(x_i)$. When $\pi(x_i)$ is small and $d_i = 1$, it is very unstable. In general using a bootstrap method for the variance estimation is preferable.

4. Dimensional reduction methods in causal inference

The propensity score encompasses multiple covariates of each individual to a single number summarizing their joint association with treatment conditions. Hansen (2008) discussed the concept of prognostic scores, as summarizes of covariates' association with potential outcomes. In the following we only discuss prognostic score for treatment outcome Y_1 only.

If $\psi(X)$ is sufficient for the potential treatment outcome Y_1 , in the sense that $Y_1 \perp X | \psi(X)$, we call $\psi(X)$ a **prognostic score**. Clearly the propensity score is a special case of the prognostic score. The prognostic score is superior to the propensity score if there is more information about Y_1 versus X than D versus X . More importantly, as the prognostic score is a summary of covariate information about Y_1 , it should lead to more efficient estimators than using a propensity score. However, finding prognostic scores is in general a challenging task since it requires full knowledge about the conditional distribution of Y_1 given X .

A few examples.

(1) Consider a location shift model

$$Y_1|x \sim f_1(y|x) = f_1(y - \mu(x)),$$

then $\mu(x)$ is a prognostic score.

(2) $Y_1|X$ follows a generalized linear model (McCullagh and Nelder 1989), then the linear predictor of Y_1 is a prognostic score.

(3) Y_1 given X is normally distributed, as a sufficient statistics, the prognostic score needs to contain both the conditional mean $E(Y_1|X)$ and the conditional variance $\text{Var}(Y_1|X)$.

(4) The propensity score $\pi(x) = P(D = 1|x)$ can be thought as a prognostic score for D since

$$D \perp X | \pi(x).$$

Hu et al. (2012) discussed the concept of double balancing score for the robust and efficient consideration. Note that

$$Y_1 \perp D | \pi(X), \psi(X)$$

if either the propensity score $\pi(X)$ is correctly specified or the prognostic score is correctly specified. Using iterated expectation,

$$E[Y_1] = E[E(Y_1|X)].$$

Cheng (1994) proposed estimating $\mu(x) = E(Y|x)$ (denoted as $\hat{\mu}(x)$) by using a kernel method and then $E(Y)$ can be estimated by $n^{-1} \sum_{i=1}^n \hat{\mu}(x_i)$. This method works well if X is low dimensional. On the other hand if X is high dimensional, the kernel smooth method may suffer from the well known curse of dimensionality problem. Hu et al. (2012) applied iterated expectation to the propensity score and balance score,

$$E(Y_1) = E[E\{Y_1|\pi(X), \psi(X)\}].$$

If we can carefully fit parametric models for $\pi(x) = P(D = 1|x\beta)$ and $E(Y_1|x) = \psi(x\gamma)$, even though the postulated models may not be correct perfectly, we can achieve dimensional reduction.

Next we need to estimate $\mu^*(\pi(x), \psi(x)) = E\{Y_1|\pi(x), \psi(x)\}$. Denote the estimator as $\hat{\mu}^*(\pi(x), \psi(x))$. Then $E(Y_1)$ can be estimated by $n^{-1} \sum_{i=1}^n \hat{\mu}^*(\pi(x_i), \psi(x_i))$. There are two ways to estimate μ^* .

- (1) Parametric approach: Y_1 is regressed on $\pi(x)$ and $\psi(x)$ by a suitable model.
- (2) Nonparametric kernel method: Note that the kernel method for estimating μ^* only involves two-dimensional covariates $\pi(x)$ and $\psi(x)$. In practical applications, $\pi(x)$ and $\psi(x)$ can be fitted using parametric models. As long as one of them is correctly specified, Hu et al. (2012) method can achieve consistency. The nice feature of their method is that both the efficient property in the parametric approach and the robust property in the nonparametric approach are combined. Moreover, their method achieves the semiparametric lower bound if both $\pi(x)$ and $\psi(x)$ are correctly specified.

5. Empirical likelihood methods

Motivated by survey sampling calibration methods, Qin and Zhang (1997) proposed an empirical likelihood calibration estimation. They started from a biased sampling likelihood with complete data only

$$L_B = \prod_{i=1}^{n_1} \frac{\pi(x_i|\beta)dF_1(x_i, y_{1i})}{\pi_1}, \quad \pi_1 = P(D=1) = \int \pi(x)dF_1(x),$$

where without loss of generality it is assumed the first n_1 observations are complete data. If β is unknown, then it can be replaced by the maximum likelihood estimator based on the parametric propensity score assumption and (D_i, X_i) , $i = 1, 2, \dots, n$. Since all X_i , $i = 1, 2, \dots, n$ are available, $dF_1(x, y)$ can be calibrated in the biased likelihood L_B such that its marginal moments $\int \int \phi(x)dF_1(x, y)$ are equal to their empirical moments based all available x_i 's, i.e., $\bar{\phi} = n^{-1} \sum_{i=1}^n \phi(x_i)$. For simplicity, denote $p_i = dF_1(x_i, y_{1i})$, $i = 1, 2, \dots, n_1$. We need to maximize the log biased sampling likelihood

$$L_B = \sum_{i=1}^{n_1} \log p_i$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^{n_1} p_i \{\phi(x_i) - \bar{\phi}\} = 0,$$

where $\phi(x) = (\phi_1(x), \phi_2(x))$, $\phi_1(x) = \pi(x)$. Note that the constraint on $\pi(x)$ is necessary since it reflects the biased sampling feature. The second constraint on $\phi_2(x)$ can be an arbitrary function of x and is optional. The purpose of using it is to increase efficiency. We will show that the best choice is the conditional mean of Y_1 given X , i.e., $\mu(x) = E(Y_1|x)$.

After profiling out p_i 's we have

$$p_i = \frac{1}{n_1} \frac{1}{1 + \lambda_1 \{\pi(x_i) - \bar{\pi}_1\} + \lambda_2 \{\phi(x_i) - \bar{\phi}\}}, \quad i = 1, 2, \dots, n_1,$$

where λ_1 and λ_2 are Lagrange multipliers determined by the equations

$$\sum_{i=1}^{n_1} \frac{\pi(x_i) - \bar{\pi}_1}{1 + \lambda_1 \{\pi(x_i) - \bar{\pi}_1\} + \lambda_2 \{\phi(x_i) - \bar{\phi}\}} = 0,$$

$$\sum_{i=1}^{n_1} \frac{\phi(x_i) - \bar{\phi}}{1 + \lambda_1 \{\pi(x_i) - \bar{\pi}_1\} + \lambda_2 \{\phi(x_i) - \bar{\phi}\}} = 0.$$

Due to biased sampling, clearly the limiting values of λ_1 and λ_2 are, $1/\pi_1$ and 0, respectively. The calibration empirical likelihood estimator is $\hat{\mu}_1 = \sum_{i=1}^{n_1} \hat{p}_i y_{1i}$.

If one component of $\phi_2(x)$ is $\mu_1(x) = E(Y_1|x)$, then

$$\begin{aligned} E\left[\sum_{i=1}^{n_1} \hat{p}_i y_{1i} | x_1, \dots, x_n\right] &= \sum_{i=1}^{n_1} \hat{p}_i \mu_1(x_i) = \bar{\mu}_1 \\ &= \frac{1}{n} \sum_{i=1}^n \mu_1(x_i), \end{aligned}$$

by the calibration constraint. Clearly this leads to an unbiased estimator of μ_1 .

Similarly if $\mu_1(x) = \sum_{i=1}^I a_i c_i(x)$, then the corresponding choice of $\phi_2(x) = (c_1(x), \dots, c_I(x))$ achieves consistency.

By extending Qin and Zhang (2007)'s results, Han and Wang (2013) imposed additional multiple constraints for the choices of $\pi(x)$, say, $\pi_1(x), \dots, \pi_k(x)$. Using the same argument, as long as one of them is correct, then the empirical likelihood method will produce consistent estimator (exercise). Theoretically we can use empirical likelihood to add as many constraints as possible for the efficient consideration, however, numerically this may be a challenging problem. It becomes more difficult to find solutions in the constrained maximization step when the number of constraints increases.

In statistical literature there exists different versions of empirical likelihood methods. In particular, the empirical likelihood can be constructed based on the joint distributions of $dF_1(y_{1i}, x_i)$, $i = 1, 2, \dots, n_1$ and $dF_0(y_{0i}, x_i)$, $i = 1, 2, \dots, n_0$. The likelihood is

$$L = \prod_{i=1}^n \pi^{d_i}(x_i, \beta) \{1 - \pi(x_i, \beta)\}^{1-d_i} \prod_{i=1}^n [dF_1(y_{1i}, x_i)]^{d_i} [dF_0(y_{0i}, x_i)]^{1-d_i}.$$

Without loss of generality, assume $d_i = 1$, $i = 1, 2, \dots, n_1$ and $d_i = 0$, $i = n_1 + 1, \dots, n$. Let $p_i = dF_1(x_i, y_i)$, $i = 1, 2, \dots, n_1$ and $q_j = dF_0(x_{n_1+j}, y_{0j})$, $j = 1, 2, \dots, n_0$. Note that the marginal distributions satisfy $F_1(\infty, x) = F_0(\infty, x)$. Then we need to maximize

$$\sum_{i=1}^{n_1} \log p_i + \sum_{j=1}^{n_0} \log q_j$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad \sum_{j=1}^{n_0} q_j = 1, \quad q_j \geq 0, \quad \sum_{i=1}^{n_1} p_i \pi(x_i) = \sum_{j=1}^{n_0} q_j \pi(x_{n_1+j}),$$

To improve efficiency, extra constraints can be imposed

$$\sum_{i=1}^{n_1} p_i \phi_k(x_i) = \sum_{j=1}^{n_0} q_j \phi_k(x_{n_1+j})$$

for “guesses” of the conditional means of $E(Y_1|x)$ and $E(Y_0|x)$ (Tan 2006). If n_0 or n_1 is small, however, it may be difficult to find a solution.

6. Imputation methods

To focus the main idea, we only discuss imputation methods for the estimation of $\mu_1 = E(Y_1)$.

Model based imputation and multiple imputation are popular methods to handle missing data. A comprehensive discussion of this topic can be found in Rubin (1987). The main idea of imputation is to replace each missing value with a set of plausible values. Then standard complete data methods can be applied to the imputed data. To find variance estimates, both within-imputation variability and between-imputation variability must be taken into account. Even though there exist model free multiple imputation methods, such as, hot deck imputation, nearest neighbour imputation etc., the most efficient multiple imputation methods are model based, i.e., a parametric model assumption is needed for the imputed data conditioning on the observed ones. If the model assumption is incorrect, however, the model based imputation would lead to biases. This motivates us to explore efficient and robust imputation or multiple imputation methods, where the model assumption does not need to be completely correct.

First we review a nonparametric imputation method discussed by Lipsitz et al. (1998). Denote the observed data as

$$(D_1, Y_1, X_1), \dots, (D_n, Y_n, X_n),$$

where D is an indicator function, equals to 1 if Y is observed and 0 otherwise. The covariate variable X is always observable.

As usual we assume a logistic regression model for the probability of observing Y as

$$P(D = 1|Y = y, X = x) = P(D = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \pi(x, \beta).$$

Based on the binomial log-likelihood, β can be estimated by maximizing

$$\ell_B = \sum_{i=1}^n d_i \log \pi(x_i, \beta) + (1 - d_i) \log\{1 - \pi(x_i, \beta)\}.$$

Denote the maximizer as $\hat{\beta}$.

Let $\theta = P(D = 1)$ be the proportion of complete observations. By observing

$$\begin{aligned} f(y, x|D = 0) &= \frac{\{1 - \pi(x)\}f(y, x)}{1 - \theta} \\ &= \frac{1 - \pi(x)}{\pi(x)} \frac{\theta}{1 - \theta} \frac{\pi(x)f(y, x)}{\theta} \\ &= \frac{1 - \pi(x)}{\pi(x)} \frac{\theta}{1 - \theta} f(y, x|D = 1), \end{aligned}$$

we find this is a biased sampling problem with weight function $\{\pi^{-1}(x) - 1\}$. Lipsitz et al. (1998) estimated $F(y, x|D = 0)$ by

$$\left[\sum_{i=1}^n \{\pi(x_i)^{-1} - 1\}^{-1} d_i I(y_i \leq y, x_i \leq x) \right] \left[\sum_{i=1}^n \{\pi(x_i)^{-1} - 1\}^{-1} d_i \right]^{-1}, i = 1, 2, \dots, n.$$

It is natural to impute the missing Y by sampling from the estimated $\hat{F}(y, x|D = 0)$. The imputation estimator is

$$\hat{\mu}_{IM} = \frac{1}{N} \sum_{i=1}^N \{D_i Y_i + (1 - D_i) \hat{Y}^i\},$$

where \hat{Y}^i is sampled from $\hat{F}(y, x|D = 0)$.

Even though this is a valid approach, there is a drawback since the covariate information X_i is not used in the imputation stage. A more efficient imputation is to sample Y^i from $F(y|x, D = 0) = F(y|x)$. If a parametric model is assumed for $F(y|x)$, we can resample from this estimated parametric conditional distribution. But this method may suffer from model misspecification bias. As an alternative, $F(y|x)$ may be estimated by using a nonparametric kernel method (Cheng 1994). Moreover Aerts et al. (2002) imputed the missing Y_i , $D_i = 0$ by resampling from the nonparametric estimated conditional distribution of Y given X_i . The local multiple imputation method works well only if the covariate is low dimensional.

7. A dimensional reduction semiparametric regression imputation method

In practical applications, high dimensional covariates appear frequently. It is desirable to reduce the dimension of $F(t|D = 0, x)$ first, and then impute the missing Y from $F(t|D = 0, S(x))$ where $S(x)$ is a judiciously chosen low dimensional function of x .

Suppose we are interested in estimating the mean μ of Y , we can use a “working mean model”

$$E(Y|x) = \mu(x, \gamma).$$

The simple imputation

$$\hat{Y}_i = D_i Y_i + (1 - D_i) \hat{E}(Y_i | S_i)$$

does not work unless the index function $S_i = E(Y_i | x_i)$.

As a dimensional reduction method, Hu et al. (2012) estimated the conditional distribution function $F(y | \pi(x), \mu(x))$ by

$$\begin{aligned} \hat{F}(t | \pi(x), \mu(x)) &= \left[\sum_{i=1}^n I(y_i \leq t) K((\pi_i - \pi)/h_n, (\mu_i - \mu)/h_n) \right] \\ &\quad \times \left[\sum_{i=1}^n I(y_i \leq t) K((\pi_i - \pi)/h_n, (\mu_i - \mu)/h_n) \right]^{-1}, \end{aligned}$$

where $K(\cdot, \cdot)$ is a bivariate kernel function, h_n is the window size, $\pi_i = \pi(x_i \hat{\beta})$, $\pi = \pi(x)$, $\mu_i = \mu(x_i \hat{\gamma})$, $\mu = \mu(x)$, and $\pi(x \hat{\beta})$ and $\mu(x \hat{\gamma})$ are fitted “working” propensity score and regression function, respectively. Consequently, an imputation estimator can be constructed, as follows:

1. Draw $Y_i^{(j)}$ from $\hat{F}(t | \pi(x_i \hat{\beta}), \mu(x_i \hat{\gamma}))$. Denote the imputed data as

$$\hat{Y}_i(j) = D_i Y_i + (1 - D_i) Y_i^{(j)}.$$

2. Formulate the imputed estimator

$$\hat{\mu}_{IM}(j) = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i(j), \quad j = 1, 2, \dots, J.$$

3. The multiple-imputation estimation of μ is

$$\tilde{\mu} = \frac{1}{J} \sum_{j=1}^J \hat{\mu}_{IM}(j).$$

Let π_i^0 be the true missing probability for the i -th individual. We can show the doubly robust property, as follows.

(1) If $\pi(x)$ is correctly specified, then D and Y are conditionally independent for given $\pi_i^0(x)$,

$$\begin{aligned} E[(1 - D_i) Y_i^{(j)}] &= E[E\{(1 - D_i) Y_i | \pi_i^0(x), \mu(x)\}] \\ &= E[(1 - \pi_i^0) E(Y_i | \pi_i^0, \mu(x))] \\ &= E[(1 - \pi_i^0) Y_i]. \end{aligned}$$

(2) If $\mu(x)$ is correctly specified,

$$\begin{aligned} E[(1 - D_i)Y_i^{(j)}] &= E[E\{(1 - D_i)E\{Y_i^{(j)}|\pi(x), \mu^0(x)\}\}] \\ &= E[P(D_i = 0|\pi(x_i), \mu^0(x_i))E\{Y_i^{(j)}|\pi(x_i), \mu^0(x_i), D_i = 0\}] \\ &= E[P\{D_i = 0|\pi(x_i), \mu^0(x_i)\}\mu^0(x_i)] \\ &= E[I(D_i = 0)\mu^0(x_i)] = E[(1 - \pi_i^0)Y_i]. \end{aligned}$$

Therefore we have shown the proposed imputation method has the doubly robust property.

8. Empirical likelihood based imputation method

Qin et al. (2008) proposed the following empirical likelihood imputation method.

Again we assume the first n_1 observations have no missing values. We pretend the last $n - n_1$ x_i 's are also not available. Based on the “working data”

$$(D_1 = 1, Y_1, X_1), \dots, (D_{n_1} = 1, Y_{n_1}, X_{n_1}), (D_{n_1+1} = 0, ?, ?), \dots, (D_n = 0, ?, ?),$$

we have likelihood

$$L = \prod_{i=1}^{n_1} [\pi(x_i, \beta) f(y_i, x_i)] [1 - \theta]^{n_0},$$

where $\theta = P(D = 1)$ is the overall complete data probability. The log-likelihood is

$$\ell = \sum_{i=1}^{n_1} \log[\pi(x_i, \beta)] + \sum_{i=1}^{n_1} \log p_i + n_0 \log(1 - \theta)$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^{n_1} p_i \{\pi(x_i, \beta) - \theta\} = 0$$

and

$$\sum_{i=1}^{n_1} p_i \{\eta(x_i, \gamma) - \bar{\eta}\} = 0, \quad \bar{\eta} = \frac{1}{n} \sum_{i=1}^n \eta(x_i, \beta),$$

where η is a known function of x , it may depend on the unknown parameter γ . Note that β can be estimated by the logistic log-likelihood

$$\ell = \sum_{i=1}^{n_1} (\beta_0 + \beta_1 x_i) - \sum_{i=1}^n \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}.$$

Profiling the p_i 's, we have

$$p_i = \frac{1}{n_1} \frac{1}{1 + \lambda_1\{\pi(x_i) - \theta\} + \lambda_2\{\eta_i - \bar{\eta}\}}, \quad i = 1, 2, \dots, n_1.$$

The profile log-likelihood is

$$\ell = - \sum_{i=1}^{n_1} \log[1 + \lambda_1\{\pi(x_i) - \theta\} + \lambda_2\{\eta_i - \bar{\eta}\}] + n_0 \log(1 - \theta).$$

Differentiating ℓ with respect to θ , we have

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n_1} \frac{\lambda_1}{1 + \lambda_1\{\pi(x_i) - \theta\} + \lambda_2\{\eta_i - \bar{\eta}\}} - \frac{n_0}{1 - \theta} = 0,$$

or

$$n_1 \lambda_1 = \frac{n_0}{1 - \theta}, \quad \lambda_1 = \frac{n_0}{n_1} \frac{1}{1 - \theta}.$$

The constraint equations become

$$\sum_{i=1}^{n_1} \frac{\eta_i - \bar{\eta}}{1 + \lambda_1\{\pi(x_i) - \theta\} + \lambda_2\{\eta_i - \bar{\eta}\}} = 0,$$

and

$$\sum_{i=1}^{n_1} \frac{\pi_i - \theta}{1 + \lambda_1\{\pi(x_i) - \theta\} + \lambda_2\{\eta_i - \bar{\eta}\}} = 0.$$

Therefore the conditional distribution $F_0(y, x)$ of (Y, X) given $D = 0$ can be estimated by

$$\hat{q}_i = \frac{\{1 - \pi(x_i)\}\hat{p}_i}{1 - \hat{\theta}}, \quad i = 1, 2, \dots, n_1, \quad \hat{\theta} = \sum_{i=1}^{n_1} \hat{p}_i \pi(x_i, \hat{\beta}).$$

Now we can construct the following estimator:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n D_i Y_i + (1 - \hat{\theta}) \sum_{i=1}^{n_1} \hat{q}_i Y_i = \frac{1}{n} \sum_{i=1}^n \left\{ D_i Y_i + (1 - D_i) \sum_{j=1}^{n_1} \hat{q}_j Y_j \right\}.$$

This estimator replaces all missing value Y by the estimated conditional expectation $E(Y|D = 0)$. In this case the observed covariate X_i is not used. This may not be the most efficient method. On the other hand the conventional imputation estimation

$$\frac{1}{n} \sum_{i=1}^n [D_i Y_i + (1 - D_i) \mu_i]$$

may be biased if the regression function $\mu(x_i) \neq E(Y_i|x_i)$. It should be corrected by using $E(Y_i - \mu_i|D_i = 0)$. As a result we use

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n \{D_i Y_i + (1 - D_i)\hat{Y}_i\}, \quad \hat{Y}_i = \mu_i + \sum_{j=1}^{n_1} q_j(Y_j - \mu_j).$$

Clearly if $\mu_i = E(Y_i|x_i)$, i.e., it is a correct guess, then $E[\tilde{\mu}|x] = n^{-1} \sum_{i=1}^n \mu_i$.

A good choice for η_i is

$$\eta_i = (1 - \pi_i)\mu_i,$$

where $\mu_i = \mu(x_i, \hat{\gamma})$, $\hat{\gamma}$ can be found by fitting this “working model” based on complete data only.

Exercise Show the doubly robust property for the imputation estimator $\tilde{\mu}$.

19.3 Inference for Average Treatment Effect for Treated

When the interest is in estimating the average treatment effect for treated, Qin and Zhang (2008) found inferential techniques developed in bias sampling problems play a key role. As usual we assume a logistic propensity score model

$$P(D = 1|x) = \pi(x, \alpha, \beta) = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)}.$$

There are two different conditional distribution assumptions.

(1) The conditional expectation assumption.

$$E(Y_0|x, D = 0) = E(Y_0|x, D = 1).$$

That is conditioning on the covariate information X , the mean of response for controls does not depend on the treatment assignment.

(2) The conditional density assumption

$$f_0(Y_0|x, D = 0) = f_0(Y_0|x, D = 1).$$

Notice that the second assumption is stronger than the first one.

The average treatment effect for treated is defined as

$$\Delta_{ATE} = E[Y_1|D = 1] - E[Y_0|D = 1].$$

When $D = 1$, Y_1 is available but not Y_0 .

Note that

$$Y_0, X|D=0 \sim \frac{\{1 - \pi(x)\}f_0(y_0|x, D=0)g(x)}{1-\theta}, \quad X \sim g(x),$$

and

$$Y_0, X|D=1 \sim \frac{\pi(x)f_0(y_0|x, D=1)g(x)}{\theta}.$$

Also marginally

$$f(x|D=1) = \frac{\pi(x)g(x)}{\theta}, \quad f(x|D=0) = \frac{\{1 - \pi(x)\}g(x)}{1-\theta},$$

$$f(x|D=1) = f(x|D=0) \frac{1-\theta}{\theta} \frac{\pi(x)}{1-\pi(x)} = f(x|D=0) \exp(\alpha^* + \beta x),$$

$$\alpha^* = \alpha + \log\{(1-\theta)/\theta\}.$$

(1) Inference under the first assumption

With Assumption (1),

$$\begin{aligned} E(Y_0|D=1) &= \frac{1}{\theta} \int \pi(x) \left\{ \int y_0 f_0(y_0|x, D=1) dy_0 \right\} g(x) dx \\ &= \frac{1}{\theta} \int \pi(x) \left\{ \int y_0 f_0(y_0|x, D=0) dy_0 \right\} g(x) dx \\ &= \frac{1-\theta}{\theta} \int \frac{\pi(x)}{1-\pi(x)} \int \frac{[1-\pi(x)]}{1-\theta} y_0 f_0(y_0|x, D=0) g(x) dy_0 dx \\ &= \frac{1-\theta}{\theta} E \left\{ \frac{\pi(x)}{1-\pi(x)} Y_0 | D=0 \right\} \\ &= E \{ \exp(\alpha^* + \beta x) Y_0 | D=0 \}. \end{aligned}$$

Let

$$p_i = dF_1(y_{1i}, x_{1i}|D_i=1), \quad i = 1, 2, \dots, n_1;$$

and

$$r_i = dF_0(y_{0i}, x_{0i}|D_i=0) \quad i = n_1 + 1, \dots, n.$$

Based on the observed data, the likelihood can be written as

$$\prod_{i=1}^{n_1} dF_1(y_{1i}, x_i) \prod_{j=n_1+1}^n dF_0(y_{0j}, x_j).$$

Note that

$$E\{\eta(x)|D=1\} = E\{\eta^*(x)|D=0\} = \eta,$$

where

$$\eta^*(x, \theta) = \eta(x)(1-\theta)\theta^{-1}w(x)\{1-w(x)\}^{-1} = \eta(x)\exp(\alpha^* + \beta x).$$

We need to maximize the log-likelihood

$$\ell = \sum_{i=1}^{n_1} \log p_i + \sum_{j=n_1+1}^n \log r_j$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad \sum_{j=n_1+1}^n r_j = 1, \quad r_j \geq 0,$$

and

$$\sum_{i=1}^{n_1} p_i \{\eta(x_i) - \eta\} = 0, \quad \sum_{i=n_1+1}^n r_i \{\eta^*(x_i) - \eta\} = 0.$$

After profiling p_i 's and r_i 's, the log-likelihood is

$$\ell = - \sum_{i=1}^{n_1} \log[1 + \lambda^T(\eta(x_i) - \eta)] - \sum_{i=n_1+1}^n \log[1 + \nu^T\{\eta^*(x_i) - \eta\}].$$

Then $\mu_0 = E(Y_0|D=1)$ and $\mu_1 = E(Y_1|D=1)$ can be estimated by

$$\sum_{i=n_1+1}^n \hat{r}_i \exp(\hat{\alpha}^* + \hat{\beta}x_{0i}) y_{0i}, \quad \sum_{i=1}^{n_1} \hat{p}_i y_{1i},$$

respectively.

A good choice for η is $\eta(x) = E(Y_1|x, D=1) = E(Y_1|x)$. Similarly for estimating $E(Y_0|D=1)$, a good choice of η is $\exp(\alpha^* + x\beta)E(Y_0|x)$. As an alternative method, we may calibrate $\sum_{i=1}^{n_1} p_i \eta(x_i) = n^{-1} \sum_{i=1}^n \eta(x_i)$ and $\sum_{i=n_1+1}^n r_i \eta(x_i) = n^{-1} \sum_{i=1}^n \eta(x_i)$.

(2) Inference under the second assumption

Under the second assumption, $f_0(y_0|x, D=0) = f_0(y_0|x, D=1)$, and we arrive at a density ratio model for (y_0, x) between groups $D=1$ and $D=0$,

$$f(y_0, x|D=0) = \frac{\theta}{1-\theta} \frac{1-\pi(x)}{\pi(x)} f(y_0, x|D=1) = \exp(-\alpha^* - \beta x) f(y_0, x|D=1).$$

Again α^* and β can be estimated by the logistic regression model. For convenience we assume they are known in the derivation below. Otherwise they can be replaced by their estimators.

The log-likelihood is

$$\begin{aligned}\ell &= \sum_{i=1}^{n_1} \log dF_1(y_{1i}, x_{1i} | D = 1) + \sum_{i=1}^{n_0} [\log dF_0(y_{0i}, x_{0i} | D = 1) - \alpha^* - \beta x_{0i}] \\ &= \sum_{i=1}^{n_1} \log p_i + \sum_{i=1}^{n_0} [\log q_i - \alpha^* - \beta x_{0i}]\end{aligned}$$

subject to the constraints

$$\begin{aligned}\sum_{i=1}^{n_1} p_i &= 1, \quad p_i \geq 0, \quad \sum_{i=1}^{n_0} q_i = 1, \quad q_i \geq 0 \\ \sum_{i=1}^{n_0} q_i [\exp(-\alpha^* - \beta x_{0i}) - 1] &= 0,\end{aligned}$$

and

$$\sum_{i=1}^{n_1} p_i \eta(x_{1i}) = \sum_{i=1}^{n_0} q_i \eta(x_{0i}).$$

Therefore Δ can be estimated by using

$$\hat{\Delta} = \sum_{i=1}^{n_1} \hat{p}_i y_{1i} - \sum_{i=1}^{n_0} \hat{q}_i y_{0i}.$$

We can show that the profile log-likelihood is

$$\begin{aligned}\ell &= - \sum_{i=1}^{n_1} \log[1 + \lambda^T \{\eta(x_{1i}) - \xi\}] - \sum_{i=1}^{n_0} \log[1 + \tau_1^T \{\eta(x_{0i}) - \xi\} \\ &\quad + \tau_2 \{\exp(-\alpha^* - \beta x_{0i}) - 1\}] + \sum_{i=1}^{n_0} (-\alpha^* - \beta x_{0i}).\end{aligned}$$

Taking derivative with respect to ξ , we have

$$\frac{\partial \ell}{\partial \xi} = \sum_{i=1}^{n_1} \frac{\lambda}{1 + \lambda^T \{\eta(x_{1i}) - \xi\}} + \sum_{i=1}^{n_0} \frac{\tau_1}{1 + \tau_1^T \{\eta(x_{0i}) - \xi\} + \tau_2 \{\exp(-\alpha^* - \beta x_{0i}) - 1\}} = 0,$$

or

$$n_1\lambda + n_0\tau_1 = 0.$$

Taking derivative with respect to α^* , we have

$$n_0 - \sum_{i=1}^{n_0} \frac{\tau_2 \exp(-\alpha^* - \beta x_{0i})}{1 + \tau_1^T \{\eta(x_{0i}) - \xi\} + \tau_2 \{\exp(-\alpha^* - \beta x_{0i}) - 1\}} = 0,$$

or

$$n_0 - n_0\tau_2 = 0, \quad \tau_2 = 1.$$

$$p_i = \frac{1}{n} \frac{D_i}{\gamma^T \eta^*}, \quad \eta^* = (1, \eta),$$

$$q_i = \frac{1}{n} \frac{1 - D_i}{1 - \gamma^T \eta^*(x_i) + n_0 n^{-1} \{\exp(-\alpha^* - \beta x_i) - 1\}}.$$

The constraint equation becomes

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{D_i}{\gamma^T \eta^*(x_i)} - \frac{1 - D_i}{1 - \gamma^T \eta(x_i) + n_0 n^{-1} \{\exp(-\alpha^* - \beta x_i) - 1\}} \right] \eta^*(x_i) = 0.$$

The limiting value of γ is $\gamma_0 = (n_1/n, 0)$.

Exercise 1 Derive the large sample results.

Exercise 2 Two obvious estimators for the average treatment effect for treated are

$$T_1 = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{0i} \exp(-\hat{\alpha}^* - x_{0i} \hat{\beta}),$$

and

$$T_2 = \frac{\sum_{i=1}^{n_0} y_{0i} \exp(-x_{0i} \hat{\beta})}{\sum_{i=1}^{n_0} \exp(-x_{0i} \hat{\beta})}$$

Show that, in general, T_2 is better than T_1 .

Matched Balance Score

Hainmueller (2011) and Zhao and Percival (2015) considered the maximum entropy likelihood based inference for the average treatment effect for treated.

Since the treatment outcomes $Y_{1i}, i = 1, 2, \dots, n_1$ are available for $D_i = 1, i = 1, 2, \dots, n_1$, $E(Y_1|D = 1)$ can be estimated directly by using $n_1^{-1} \sum_{i=1}^{n_1} y_{1i}$. We only need to consider the estimation problem for $E(Y_0|D = 1)$. The key is to balance covariates between treatment and control groups. Based on control data $(Y_{0i}, X_{0i}), i = 1, 2, \dots, n_0$, the Kullback–Leibler likelihood is given by

$$\max \sum_{i=1}^{n_0} p_i \log p_i$$

subject to constraints

$$\sum_{i=1}^{n_0} p_i = 1, \quad p_i \geq 0,$$

and

$$\sum_{i=1}^{n_0} p_i h(x_{0i}) = \bar{h}$$

where $\bar{h} = n_1^{-1} \sum_{i=1}^{n_1} h(x_{1i})$ is the average based on observed covariates from treatment group. In other words this method tries to balance covariate $h(x)$ between control and treatment groups.

Using the Lagrange multiplier method (Chap. 9), we have the entropy solution

$$p_i = \frac{\exp\{\lambda^T h(x_{0i})\}}{\sum_{j=1}^{n_0} \exp\{\lambda^T h(x_{0j})\}}, \quad i = 1, 2, \dots, n_0,$$

where the Lagrange multiplier λ is determined by the constraint equation

$$\sum_{i=1}^{n_0} \exp\{\lambda^T h(x_{0i})\}[h(x_{0i}) - \bar{h}] = 0.$$

Equivalently,

$$\frac{1}{n} \sum_{i=1}^n (1 - d_i) \exp\{\lambda^T h(x_i)\}[h(x_i) - \bar{h}] = 0.$$

It is seen that as $n_1 \rightarrow \infty$, $\bar{h} \rightarrow h_1 = E[h(X)|D = 1]$.

If the true propensity score is the logistic regression model

$$\pi(x) = \frac{\exp(\beta^T h(x))}{1 + \exp(\beta^T h(x))},$$

we can use the following heuristical argument to show that the matched balance score estimate of the average treatment effect for treated is consistent.

Note that the constraint equation converges to

$$E[(1 - \pi(X)) \exp(\lambda^T h(X))(h(X) - h_1)] = 0,$$

Clearly if $\lambda = \beta$, then the above equation becomes

$$E[\pi(X)(h(X) - h_1)] = 0.$$

Thus we have shown that the solution of the Lagrange multiplier $\hat{\lambda} \rightarrow \beta$. Furthermore

$$\begin{aligned} \sum_{i=1}^{n_0} \hat{p}_i y_{0i} &= \frac{\sum_{i=1}^n (1 - d_i) y_{0i} \exp\{\hat{\lambda}^T h(x_i)\}}{\sum_{j=1}^n (1 - d_j) \exp\{\hat{\lambda}^T h(x_j)\}} \\ &\rightarrow \frac{E[\{1 - \pi(X)\} Y_0 \exp\{\beta^T h(X)\}]}{E[\{1 - \pi(X)\} \exp\{\beta^T h(X)\}]} \\ &= \frac{E\{\pi(X) Y_0\}}{E\{\pi(X)\}} = E(Y_0 | D = 1). \end{aligned}$$

This intuitive argument shows that the matched balance score method indeed can produce consistent estimator. Along this line, we can derive large sample results.

Finally the treatment effect for treated can be estimated by

$$\hat{\Delta}_{ATET} = n_1^{-1} \sum_{i=1}^{n_1} y_{1i} - \sum_{i=1}^{n_0} \hat{p}_i y_{0i}.$$

Exercise 3 Can the matched balance score method produce a consistent estimator for the average treatment effect for treated if the propensity score does not follow a logistic regression model?

Exercise 4 If the interest is in using the Wilcoxon statistic to assess treatment effects, then a simple inverse probability weighted estimator is given by

$$MW_I = \frac{2}{n(n-1)} \sum_{i \neq j} \frac{D_i(1 - D_j) I(Y_{1i} \leq Y_{0j})}{\pi(x_i)\{1 - \pi(x_j)\}}.$$

Extend this to the augmented inverse probability weighted estimator. Moreover study the dimensional reduction version of this estimator. For more details, see Chen et al. (2013).

19.4 Regression Generalizations

Robins et al. (1992) considered a regression model for the responses in treatment and control through

$$Y_i = \beta D_i + h(x_i) + \epsilon_i, \quad E(\epsilon_i | D_i, X_i) = 0,$$

where $\beta = E[Y|D = 1, x] - E[Y|D = 0, x]$ is the difference between treatment and control outcome for a given covariate X , which is a constant, and $h(x) = E(Y|D = 0, x)$ is the mean conditional response for a control. Assume the choice of treatment follows the logistic regression model

$$P(D = 1|x) = \frac{\exp(\alpha_1 + \alpha_2 x)}{1 + \exp(\alpha_1 + \alpha_2 x)} = \pi(x\alpha).$$

Let $\hat{\alpha}$ be the maximum likelihood estimator of α .

If we perform a simple least squares to estimate β , then the knowledge of $E[Y|D = 0, x] = h(x)$ is crucial. Otherwise we have biased results for β . Instead, Robins et al. (1992) studied an E-type estimator, which solves

$$\sum_{i=1}^n \{y_i - \beta D_i - h(x_i \hat{\gamma})\} \{D_i - \pi(x_i \hat{\alpha})\} = 0, \quad (19.4.9)$$

where $\hat{\gamma}$ minimizes

$$\sum_{i=1}^n (1 - D_i)[y_i - h(x_i \gamma)]^2. \quad (19.4.10)$$

A nice feature of this estimator is the doubly robust property.

(1) If $h(x_i \gamma)$ is correctly specified, then

$$E[\{y_i - \beta D_i - h(x_i \gamma)\} \{D_i - \pi(x_i \alpha)\}] = E[\epsilon_i \{D_i - \pi(x_i \alpha)\}] = 0.$$

(2) Denote $h_0(x) = E(Y|D = 0, x)$. On the other hand side, if $\pi(x\alpha)$ is correctly specified, then

$$\begin{aligned} E[\{y_i - \beta D_i - h(x_i \gamma)\} \{D_i - \pi(x_i \alpha)\}] &= E[\{\epsilon_i + h_0(x_i) - h(x_i \gamma)\} \{D_i - \pi(x_i \alpha)\}] \\ &= E[\{h_0(x_i) - h(x_i \gamma)\} \{D_i - \pi(x_i \alpha)\}] = 0 \end{aligned}$$

since $E\{D_i - \pi(x_i \alpha)|x_i\} = 0$.

The optimal choice of $h(x, \gamma)$ is $E[Y|D = 0, x]$. In practice, $E[Y|D = 0, x]$ is unknown, and it would be difficult to make a correct guess of $E[Y|D = 0, x]$. Instead of estimating γ through (19.4.10), alternatively, for fixed γ , we may solve

$$\sum_{i=1}^n \{y_i - \beta D_i - h(x_i, \gamma)\} \{D_i - \pi(x_i \hat{\alpha})\} = 0,$$

for $\hat{\beta}(\gamma)$. Then γ can be determined by minimizing the asymptotic variance of $\hat{\beta}(\gamma)$.

Exercise If $h(x) \neq E(Y|D=0, x)$, then another promising method is to consider the estimating equation

$$\sum_{i=1}^n \{y_i - \beta D_i - \xi h(x_i)\} \{D_i - \pi(x_i \hat{\alpha})\} = 0.$$

Discuss the optimal choice of ξ . Is it possible to make it doubly robust?

Regression Model for Average Treatment for Treated

Instead of considering the overall treatment effect, recently personalized medicine has become popular. Patients constantly ask the question, “Is this therapy going to work for me?” In individualized treatment regime, it is of interest to fit a conditional model for the causal effect for a given covariate. Specifically we can assume a regression model

$$E(Y_1 - Y_0|X = x, D = 1) = \psi(x\beta),$$

where β quantifies the benefits a patient gets between treatment and no treatment. Observe that

$$E\left[\frac{DY_1}{\pi(x)} - \frac{(1-D)Y_0}{1-\pi(x)}|x\right] = E[Y_1 - Y_0|x] = E[Y_1 - Y_0|x, D = 1].$$

Note that the observed response is $Y = DY_1 + (1-D)Y_0$. Therefore $DY = DY_1$, $(1-D)Y = (1-D)Y_0$, and

$$\frac{DY_1}{\pi(x)} - \frac{(1-D)Y_0}{1-\pi(x)} = \frac{DY}{\pi(x)} - \frac{(1-D)Y}{1-\pi(x)} = Y \frac{D - \pi(x)}{\pi(x)\{1 - \pi(x)\}}.$$

Define

$$\rho(x) = \frac{D - \pi(x)}{\pi(x)\{1 - \pi(x)\}}. \quad (19.4.11)$$

Then $E[Y\rho(x)|x] = E[Y_1 - Y_0|x] = \psi(x\beta)$. More generally

$$E[\{Y - h(x)\}\rho(x) - \psi(x\beta)|x] = 0$$

for any function $h(x)$. Since Y_i , D_i , X_i are available for each individual, we can solve

$$\sum_{i=1}^n a(x_i)[\{y_i - h(x_i)\}\rho(x_i) - \psi(x_i\beta)] = 0,$$

for β estimation, where $a(\cdot)$ is a given vector function with the same dimension as β . By the optimal estimating theory discussed in Sect. 6.2, the optimal choice of a is

$$a(x_i) = \frac{\partial \psi(x_i|\beta)}{\partial \beta} \text{Var}^{-1}(g_i|x_i), \quad g_i = \{y_i - h(x_i)\}\rho(x_i) - \psi(x_i|\beta).$$

Exercise Discuss the optimal choice of $h(x)$.

19.5 Projection Methods in Missing Data Problems

So far we have mainly focused on discussions for the mean response estimation. Next we study estimation problems in general regression model setups. For illustration we consider a missing covariate problem.

Let $(y_1, x_1, z_1, d_1), \dots, (y_n, x_n, z_n, d_n)$ be n independent observations of (Y, X, Z, D) , where Y is the response variable, (X, Z) is a vector of random covariates, and D is an indicator variable, which equals 1 if Z is observed and 0 otherwise. We assume $(Y_i, X_i), i = 1, 2, \dots, n$ are always available. Suppose Y and (X, Z) are related by a regression model

$$Y = \mu(X, Z, \beta) + \epsilon,$$

where $\mu(X, Z, \beta)$ is a (possibly nonlinear) link function indexed by an unknown $p \times 1$ vector parameter β and the random error ϵ satisfies $E(\epsilon|X, Z) = 0$ so that $E(Y|X, Z) = \mu(X, Z, \beta)$. The missing data mechanism associated with the missingness of the Z is characterized by the conditional distribution of D given (Y, X, Z) , which is assumed to satisfy

$$P(D = 1|Y, X, Z) = P(D = 1|Y, X) = w(Y, X, \eta),$$

where w is a specific probability distribution function for given η , a $q \times 1$ unknown vector parameter. Typically a logistic regression is used.

Based on the observed data $\{(y_i, x_i, z_i, d_i), i = 1, \dots, n\}$, we are interested in estimating the regression parameter β . Let $U(Y, X, Z, \beta)$ be a set of unbiased estimating functions for β . In the absence of missing data, so that $d_i = 1, i = 1, \dots, n$, β can be estimated by solving $\sum_{i=1}^n U(y_i, x_i, z_i, \beta) = 0$, a common choice for $U(Y, X, Z, \beta)$ is $a(X, Z, \beta)\{Y - \mu(X, Z, \beta)\}$, where $a(X, Z, \beta)$ is a vector of known functions, up to an unknown parameter β .

In general η is unknown. Naturally the maximum binomial likelihood can be used to estimate η . Let $\hat{\eta}$ maximize the binomial likelihood

$$L_B(\eta) = \prod_{i=1}^n \{w(y_i, x_i, \eta)\}^{d_i} \{1 - w(y_i, x_i, \eta)\}^{1-d_i}.$$

Equivalently $\hat{\eta}$ is a solution to the following system of score equations

$$\frac{\partial \log L_B(\eta)}{\partial \eta} = \sum_{i=1}^n g_3(d_i, y_i, x_i, \eta) = 0, \quad g_3(D, Y, X, \eta) = \frac{\{d - w(Y, X, \eta)\}\partial w(Y, X, \eta)/\partial \eta}{w(Y, X, \eta)\{1 - w(Y, X, \eta)\}}. \quad (19.5.12)$$

In the presence of missing data, the commonly used Horvitz–Thompson estimator $\hat{\beta}_{HT}$ of β is the solution to

$$\sum_{i=1}^n g_1(d_i, y_i, x_i, z_i, \beta, \hat{\eta}) = 0, \quad g_1(d_i, y_i, x_i, z_i, \beta, \hat{\eta}) = \frac{d_i U(y_i, x_i, z_i, \beta)}{w(y_i, x_i, \hat{\eta})}. \quad (19.5.13)$$

We note that $(y_i, x_i, d_i = 0)$'s are not used in (19.5.13).

As a remedy, we define

$$g_2(D, Y, X, \beta, \eta) = \frac{D - w(Y, X, \eta)}{w(Y, X, \eta)} h(Y, X, \gamma, \beta), \quad (19.5.14)$$

where $h(Y, X, \gamma, \beta)$ is a specified function which may depend on an additional parameter γ in the regression model $E(Z|y, x) = h(y, x, \gamma)$. Since $E\{D - w(Y, X, \eta)|Y, X\} = 0$, the large-sample inference for β based on $g_1(D, Y, X, Z, \beta, \eta)$ and $g_2(D, Y, X, \beta, \eta)$ produce the same result whether γ or $\hat{\gamma}$ is used in $h(Y, X, \beta, \gamma)$. Thus, in theory we may proceed as if γ were known and simply write $h(Y, X, \beta) \equiv h(Y, X, \beta, \gamma)$. Now since $E\{g_1(D, Y, X, Z, \beta, \eta)\} = 0$ and $E\{g_2(D, Y, X, \beta, \eta)\} = 0$ for any function $h(Y, X, \beta)$, we can also estimate β based on the following system of estimating equations

$$\sum_{i=1}^n g_1(d_i, y_i, x_i, z_i, \beta, \hat{\eta}) = 0, \quad \sum_{i=1}^n g_2(d_i, y_i, x_i, \beta, \hat{\eta}) = 0.$$

The combined estimating functions $g_1(D, Y, X, Z, \beta, \eta)$ and $g_2(D, Y, X, \beta, \eta)$ utilize both the complete and the incomplete data, whereas the single set of estimating function $g_1(D, Y, X, Z, \beta, \eta)$ uses complete data only. Consequently, we anticipate that the estimators of β based on the combined estimating equations to be better than the Horvitz–Thompson estimator $\hat{\beta}_{HT}$.

Since the number of combined estimating equations is $2p$, which is greater than the dimension p of β , the question arises as how to combine them optimally.

To simplify notations, we write $g_1(\beta, \eta) \equiv g_1(D, Y, X, Z, \beta, \eta)$, $g_2(\beta, \eta) \equiv g_2(D, Y, X, \beta, \eta)$, $g_3(\eta) \equiv g_3(D, Y, X, \eta)$, $w(\eta) \equiv w(Y, X, \eta)$, $h(\beta) \equiv h(Y, X, \beta)$, $U(\beta) \equiv U(Y, X, Z, \beta)$, $g_{1i}(\beta, \eta) \equiv g_1(d_i, y_i, x_i, z_i, \beta, \eta)$, $g_{2i}(\beta, \eta) \equiv g_2(d_i, y_i, x_i, \beta, \eta)$, $g_{3i}(\eta) \equiv g_3(d_i, y_i, x_i, \beta, \eta)$, $w_i(\eta) \equiv w(y_i, x_i, \eta)$, $h_i(\beta) \equiv h(y_i, x_i, \beta)$, $U_i(\beta) \equiv U(y_i, x_i, z_i, \beta)$, $w^\eta(\eta) = \partial w(\eta)/\partial \eta$, $w_i^\eta(\eta) = \partial w_i(\eta)/\partial \eta$. Furthermore, define

$$g_{12}(\beta, \eta) = g_1(\beta, \eta) - g_2(\beta, \eta), \quad g_{23}(\beta, \eta) = \{g_2(\beta, \eta), g_3(\eta)\}^T. \quad (19.5.15)$$

The augmented inverse-probability weighted estimator $\hat{\beta}_{\text{AI}}$ is defined by the solution of

$$g_{\text{AI}}(\beta, \hat{\eta}) \equiv \sum_{i=1}^n \{g_{1i}(\beta, \hat{\eta}) - g_{2i}(\beta, \hat{\eta})\} \equiv \sum_{i=1}^n g_{12i}(\beta, \hat{\eta}) = 0. \quad (19.5.16)$$

Exercise Similar to the argument in Sect. 19.2, show that g_{AI} has the doubly robust property.

Analogous to the process of Rao–Blackwellization, the projection method (Small and McLeish 1989) reduces variation effectively, and may be used either to increase the power of a test, the efficiency of a point estimator, or to render an inference function insensitive to the value of a nuisance parameter. In terms of estimation of β , when the probability of missingness $1 - w(\eta)$ is correctly specified, $E_{(\beta^*, \eta)}\{g_{23}(\beta, \eta)\} = 0$ for all values of β^* , not just for the true value β . Thus, the estimating functions $g_{23}(\beta, \eta)$ are E-ancillary estimating functions in terms of parameter β (Sects. 5.5–5.6 in Chap. 5). This fact, together with the projection method of Small and McLeish (1989), motivates the projection of $g_{12}(\beta, \eta)$ onto the linear subspace spanned by the E-ancillary estimating functions $g_{23}(\beta, \eta)$. To this end, let

$$h_i^*(\hat{\beta}_{\text{HT}}, \hat{\eta}) = [h_i(\hat{\beta}), w_i^\eta(\hat{\eta})/\{1 - w(\hat{\eta})\}]^T.$$

Define the projection estimation function

$$g_{\text{P}}(\beta, \eta) = \sum_{i=1}^n \{g_{12i}(\beta, \eta) - \hat{H}_{12,23}g_{23i}(\beta, \eta)\}, \quad (19.5.17)$$

where

$$\hat{H}_{12,23} = \hat{B}_1 \hat{B}_2^{-1}, \quad (19.5.18)$$

$$\begin{aligned} \hat{B}_1 &= \frac{1}{n} \sum_{i=1}^n \frac{d_i(1 - w_i(\hat{\eta}))(U_i(\hat{\beta}_{\text{HT}}) - h_i(\hat{\beta}_{\text{HT}}))(h_i^*(\hat{\beta}_{\text{HT}}, \hat{\eta}))^T}{w_i^2(\hat{\eta})}, \\ \hat{B}_2 &= \frac{1}{n} \sum_{i=1}^n \frac{(d_i - w_i(\hat{\eta}))^2}{w_i(\hat{\eta})^2} h_i^*(\hat{\beta}_{\text{HT}}, \hat{\eta})(h_i^*(\hat{\beta}_{\text{HT}}, \hat{\eta}))^T. \end{aligned}$$

The quantity $\hat{H}_{12,23} = \hat{B}_1 \hat{B}_2^{-1}$ is a consistent estimator of

$$H_{12,23}(\beta, \eta) \equiv \text{Cov}\{g_{12}(\beta, \eta), g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1},$$

which is the coefficient in the population regression of $g_{12}(\beta, \eta)$ on $g_{23}(\beta, \eta)$. Let $\hat{\beta}_P$ be the solution of $g_P(\beta, \hat{\eta}) = 0$. We call $\hat{\beta}_P$ the projection estimator of β .

Theorem 19.1 *The projection estimator $\hat{\beta}_P$ is a consistent estimator of β if either (a) the observing probability function $w(\eta)$ is correctly specified, or (b) the “working regression function” $h(\beta) = E\{U(\beta)|Y, X\}$.*

Proof For part (a), if $w(\eta)$ is correctly specified, then $E\{g_{12}(\beta, \eta)\} = 0$ and $E\{g_{23}(\beta, \eta)\} = 0$. Since $\hat{\eta}$ is a consistent estimator of η , $n^{-1}g_P(\beta, \hat{\eta})$ converges to zero in probability. Thus, $\hat{\beta}_P$ is consistent for β . For part (b), when the functional form of $w(\eta)$ is misspecified, but $h(\beta) = E\{U(\beta)|Y, X\}$, since $E\{U(\beta) - h(\beta)|Y, X\} = 0$, we can show that $\hat{B}_1 \rightarrow 0$ in probability. As a result, $g_P(\beta, \eta)$ is asymptotically equivalent to $g_{12}(\beta, \eta)$ and converges in probability to $E\{g_{12}(\beta, \eta)\} = 0$. Consequently, the estimator $\hat{\beta}_P$ is still consistent for β in this case. The proof is complete.

Next we study the asymptotic properties of the projection estimator $\hat{\beta}_P$, Horvitz–Thompson estimator $\hat{\beta}_{HT}$ and augmented inverse-probability weighted estimator $\hat{\beta}_{AI}$.

Theorem 19.2 *For a correctly specified missing probability function $1 - w(\eta)$ and a given “working regression function” $h(\beta)$, where it is not necessary $h(\beta) = E\{U(\beta)|Y, X\}$, the projection estimator $\hat{\beta}_P$ has smaller asymptotic variance than both the Horvitz–Thompson estimator $\hat{\beta}_{HT}$ and the augmented inverse-probability weighted estimator $\hat{\beta}_{AI}$. However in general, there is no definitive result on the comparison of the asymptotic variances between $\hat{\beta}_{HT}$ and $\hat{\beta}_{AI}$.*

Proof We first study the asymptotic property of $\hat{\beta}_P$. By expanding $\sum_{i=1}^n g_{1i}(\beta, \hat{\eta})$ at η , we have

$$\sum_{i=1}^n g_{1i}(\beta, \hat{\eta}) = \sum_{i=1}^n \frac{d_i U_i(\beta)}{w_i(\eta)} - \frac{d_i U_i(\beta)}{w_i^2(\eta)} w_i^\eta(\eta)(\hat{\eta} - \eta) + o_p(n^{1/2}),$$

where $w_i^\eta = \partial w_i / \partial \eta$. Since

$$\frac{1}{n} \sum_{i=1}^n \frac{d_i U_i(\beta)}{w_i^2(\eta)} w_i^\eta(\eta) \rightarrow E \left\{ \frac{U(\beta)}{w(\eta)} w_i^\eta(\eta) \right\} \equiv E\{g_1(\beta, \eta) g_3^T(\eta)\},$$

we have

$$\sum_{i=1}^n g_{1i}(\beta, \hat{\eta}) = \sum_{i=1}^n g_{1i}(\beta, \eta) - E\{g_1(\beta, \eta) g_3^T(\eta)\} n(\hat{\eta} - \eta) + o_p(n^{1/2}).$$

Similarly, we can show that

$$\sum_{i=1}^n g_{2i}(\beta, \hat{\eta}) = \sum_{i=1}^n g_{2i}(\beta, \eta) - E\{g_2(\beta, \eta) g_3^T(\eta)\} n(\hat{\eta} - \eta) + o_p(n^{1/2}),$$

$$\sum_{i=1}^n g_{3i}(\hat{\eta}) = \sum_{i=1}^n g_{3i}(\eta) - E\{g_3(\eta)g_3^T(\eta)\}n(\hat{\eta} - \eta) + o_p(n^{1/2}).$$

Now expanding $g_P(\beta, \hat{\eta})$ at η yields

$$\begin{aligned} g_P(\beta, \hat{\eta}) &= \sum_{i=1}^n \{g_{12i}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23i}(\beta, \eta)\} \\ &\quad - E[\{g_{12}(\beta, \eta) - H_{12,23}g_{23}(\beta, \eta)\}g_3^T(\eta)]n(\hat{\eta} - \eta) + o_p(n^{1/2}). \end{aligned}$$

Since $H_{12,23}(\beta, \eta)$ is the regression coefficient in the population regression of $g_{12}(\beta, \eta)$ on $g_{23}(\beta, \eta)$, therefore, $E[\{g_{12}(\beta, \eta) - H_{12,23}g_{23}(\beta, \eta)\}g_3^T(\eta)] = 0$. As a result, $g_P(\beta, \hat{\eta}) = \xi(\beta, \eta) + o_p(n^{1/2})$, where

$$\xi(\beta, \eta) = \sum_{i=1}^n \{g_{12i}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23i}(\beta, \eta)\}.$$

Note that $H_{12,23}(\beta, \eta)$ can be written as

$$\begin{aligned} H_{12,23}(\beta, \eta) &= \text{Cov}\{g_1(\beta, \eta), g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1} \\ &\quad - \text{Cov}\{g_2(\beta, \eta), g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1}. \end{aligned}$$

Since $\text{Var}\{g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1} = I_{p+q}$, we have

$$\begin{aligned} [\text{Var}\{g_2(\beta, \eta)\}, \text{Cov}\{g_2, g_3(\eta)\}] &\left(\begin{array}{cc} \text{Var}\{g_2(\beta, \eta)\} & \text{Cov}\{g_2(\beta, \eta), g_3(\eta)\} \\ \text{Cov}\{g_3(\eta), g_2(\beta, \eta)\} & \text{Var}\{g_3(\eta)\} \end{array} \right)^{-1} \\ &= (I_p, 0_{p \times q}). \end{aligned}$$

As a result,

$$\begin{aligned} \text{Cov}\{g_2(\beta, \eta), g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1}g_{23}(\beta, \eta) &= (I_p, 0_{p \times q}) \begin{pmatrix} g_2(\beta, \eta) \\ g_3(\eta) \end{pmatrix} \\ &= g_2(\beta, \eta). \end{aligned}$$

Therefore,

$$\xi(\beta, \eta) = \sum_{i=1}^n \{g_{1i}(\beta, \eta) - H_{1,23}(\beta, \eta)g_{23}(\beta, \eta)\},$$

where $H_{1,23}(\beta, \eta) = \text{Cov}\{g_1(\beta, \eta), g_{23}(\beta, \eta)\}[\text{Var}\{g_{23}(\beta, \eta)\}]^{-1}$ is the regression coefficient in the population regression of $g_1(\beta, \eta)$ on $g_{23}(\beta, \eta)$.

To study the Horvitz–Thompson estimator $\hat{\beta}_{\text{HT}}$, we can write

$$\hat{\eta} - \eta = [\text{Var}\{g_3(\eta)\}]^{-1} \left\{ n^{-1} g_{3i}(\eta) \right\} + o_p(n^{-1/2}),$$

which implies that

$$\sum_{i=1}^n g_{1i}(\beta, \hat{\eta}) = \xi_{\text{HT}}(\beta, \eta) + o_p(n^{1/2}),$$

where

$$\xi_{\text{HT}}(\beta, \eta) = \sum_{i=1}^n \{g_{1i}(\beta, \eta) - H_{1,3}(\beta, \eta)g_{3i}(\eta)\},$$

and $H_{1,3}(\beta, \eta) = \text{Cov}\{g_1(\beta, \eta), g_3(\eta)\}[\text{Var}\{g_3(\eta)\}]^{-1}$ being the regression coefficient in the population regression of $g_1(\beta, \eta)$ on $g_3(\eta)$. Since the residual variance in a population regression decreases as the number of covariates increases, $\xi_{\text{HT}}(\beta, \eta)$ has larger variance than $\xi(\beta, \eta)$.

Similarly, we can study the augmented inverse-probability weighted estimator $\hat{\beta}_{\text{AI}}$. It can be shown after some algebra that

$$\sum_{i=1}^n g_{12i}(\beta, \hat{\eta}) = \xi_{\text{AI}}(\beta, \eta) + o_p(n^{1/2}),$$

where

$$\xi_{\text{AI}}(\beta, \eta) = \sum_{i=1}^n \{g_{12i}(\beta, \eta) - H_{12,3}(\beta, \eta)g_{3i}(\beta, \eta)\},$$

and $H_{12,3}(\beta, \eta) = \text{Cov}\{g_{12}(\beta, \eta), g_3(\eta)\}[\text{Var}\{g_3(\eta)\}]^{-1}$ is the regression coefficient in the population regression of $g_{12}(\beta, \eta)$ on $g_3(\eta)$. Clearly, this residual has larger variation than the residual by projecting $g_1(\beta, \eta)$ onto the space spanned by $g_2(\beta, \eta)$ and $g_3(\eta)$. However, there does not exist a definitive relationship on the variances between the residuals from projecting $g_1(\beta, \eta)$ onto $g_3(\eta)$ and from projecting $g_{12}(\beta, \eta)$ onto $g_3(\eta)$.

When the functional form of $w(\eta)$ is correctly specified, $n^{-1}\partial\xi(\beta, \eta)/\partial\beta^T$, $n^{-1}\partial\xi_{\text{HT}}(\beta, \eta)/\partial\beta^T$, and $n^{-1}\partial\xi_{\text{AI}}(\beta, \eta)/\partial\beta^T$ all converge in probability to the same limit $E\{\partial U(\beta)/\partial\beta^T\}$. As a result, the asymptotic variances of $\hat{\beta}_{\text{P}}$, $\hat{\beta}_{\text{HT}}$ and $\hat{\beta}_{\text{AI}}$ are given, respectively, by

$$E \left\{ \frac{U(\beta)}{\partial\beta^T} \right\} \text{Var}\{g_{12}(\beta, \eta) - H_{12,3}(\beta, \eta)g_{23}(\beta, \eta)\} E \left\{ \frac{U(\beta)}{\partial\beta^T} \right\}^T,$$

$$E \left\{ \frac{U(\beta)}{\partial\beta^T} \right\} \text{Var}\{g_1(\beta, \eta) - H_{1,3}(\beta, \eta)g_3(\beta, \eta)\} E \left\{ \frac{U(\beta)}{\partial\beta^T} \right\}^T,$$

$$E \left\{ \frac{\partial U(\beta)}{\partial \beta^T} \right\} \text{Var}\{g_{12}(\beta, \eta) - H_{12,3}(\beta, \eta)g_3(D, Y, X, \beta, \eta)\} E \left\{ \frac{\partial U(\beta)}{\partial \beta^T} \right\}^T.$$

Therefore, $\hat{\beta}_P$ is consistent and more efficient than the Horvitz–Thompson estimator $\hat{\beta}_{HT}$ and the augmented inverse-probability weighted estimator $\hat{\beta}_{AI}$. The proof is complete.

Exercise Suppose η is known in the observing probability function $w(y, x, \eta) = w(y, x, \eta_0)$. Show that the Horvitz–Thompson estimator $\hat{\beta}_{HT}$ derived from $\sum_{i=1}^n g_1(d_i, y_i, x_i, z_i, \beta, \hat{\eta}) = 0$ has smaller variance than that derived from $\sum_{i=1}^n g_1(d_i, y_i, x_i, z_i, \beta, \eta_0) = 0$. Henmi and Eguchi (2004) called this is “a paradox concerning nuisance parameters and projected estimating equations”.

Since $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ depend on the choice of the “working regression function” $h(\beta)$, the following theorem addresses the issue of the optimal choice of $h(\beta)$.

Theorem 19.3 Suppose $w(\eta)$ is correctly specified. Then $h^0(\beta) = E\{U(\beta)|Y, X\}$ is the optimal choice of $h(\beta)$ for $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ in the sense of having the smallest asymptotic variance. Moreover, $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ are asymptotically equivalent for the optimal choice $h^0(\beta)$.

Proof Let $g_{12}^0(\beta, \eta)$ denote the version of $g_{12}(\beta, \eta)$ with the optimal choice $h^0(\beta)$ in place of $h(\beta)$, namely,

$$g_{12}^0(\beta, \eta) = \frac{DU(\beta)}{w(\eta)} - \frac{D - w(\eta)}{w(\eta)} h^0(\beta).$$

To see the asymptotic equivalence between $\hat{\beta}_P$ and $\hat{\beta}_{AI}$, it can be shown that

$$E\{g_{12}^0(\beta, \eta)g_{23}^T(\beta, \eta)\} = 0,$$

which implies that $H_{12,23}(\beta, \eta) = 0$. As a result, $\hat{\beta}_P$ is asymptotically equivalent to $\hat{\beta}_{AI}$. Note, however, that $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ are not asymptotically equivalent if $h(\beta) \neq h^0(\beta)$.

In order to show that the optimal choice of $h(\beta)$ for $\hat{\beta}_P$ is $h^0(\beta)$, we only need to show that

$$\text{Var}\{g_{12}^0(\beta, \eta)\} \leq \text{Var}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\}$$

for all possible choices of $h(\beta)$. Indeed we have

$$\begin{aligned} \text{Var}\{g_{12}^0(\beta, \eta) - g_{12}(\beta, \eta) + H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\} &= \text{Var}\{g_{12}^0(\beta, \eta)\} \\ &+ \text{Var}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\} - 2\text{Cov}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta), g_{12}^0(\beta, \eta)\}. \end{aligned}$$

Since $E\{g_{12}^0(\beta, \eta)g_{23}^T(\beta, \eta)\} = 0$ as shown before, we have

$$\begin{aligned}\text{Cov}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta), g_{12}^0(\beta, \eta)\} \\ = E[g_1(\beta, \eta)\{g_{12}^0(\beta, \eta)\}^T] \\ = E[g_{12}^0(\beta, \eta)\{g_{12}^0(\beta, \eta)\}^T] \equiv \text{Var}\{g_{12}^0(\beta, \eta)\},\end{aligned}$$

therefore,

$$\begin{aligned}0 &\leq \text{Var}\{g_{12}^0(\beta, \eta) - g_{12}(\beta, \eta) + H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\} \\ &= \text{Var}\{g_{12}^0(\beta, \eta)\} + \text{Var}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\} - 2\text{Var}\{g_{12}^0(\beta, \eta)\},\end{aligned}$$

which implies that

$$\text{Var}\{g_{12}(\beta, \eta) - H_{12,23}(\beta, \eta)g_{23}(\beta, \eta)\} \geq \text{Var}\{g_{12}^0(\beta, \eta)\}.$$

This completes the proof.

Remark 1: Although $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ are asymptotically equivalent for the optimal choice $h(\beta) = h^0(\beta) \equiv E\{U(\beta)|Y, X\}$, it is rare in practice to know the conditional score function $h^0(\beta)$, especially when the dimension of X is high. In general, $\hat{\beta}_P$ and $\hat{\beta}_{AI}$ are not asymptotically equivalent. Similar to the augmented inverse-probability weighted estimator $\hat{\beta}_{AI}$, the projection estimator $\hat{\beta}_P$ is also doubly robust, i.e., $\hat{\beta}_P$ is consistent if either the missing probability function $1 - w(\eta)$ is correctly specified or $h(\beta) = E\{U(\beta)|Y, X\}$. On the other hand, the Horvitz–Thompson estimator $\hat{\beta}_{HT}$ is consistent only if the missing probability function $w(\eta)$ is correctly specified.

Simulation studies for projection methods discussed in this section can be found in Qin et al. (2017b).

19.6 Optimal Estimating Function Based Inferences

In this section we will use Godambe's optimality theory for estimating functions to directly combine different estimating equations. To implement this, we also will make sure the optimal combination estimating equations have the doubly robust property. It is instructive to conduct the following derivations for graduate students.

Recall that β and η are $p \times 1$ and $q \times 1$ unknown vector parameters, respectively. We can construct optimal estimating equations for (β, η) based on $g_1(d, y, x, z, \beta, \eta)$ and $g_{23}(d, y, x, \beta, \eta)$ or, equivalently, on $g_{12}(d, y, x, z, \beta, \eta)$ and $g_{23}(d, y, x, \beta, \eta)$, where g_1, g_{12}, g_{23} are defined in (19.5.12)–(19.5.15).

According to the estimating function theories discussed in Chaps. 5 and 6, the optimal estimating equations for (β, η) are given by

$$m_n(\beta, \eta) = \frac{1}{n} \sum_{i=1}^n A(\beta, \eta) \Sigma^{-1}(\beta, \eta) g(D_i, Y_i, X_i, Z_i, \beta, \eta) = 0, \quad (19.6.19)$$

where

$$\begin{aligned} g(d, y, x, \beta, \eta) &= \begin{pmatrix} g_{12}(d, y, x, z, \beta, \eta) \\ g_{23}(d, y, x, \beta, \eta) \end{pmatrix}, \\ A(\beta, \eta) &= \begin{pmatrix} A_{11}(\beta, \eta) & A_{12}(\beta, \eta) \\ A_{21}(\beta, \eta) & A_{22}(\beta, \eta) \end{pmatrix} = \begin{pmatrix} E \left(\frac{\partial g_{12}(D, Y, X, Z, \beta, \eta)}{\partial \beta^\top} \right)^\top & E \left(\frac{\partial g_{23}(D, Y, X, \beta, \eta)}{\partial \beta^\top} \right)^\top \\ E \left(\frac{\partial g_{12}(D, Y, X, Z, \beta, \eta)}{\partial \eta^\top} \right)^\top & E \left(\frac{\partial g_{23}(D, Y, X, \beta, \eta)}{\partial \eta^\top} \right)^\top \end{pmatrix}, \\ \Sigma(\beta, \eta) &= \begin{pmatrix} \Sigma_{11}(\beta, \eta) & \Sigma_{12}(\beta, \eta) \\ \Sigma_{21}(\beta, \eta) & \Sigma_{22}(\beta, \eta) \end{pmatrix}, \\ \Sigma_{11}(\beta, \eta) &= E\{g_{12}(D, Y, X, Z, \beta, \eta) g_{12}^\top(D, Y, X, Z, \beta, \eta)\}, \\ \Sigma_{12}(\beta, \eta) &= \Sigma_{21}^\top(\beta, \eta) = E\{g_{12}(D, Y, X, Z, \beta, \eta) g_{23}^\top(D, Y, X, \beta, \eta)\}, \\ \Sigma_{22}(\beta, \eta) &= E\{g_{23}(D, Y, X, \beta, \eta) g_{23}^\top(D, Y, X, \beta, \eta)\}. \end{aligned}$$

Write

$$\Sigma^{-1}(\beta, \eta) = \begin{pmatrix} \Sigma^{11}(\beta, \eta) & \Sigma^{12}(\beta, \eta) \\ \Sigma^{21}(\beta, \eta) & \Sigma^{22}(\beta, \eta) \end{pmatrix}.$$

Since $A_{12}^\top(\beta, \eta) = E\{\partial g_{23}(D, Y, X, \beta, \eta)/\partial \beta^\top\} = 0$, it can be shown after some algebra that $m_n(\beta, \eta) = 0$ is equivalent to

$$\begin{aligned} A_{11}(\beta, \eta) \left[\frac{1}{n} \sum_{i=1}^n \{\Sigma^{11}(\beta, \eta) g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \Sigma^{12}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)\} \right] &= 0, \\ A_{21}(\beta, \eta) \left[\frac{1}{n} \sum_{i=1}^n \{\Sigma^{11}(\beta, \eta) g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \Sigma^{12}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)\} \right] \\ + A_{22}(\beta, \eta) \left[\frac{1}{n} \sum_{i=1}^n \{\Sigma^{21}(\beta, \eta) g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \Sigma^{22}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)\} \right] &= 0. \end{aligned}$$

Assume that $A_{11}(\beta, \eta) = E \left(\frac{\partial g_{12}(D, Y, X, Z, \beta, \eta)}{\partial \beta^\top} \right)^\top$ is a $p \times p$ invertible matrix, $m_n(\beta, \eta) = 0$ is equivalent to

$$\begin{aligned} \sum_{i=1}^n [g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \{\Sigma^{11}(\beta, \eta)\}^{-1} \Sigma^{12}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)] &= 0, \\ A_{22}(\beta, \eta) \sum_{i=1}^n [\{\Sigma^{21}(\beta, \eta) g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \Sigma^{22}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)\}] &= 0. \end{aligned}$$

For simplicity let

$$\begin{aligned}\bar{g}_{12} &= \frac{1}{n} \sum_{i=1}^n g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta), \quad \bar{g}_{23} = \frac{1}{n} \sum_{i=1}^n g_{23}(D_i, Y_i, X_i, \beta, \eta), \quad \bar{g}_3 = \frac{1}{n} \sum_{i=1}^n g_3(D_i, Y_i, X_i, \eta), \\ \Sigma_{11} &= \Sigma_{11}(\beta, \eta), \quad \Sigma_{12} = \Sigma_{12}(\beta, \eta), \quad \Sigma_{21} = \Sigma_{21}(\beta, \eta), \quad \Sigma_{22} = \Sigma_{22}(\beta, \eta), \\ \Sigma^{11} &= \Sigma^{11}(\beta, \eta), \quad \Sigma^{12} = \Sigma^{12}(\beta, \eta), \quad \Sigma^{21} = \Sigma^{21}(\beta, \eta), \quad \Sigma^{22} = \Sigma^{22}(\beta, \eta).\end{aligned}$$

It can be shown that $m_n(\beta, \eta) = 0$ is equivalent to

$$\bar{g}_{12} + (\Sigma^{11})^{-1} \Sigma^{12} \bar{g}_{23} = 0, \quad A_{22} \{\Sigma^{22} - \Sigma^{21} (\Sigma^{11})^{-1} \Sigma^{12}\} \bar{g}_{23} = 0.$$

Moreover we will show that this is equivalent to

$$\begin{aligned}\sum_{i=1}^n [g_{12}(D_i, Y_i, X_i, Z_i, \beta, \eta) + \{\Sigma^{11}(\beta, \eta)\}^{-1} \Sigma^{12}(\beta, \eta) g_{23}(D_i, Y_i, X_i, \beta, \eta)] &= 0, \\ \sum_{i=1}^n g_3(D_i, Y_i, X_i, \eta) &= 0.\end{aligned}\tag{19.6.20}$$

We can use the following arguments. Note that

$$A_{22} = A_{22}(\beta, \eta) = E \left(\frac{\partial g_{23}(D, Y, X, \beta, \eta)}{\partial \eta^\tau} \right)^\tau = E \{g_3(D, Y, X, \eta) g_{23}^\tau(D, Y, X, \beta, \eta)\}$$

and

$$\begin{aligned}\Sigma_{22} &= \Sigma_{22}(\beta, \eta) = E \{g_{23}(D, Y, X, \beta, \eta) g_{23}^\tau(D, Y, X, \beta, \eta)\} \\ &= \begin{pmatrix} E\{g_2(D, Y, X, \beta, \eta) g_{23}^\tau(D, Y, X, \beta, \eta)\} \\ E\{g_3(D, Y, X, \eta) g_{23}^\tau(D, Y, X, \beta, \eta)\} \end{pmatrix} = \begin{pmatrix} E\{g_2(D, Y, X, \beta, \eta) g_{23}^\tau(D, Y, X, \beta, \eta)\} \\ A_{22} \end{pmatrix}.\end{aligned}$$

Since

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix} = I,$$

we have

$$\Sigma_{21} \Sigma^{11} + \Sigma_{22} \Sigma^{21} = 0, \quad \Sigma_{21} = -\Sigma_{22} \Sigma^{21} (\Sigma^{11})^{-1}, \quad \Sigma_{21} \Sigma^{12} + \Sigma_{22} \Sigma^{22} = I_{(p+q) \times (p+q)}.$$

Therefore,

$$\Sigma_{22} \{\Sigma^{22} - \Sigma^{21} (\Sigma^{11})^{-1} \Sigma^{12}\} = \Sigma_{22} \Sigma^{22} + \Sigma_{21} \Sigma^{12} = I_{(p+q) \times (p+q)}.$$

As a result,

$$A_{22} \{\Sigma^{22} - \Sigma^{21} (\Sigma^{11})^{-1} \Sigma^{12}\} = (0_{q \times p}, I_{q \times q})$$

so that

$$A_{22}\{\Sigma^{22} - \Sigma^{21}(\Sigma^{11})^{-1}\Sigma^{12}\}\bar{g}_{23} = (0_{q \times p}, I_{q \times q})\bar{g}_{23} = \bar{g}_3.$$

The proof is complete.

This shows that even in the joint optimal estimating equations for (β, η) , asymptotically η can be separately estimated by the marginal binomial likelihood. As a consequence, the computation is simplified.

Next we establish the double robustness of the above estimating equations.

For notational simplicity, write

$$h^*(y, x, \beta, \eta) = \begin{pmatrix} h(y, x, \beta) \\ \frac{\partial w(y, x, \eta)}{\partial \eta} \end{pmatrix}, \quad g_{23}(d, y, x, \beta, \eta) = \frac{d - w(y, x, \eta)}{w(y, x, \eta)} h^*(y, x, \beta, \eta).$$

For $i = 1, \dots, n$, let

$$\begin{aligned} u_i &= u(Y_i, X_i, Z_i, \hat{\beta}_{\text{HT}}, \hat{\eta}_{\text{ML}}), & w_i &= w(Y_i, X_i, \hat{\eta}_{\text{ML}}), \\ h_i &= h(Y_i, X_i, \hat{\beta}_{\text{HT}}), & h_i^* &= h^*(Y_i, X_i, \hat{\beta}_{\text{HT}}, \hat{\eta}_{\text{ML}}), \\ g_{12i} &= g_{12}(D_i, Y_i, X_i, Z_i, \hat{\beta}_{\text{HT}}, \hat{\eta}_{\text{ML}}), & g_{23i} &= g_{23}(D_i, Y_i, X_i, \hat{\beta}_{\text{HT}}, \hat{\eta}_{\text{ML}}), \end{aligned}$$

where $\hat{\eta}_{\text{ML}}$ is the MLE for η . Furthermore, let

$$\begin{aligned} \hat{\Sigma}_{11} &= \frac{1}{n} \sum_{i=1}^n g_{12i} g_{12i}^\top, & \hat{\Sigma}_{12} = \hat{\Sigma}_{21}^\top &= \frac{1}{n} \sum_{i=1}^n \frac{D_i(1-w_i)(u_i-h_i)(h_i^*)^\top}{w_i^2}, & \hat{\Sigma}_{22} &= \frac{1}{n} \sum_{i=1}^n g_{23i} g_{23i}^\top, \\ \hat{\Sigma} &= \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{pmatrix}, & \hat{\Sigma}^{-1} &= \begin{pmatrix} \hat{\Sigma}^{11} & \hat{\Sigma}^{12} \\ \hat{\Sigma}^{21} & \hat{\Sigma}^{22} \end{pmatrix}. \end{aligned}$$

When $w(y, x, \eta)$ is correctly specified, since $E[\{D - w(Y, X, \eta)\}^2 | Y = y, X = x] = E[D\{1 - w(Y, X, \eta)\} | Y = y, X = x]$, it can be shown that $\hat{\Sigma}$ is a consistent estimator of Σ . Suppose $\hat{\beta}_O$ is the solution of

$$\frac{1}{n} \sum_{i=1}^n \{g_{12}(D_i, Y_i, X_i, Z_i, \beta, \hat{\eta}_{\text{ML}}) + (\hat{\Sigma}^{11})^{-1} \hat{\Sigma}^{12} g_{23}(D_i, Y_i, X_i, \beta, \hat{\eta}_{\text{ML}})\} = 0. \quad (19.6.21)$$

Then asymptotically $(\hat{\beta}_O, \hat{\eta}_{\text{ML}})$ is a solution to the optimal estimating equations in (19.6.20) with $(\Sigma^{11}(\beta, \eta), \Sigma^{12}(\beta, \eta))$ replaced by $(\hat{\Sigma}^{11}, \hat{\Sigma}^{12})$. According to the optimal estimating function theory of Godambe (1960) and Godambe and Thompson (1989), $(\hat{\beta}_O, \hat{\eta}_{\text{ML}})$ are the optimal estimators derived from the class of unbiased estimating equations

$$\sum_{i=1}^n C(\beta, \eta) g(D_i, Y_i, X_i, Z_i, \beta, \eta) = 0$$

for any choice of $h(y, x, \beta)$, where $C(\beta, \eta)$ is a $(p+q) \times (2p+q)$ constant matrix.

Finally, when the propensity score $w(y, x, \eta)$ is misspecified, but $h(y, x, \beta) = E\{u(Y, X, Z, \beta)|Y = y, X = x\}$, we can show that $\hat{\Sigma}_{12} \rightarrow 0$ in probability and hence $\hat{\Sigma}^{12} \rightarrow 0$ in probability. As a result, the left-hand side of the first set of estimating equations in (19.6.20) converges in probability to

$$\begin{aligned} E\{g_{12}(D, Y, X, Z, \beta, \eta)\} &= E\{g_1(D, Y, X, Z, \beta, \eta)\} - E\{g_2(D, Y, X, Z, \beta, \eta)\} \\ &= E\{u(Y, X, Z, \beta)\} = 0. \end{aligned}$$

Consequently, even if the model for $w(y, x, \eta)$ is misspecified, the resulting estimator $\hat{\beta}_O$ is still consistent for β ; the price to pay is the possible loss of efficiency.

As an illustration, the Horvitz–Thompson estimator $\hat{\beta}_{HT}$ and $\hat{\eta}_{ML}$ solves

$$\frac{1}{n} \sum_{i=1}^n g_1(D_i, Y_i, X_i, Z_i, \beta, \eta) = 0, \quad \frac{1}{n} \sum_{i=1}^n g_3(D_i, Y_i, X_i, \eta) = 0.$$

It can be rewritten as solving

$$\sum_{i=1}^n C_{HT}(\beta, \eta) g(D_i, Y_i, X_i, Z_i, \beta, \eta) = 0, \quad C_{HT} = \begin{pmatrix} I_{p \times p} & 0_{p \times p} & 0_{p \times q} \\ 0_{q \times p} & 0_{q \times p} & I_{q \times q} \end{pmatrix}.$$

This implies that the optimal estimator $\hat{\beta}_O$ is asymptotically more efficient than the Horvitz–Thompson estimator $\hat{\beta}_{HT}$, provided that $w(y, x, \eta)$ is correctly specified.

Similarly the augmented inverse probability weighted estimator $\hat{\beta}_{AI}$ is the solution of

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n g_1(D_i, Y_i, X_i, Z_i, \beta, \eta) - g_2(D_i, Y_i, X_i, Z_i, \beta, \eta) &= 0, \\ \frac{1}{n} \sum_{i=1}^n g_3(D_i, Y_i, X_i, \eta) &= 0. \end{aligned}$$

These equations can be written as

$$\sum_{i=1}^n C_{AI}(\beta, \eta) g(D_i, Y_i, X_i, Z_i, \beta, \eta) = 0, \quad C_{AI} = \begin{pmatrix} I_{p \times p} & -I_{p \times p} & 0_{p \times q} \\ 0_{q \times p} & 0_{q \times p} & I_{q \times q} \end{pmatrix}.$$

As a result, the optimal estimation $\hat{\beta}_O$ is asymptotically more efficient than $\hat{\beta}_{AI}$.

In summary we have proven the following theorem:

Theorem 19.4 (1) The optimal estimator defined by (19.6.20) is consistent under either (a) The missing probability function is correctly specified or (b) The “working regression function” is the conditional score function $h = E[u|y, x]$.

(2) The optimal estimator $\hat{\beta}_O$ is asymptotically more efficient than both the Horvitz and Thompson estimator $\hat{\beta}_{HT}$ and the augmented inverse probability weighted estimator $\hat{\beta}_{AI}$.

Corollary. In general $\hat{\beta}_O$ is not asymptotically equivalent to $\hat{\beta}_{AI}$. However, if $h = E[u|y, x]$, then they are equivalent to each other.

In fact, we can observe that

$$E[g_{12}g_{23}^T] = 0,$$

if $h = E[u|y, x]$. As a result, $\Sigma^{12} = 0$ and (19.6.20) is equivalent to

$$g_1 - g_2 = 0, \quad g_3 = 0.$$

Empirical Likelihood Based Combination

Qin et al. (2009) used empirical likelihood method to combine estimating equations g_1 and g_2 , or equivalently to combine $g_1 - g_2$ and g_2 . They maximize the empirical likelihood $\prod_{i=1}^n p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i g_{1i} = 0, \quad \sum_{i=1}^n p_i g_{2i} = 0.$$

They showed many nice properties for empirical likelihood based estimators.

19.7 Parameter Estimation for the “Working Regression Model”

Finally we discuss the parameter estimation problems for the “working regression model”.

As observed in Sect. 19.2 on the response mean estimation, the parameters in the “working regression model” should be estimated by the weighted least squares method with optimal weights $d_i(1 - \pi_i)/\pi_i^2$, where $\pi_i = \pi(x_i)$, $i = 1, 2, \dots, n$. Duan et al. (2010) generalized this result to the regression problem with missing at random data.

Consider a regression model

$$E(Y|x) = \psi(x, \beta), \quad \text{Var}(Y|x) < \infty,$$

where β is a vector parameter and ψ is a known function of x and β . We assume that the response Y may be missing at random. However, the covariate X and a surrogate variable S of Y are always available. Again the propensity score is specified by

$$P(D = 1|y, s, x) = P(D = 1|s, x) = \pi(s, x, \alpha),$$

where $D = 1$ implies Y is observed. In the presence of missing data, the inverse probability weighted estimating equation is given by

$$n^{-1} \sum_{i=1}^n \frac{D_i}{\pi(s_i, x_i)} A(x_i) \{y_i - \psi(x_i, \beta)\} = 0,$$

where $A(x)$ is a given vector function of x . For the time being we assume the propensity score is known completely. The augmented inverse probability weighted estimating equation is

$$\sum_{i=1}^n A(x_i) \left[\frac{D_i}{\pi_i} \{y_i - \psi(x_i, \beta)\} - \frac{D_i - \pi_i}{\pi_i} \{\phi(s_i, x_i, \gamma) - \psi(x_i, \beta)\} \right] = 0,$$

where $\phi(s, x, \gamma)$ is a “working model” for $E(Y|s, x)$. For fixed γ , denote $\Sigma(\gamma)$ as the asymptotic variance matrix of $\hat{\beta}$, i.e., the solution of the above estimating equation. Ideally if we can find γ^* such that $\Sigma(\gamma) - \Sigma(\gamma^*) \geq 0$, then γ^* is the optimal choice, where \geq implies non-negative definite matrix. However, in general there is no solution. Instead, we may find γ such that the determinant of Σ has a minimum value, or, the trace of Σ has a minimum value. Duan et al. (2010) used the minimum trace as the criterion.

Exercise Show that for fixed γ the asymptotic variance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is

$$\begin{aligned} \Sigma(\gamma) &= E \left[\frac{1 - \pi(S, X)}{\pi(S, X)} \{Y - \phi(S, X, \gamma)\}^2 V^{-1} A(x) A^T(X) (V^{-1})^T \right] \\ &\quad + \text{Var}[V^{-1} A(X) \{Y - \psi(x, \beta)\}], \end{aligned}$$

where

$$V = E[A(X)\psi_\beta^T(X, \beta)], \quad \psi_\beta = \frac{\partial \psi(x, \beta)}{\partial \beta}.$$

The optimal estimation of γ with minimum trace solves the equation

$$\sum_{i=1}^n \frac{D_i}{\pi_i} \frac{1 - \pi_i}{\pi_i} \{Y_i - \phi(s_i, x_i, \gamma)\} \phi_\gamma(s_i, x_i, \gamma) A^T(x_i) [\hat{V}^{-1}]^T \hat{V}^{-1} A(x_i) = 0,$$

where \hat{V} is the sample version of V . Some adjustments are needed if α in the propensity score is unknown. Fortunately, numerical result shows this is not necessary. Simulation results can be found in Duan et al. (2010). Moreover Chen and Qin (2013) extended this result to longitudinal missing data problems.

19.8 Instrument Variable Approach in Casual Inferences

In the presence of non-compliers, using instrumental variables for estimating causal effects has become very popular in economics and clinical trials. Many references can be found in statistics, econometrics and epidemiology literature, among others, including the works by Angrist and Krueger (2001), Angrist et al. (1996), Imbens and Rubin (1997a,b) and Abadie (2003). Baiocchi et al. (2014) reviewed this topic comprehensively. A major challenge to the validity of observational studies is the possibility of unmeasured confounding (i.e., unobservable baseline variables that differ in the treatment and control groups and that affect the outcome). Instrumental variables analysis is a valid method for controlling for unmeasured confounding. This type of analysis requires the measurement of a valid instrumental variable, which is a variable that (i) is independent of the unmeasured confounding; (ii) affects the treatment; and (iii) affects the outcome only indirectly through its effect on the treatment.

Consider a simple linear regression model

$$Y = \beta_0 + \beta_1 D + \epsilon,$$

where D is the treatment indicator. If $D \perp \epsilon$ or $E(\epsilon|D=0) = E(\epsilon|D=1)$, then the least squares estimator of β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n (1 - D_i) Y_i}{\sum_{i=1}^n (1 - D_i)}$$

is unbiased.

On the other hand if $E(\epsilon|D=0) \neq E(\epsilon|D=1)$, in general the least squares estimator is biased. In fact from

$$E(Y|D=1) - E(Y|D=0) = E(Y_1 - Y_0|D=1) + E(Y_0|D=1) - E(Y_0|D=0),$$

we can observe the first term is the average treatment effect for treated (ATET), and the second term is the selection bias effect. For the linear regression model it becomes

$$\beta_1 + E(\epsilon|D=1) - E(\epsilon|D=0) \neq \beta_1$$

if the choice of treatment depends on the outcome.

An instrumental variable (IV) Z is a $0 - 1$ variable that is independent of ϵ , i.e.,

$$E(\epsilon|Z=1) = E(\epsilon|Z=0),$$

and

$$P(D=1|Z=1) \neq P(D=1|Z=0).$$

Then the IV estimator of β_1 is

$$\hat{\beta}_{1,IV} = A/B, \quad A = \frac{\sum_{i=1}^n Y_i Z_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n Y_i (1 - Z_i)}{\sum_{i=1}^n (1 - Z_i)}, \quad B = \frac{\sum_{i=1}^n D_i Z_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n D_i (1 - Z_i)}{\sum_{i=1}^n (1 - Z_i)}.$$

It can be shown that in probability 1 $\hat{\beta}_{1,IV}$ converges to

$$\frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \frac{\beta_1 E(D|Z = 1) - \beta_1 E(D|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \beta_1.$$

1. Assumptions and local average treatment effect

Let Z be the treatment assignment indicator. Let $D_0 = D(0)$ be the indicator for the treatment that is actually taken, if assigned to the control group ($Z = 0$). Similarly definition applies to $D_1 = D(1)$. Let $Y_0 = Y(0)$ and $Y_1 = Y(1)$ be potential control and treatment outcomes, respectively. The observed data are Y, D, Z , where

$$Y = DY_1 + (1 - D)Y_0, \quad D = ZD_1 + (1 - Z)D_0.$$

Note that $Z = 0$ or 1 determines whether $D_0 = D(0)$ or $D_1 = D(1)$ is observed, and $D(Z) = 0$ or 1 indicates whether the true control outcome $Y_0 = Y(0)$ or treatment outcome $Y_1 = Y(1)$ is observed. If a subject satisfies $D(0) = 0$ and $D(1) = 1$, i.e., takes control or treatment based on the assignment of Z , then this subject is called a complier.

Assumptions

- (1) Independence of the instrument: $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ is independent of Z_i .
- (2) The non-zero average causal effect of randomization on treatment received, $0 < P(Z_i = 1) < 1$ and $P(D_i(1) = 1) > P(D_i(0) = 1)$.
- (3) Monotonicity: $P(D_i(1) \geq D_i(0)) = 1$.

In the presence of covariates, Assumptions (1)–(3) need to condition on the covariate X .

For all members $i = 1, 2, \dots, n$ of a sampling population, let $D_{1i} = 1$ or 0 if the i -th individual is assigned to the treatment group and he/she actually takes the treatment or not, respectively. The same notation applies to D_{0i} . The monotonicity assumption implies

$$D_{1i} \geq D_{0i}, \quad i = 1, 2, \dots, n.$$

In other words, this is equivalent to

- (1) If i chooses non-treatment when assigned treatment, i.e., $D_{1i} = 0$, then i chooses non-treatment when not assigned treatment, i.e., $D_{0i} = 0$.
- (2) If i chooses treatment when not assigned treatment, i.e., $D_{0i} = 1$, then i chooses treatment when assigned treatment, i.e., $D_{1i} = 1$.

For each individual it is possible marginally that $Z = 0, D(0) = 1$ or $Z = 1, D(1) = 0$, but this cannot be true jointly. The monotonic condition is not testable since we cannot simultaneously observe $D(0)$ and $D(1)$ for each individual.

Example In economics, selection of treatment or control is often modelled by a latent index crossing a threshold, where the latent index is interpreted as the expected utility of choosing treatment or control.

$$Y_0 = \mu_0 + \epsilon_0, \quad Y_1 = \mu_1 + \epsilon_1,$$

$$D^* = \alpha + Z\beta + \epsilon, \quad D = I(D^* > 0),$$

where (ϵ_0, ϵ) and (ϵ_1, ϵ) are correlated. The 0-1 variable Z is independent of $(\epsilon_0, \epsilon_2, \epsilon)$. Suppose assignment is based on different levels of Z

$$D(z) = I(-\epsilon \leq \alpha + z\beta).$$

If $\beta > 0$, then in the latent variable model the monotonicity always holds.

Based on a subject's joint values of potential treatment received ($D_i(0), D_i(1)$), the subject in a two-arm trial can be classified into one of four latent compliance classes (Angrist et al. 1996):

$$C_i = \begin{cases} \text{never-taker} & \text{if } (D_i(0), D_i(1)) = (0, 0) \\ \text{complier} & \text{if } (D_i(0), D_i(1)) = (0, 1) \\ \text{always-taker} & \text{if } (D_i(0), D_i(1)) = (1, 1) \\ \text{defier} & \text{if } (D_i(0), D_i(1)) = (1, 0) \end{cases}$$

The monotonicity implies no defier. As a consequence $D_i(0) = 1$ is equivalent to $D_i(0) = 1, D_i(1) = 1$, i.e., taker group is always observed directly. Similarly the never taker group is always observed directly ($\{D_i(1) = 0\} = \{D_i(0) = 0, D_i(1) = 0\}$). Note that

$$\{D_i(0) = 0\} = \{D_i(0) = 0, D_i(1) = 0\} \cup \{D_i(0) = 0, D_i(1) = 1\},$$

implies the no treatment takers in the control arm ($Z = 0$) consist of never takers and compliers. Similarly

$$\{D_i(1) = 1\} = \{D_i(0) = 1, D_i(1) = 1\} \cup \{D_i(0) = 0, D_i(1) = 1\},$$

so the actual treatment takers in the treatment arm ($Z = 1$) consist of always takers and compliers.

Denote

$$\phi_n = P(D(Z) = 0|Z = 1), \quad \phi_a = P(D(Z) = 1|Z = 0),$$

where ϕ_n is the probability of never takers in the treatment group ($Z = 1$) by randomization, and ϕ_a is the probability of always takers in the control group ($Z = 0$). Due to Assumption (1), the probabilities of never-taker, always-taker and complier ($\phi_c = 1 - \phi_n - \phi_a$) should be the same between groups $Z = 0$ and $Z = 1$.

The local average treatment effect $E[Y(1) - Y(0)|D(0) = 0, D(1) = 1]$ can be estimated by taking the ratio of the average difference in Y by instrument and the average difference in D by instrument

$$E[Y(1) - Y(0)|D(0) = 0, D(1) = 1] = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E(D|Z = 1) - E(D|Z = 0)}.$$

In fact we can show this by using the following approach:

Note

$$\begin{aligned} E(D|Z = 1) - E(D|Z = 0) &= E[D(1)|Z = 1] - E[D(0)|Z = 0] \\ &= P(D(1) = 1|Z = 1) - P(D(0) = 1|Z = 0) \\ &= 1 - P(D(1) = 0|Z = 1) - P(D(0) = 1|Z = 0) \\ &= 1 - \phi_n - \phi_a = \phi_c. \end{aligned}$$

Assumption (2) on the non-zero average causal effect of randomization on treatment received implies $\phi_c > 0$.

Let g_a, g_n, g_{c1}, g_{c0} be the densities of outcome Y for always taker, never taker, complier with treatment and complier without treatment, respectively. We can observe that individuals assigned to treatment consist of three types (1) the true compliers, (2) always takers, and (3) never takers. Therefore

$$Y|Z = 1 \sim \phi_c g_{c1} + \phi_a g_a + \phi_n g_n.$$

Similarly

$$Y|Z = 0 \sim \phi_a g_a + \phi_c g_{c0} + \phi_n g_n.$$

Note that

$$E(Y|Z = 1) - E(Y|Z = 0) = \phi_c \{E[Y|g_{c1}] - E[Y|g_{c0}]\}.$$

Therefore,

$$\frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = E[Y|g_{c1}] - E[Y|g_{c0}].$$

Moreover we have found that the so-called “intention to treatment analysis” by comparing outcome Y based on the assignment of Z is a valid method to assess treatment effects in the instrument variable approach.

2. Mixture structure in the instrument variable approach

Next we identify the mixture structure in the instrument variable approach.

Let $P(Z = 1) = \delta$ be the probability of randomization to the treatment arm.

Denote

$$f_{ij}(y) = P(Y = y|Z = i, D = j), \quad i = 0, 1; j = 0, 1.$$

Let

$$g_n(y) = f_{10}(y), \quad g_a(y) = f_{01}(y)$$

be the densities of the never-taker and always-taker, respectively.

We can write the probability density functions as follows:

(1) $Z = 0, D = 1$

$$P(Z = 0)P(D = 1|Z = 0)f(y|D = 1, Z = 0) = (1 - \delta)\phi_a f_{01}(y) = (1 - \delta)\phi_a g_a(y).$$

(2) $Z = 0, D = 0,$

$$\begin{aligned} P(Z = 0)P(D = 0|Z = 0)f(y|D = 0, Z = 0) &= (1 - \delta)(1 - \phi_a)f_{00}(y) \\ &= (1 - \delta)(1 - \phi_a) \\ &\quad \times \left[\frac{\phi_c}{\phi_c + \phi_n} g_{c0}(y) + \frac{\phi_n}{\phi_c + \phi_n} g_n(y) \right]. \end{aligned}$$

(3) $Z = 1, D = 0,$

$$P(Z = 1)P(D = 0|Z = 1)f(y|D = 0, Z = 1) = \delta\phi_n f_{10}(y) = \delta\phi_n g_n(y).$$

(4) $Z = 1, D = 1,$

$$\begin{aligned} P(Z = 1)P(D = 1|Z = 1)f(y|D = 1, Z = 1) &= \delta(1 - \phi_n)f_{11}(y) \\ &= \delta(1 - \phi_n) \\ &\quad \times \left[\frac{\phi_c}{\phi_c + \phi_a} g_{c1}(y) + \frac{\phi_a}{\phi_c + \phi_a} g_a(y) \right]. \end{aligned}$$

Note that f_{00} is a mixture of the distribution of $Y(0)$ for never-takers, $g_n(y)$, and for compliers, g_{c0}

$$f_{00} = \frac{\phi_c}{\phi_c + \phi_n} g_{c0}(y) + \frac{\phi_n}{\phi_c + \phi_n} g_n(y).$$

Analogously f_{11} is a mixture of distribution of $Y(1)$ for compliers, g_{c1} and for always-takers, $g_a(y)$

$$f_{11}(y) = \frac{\phi_c}{\phi_c + \phi_a} g_{c1}(y) + \frac{\phi_a}{\phi_c + \phi_a} g_a(y).$$

Solving g_{c0} we have

$$g_{c0} = \frac{(\phi_c + \phi_n)f_{00} - \phi_n f_{10}}{\phi_c} = \frac{\{1 - P(D = 1|Z = 0)\}f_{00} - P(D = 0|Z = 1)f_{10}}{\phi_c}.$$

Therefore the expectation of the response Y for those compliers in the control group is

$$\begin{aligned} E[Y(0)|C = c] &= \frac{E[1 - D|Z = 0]E[Y|Z = 0, D = 0] - E[1 - D|Z = 1]E[Y|Z = 1, D = 0]}{E[D|Z = 1] - E[D|Z = 0]} \\ &= \frac{E[Y(1 - D)|Z = 0] - E[Y(1 - D)|Z = 1]}{E[D|Z = 1] - E[D|Z = 0]}. \end{aligned}$$

This can also be understood by observing that

- (1) $Y(1 - D) = 0$ if $D = 1$.
- (2) $E[Y(1 - D)|Z = 0]$ is the expectation of the response Y with respect to compliers and never takers in randomized to the control group.
- (3) $E[Y(1 - D)|Z = 1]$ is the expectation of the response Y with respect to never takers in randomized to the treatment group.

Based on Assumption (1) ($Y_i(0)$, $Y_i(1)$, $D_i(0)$, $D_i(1)$) is independent of Z_i , the expectation with respect to never takers is cancelled out in the evaluation of the difference.

Similarly it can be shown that

$$E[Y(1)|C = c] = \frac{E[YD|Z = 1] - E[YD|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]}.$$

Note that the right hand side can be estimated using the observed data.

In the presence of covariate x , it is convenient to reparametrize ϕ_n , ϕ_a as

$$\phi_a = \frac{\exp(x\beta_a)}{1 + \exp(x\beta_n) + \exp(x\beta_a)}, \quad \phi_n = \frac{\exp(x\beta_n)}{1 + \exp(x\beta_n) + \exp(x\beta_a)}.$$

Then

$$\phi_c = \frac{1}{1 + \exp(x\beta_n) + \exp(x\beta_a)}, \quad \frac{\phi_n}{\phi_c + \phi_n} = \frac{\exp(x\beta_n)}{1 + \exp(x\beta_n)}, \quad \frac{\phi_a}{\phi_c + \phi_a} = \frac{\exp(x\beta_a)}{1 + \exp(x\beta_a)}$$

still have the logistic probability forms.

3. Understand Abadie's approach (2003)

We will use notations $D(0) = D_0$ or $D(1) = D_1$ interchangeably below, which denotes the actual treatment status when $Z = 0$ or 1. Clearly the complier $\{D(0) = 0, D(1) = 1\}$ can be written as $D_1 > D_0$.

Note that the proportions of compliers in $Z = 0$ and 1 groups are the same and equal to ϕ_c . The overall proportion of compliers $P(D(1) > D(0))$ is $(1 - \delta)\phi_c + \delta\phi_c = \phi_c$ again. Among all compliers, the proportions for g_{c0} and g_{c1} are $(1 - \delta)$ and δ , respectively. Abadie (2003) made comparisons directly for compliers between the treatment and control groups. He defined the following quantities

$$k = 1 - \frac{D(1 - Z)}{P(Z = 0|x)} - \frac{(1 - D)Z}{P(Z = 1|x)},$$

$$k_0 = (1 - D) \frac{(1 - Z) - P(Z = 0|x)}{P(Z = 0|x)P(Z = 1|x)}, \quad k_1 = D \frac{Z - P(Z = 1|x)}{P(Z = 0|x)P(Z = 1|x)}.$$

Then the following results hold.

(a)

$$E[\psi(Y, D, X)|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[k\psi(Y, D, X)],$$

(b)

$$E[\psi(Y_0, X)|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[k_0\psi(Y, X)],$$

(c)

$$E[\psi(Y_1, X)|D_1 > D_0] = \frac{1}{P(D_1 > D_0)} E[k_1\psi(Y, X)].$$

Let $H(x)$ be the marginal distribution function of X . In the presence of covariate X , all definitions before should be understood as conditioning on X , for example

$$\begin{aligned} \delta(x) &= P(Z = 1|x), \quad \phi_a(x) = P(D(Z = 1|Z = 0, x)), \\ f_{ij}(y|x) &= P(Y = y|D = i, Z = j, x). \end{aligned}$$

Based on $D = i, Z = j, i = 0, 1; j = 0, 1$,

$$k(D = 0, Z = 0) = 1, \quad k(D = 0, Z = 1) = 1 - \frac{1}{P(Z = 1|x)},$$

$$k(D = 1, Z = 0) = 1 - \frac{1}{P(Z = 0|x)}, \quad k(D = 1, Z = 1) = 1.$$

The right hand side of (a) is

$$\begin{aligned} &E[k\psi(Y, D, X)] \\ &= (1 - \delta) \int \{\phi_c g_{c0} + \phi_n g_n\} \psi(y, 0, x) dy dH(x) \end{aligned}$$

$$\begin{aligned}
& + \delta \int \{1 - 1/P(Z = 1|x)\} \phi_n g_n \psi(y, 0, x) dy dH(x) \\
& + (1 - \delta) \int \{1 - 1/P(Z = 0|x)\} \phi_a g_a \psi(y, 1, x) dy dH(x) \\
& + \delta \int \{\phi_c g_{c1} + \phi_a g_a\} \psi(y, 1, x) dy dH(x) \\
& = \phi_c (1 - \delta) \int g_{c0} \psi(y, 0, x) dy dH(x) + \phi_c \delta \int g_{c1} \psi(y, 1, x) dy dH(x).
\end{aligned}$$

Equivalently

$$\frac{E[k\psi(Y, D, X)]}{\phi_c} = (1 - \delta) \int g_{c0} \psi(y, 0, x) dy dH(x) + \delta \int g_{c1} \psi(y, 1, x) dy dH(x).$$

The first term on the right hand side is the expectation with respect to compliers within the control arm (with probability $(1 - \delta)$), and the second term is the expectation with respect to compliers within the treatment arm (with probability δ). Thus we have shown (a).

Similarly

$$k_1 = \frac{DZ}{P(Z = 0|x)P(Z = 1|x)} - \frac{D}{P(Z = 0|x)} = \frac{DZ}{P(Z = 1|x)} - \frac{D(1 - Z)}{P(Z = 0|x)}.$$

Noting the definitions

$$\phi_n(x) = P(D = 0|Z = 1, x), \quad \phi_a(x) = P(D = 1|Z = 0, x), \quad \phi_c(x) = 1 - \phi_n(x) - \phi_a(x),$$

and

$$Y|Z = 1, D = 1, x \sim \frac{\phi_c(x)g_{c1}(y|x) + \phi_a(x)g_a(y|x)}{\phi_c(x) + \phi_a(x)},$$

therefore,

$$\begin{aligned}
E[k_1\psi(Y_1, X)|x] &= E_{11}[P(D = 1|Z = 1, x)\psi(Y_{11}, X)] - E_{01}[P(D = 1|Z = 0, x)\psi(Y_{01}, X)] \\
&= \int \{\phi_c g_{c1} + \phi_a g_a - \phi_a g_a\} \psi(y, x) dy = \phi_c E_c[\psi(y_1, x)].
\end{aligned}$$

Therefore (c) is true. Similarly we can show (b) is also true.

Abadie (2003) directly postulated a model for $(1 - \delta)g_{c0}(y) + \delta g_{c1}(y)$. In general k can take negative values if $D \neq Z$. Abadie et al. (2002) showed that

$$E[k|Y, D, X] = P(D_1 > D_0|Y, D, X).$$

In fact by using the monotonic property, $D_0 = 1$, implies $D_1 = 1$.

$$\begin{aligned}
E[D(1 - Z)|Y, D, X] &= P(D_0 = 1, Z = 0|Y, D, X) = P(D_0 = 1, D_1 = 1, Z = 0|Y, D, X) \\
&= P(D_1 = D_0 = 1|Y, D, X)P(Z = 0|D_1 = D_0 = 1, Y_1, X) \\
&= P(D_1 = D_0 = 1|Y, D, X)P(Z = 0|X).
\end{aligned}$$

Similarly

$$E[(1 - D)Z|Y, D, X] = P(D_1 = D_0|Y, D, X)P(Z = 1|X),$$

$$\begin{aligned}
E[k|Y, D, X] &= 1 - P(D_1 = D_0 = 1|Y, D, X) - P(D_1 = D_0 = 0|Y, D, X) \\
&= P(D_1 > D_0|Y, D, X).
\end{aligned}$$

If the following model is postulated

$$P(Y|X, D, D_1 > D_0) = \psi(D, X\theta),$$

then θ can be estimated through

$$\theta = \operatorname{argmin}_\theta E[\{Y - \psi(D, X\theta)\}^2|D_1 > D_0] = \operatorname{argmin}_\theta E[k\{Y - \psi(D, X\theta)\}^2].$$

The right hand side can be implemented readily since it depends only on the observed data.

Imbens and Rubin (1997a, b) pointed out, however, the above approaches did not use the mixture model structure explicitly. Cheng et al. (2009) explored this structure and proposed a more intuitive estimation method in the no covariate case.

4. Application of density ratio models in causal inference

Next we show that the density ratio model or exponential tilting model can be utilized effectively to solve the causal inference problem discussed above.

Let

$$\lambda = \frac{\phi_c}{\phi_c + \phi_n} = \frac{1 - \phi_a - \phi_n}{1 - \phi_a}, \quad \tau = \frac{\phi_c}{\phi_c + \phi_a} = \frac{1 - \phi_a - \phi_n}{1 - \phi_n}.$$

For notational simplicity, we assume

$$g_{c0}(y) = h_0(y), \quad g_n(y) = h_1(y), \quad g_{c1}(y) = h_2(y), \quad g_a(y) = h_3(y).$$

Then

$$f_{00}(y) = \lambda h_0(y) + (1 - \lambda)h_1(y), \quad f_{01}(y) = h_3(y),$$

$$f_{10}(y) = h_1(y), \quad f_{11}(y) = \tau h_2(y) + (1 - \tau)h_3(y).$$

Cheng et al. (2009) assumed

$$h_i(y)/h_0(y) = \exp(\alpha_i + \beta_i y), i = 1, 2, 3. \quad (19.8.22)$$

As a consequence the likelihood based inference can be applied, which produces the most efficient estimate.

Let $n_{jk} = \sum_{i=1}^n I(Z_i = j, D_i = k)$, $j, k = 0, 1$. The log-likelihood is

$$\begin{aligned} \ell &= n_{01} \log \phi_a + n_{00} \log(1 - \phi_a) + n_{10} \log \phi_n + n_{11} \log(1 - \phi_n) \\ &+ \sum_{i=1}^n [I(Z_i = 0, D_i = 1) \log h_3(y_i) + I(Z_i = 0, D_i = 0) \log\{\lambda h_0(y_i) + (1 - \lambda)h_1(y_i)\}] \\ &+ \sum_{i=1}^n [I(Z_i = 1, D_i = 0) \log h_1(y_i) + I(Z_i = 1, D_i = 1) \log\{\tau h_2(y_i) + (1 - \tau)h_3(y_i)\}]. \end{aligned}$$

Using the density ratio model (19.8.22), the log-likelihood can be written as

$$\begin{aligned} \ell &= n_{01} \log \phi_a + n_{00} \log(1 - \phi_a) + n_{10} \log \phi_n + n_{11} \log(1 - \phi_n) \\ &+ \sum_{i=1}^n [I(Z_i = 0, D_i = 1)(\alpha_3 + \beta_3 y_i) + I(Z_i = 0, D_i = 0) \log[\lambda + (1 - \lambda) \exp(\alpha_1 + \beta_1 y_i)]] \\ &+ \sum_{i=1}^n [I(Z_i = 1, D_i = 0)(\alpha_1 + \beta_1 y_i) \\ &+ I(Z_i = 1, D_i = 1) \log[\tau \exp(\alpha_2 + \beta_2 y_i) + (1 - \tau) \exp(\alpha_3 + \beta_3 y_i)] + \sum_{i=1}^n \log h_0(y_i)]. \end{aligned}$$

After profiling the $h_0(y_i)$'s, the profile log-likelihood is

$$\begin{aligned} \ell &= n_{01} \log \phi_a + n_{00} \log(1 - \phi_a) + n_{10} \log \phi_n + n_{11} \log(1 - \phi_n) \\ &+ \sum_{i=1}^n [I(Z_i = 0, D_i = 1)(\alpha_3 + \beta_3 y_i) + I(Z_i = 0, D_i = 0) \log[\lambda + (1 - \lambda) \exp(\alpha_1 + \beta_1 y_i)]] \\ &+ \sum_{i=1}^n [I(Z_i = 1, D_i = 0)(\alpha_1 + \beta_1 y_i)] \\ &+ I(Z_i = 1, D_i = 1) \log[\tau \exp(\alpha_2 + \beta_2 y_i) + (1 - \tau) \exp(\alpha_3 + \beta_3 y_i)] \\ &- \sum_{i=1}^n \log[1 + \xi_1 \{\exp(\alpha_1 + \beta_1 y_i) - 1\} + \xi_2 \{\exp(\alpha_2 + \beta_2 y_i) - 1\} + \xi_3 \{\exp(\alpha_3 + \beta_3 y_i) - 1\}], \end{aligned}$$

where ξ_j , $j = 1, 2, 3$ are Lagrange multipliers determined by

$$\frac{1}{n} \sum_{i=1}^n \frac{\exp(\alpha_j + \beta_j y_i) - 1}{1 + \sum_{j=1}^3 \xi_j \{\exp(\alpha_j + \beta_j y_i) - 1\}} = 0, \quad j = 1, 2, 3.$$

It can be shown that the limiting value of ξ is

$$\xi_0 = (\xi_{10}, \xi_{20}, \xi_{30}),$$

where

$$\xi_{10} = (1 - \delta)\phi_a(1 - \lambda) + \delta\phi_n, \quad \xi_{20} = \tau\delta(1 - \phi_n), \quad \xi_{30} = (1 - \delta)\phi_a + \delta(1 - \phi_n)(1 - \tau).$$

Let

$$\eta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \phi_a, \phi_n).$$

The large sample results can be found in Cheng et al. (2009). If we wish to test the equality of $h_0(y) = h_2(y)$, or

$$H_0 : \alpha_2 = \beta_2 = 0,$$

then the semiparametric empirical likelihood ratio statistic is

$$R = 2\{\max_{\eta} \ell_F(\eta) - \max_{\eta_1} \ell_F(\eta_1, 0, 0)\}, \quad \eta_1 = (\alpha_1, \beta_1, \alpha_3, \beta_3, \phi_a, \phi_n).$$

Under some regularity conditions, and under $H_0 : h_0(y) = h_2(y)$, in distribution

$$R \rightarrow \chi^2(1),$$

where $\chi^2(1)$ is a chi-square distribution with one degree of freedom.

To implement the maximum semiparametric likelihood estimation, naturally the EM algorithm can be applied.

We need to introduce a new indicator variable C , being 1 if an individual is a complier and 0 otherwise. If we can observe C , then the full log-likelihood is

$$\begin{aligned} \ell_F = & \sum_{i=1}^n [I(Z_i = 0, D_i = 1)\{\log \phi_a + \log h_3(y_i)\} + I(Z_i = 0, D_i = 0, C_i = 1)\{\log \phi_c + \log h_0(y_i)\} \\ & + I(Z_i = 0, D_i = 0, C_i = 0)\{\log \phi_n + \log h_1(y_i)\} + I(Z_i = 1, D_i = 0)\{\log \phi_n + \log h_1(y_i)\} \\ & + I(Z_i = 1, D_i = 1, C_i = 1)\{\log \phi_c + \log h_2(y_i)\} + I(Z_i = 1, D_i = 1, C_i = 0)\{\log \phi_a + \log h_3(y_i)\}]. \end{aligned}$$

By using the model assumptions, we can write ℓ_F as

$$\begin{aligned} \ell_F = & \sum_{i=1}^n [I(Z_i = 0, D_i = 1)\{\log \phi_a + \alpha_3 + \beta_3 y_i\} + I(Z_i = 0, D_i = 0, C_i = 1)\log \phi_c \\ & + I(Z_i = 0, D_i = 0, C_i = 0)\{\log \phi_n + \alpha_1 + \beta_1 y_i\}] + I(Z_i = 1, D_i = 0)\{\log \phi_n + \alpha_1 + \beta_1 y_i\} \\ & + I(Z_i = 1, D_i = 1, C_i = 1)\{\log \phi_c + \alpha_2 + \beta_2 y_i\} + I(Z_i = 1, D_i = 1, C_i = 0)\{\log \phi_a + \alpha_3 + \beta_3 y_i\} \\ & - \sum_{i=1}^n \log[1 + \xi_1\{\exp(\alpha_1 + \beta_1 y_i) - 1\} + \xi_2\{\exp(\alpha_2 + \beta_2 y_i) - 1\} + \xi_3\{\exp(\alpha_3 + \beta_3 y_i) - 1\}]. \end{aligned}$$

Next we find the conditional expectation

$$\begin{aligned}
& E(\ell_F | Z, D, Y = y) \\
&= \sum_{i=1}^n [I(Z_i = 0, D_i = 1)\{\log \phi_a + \alpha_3 + \beta_3 y_i\} + I(Z_i = 0, D_i = 0)w_1 \log \phi_c \\
&\quad + I(Z_i = 0, D_i = 0)(1 - w_1)\{\log \phi_n + \alpha_1 + \beta_1 y_i\}] + I(Z_i = 1, D_i = 0)\{\log \phi_n + \alpha_1 + \beta_1 y_i\} \\
&\quad + I(Z_i = 1, D_i = 1)w_2\{\log \phi_c + \alpha_2 + \beta_2 y_i\} + I(Z_i = 1, D_i = 1)(1 - w_2)\{\log \phi_a + \alpha_3 + \beta_3 y_i\} \\
&\quad - \sum_{i=1}^n \log[1 + \xi_1\{\exp(\alpha_1 + \beta_1 y_i) - 1\} + \xi_2\{\exp(\alpha_2 + \beta_2 y_i) - 1\} + \xi_3\{\exp(\alpha_3 + \beta_3 y_i) - 1\}],
\end{aligned}$$

where

$$w_1 = \frac{\phi_c h_0}{\phi_c h_0 + \phi_n h_1} = \frac{\phi_c}{\phi_c + \phi_n \exp(\alpha_1 + \beta_1 y)},$$

$$w_2 = \frac{\phi_c h_2}{\phi_c h_2 + \phi_a h_3} = \frac{\phi_c \exp(\alpha_2 + \beta_2 y)}{\phi_c \exp(\alpha_2 + \beta_2 y) + \phi_a \exp(\alpha_3 + \beta_3 y)}.$$

Differentiating $E[\ell_F | Z, D]$ with respect to α_1 , we have

$$\begin{aligned}
& \sum_{i=1}^n [I(Z_i = 0, D_i = 0)(1 - w_1) + I(Z_i = 1, D_i = 0)] \\
& - \sum_{i=1}^n \frac{\xi_1 \exp(\alpha_1 + \beta_1 y_i)}{1 + \xi_1\{\exp(\alpha_1 + \beta_1 y_i) - 1\} + \xi_2\{\exp(\alpha_2 + \beta_2 y_i) - 1\} + \xi_3\{\exp(\alpha_3 + \beta_3 y_i) - 1\}} = 0,
\end{aligned}$$

or

$$\xi_1 = n^{-1} \sum_{i=1}^n [I(Z_i = 0, D_i = 0)(1 - w_1) + I(Z_i = 1, D_i = 0)].$$

Similarly

$$\xi_2 = n^{-1} \sum_{i=1}^n w_2 I(Z_i = 1, D_i = 1),$$

$$\xi_3 = n^{-1} \sum_{i=1}^n [I(Z_i = 0, D_i = 1) + (1 - w_2)I(Z_i = 1, D_i = 1)].$$

Replacing ξ by the estimated value, we can maximize $E(\ell_F | Z, D)$ or minimize $-E(\ell_F | Z, D)$ with respect to the underlying parameters.

Consequently, the local average treatment effect can be estimated using

$$\hat{\Delta} = \int y[\exp(\hat{\alpha}_2 + \hat{\beta}_2 y) - 1]d\hat{H}_0(y).$$

Moreover the median difference for the compliers in treatment and control groups can be assessed by

$$\hat{H}_2^{-1}(0.5) - \hat{H}_0^{-1}(0.5).$$

Exercise Under some regularity conditions, find the limiting distributions

$$\sqrt{n}(\hat{\Delta} - \Delta) \rightarrow N(0, \sigma^2),$$

$$\sqrt{n}\{\hat{H}_2^{-1}(0.5) - \hat{H}_0^{-1}(0.5)\} \rightarrow N(0, \sigma_1^2).$$

Regression Model Assumption for Compliers

We can generalize the above approach to regression problems. More specifically, we can assume regression models

$$g_{c0}(y|x) = g_{c0}(y|x\beta_0), \quad g_{c1}(y|x) = g_{c1}(y|x\beta_1),$$

and

$$g_n(y|x) = g_n(y|x\beta_n), \quad g_a(y|x) = g_a(y|x\beta_a)$$

for the compliers in the control group, compliers in the treatment group, never taker group and always taker group, respectively. The mixture structure can be used to identify the underlying parameters.

Chapter 20

Inference in Finite Populations

Most statistical methods such as imputation methods, inverse probability weighted methods, regression calibration methods developed in missing data and biased sampling problems originated from survey sampling studies. In this chapter, we briefly review some important survey sampling problems. There are many excellent survey sampling books, among others, for example, Cochran (1977), Sarndal et al. (1991) and Thompson (1997). We focus on the use of auxiliary information in finite populations.

20.1 Basic Concepts in Finite Sampling

Consider a finite population $\mathcal{P} = \{i, i = 1, 2, \dots, N\}$. Corresponding to each i , there is a quantity Y_i of interest. Let $\mathcal{S} = \{s, s \subset \mathcal{P}\}$ be all possible subsets of \mathcal{P} . A sampling design is a function $p : \mathcal{S} \rightarrow [0, 1]$, with properties

$$p(s) \geq 0, \quad s \in \mathcal{S}, \quad \sum_{s \in \mathcal{S}} p(s) = 1.$$

The inclusion probability π_i is the probability that the i -th unit is in the sample. It is the summation of all probabilities of subsets $s \in \mathcal{S}$ such that $i \in s$, i.e.,

$$\pi_i = P(s \ni i) = \sum_{s \ni i} p(s).$$

Example 1 Simple random sampling without replacement.

Suppose n elements are chosen from \mathcal{P} without replacement and with equal probability. Let \mathcal{S}_n be all subsets of \mathcal{P} with n units. It can be shown that

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & s \in \mathcal{S}_n \\ 0 & \text{otherwise.} \end{cases}$$

The inclusion probability is

$$\pi_i = \sum_{s \ni i} p(s) = \binom{N-1}{n-1} \binom{N}{n}^{-1} = n/N.$$

Example 2 Suppose two units are selected without replacement from \mathcal{P} . In this case \mathcal{S} contains all two unit subsets. Suppose the first unit is drawn with probability p_j , where p_1, \dots, p_N are given probabilities that are related to some measurements of population size. If the i -th unit is selected in the first draw, naturally the second unit will be drawn with probabilities $p_j/(1 - p_i)$ from the remaining units $\mathcal{P} - \{i\}$. In this example $p(s) = 0$ if s contains not exactly two units. The inclusion probability is

$$\pi_i = p_i + \sum_{j \neq i}^N \frac{p_j p_i}{1 - p_j} = p_i \left(1 + \sum_{j \neq i}^N \frac{p_j}{1 - p_i} \right).$$

The probability that both i and j are included is

$$\pi_{ij} = p_i \frac{p_j}{1 - p_i} + p_j \frac{p_i}{1 - p_j}.$$

In general if a sampling design with fixed n ($n \leq N$) units being selected without replacement, then the inclusion probabilities satisfy

$$\sum_{i=1}^N \pi_i = n, \quad \sum_{j \neq i}^N \pi_{ij} = (n-1)\pi_i, \quad \sum_{j > i} \pi_{ij} = n(n-1)/2.$$

To prove this, we denote the selected sample as S . In the fixed sample size design

$$n = \sum_{i=1}^N I(i \in S).$$

Taking expectation with sampling design, then we have

$$n = \sum_{i=1}^N E[I(i \in S)] = \sum_{i=1}^N \pi_i.$$

Moreover

$$\begin{aligned}\sum_{j \neq i} \pi_{ij} &= \sum_{j \neq i} E[I(i \in S)I(j \in S)] = E \left[\sum_{j \neq i} I(i \in S)I(j \in S) \right] \\ &= E \left[I(i \in S) \sum_{j \neq i} I(j \in S) \right] \\ &= E[I(i \in S)(n - I(i \in S))] = nE[I(i \in S)] - E[I(i \in S)] = (n - 1)\pi_i.\end{aligned}$$

Finally taking expectation on the both sides of

$$n^2 = \left[\sum_{i=1}^N I(i \in S) \right]^2,$$

we have $\sum_{i < j} \pi_{ij} = n(n - 1)/2$.

Given a sampling design or inclusion probability, we are interested in finding an unbiased estimator of the population total $T = \sum_{i=1}^N y_i$. Let

$$\hat{T} = \sum_{i \in S} \phi(y_i) = \sum_{i=1}^N \phi_i(y_i)I(i \in S).$$

Taking expectation

$$E[\hat{T}] = \sum_{i=1}^N \phi_i(y_i)P(i \in S) = \sum_{i=1}^N \phi_i(y_i)\pi_i.$$

In order for \hat{T} to be an unbiased estimator of T , we can choose $\phi_i(y_i)\pi_i = y_i$, i.e. $\phi_i(y_i) = y_i/\pi_i$. This is the well known Horvitz and Thompson (1952) estimator

$$\hat{T}_{HT} = \sum_{i=1}^N \frac{y_i I(i \in S)}{\pi_i}.$$

In contrast to the inverse probability weighted estimator discussed in infinite populations (Chap. 19), however, $I(i \in S)$ and $I(j \in S)$ are correlated for $i \neq j$ in finite sampling problems. This is the fundamental difference between the missing data problem in super-populations and survey sampling data in finite populations. Note that

$$\text{cov}[I(i \in S), I(j \in S)] = E[I(i \in S)I(j \in S)] - E[I(i \in S)]E[I(j \in S)] = \pi_{ij} - \pi_i\pi_j,$$

we have

$$\begin{aligned}\text{Var}(\hat{T}_{HT}) &= \sum_{i=1}^N \frac{y_i^2}{\pi_i^2} \text{Var}[I(i \in S)] + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \text{cov}[I(i \in S), I(j \in S)] \\ &= \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \\ &= \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.\end{aligned}$$

Exercise Show the last equality.

20.2 Poisson and Binomial Sampling

Before moving to inferential problems for finite sample population we consider a special sampling design called Poisson and Binomial sampling. It was briefly discussed in Chap. 16. Our main motivation to investigate this design is its popularity in survey sampling. Individual units are not necessarily drawn with equal probabilities. A problem often considered in the literature, sometimes called “probability-proportional-to-size (PPS)” sampling, is a sampling scheme that achieves pre-specified marginal probabilities π_i for the i -th population unit to be included in a survey sample, where $0 < \pi_i < 1$ for $i = 1, \dots, N$, and $\sum_{i=1}^N \pi_i = n$.

In general, given a sampling design $p(s)$, $s \in \mathcal{S}$, it is not hard to find the inclusion probability π_i or π_{ij} . On the contrary, given the inclusion probability π_i , it is challenging to find a sampling design that satisfies the marginal inclusion probability constraint.

Suppose Z_1, \dots, Z_N are independent Bernoulli random variables, each with success probability p_i , $i = 1, 2, \dots, N$. Denote $Z = (Z_1, \dots, Z_N)$ and $p = (p_1, \dots, p_N)$. Then

$$S = Z_1 + \dots + Z_N$$

is called a Poisson-Binomial random variable with parameter p . When all the p_i 's are equal, this reduces to the Binomial distribution. When N is large and all the p_i 's are small but not necessarily equal, the distribution of S is well approximated by a Poisson distribution due to the Law of Small Numbers. We are mainly interested in exact computation of the distribution of S . A closely related sampling model is the conditional Bernoulli model denoted as the conditional distribution of Z given that $S = n$. This model is very useful in a number of different areas, including sample survey and biomedical applications.

Note that

$$P(Z_1, \dots, Z_N) = \prod_{i=1}^N p_i^{z_i} (1 - p_i)^{1-z_i} = \prod_{i=1}^N (1 - p_i) \prod_{i=1}^N \left(\frac{p_i}{1 - p_i} \right)^{z_i},$$

$$\begin{aligned} P(S = n) &= \sum_{z_1 + \dots + z_N = n} \prod_{j=1}^N p_j^{z_j} (1 - p_j)^{1-z_j} = \prod_{i=1}^N (1 - p_i) \sum_{z_1 + \dots + z_N = n} \prod_{j=1}^N w_j^{z_j} \\ &= \prod_{i=1}^N (1 - p_i) \sum_{i_1 < i_2 < \dots < i_n} w_{i_1} w_{i_2} \dots w_{i_n}, \end{aligned}$$

where $w_j = p_j / (1 - p_j)$, and the summation is over all possible combinations of distinct i_1, \dots, i_n from $\{1, \dots, N\}$. A naive way of computing the summation on the right hand side of the equation requires summing $N!/[n!(N-n)!]$ terms, which is impractical even when n and N are of moderate sizes.

We can easily show that the conditional probability is

$$P(Z_1, \dots, Z_N | S = n) = \frac{\prod_{i=1}^N w_i^{z_i}}{\sum_{z_1 + \dots + z_N = n} \prod_{j=1}^N w_j^{z_j}}, \quad z_1 + \dots + z_N = n.$$

Let $D = (D_1, \dots, D_N)$ be a random vector on space D^n , where D_i takes the values 1 or 0 according to whether the i -th unit is in or out of the sample, and $D^n = \{d = (d_1, \dots, d_N) : d_i = 0 \text{ or } 1, \text{ and } d_1 + \dots + d_N = n\}$. Note that if $p_1 = p_2 = \dots = p_N$, then $P_0(D = d) = n!(N-n)!/N!$.

We are seeking a probability mass function P^* such that it has fixed margins $P^*(D_i = 1) = \pi_i, i = 1, 2, \dots, N$ and is as close to P_0 as possible. It can be easily shown that the maximum entropy model discussed in Chap. 9 is just a conditional Bernoulli model with w_i proportional to $p_i / (1 - p_i)$. Chen et al. (1994) studied this maximum entropy distribution for the sampled units. Then the maximum entropy model for D , on space D^n , has the form

$$P(D = d) = \prod_{i=1}^N w_i^{d_i} / \left[\sum_{d \in D^n} \prod_{j=1}^N w_j^{d_j} \right]^{-1} \propto \exp \left(\sum_{i=1}^N \theta_i d_i \right), \quad d \in D^n, \quad \theta_i = \log w_i,$$

where (w_1, \dots, w_N) is chosen to satisfy the constraints

$$\pi_i = E(D_i) = \sum_{d \in D^n} d_i P(D = d).$$

Chen et al. (1994) showed that for any given π_1, \dots, π_N satisfying $\sum_{i=1}^N \pi_i = n$, there are $w_i, i = 1, 2, \dots, N$ such that the corresponding P^* exists. Moreover

they proposed recursive formulas that require $O(nN)$ operations for computing the summation.

20.3 Inferences in Super-population and Survey Population

In survey sampling, population units $\{y_1, \dots, y_N\}$ are assumed fixed. A sample is used to make inferences for the population total $\sum_{i=1}^N y_i$, mean $N^{-1} \sum_{i=1}^N y_i$, median or other quantities of y_i 's. It is more convenient to assume that $y_i, i = 1, 2, \dots, N$ are generated from a super-population with a distribution function $F(y)$. For parameters defined by a super population model but with available data from a finite population, Godambe and Thompson (1986) developed a unified optimal estimation theory.

Suppose the super-parameter θ is defined by

$$\bar{\phi}_N = \sum_{i=1}^N \phi_i(y_i, \theta),$$

where ϕ_i is a given real function such that $\mathcal{E}(\phi_i) = 0, i = 1, 2, \dots, N$, and the expectation is taken with respect to the super-population model. When a finite sample s is drawn from the finite population $\{y_1, \dots, y_N\}$, the observed data is

$$\mathcal{X}_s = \{(i, y_i) : i \in s\}.$$

It seems natural to consider a class of estimating functions

$$\mathcal{H} = \left\{ h(\mathcal{X}_s) : E[h(\mathcal{X}_s)] = \sum_{i=1}^N \phi_i(y_i, \theta) \right\}, \quad (20.3.1)$$

where E is the expectation with respect to the finite population. Basically this requires that the finite sample estimating function is an unbiased estimation of the super-population estimating function. Denote

$$\pi_i = \sum_{s:i \in s} p(s) > 0, \quad i = 1, 2, \dots, N$$

as the inclusion probability. Clearly

$$h^*(\mathcal{X}_s, \theta) = \sum_{i \in s} \frac{\phi_i(y_i, \theta)}{\pi_i} \in \mathcal{H}. \quad (20.3.2)$$

A slight generalization of Godambe's (1960) optimal estimating function criteria is to find a $h \in \mathcal{H}$ such that

$$\mathcal{E}[E\{h^2(\mathcal{X}_s)\}]/[\mathcal{E}E\{\partial h/\partial \theta\}]^2$$

is the minimum.

Theorem 20.1 h^* defined in (20.3.2) is optimal in the class of \mathcal{H} .

Proof For any $h \in \mathcal{H}$, the denominator becomes

$$\left[\mathcal{E} \left\{ \frac{\partial E(h)}{\partial \theta} \right\} \right]^2 = \left[\mathcal{E} \left\{ \frac{\partial \sum_{i=1}^N \phi_i(y_i, \theta)}{\partial \theta} \right\} \right]^2,$$

which is invariant for given $\phi_i, i = 1, 2, \dots, N$. We only need to minimize the numerator.

Let

$$\alpha(\mathcal{X}_s, \theta) = h(\mathcal{X}_s, \theta) - h^*(\mathcal{X}_s, \theta).$$

The constraint on unbiasedness in (20.3.1) implies

$$E\{\alpha(\mathcal{X}_s, \theta)\} = \sum_{s \in \mathcal{S}} p(s)\alpha(\mathcal{X}_s, \theta) = 0. \quad (20.3.3)$$

Using $h = \alpha(\mathcal{X}_s, \theta) + h^*$,

$$\mathcal{E}E(h^2) = \mathcal{E}E(\alpha^2) + \mathcal{E}E\{(h^*)^2\} + 2\mathcal{E}E(\alpha h^*).$$

We would like to show the last term is 0. In fact

$$\begin{aligned} \mathcal{E}E(\alpha h^*) &= \mathcal{E}E \left\{ \sum_{i \in s} \frac{\phi_i}{\pi_i} \alpha(s) \right\} \\ &= \mathcal{E} \left\{ \sum_{i=1}^N \frac{\phi_i}{\pi_i} \sum_{s:i \in s} p(s) \alpha(s) \right\} \\ &= -\mathcal{E} \left\{ \sum_{i=1}^N \frac{\phi_i}{\pi_i} \sum_{s:i \notin s} p(s) \alpha(s) \right\} \\ &= -\sum_{i=1}^N \frac{\mathcal{E}\phi_i}{\pi_i} \sum_{s:i \notin s} p(s) \mathcal{E}\alpha(s) \\ &= 0. \end{aligned}$$

We have used two facts.

(1) From (20.3.3),

$$\sum_{s:i \in s} p(s) \alpha(s) = -\sum_{s:i \notin s} p(s) \alpha(s),$$

which implies the third equality.

(2) The independence among y_i 's and $\mathcal{E}(\phi_i) = 0$ imply the fourth equality.

Godambe and Thompson (1986) showed that the Horvitz–Thompson type estimating function (20.3.2) is optimal in the class (20.3.1). However, their result does not cover the case when the covariate information for every unit $i = 1, 2, \dots, N$ is available since their estimating functions depend only on those sampled data \mathcal{X}_s .

20.4 Utilizing Auxiliary Information

Consider a finite population

$$(Y_1, X_1, Z_1), \dots, (Y_N, X_N, Z_N)$$

drawn from a super population with probability density $f(y|x, z)g(x, z)$. Given the super population, let p be the sampling design

$$p(s) = p\{s|(Y_1, X_1, Z_1), \dots, (Y_N, X_N, Z_N), s \subset \mathcal{P}\}, \quad \mathcal{P} = \{1, 2, \dots, N\},$$

which in general depends on all $(Y_1, X_1, Z_1), \dots, (Y_N, X_N, Z_N)$. This makes likelihood based inference intractable. For example, the likelihood is

$$\int p(s) \prod_{i=1}^N f(y_i, x_i, z_i) \prod_{j \notin s} dy_j,$$

where it is required to integrate out the non-sampled units. However, it is difficult to have a closed form solution. To illustrate this we will use the sampling proportional to the size or successive sampling as an example.

Geologists often evaluate aggregate volumes of discovered plus undiscovered oil and/or gas in a petroleum basin by use of geologic-volumetric methods. Essentially the larger units are more likely to be discovered first. Suppose $Y_1, \dots, Y_N \sim i.i.d$ from a super population with density $f(y)$. Suppose the sampling design is proportional to the size of Y . After n draws, without loss of generality we assume the observed data are $Y_1 = y_1, \dots, Y_n = y_n$. The probability of observing the ordered sample of $Y_1 = y_1, \dots, Y_n = y_n$ is

$$P(Y_1 = y_1, \dots, Y_n = y_n | y_1, \dots, y_N) \propto \prod_{i=1}^n \frac{w(y_i)}{b_i + w(y_{n+1}) + \dots + w(y_N)},$$

where

$$b_i = w(y_i) + \dots + w(y_n),$$

and $w(y)$ is a known function, typically $w(y) = y$. In the full likelihood we need to integrate out those y_j 's not in the sample. The likelihood is

$$L \propto \left(\prod_{i=1}^n \frac{w(y_i) f(y_i)}{b_i} \right) E \left(\prod_{i=1}^n \frac{b_i}{b_i + w(Y_{n+1}) + \dots + w(Y_N)} \right),$$

where the expectation is taken with respect to unobserved y_{n+1}, \dots, y_N . If $N \rightarrow \infty$ and $n/N \rightarrow 0$, the likelihood function will approximate

$$L_w \propto \left(\prod_{i=1}^n \frac{w(y_i) f(y_i)}{\int w(y) f(y) dy} \right),$$

which is a biased sampling problem discussed in Chap. 10. On the other hand if $n = N$, then all $Y_i, i = 1, 2, \dots, N$ can be observed. Naturally $\hat{F}_N(t) = \sum_{i=1}^N I(y_i \leq t)$ is an unbiased estimate of the population distribution function F . In this case, the sampling design becomes irrelevant. Note that Horvitz–Thompson-type estimators are not applicable in this example, as they are functions of inclusion probabilities depending on all unobserved Y . This is the fundamental difficulty in informative missing data or non-ignorable missing data problems.

Nair and Wang (1989) assumed a parametric model for $f(y) = f(y, \theta)$. Denote $R = w(Y_{n+1}) + \dots + w(Y_N)$, and $T = \sum_{i=1}^n \epsilon_i/b_i$, where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. standard exponential random variables independent of the Y_i s. Then the expectation can be written as

$$E_R \left[\prod_{i=1}^n \frac{b_i}{b_i + R} \right] = E_R E_T [\exp(-RT)].$$

Note that T is a weighted summation of standard exponential random variables, it has the so called Erlang density

$$g_n(t) = \sum_{i=1}^n c_i \{b_i \exp(-tb_i)\}, \quad t > 0,$$

where $c_i = \prod_{k \neq i} b_k / (b_k - b_i)$.

Moreover we define the Laplace transform

$$\phi(t, \theta) = E_\theta [\exp\{-tw(Y_1)\}].$$

The likelihood is proportional to

$$\left(\prod_{i=1}^n \frac{w(y_i) f(y_i)}{b_i} \right) \int_0^\infty \{\phi(t, \theta)\}^{N-n} g_n(t) dt.$$

By using missing information principle, Nair and Wang (1989) used the EM algorithm to find the maximum likelihood estimate. Details are given in their paper.

Exercise 1 Let T_1, \dots, T_M be i.i.d. samples from $g_n(t)$. Use Monte Carlo method to evaluate the integral in the likelihood through

$$\frac{1}{M} \sum_{i=1}^M \phi^{N-n}(T_i, \theta).$$

Exercise 2 Is it possible to find the nonparametric MLE for F under the super-population assumption?

Instead of maximum likelihood estimation, an alternative method is the estimating equations based approach. We assume the inclusion probabilities only depend on the observable quantities, or the so called missing at random problem. The estimating functions \mathcal{H} defined in (20.3.1) depend only on the sampled data. In other words, $x_i, i \notin s$ are not used at all. In applications, however, covariate information is available for all individuals. Hence we may consider a larger class of estimating functions.

Define a new class of estimating functions

$$\mathcal{H}_N = \left\{ h = \sum_{i=1}^N \frac{\phi_i I(i \in s)}{\pi_i} - \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i \right\}, \quad (20.4.4)$$

where $\psi_i(x_i, z_i, \beta)$ is a specified function of (x_i, z_i) and β . Note that the first term is the Horvitz and Thompson type estimating function, which is optimal in the class defined in (20.3.1). The second term has mean zero under finite sample design. Essentially, h is the popular difference estimator used in survey sampling. It is design unbiased, i.e., $E(h) = \sum_{i=1}^N \phi_i$ if π_i is correctly specified. On the other hand it is also model unbiased, i.e., if $\psi_i = \mathcal{E}[\phi_i | x_i, z_i]$,

$$\mathcal{E}(h) = \sum_{i=1}^N \frac{\mathcal{E}(\phi_i | x_i, z_i) I(i \in s)}{\pi_i} - \sum_{i=1}^N \frac{I(i \in s) \psi_i}{\pi_i} + \sum_{i=1}^N \psi_i = \sum_{i=1}^N \psi_i.$$

In both cases $E\{\mathcal{E}(h)\} = \mathcal{E}\{E(h)\} = 0$.

Rao et al. (1990) proposed this type of estimator in the linear regression model for the distribution function or quantile estimation.

The finite sample variance of the Horvitz–Thompson estimator $T_{HT} = \sum_{i=1}^N I(i \in s) \phi_i / \pi_i$ under the sampling design is

$$\text{Var}_p(T_{HT}) = \sum_{i,j=1}^N (\pi_{ij} - \pi_i \pi_j) \frac{\phi_i \phi_j^T}{\pi_i \pi_j}.$$

We want to improve upon the efficiency of the Horvitz–Thompson estimator by using auxiliary information when $(x_i, z_i), i = 1, 2, \dots, N$ are available for all individuals. Let $\psi_i^0 = \mathcal{E}[\phi_i | x_i, z_i]$ be the true conditional mean, the corresponding h in (20.4.4)

is denoted as

$$h_0 = \sum_{i=1}^N \frac{\phi_i I(i \in s)}{\pi_i} - \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i^0.$$

Then h_0 is optimal in the class \mathcal{H}_N in the sense,

$$\begin{aligned} & [\mathcal{E}E(\partial h / \partial \beta)]^{-1} \mathcal{E}E[h h^T] [\mathcal{E}E(\partial h / \partial \beta^T)]^{-1} \\ & - [\mathcal{E}E(\partial h_0 / \partial \beta)]^{-1} \mathcal{E}E[h_0 h_0^T] [\mathcal{E}E(\partial h_0 / \partial \beta^T)]^{-1} \geq 0. \end{aligned}$$

In fact for any

$$h = \sum_{i=1}^N \frac{\phi_i I(i \in s)}{\pi_i} - \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i \in \mathcal{H}_N,$$

$$E[h] = \sum_{i=1}^N \phi_i(y_i, x_i, \beta).$$

This implies

$$\mathcal{E}E[\partial h / \partial \beta] = \mathcal{E}\left[\sum_{i=1}^N \partial \phi_i / \partial \beta\right],$$

which is independent of the choice of ψ_i . Therefore we only need to show that

$$\mathcal{E}E[h h^T] \geq \mathcal{E}E[h_0 h_0^T].$$

Note that

$$\begin{aligned} \mathcal{E}E[h_0 h^T] &= \mathcal{E}E\left[\sum_{i=1}^N \frac{I(i \in s)\phi_i}{\pi_i} \sum_{i=1}^N \frac{I(i \in s)\phi_i^T}{\pi_i}\right] \\ &\quad - \mathcal{E}\left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i^0 \phi_i^T + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i^0 \phi_j^T\right] \\ &\quad - \mathcal{E}\left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \phi_i \psi_i^T + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \phi_i \psi_j^T\right] \\ &\quad + \mathcal{E}\left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i^0 \psi_i^T + \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i^0 \psi_j^T\right]. \end{aligned}$$

Using the fact that

$$\mathcal{E}[\phi_i | x, z] = \mathcal{E}[\phi_i | x_i, z_i] = \psi_i^0,$$

it can be shown that

$$\mathcal{E}E[h_0 h^T] = \mathcal{E}E[h_0 h_0^T].$$

As a result

$$\mathcal{E}E[(h_0 - h)(h_0 - h)^T] = \mathcal{E}E[h_0 h_0^T] - \mathcal{E}E[hh_0^T] - \mathcal{E}E[h_0 h^T] + \mathcal{E}E[hh^T],$$

or

$$\mathcal{E}E[hh^T] \geq \mathcal{E}E[h_0 h_0^T].$$

This shows that h_0 is optimal in the class \mathcal{H}_N .

If we know the form $\psi_i^0 = \mathcal{E}[\phi_i | x_i, z_i]$, then we can use h_0 as an estimating function. In applications, however, without specifying the joint distribution of x_i, z_i, y_i , the form of ψ_i^0 is unknown. In this case we can make a guess of the conditional expectation $\mathcal{E}[\phi_i | x_i, z_i]$. Denote the postulated conditional expectation as $\psi_i(x_i, z_i, \beta)$. However, the optimality of using

$$h = \sum_{i=1}^N \frac{\phi_i I(i \in s)}{\pi_i} - \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i \in \mathcal{H}_N$$

is not guaranteed. Next we define a class of estimating functions based on the “working regression models”

$$\mathcal{H}_W = \left\{ h = \sum_{i=1}^N \frac{\phi_i I(i \in s)}{\pi_i} - \gamma \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i : | \gamma \text{ is a } p \times p \text{ no random allowed matrix} \right\}. \quad (20.4.5)$$

Note that $\{I(i \in s) - \pi_i\}\psi_i/\pi_i$ has 0 mean for any $\psi_i = \psi_i(x_i, z_i)$. It is an E-ancillary discussed by Small and McLeish (1989) and also in Sect. 5.5. A good estimating function should be orthogonal to this E-ancillary statistic. This can be achieved by projecting $\sum_{i=1}^N \phi_i I(i \in s)/\pi_i$ onto the space spanned by $\sum_{i=1}^N \{I(i \in s) - \pi_i\}/\pi_i$ (Chap. 5). In other words we need to choose γ such that

$$\mathcal{E}E \left[h \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i^T \right] = 0.$$

Note that

$$\begin{aligned} h \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i^T &= \sum_{i=1}^N \frac{I(i \in s)}{\pi_i^2} \{I(i \in s) - \pi_i\} \phi_i \psi_i^T + \sum_{i \neq j}^N \frac{I(i \in s)\{I(j \in s) - \pi_j\}}{\pi_i \pi_j} \phi_i \psi_j^T \\ &\quad - \gamma \sum_{i=1}^N \frac{\{I(i \in s) - \pi_i\}^2}{\pi_i^2} \psi_i \psi_i^T - \gamma \sum_{i \neq j}^N \frac{\{I(i \in s) - \pi_i\}\{I(j \in s) - \pi_j\}}{\pi_i \pi_j} \psi_i \psi_j^T. \end{aligned}$$

Therefore

$$\begin{aligned} \mathcal{E}E & \left[h \sum_{i=1}^N \frac{I(i \in s) - \pi_i}{\pi_i} \psi_i^T \right] \\ & = \mathcal{E} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \phi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \phi_i \psi_j^T \right] \\ & \quad - \mathcal{E} \left[\gamma \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i \psi_i^T + \gamma \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i \psi_j^T \right]. \end{aligned}$$

As a result, we may choose

$$\gamma = \mathcal{E} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \phi_i \psi_i + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \phi_i \psi_j^T \right] \mathcal{E}^{-1} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i \psi_j^T \right].$$

Clearly if $\gamma = 0$, then we end up with the estimating function defined in Godambe and Thompson (1986).

In general, γ is unknown but it can be estimated by

$$\hat{\gamma} = \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i^2} I(i \in s) \phi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I(i, j \in s)}{\pi_{ij}} \phi_i \psi_j^T \right] \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i^2} I(i \in s) \psi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I(i, j \in s)}{\pi_{ij}} \psi_i \psi_j^T \right]^{-1}.$$

Example 1 If $\psi_i = \mathcal{E}(\phi_i | x_i, z_i) = \psi_i^0$, then

$$\gamma = \mathcal{E} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i \psi_j^T \right] \mathcal{E}^{-1} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i \psi_j^T \right] = I_{p \times p}.$$

In other words, corresponding to the optimal choice, γ is an identity matrix. The difference estimator in the class (20.4.4) is also optimal in the class (20.4.5) if ψ_i is the true regression function $\mathcal{E}[\phi_i | x_i, z_i]$. Furthermore, if $\psi_i = \psi_i^0 = \mathcal{E}[\phi_i | x_i, z_i]$, and even if $\pi'_i = P(i \in s) \neq \pi_i$, the corresponding estimating function h_0 satisfies

$$\begin{aligned} \mathcal{E}E[h_0] & = \mathcal{E} \left[\sum_{i=1}^N \frac{\pi'_i}{\pi_i} \phi_i - \sum_{i=1}^N \frac{\pi'_i - \pi_i}{\pi_i} \psi_i^0 \right] \\ & = \mathcal{E} \left[\sum_{i=1}^N \frac{\pi'_i}{\pi_i} \psi_i - \sum_{i=1}^N \frac{\pi'_i - \pi_i}{\pi_i} \psi_i^0 \right] \end{aligned}$$

$$= \mathcal{E} \left[\sum_{i=1}^N \psi_i^0 \right] = 0.$$

Hence, if the conditional expectation $\mathcal{E}[\phi_i | x_i, z_i]$ is correctly specified, the resulting estimation function h_0 is unbiased regardless of whether design probability is correctly specified. This is well known “model based unbiasedness” property.

Example 2 Let

$$\phi_i = \phi_i(y_i, x_i, \beta), \quad \mathcal{E}[\phi_i | x_i] = 0.$$

Here, the parameter β involves the conditional distribution of Y for given X . If $\pi_i = P(i \in s | x)$, then

$$\gamma = \mathcal{E} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \phi_i \psi_i^T \right] \mathcal{E}^{-1} \left[\sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} \psi_i \psi_i^T + \sum_{i \neq j}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \psi_i \psi_j^T \right].$$

Since

$$\mathcal{E}[\phi_i \psi_i^T | x_i] = \mathcal{E}[\phi_i | x_i] \psi_i^T = 0,$$

then $\gamma = 0_{p \times p}$. This shows that Godambe and Thompson's (1986) estimation is optimal for the regression parameter β when all x_i 's are available.

Example 3 Let

$$\phi_i = y_i - \beta.$$

Then the optimal estimating equation is

$$\sum_{i=1}^N \frac{I(i \in s)(y_i - \beta)}{\pi_i} - \frac{I(i \in s) - \pi_i}{\pi_i} [\mathcal{E}(Y_i | x_i) - \beta] = 0,$$

or

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^N \left[\frac{I(i \in s)y_i}{\pi_i} - \frac{I(i \in s) - \pi_i}{\pi} \mathcal{E}(y_i | x_i) \right].$$

In this case $\hat{\beta}$ estimates the population mean $Y_N = N^{-1} \sum_{i=1}^N y_i$.

In most applications, x_i 's are not available as individual observations, rather they appeared in summarized form such as $\bar{X} = N^{-1} \sum_{i=1}^N x_i$. In this case the population mean estimation of \bar{Y}_N is

$$\tilde{\beta} = \frac{1}{N} \sum_{i=1}^N \left[\frac{I(i \in s)y_i}{\pi_i} - \hat{\gamma} \frac{I(i \in s) - \pi_i}{\pi} x_i \right]$$

$$= \frac{1}{N} \sum_{i=1}^N \frac{I(i \in s)y_i}{\pi_i} - \hat{\gamma} \left[\frac{1}{N} \sum_{i=1}^N \frac{I(i \in s)x_i}{\pi_i} - \bar{X}_N \right].$$

This becomes the familiar population regression estimator in linear regression model, see for example, Cochran (1977).

Example 4 Distribution function estimation.

Define

$$\phi_i = I(y_i \leq t) - \beta.$$

Since the range of a distribution is between 0 and 1; to find $\mathcal{E}[I(Y_i \leq t)|x_i]$, we may postulate a logistic or Probit model (Wu, 2003)

$$\mathcal{E}[I(Y_i \leq t)|x_i] = \frac{\exp(\alpha_t + \beta_t x_i)}{1 + \exp(\alpha_t + \beta_t x_i)},$$

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{I(i \in s)I(y_i \leq t)}{\pi_i} - \frac{I(i \in s) - \pi_i}{\pi_i} \frac{\exp(\alpha_t + \beta_t x_i)}{1 + \exp(\alpha_t + \beta_t x_i)} \right].$$

In the linear model

$$y_i = \beta x_i + v(x_i)\epsilon_i, \quad \epsilon \sim G(s),$$

$$\mathcal{E}[I(Y_i \leq t)|x_i] = \mathcal{E}[I((y_i - x_i\beta)/v(x_i) \leq (t - x_i\beta)/v(x_i))|x_i] := G\left(\frac{t - x_i\beta}{v(x_i)}\right).$$

$G(s)$ can be estimated by

$$\hat{G}(s) = N^{-1} \sum_{i=1}^N \frac{I(i \in s)}{\pi_i} I((y_i - x_i\beta)/v(x_i) \leq s),$$

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{I(i \in s)I(y_i \leq t)}{\pi_i} - \frac{I(i \in s) - \pi_i}{\pi_i} \hat{G}((t - x_i\beta)/v(x_i)) \right].$$

More details can be found in Rao et al. (1990) and Wu and Sitter (2001).

Example 5 Correlation coefficient estimation. Let

$$\phi_i = \begin{pmatrix} x_i - \mu_x \\ y_i - \mu_y \\ (x_i - \mu_x)^2 - \sigma_x^2 \\ (y_i - \mu_y)^2 - \sigma_y^2 \\ (x_i - \mu_x)(y_i - \mu_y) - \rho\sigma_x\sigma_y \end{pmatrix}.$$

We need to guess

$$\mathcal{E}[Y_i|x_i], \quad \mathcal{E}[Y_i^2|x_i].$$

General Choice of ψ_i

Generalized linear model, polynomial function, local polynomial regression, or regression spline function can be used to model $E[y|x_i, z_i] = \psi_i(x_i, z_i, \theta)$. Based on complete data, least squares or weighted least squares may be used to find $\hat{\theta}$. For parameters defined by general estimating equations, the situation is a little more complex, especially when the dimension of the covariate is high. We are interested in modeling $E[\phi_i(y_i, x_i, \beta)|x_i, z_i] = \psi(x_i, z_i, \beta, \gamma)$, where γ is an additional parameter involving in $f(y_i|x_i, z_i, \gamma)$. Since ϕ_i is a vector function, the components of ϕ_i may be fitted individually or simultaneously to a postulated model.

Similar to Rubin and van der Laan (2008) and Cao et al. (2009), the variance $\mathcal{E}\{\text{Var}_E(h)\}$ may be minimized with respect to a “working model” parameter γ . As pointed out before, the variance calculation involves the cross-product term which is different from the missing data problem in superpopulations.

20.5 Pseudo Likelihoods Method in Finite Sampling Problems

The applications of pseudo empirical likelihood methods for the estimation of finite population total or mean were discussed by Chen and Qin (1993), then subsequently by Wu and Sitter (2001), Chen et al. (2002), Wu (2003) and many others. Next we consider the situation where parameters are defined by general estimating equations.

Let

$$(X_1, Y_1), \dots, (X_N, Y_N)$$

be a random sample from a super-population with density $f(y, x)$. From this super-population we randomly select a finite sample s . Denote the inclusion probabilities as

$$\pi_i = E[I(i \in S)], \quad \pi_{ij} = E[I(i \in S, j \in S)].$$

Suppose the super-population parameter β is defined by $E[U(Y, X, \beta)] = 0$. Given the super-population, the finite population expectation is

$$E \left\{ \sum_{i=1}^N \frac{I(i \in S)U(y_i, x_i, \beta)}{\pi_i} \right\} = \sum_{i=1}^N U(y_i, x_i, \beta).$$

Empirical likelihood may be formed based on the super-population

$$\prod_{i=1}^N p_i, \quad p_i = dF(y_i, x_i), \quad i = 1, 2, \dots, N,$$

subject to the constraints

$$\sum_{i=1}^N p_i = 1, \quad p_i \geq 0,$$

and

$$\sum_{i=1}^N p_i \frac{I(i \in s)U(y_i, x_i, \beta)}{\pi_i} = 0, \quad \sum_{i=1}^N p_i \frac{I(i \in s) - \pi_i}{\pi_i} h(x_i) = 0.$$

Note that the above estimating functions depend only on the observed data. In the Bernoulli sample case, i.e., $P(i \in s) = \pi(x_i)$ only depends on the i -th individual and $\pi_{ij} = \pi_i \pi_j$, Qin et al. (2009) discussed this special case in detail. In general sampling design, however, $\pi_{ij} \neq \pi_i \pi_j$. The large sample results remain open. Some further research is warranted.

Equivalently we can also use the following constraints

$$\sum_{i=1}^N p_i \left\{ \frac{I(i \in s)U(y_i, x_i, \beta)}{\pi_i} - \frac{I(i \in s) - \pi_i}{\pi_i} h(x_i) \right\} = 0, \quad \sum_{i=1}^N p_i \frac{I(i \in s) - \pi_i}{\pi_i} h(x_i) = 0.$$

In other words we may start from the doubly robust estimating function and then find the optimal combination with the E-ancillary estimating function $h(x_i)$ $\{I(i \in s) - \pi_i\}/\pi_i$. The general theoretical results are still open problems.

Chapter 21

Inference for Density Ratio Model with Continuous Covariates

We have discussed different density ratio models for two-sample or multiple-sample problems in Chaps. 10, 11, 17 and 18. A natural generalization is to study a density ratio model for continuous covariates. In this chapter we discuss two approaches. (1) the pairwise conditional likelihood method to eliminate the baseline “carrier density” or nuisance parameters, and (2) the profile maximum likelihood method. Also we will consider conditional independent tests in general semiparametric models and in partially specified exponential graphical models with high dimensional parameters.

21.1 Generalized Odds Ratio Model and Pairwise Conditional Likelihood

Consider a partially specified exponential family model or density ratio model

$$f(y|x) = \frac{\exp(yx\theta)f(y)}{\int \exp(yx\theta)f(y)dy}, \quad (21.1.1)$$

where $f(y)$ is the baseline “carrier density”. Inference for θ is straightforward if $f(y)$ has a known form. However it is possible to identify θ even if $f(y)$ is unknown. The pioneer work for estimating θ was discussed by Kalbfleisch (1978). In his approach the carrier density is eliminated by conditioning on all covariates and order statistics $y_{(1)} \leq \dots \leq y_{(n)}$. As discussed in Chap. 3, this approach may not be feasible due to the computational challenge when the sample size n is large.

Qin and Liang (1999) and Liang and Qin (2000) considered a generalized odds ratio model

$$\frac{f(y_j|x_i)f(y_i|x_j)}{f(y_j|x_j)f(y_i|x_i)} = R(y_i, x_i, y_j, x_j; \theta), \quad (21.1.2)$$

where the form of R is known but θ is an unknown parameter. Under the density ratio model (21.1.1)

$$R(x_i, y_i, x_j, y_j; \theta) = \exp\{-(y_i - y_j)(x_i - x_j)\theta\}.$$

Furthermore, for any function $\eta(y_1, x_1, y_2, x_2; \theta)$, we have estimating equation

$$E\{\eta(y_1, x_1, y_2, x_2; \theta) - \eta(y_2, x_1, y_1, x_2; \theta)R(y_1, x_1, y_2, x_2; \theta)|x_1, x_2\} = 0. \quad (21.1.3)$$

In fact, the left hand side is

$$\begin{aligned} & \int \int \eta(y_1, x_1, y_2, x_2, \theta) f(y_1|x_1) f(y_2|x_2) dy_1 dy_2 \\ & - \int \int \eta(y_2, x_1, y_1, x_2, \theta) R(y_1, x_1, y_2, x_2, \theta) f(y_1|x_1) f(y_2|x_2) dy_1 dy_2 \\ & = \int \int \eta(y_1, x_1, y_2, x_2; \theta) f(y_1|x_1) f(y_2|x_2) dy_1 dy_2 \\ & - \int \int \eta(y_2, x_1, y_1, x_2; \theta) f(y_2|x_1) f(y_1|x_2) dy_1 dy_2 = 0. \end{aligned}$$

The sample version of this equation is

$$\sum_{i < j} \{\eta(y_i, x_i, y_j, x_j; \theta) - \eta(y_j, x_i, y_i, x_j; \theta)R(y_i, x_i, y_j, x_j; \theta)\} = 0, \quad (21.1.4)$$

which can be used to estimate θ .

Next we construct the pairwise conditional likelihood. For any two individuals, say, 1 and 2, we can construct a pairwise conditional likelihood

$$f(y_1, y_2|y_{(1)}, y_{(2)}, x_1, x_2) = \frac{f(y_1|x_1)f(y_2|x_2)}{f(y_1|x_1)f(y_2|x_2) + f(y_2|x_1)f(y_1|x_2)} = \frac{1}{1 + R(z_1, z_2; \theta)},$$

where $y_{(1)}$ and $y_{(2)}$ are order statistics of y_1 and y_2 , and for simplicity we have written $z_i = (y_i, x_i)$, $i = 1, 2$. The pairwise conditional log-likelihood is

$$\ell_p(\theta) = - \sum_{i < j} \log\{1 + R(z_i, z_j; \theta)\}, \quad (21.1.5)$$

with corresponding score estimating equation

$$\frac{\partial \ell_p(\theta)}{\partial \theta} = \sum_{i < j} \psi_{ij}(\theta), \quad \psi_{ij} = - \sum_{i < j} \frac{1}{1 + R(z_i, z_j; \theta)} \frac{\partial R(z_i, z_j; \theta)}{\partial \theta} = 0.$$

Next we present some large sample results.

Theorem 21.1 Suppose (y_i, x_i) , $i = 1, 2, \dots, n$, are independent and identically distributed from the density ratio model (21.1.1). Let $\hat{\theta}$ be the solution of $\partial\ell_p(\theta)/\partial\theta = 0$. Then under some suitable conditions, with probability one, $\hat{\theta}$ exists and is a consistent estimator of θ_0 . Furthermore

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, V),$$

where $V = V_1^{-1}V_2V_1^{-1}$ with

$$V_1 = -0.5 E(\partial\psi_{12}/\partial\theta), \quad V_2 = \text{cov}(\psi_{12}, \psi_{13}).$$

Proof Using the large sample property of U -statistics, we can show in probability

$$\ell_p(\theta) \binom{n}{2}^{-1} \rightarrow E[-\log\{1 + R(z_1, z_2; \theta)\}].$$

Note that

$$E[-\log\{1 + R(z_1, z_2; \theta)\}] = E(E[-\log\{1 + R(z_1, z_2; \theta)\}|y_{(1)}, y_{(2)}, x_1, x_2]).$$

Since $-\log\{1 + R(z_1, z_2; \theta)\}$ is the likelihood of y_1, y_2 conditional on $y_{(1)}, y_{(2)}, x_1, x_2$, $E[-\log\{1 + R(z_1, z_2; \theta)\}|y_{(1)}, y_{(2)}, x_1, x_2]$ achieves the maximum value when $\theta = \theta_0$ by the conditional Kullback–Leibler information inequality, see, for example, Andersen (1970). Therefore, the maximum pairwise conditional likelihood estimate is consistent.

The projection method for U -statistics can be used here for establishing the Central Limit Theorem for $\hat{\theta}$. A Taylor's expansion on $\partial\ell_p(\hat{\theta})/\partial\theta$ gives

$$\sqrt{n}(\hat{\theta} - \theta) = \left(\frac{-\sum_{i < j} \partial\psi_{ij}/\partial\theta}{n^2} \right)^{-1} \sum_{i < j} \psi_{ij}(\theta)/n^{3/2} + o_p(1).$$

The first term on the right-hand side is

$$\frac{-\sum_{i < j} \partial\psi_{ij}/\partial\theta}{n^2} = \frac{-\binom{n}{2}}{n^2} \frac{\sum_{i < j} \partial\psi_{ij}/\partial\theta}{\binom{n}{2}} \rightarrow V_1,$$

where

$$\begin{aligned} V_1 &= -0.5E(\partial\psi_{12}/\partial\theta) \\ &= -\frac{1}{2}E\left[-\frac{\partial^2 \log\{1 + R(z_1, z_2; \theta)\}}{\partial\theta\partial\theta^T}\right] \end{aligned}$$

$$= -0.5E \left(Var \left[\frac{\partial \log\{1 + R(z_1, z_2; \theta)\}}{\partial \theta} |_{y(1), y(2), x_1, x_2} \right] \right).$$

The second term may be re-expressed as

$$\left\{ \frac{\text{cov}(\psi(\theta))}{n^3} \right\}^{1/2} \text{cov}^{-1/2}\{\psi(\theta)\} \cdot \psi(\theta).$$

Following Lehmann (1975, p. 337, 367), we have

$$\begin{aligned} \frac{\text{cov}(\psi(\theta))}{n^3} &= \frac{0.5n(n-1)\text{cov}(\psi_{12}) + n(n-1)(n-2)\text{cov}(\psi_{12}, \psi_{13})}{n^3} \\ &\rightarrow V_2 = \text{cov}(\psi_{12}, \psi_{13}) \end{aligned}$$

and $\text{cov}^{-1/2}(\psi(\theta)) \cdot \psi(\theta)$ converges to a standard multivariate normal distribution. This completes the proof.

Remark 1 V_1 and V_2 can be consistently estimated by

$$\begin{aligned} \hat{V}_1 &= -\frac{\sum_{i < j} \partial \psi_{ij}(\hat{\theta}) / \partial \theta}{n^2} \\ \hat{V}_2 &= \sum_{i=1}^n \sum_{j < k} \psi_{ij}(\hat{\theta}) \psi_{ik}(\hat{\theta}) / n^3. \end{aligned}$$

Remark 2 If we are interested in testing $\theta = 0$, the independence between Y and X , the pairwise conditional likelihood based score statistic has a very simple form of

$$Q = \sum_{i < j} \frac{(y_i - y_j)(x_i - x_j)}{2}.$$

Note that Q has zero expectation when X and Y are independent. In general $E(Q) = 0$ does not imply independence. However, for the bivariate normal distribution $E(Q) = n(n-1)\rho\sigma_1\sigma_2$, where ρ is the correlation coefficient and σ_1^2 and σ_2^2 are variances of Y and X , respectively. Therefore $E(Q) = 0$ is equivalent to X and Y being independent under the bivariate normality assumption. In contrast to Kendall's tau and Spearman's rank correlation coefficient tests, Q is an alternative test statistic for testing independence.

Remark 3 Kalbfleisch (1978) considered a generalized linear model with a conditional density function

$$f(y_i | x_i; \theta_i, \phi) = \exp[a(\phi)\{y_i \theta_i - b(\theta_i)\} + c(y_i, \phi)], \quad (21.1.6)$$

where $\theta_i = U(\beta^T x_i)$, U is a known link function. Even if the forms of $b(\cdot)$ and $c(\cdot)$ are not specified, it is possible to test $\beta = 0$ in this model without using Kalbfleisch's

conditional likelihood approach. In fact, the log-likelihood is

$$l_F = \sum_{i=1}^n [a(\phi)\{y_i U(\beta_0 + \beta_1^T x_i) - b(U(\beta_0 + \beta_1^T x_i))\} + c(y_i, \phi)].$$

Taking derivative with respect to β_1 and β_0 , respectively,

$$\frac{\partial l_F}{\partial \beta_1} = a(\phi) \sum_{i=1}^n [y_i U'(\beta_0 + \beta_1^T x_i) x_i - b'(U(\beta_0 + \beta_1^T x_i)) U'(\beta_0 + \beta_1^T x_i) x_i]$$

and

$$\frac{\partial l_F}{\partial \beta_0} = a(\phi) \sum_{i=1}^n [y_i U'(\beta_0 + \beta_1^T x_i) - b'(U(\beta_0 + \beta_1^T x_i)) U'(\beta_0 + \beta_1^T x_i)].$$

Plugging $\beta_1 = 0$ in the above equations, the score functions are

$$\frac{\partial l_F}{\partial \beta_1}|_{\beta_1=0} = a(\phi) \sum_{i=1}^n [y_i U'(\beta_0) x_i - b'(U(\beta_0)) U'(\beta_0) x_i]$$

and

$$\frac{\partial l_F}{\partial \beta_0}|_{\beta_1=0} = a(\phi) \sum_{i=1}^n [y_i U'(\beta_0) - b'(U(\beta_0)) U'(\beta_0)].$$

From $\partial l_F / \partial \beta_0|_{\beta_1=0} = 0$, the score equation given above is equivalent to

$$\frac{\partial l_F}{\partial \beta_1}|_{\beta_1=0} = a(\phi) U'(\beta_0) \left\{ \sum_{i=1}^n y_i x_i - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n x_i \right\}.$$

This is exactly the same score equation based on Kalbfleisch's (1978) conditional likelihood (Chap. 3 Eq. 3.4) for testing $\beta_1 = 0$. When β_1 is close to zero, the conditional likelihood and the pairwise log conditional likelihood carry most of information for β_1 . This will be demonstrated in the simulation studies.

Next we extend the pairwise conditional likelihood to the situation where θ satisfies a constraint $\xi(\theta) = 0$. To do this, we maximize $\ell_p(\theta)$ subject to the constraint $\xi(\theta) = 0$. This may be done by considering the function

$$M = \ell_p(\theta) \binom{n}{2}^{-1} + \nu^T \xi(\theta),$$

where ν is a vector of Lagrange multipliers. Differentiating M with respect to θ and ν ,

$$\frac{\partial \ell_p(\theta)}{\partial \theta} \binom{n}{2}^{-1} + \nu^T \frac{\partial \xi(\theta)}{\partial \theta} = 0, \quad \xi(\theta) = 0.$$

Let $\tilde{\theta}$ and $\tilde{\nu}$ be the solution to the above equations.

Theorem 21.2 Under suitable conditions, if θ_0 satisfies $\xi(\theta) = 0$, then the constrained maximum pairwise conditional likelihood estimator $(\tilde{\theta}, \tilde{\nu})$ satisfies

$$\sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ \tilde{\nu} \end{pmatrix} \rightarrow N(0, V_c),$$

where $V_c = V_{c1}^{-1} V_{c2} V_{c1}^{-1}$ and

$$V_{c1} = \begin{pmatrix} -E \frac{\partial \psi_{12}(\theta_0)}{\partial \theta} - \frac{\partial \xi(\theta_0)}{\partial \theta} \\ -\frac{\partial \xi(\theta_0)}{\partial \theta^T} \end{pmatrix}, \quad V_{c2} = \begin{pmatrix} \text{cov}(\psi_{12}, \psi_{13}) & 0 \\ 0 & 0 \end{pmatrix}.$$

Proof The existence and consistency of the constrained maximum pairwise conditional likelihood estimator can be shown using a similar argument as Aitchison and Silvey (1958).

Expanding

$$\frac{\partial \ell_p(\tilde{\theta})}{\partial \theta} \binom{n}{2}^{-1} + \nu^T \frac{\partial \xi(\tilde{\theta})}{\partial \theta} \quad \text{and} \quad \xi(\tilde{\theta})$$

at $(\theta_0, 0)$, we have

$$\begin{pmatrix} \tilde{\theta} - \theta_0 \\ \tilde{\nu} \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \frac{2}{n(n-1)} - \frac{\partial \xi(\theta_0)}{\partial \theta} \\ -\frac{\partial \xi(\theta_0)}{\partial \theta^T} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \ell_p(\theta_0)}{\partial \theta} \frac{2}{n(n-1)} \\ 0 \end{pmatrix} + o_p(n^{-1/2}).$$

It can be shown that

$$\begin{pmatrix} -\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \frac{2}{n(n-1)} - \frac{\partial \xi(\theta_0)}{\partial \theta} \\ -\frac{\partial \xi(\theta_0)}{\partial \theta^T} \end{pmatrix}^{-1} \rightarrow \begin{pmatrix} -E \frac{\partial \psi_{12}(\theta_0)}{\partial \theta} - \frac{\partial \xi(\theta_0)}{\partial \theta} \\ -\frac{\partial \xi(\theta_0)}{\partial \theta^T} \end{pmatrix}^{-1} = V_{c1}^{-1}$$

and

$$\sqrt{n} \begin{pmatrix} \frac{\partial \ell_p(\theta_0)}{\partial \theta} \frac{2}{n(n-1)} \\ 0 \end{pmatrix} \rightarrow N(0, V_c), \quad V_{c2} = \begin{pmatrix} \text{cov}(\psi_{12}, \psi_{13}) & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence

$$\sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ \tilde{\nu} \end{pmatrix} \rightarrow N(0, V_{c1}^{-1} V_{c2} V_{c1}^{-1}).$$

While $\hat{\theta}$ and its estimated standard errors can be used as bases for inferences, it is well known that this Wald-based procedure may not be desirable with modest

sample size. An alternative is to use the likelihood-ratio-based approach. For testing $H_0 : \theta = \theta_0$, define

$$T(\theta_0) = 2n^{-1}[\max_{\theta} \ell_p(\theta) - \ell_p(\theta_0)].$$

Theorem 21.3 Under suitable conditions, if θ_0 is the true value of θ , then the pairwise conditional likelihood ratio test statistic $T(\theta_0)$ converges to a weighted chi-square variable, where the weights are eigenvalues of the matrix $V_2^{1/2}V_1^{-1}V_2^{1/2}$.

Proof Expanding $\ell_p(\theta_0)$ at $\hat{\theta}$, we have

$$\ell_p(\theta_0) = \ell_p(\hat{\theta}) + \frac{\partial \ell_p(\hat{\theta})}{\partial \theta}(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^T \frac{\partial^2 \ell_p(\hat{\theta})}{\partial \theta \partial \theta^T}(\theta_0 - \hat{\theta}) + o_p(n).$$

Therefore the pairwise conditional likelihood ratio statistic is

$$\begin{aligned} T &= 2n^{-1} \sum_{i < j} [\log\{1 + R(z_i, z_j; \theta_0)\} - \log\{1 + R(z_i, z_j; \hat{\theta})\}] \\ &= n^{-1}(\theta_0 - \hat{\theta})^T \frac{\partial^2 \ell_p(\hat{\theta})}{\partial \theta \partial \theta^T} (\theta_0 - \hat{\theta}) + o_p(1). \end{aligned}$$

Note

$$\frac{\partial^2 \ell_p(\theta)}{\partial \theta \partial \theta^T} \binom{n}{2}^{-1} \rightarrow E \left(-\frac{\partial \psi_{12}(\theta_0)}{\partial \theta} \right).$$

Since $\hat{\theta}$ is the solution of $\partial \ell_p(\theta)/\partial \theta = 0$,

$$\hat{\theta} - \theta_0 = \left(\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \frac{\partial \ell_p(\theta_0)}{\partial \theta} + o_p(n^{-1/2}).$$

Therefore

$$T = n^{-1} \left(\frac{\partial \ell_p(\theta_0)}{\partial \theta} \right)^T \left(\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \frac{\partial \ell_p(\theta_0)}{\partial \theta} + o_p(1) \rightarrow Z^T (V_2^{1/2} V_1^{-1} V_2^{1/2}) Z,$$

where Z is a standard normal variable. Consequently, T converges to a weighted chi-square variable with weights being the eigenvalues of the matrix $V_2^{1/2} V_1^{-1} V_2^{1/2}$.

The same strategy can be used to test the components of θ . A profile pairwise conditional likelihood ratio statistic can be constructed in the same way as the full likelihood; details are omitted.

Model fitting is a very important issue. We can employ nonparametric method to estimate the regression function $P(y_1 = y_{(1)}, y_2 = y_{(2)} | y_{(1)}, y_{(2)}, x_1, x_2)$. The difference between the nonparametric regression estimation and the proposed generalized odds ratio model $R(z_1, z_2, \theta)$ can be used to assess the adequacy of the proposed model. Here, however, we use the smooth goodness of fit test technique to test model

fit. We embed $h(y, x; \theta)$ or $R(y_1, x_1, y_2, x_2)$ in a large parametric family and then test some of the parameters to be zero. Alternatively, we test the generalized odds ratio $R(z_1, z_2, \theta)$ or function h . We can choose different η in estimating equations (21.1.4).

Theorem 21.4 Suppose $\eta_{r \times 1}$ is a vector valued function that is independent of the pairwise conditional likelihood based score function ψ . Let

$$Q(\hat{\theta}) = \frac{2}{n(n-1)} \left[\sum_{i < j} \{\eta(y_i, x_i, y_j, x_j, \hat{\theta}) - \eta(y_j, x_i, y_i, x_j, \hat{\theta})\} R(y_i, x_i, y_j, x_j, \hat{\theta}) \right].$$

Under some regularity conditions, if the generalized odds ratio model is correct, then

$$\sqrt{n} Q(\hat{\theta}) \rightarrow N(0, \Sigma),$$

where Σ is defined in the proof below. A chi-squared based test statistic is

$$T_1 = n Q^T(\hat{\theta}) \hat{\Sigma}^{-1} Q(\hat{\theta}),$$

where $\hat{\Sigma}$ is the sample version of Σ . It can be shown that $T_1 \rightarrow \chi_r^2$.

Proof Denoting

$$q_{ij} = \eta(y_i, x_i, y_j, x_j, \hat{\theta}) - \eta(y_j, x_i, y_i, x_j, \hat{\theta}) R(y_i, x_i, y_j, x_j, \hat{\theta}),$$

we have

$$Q(\hat{\theta}) = \frac{2}{n(n-1)} \sum_{i < j} q_{ij}(\hat{\theta}).$$

We can obtain the following representation by Taylor's expansion,

$$Q(\hat{\theta}) = Q(\theta_0) + \frac{\partial Q(\theta_0)}{\partial \theta} (\hat{\theta} - \theta_0) + o_p(n^{-1/2}).$$

Therefore

$$\begin{aligned} Q(\hat{\theta}) &= Q(\theta_0) - \frac{\partial Q(\theta_0)}{\partial \theta} \left(\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \frac{\partial \ell_p(\theta_0)}{\partial \theta} + o_p(n^{-1/2}) \\ &= \left(I_{s \times s}, -\frac{\partial Q(\theta_0)}{\partial \theta} \left(\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \right) \left(\frac{\partial Q(\theta_0)}{\partial l_p(\theta_0)/\partial \theta} \right) + o_p(n^{-1/2}). \end{aligned}$$

We can prove that

$$\left(I_{s \times s}, -\frac{\partial Q(\theta_0)}{\partial \theta} \left(\frac{\partial^2 \ell_p(\theta_0)}{\partial \theta \partial \theta^T} \right)^{-1} \right) \rightarrow V_{s1} = \left(I_{s \times s}, -E \left(\frac{\partial q_{12}}{\partial \theta} \right) \left[E \left(\frac{\partial \psi_{12}}{\partial \theta} \right) \right]^{-1} \right)$$

and

$$\sqrt{n} \left(\frac{Q(\theta_0)}{\partial l_p(\theta_0)/\partial\theta} \right) = \sqrt{n} \frac{2}{n(n-1)} \sum_{i < j} \begin{pmatrix} q_{ij}(\theta_0) \\ \psi_{ij}(\theta_0) \end{pmatrix} \rightarrow N(0, V_{s2}),$$

where

$$V_{s2} = cov(\phi_{12}, \phi_{13}), \quad \phi_{12} = \begin{pmatrix} q_{12} \\ \psi_{12} \end{pmatrix}.$$

As a result

$$n^{1/2} Q(\hat{\theta}) \rightarrow N(0, \Sigma), \quad \Sigma = V_{s1} V_{s2} V_{s1}^T$$

and

$$n Q^T(\hat{\theta}) \Sigma^{-1} Q(\hat{\theta}) \rightarrow \chi_r^2.$$

As pointed out earlier, the pairwise conditional likelihood procedure is less efficient than the full likelihood procedure due to conditioning. Next we study the asymptotic relative efficiency (ARE) of the two procedures using simulations. The ARE is defined as the ratio of the asymptotic variances of estimators,

$$ARE = V^{-1} E[-\partial^2 \log f(y|x, \theta) / \partial\theta\partial\theta^T]^{-1},$$

where $-E[\partial^2 \log f(y|x, \theta) / \partial\theta\partial\theta^T]^{-1}$ is the asymptotic variance of θ based on the full likelihood and V is defined in Theorem 21.1 of Sect. 2.

Let $l = \sum_{i=1}^n \log f(y_i|x_i; \theta)$ be the full parametric log-likelihood. It can be decomposed as

$$\begin{aligned} (n-1)l &= \sum_{i < j} \log\{f(y_i, y_j|x_i, x_j; \theta)\} \\ &= \sum_{i < j} \log\{f(y_i, y_j|y_{(i)}, y_{(j)}, x_i, x_j)\} + \sum_{i < j} \log\{f(y_{(i)}, y_{(j)}|x_i, x_j)\}, \end{aligned} \quad (21.1.7)$$

where the first term is the logarithm of the pairwise conditional likelihood.

Example 1 Consider a regression model

$$y|x \sim N(x\beta, 1),$$

where for simplicity we assume that β is a scale parameter with the intercept set at zero. Then the generalized odds ratio is

$$R(y_1, x_1, y_2, x_2) = \exp\{(y_2 - y_1)(x_1 - x_2)\beta\}.$$

The full likelihood based score equation is

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^n x_i(y_i - x_i\beta) = 0.$$

The MLE is $\hat{\beta} = \sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2$ and the Fisher information is

$$\frac{\partial^2 l(\beta)}{\partial \beta^2} = - \sum_{i=1}^n x_i^2.$$

On the other hand the pairwise log-likelihood is

$$\ell_p(\beta) = - \sum_{i < j} \log\{1 + \exp[(y_j - y_i)(x_i - x_j)\beta]\}$$

and the pairwise score equation is

$$\frac{\partial \ell_p(\beta)}{\partial \beta} = - \sum_{i < j} \frac{(y_j - y_i)(x_i - x_j) \exp[(y_j - y_i)(x_i - x_j)\beta]}{1 + \exp[(y_j - y_i)(x_i - x_j)\beta]} = 0.$$

The second derivative is

$$\frac{\partial^2 \ell_p(\beta)}{\partial \beta^2} = - \sum_{i < j} \{(y_j - y_i)(x_i - x_j)\}^2 \frac{\exp\{(y_j - y_i)(x_i - x_j)\beta\}}{[1 + \exp\{(y_j - y_i)(x_i - x_j)\beta\}]^2} < 0.$$

Therefore $\partial \ell_p / \partial \beta$ is a monotonic function and has a unique root.

In Table 21.1 we report simulation results based on various sample sizes. We generated 1000 pseudo random samples of sizes 60 and 100 from $y|x \sim N(x\beta, 1)$ and $x \sim N(0, 1)$, for three values of $\beta = 0.2, 0.5, 1.0$. For each sample, we obtained two estimates of β : the full maximum likelihood estimator (method1) and the pairwise conditional likelihood estimator (method2). Table 21.1 shows the estimated mean and variance of each estimator, obtained from the simulation. We observe that the relative efficiency of the maximum pairwise conditional likelihood estimate to the maximum full likelihood estimate ranges from 24 to 79%. The efficiency decreases as the value of β increases.

Example 2 Consider the exponential model

$$f(y|x) = (\alpha + x\beta) \exp\{-(\alpha + x\beta)y\}, \quad y > 0$$

where $\alpha + x\beta > 0$. The generalized odds ratio is

$$R(y_1, x_1, y_2, x_2) = \exp[-(y_2 - y_1)(x_1 - x_2)\beta],$$

where the parameter α is cancelled out. The pairwise conditional likelihood based score equation has a similar form as in Example 1. In our simulation study, x was

Table 21.1 Summary of 1,000 simulations under Example 1: Mean: average estimates and Var: average estimated variance

Sample sizes	β	Method1		Method2	
		Mean1	Var1	Mean2	Var2
$n = 60$	0.2	0.205	0.017	0.226	0.024
$n = 60$	0.5	0.505	0.017	0.544	0.035
$n = 60$	1.0	1.005	0.017	1.069	0.070
$n = 100$	0.2	0.204	0.010	0.216	0.013
$n = 100$	0.5	0.504	0.010	0.526	0.018
$n = 100$	1.0	1.004	0.010	1.042	0.037

Table 21.2 Summary of 1,000 simulations under Example 2: Mean: average estimates and Var: average estimated variance

Sample sizes	β	Method1		Method2	
		Mean1	Var1	Mean2	Var2
$n = 60$	0.2	0.174	0.262	0.194	0.340
$n = 60$	0.5	0.469	0.338	0.527	0.439
$n = 60$	1.0	1.045	0.505	1.176	0.725
$n = 100$	0.2	0.193	0.150	0.207	0.180
$n = 100$	0.5	0.491	0.193	0.527	0.235
$n = 100$	1.0	1.015	0.286	1.091	0.378

generated from uniform(0, 1) and set $\alpha = 1$. For different choices of β , we tabulated the simulation results in Table 21.2. Again the replication is 1000. From this table, we see the relative efficiency of the maximum pairwise conditional likelihood estimate to the maximum full likelihood estimate ranges from 70 to 83%.

Example 3 Consider a discrete Poisson model

$$P(y|x) = (\alpha + x\beta)^y \exp[-(\alpha + x\beta)]/y! = \alpha^y (1 + x\gamma)^y \exp\{-\alpha(1 + x\gamma)\}, \quad y = 0, 1, 2, \dots$$

where $\gamma = \alpha/\beta$ and $\alpha(1 + x\gamma) > 0$. The full log-likelihood is

$$l(\alpha, \gamma) = \sum_{i=1}^n \{y_i \log \alpha + y_i \log(1 + x_i\gamma) - \alpha(1 + x_i\gamma)\}$$

and the score estimating equation is

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^n \frac{y_i}{\alpha} - (1 + x_i\gamma) = 0, \quad \frac{\partial l}{\partial \gamma} = \sum_{i=1}^n \frac{y_i x_i}{1 + x_i\gamma} - \alpha x_i = 0.$$

Table 21.3 Summary of 1,000 simulations under Example 3: Mean: average estimates and Var: average estimated variance

Sample sizes	γ	Method1		Method2	
		Mean1	Var1	Mean2	Var2
$n = 200$	0.2	0.235	0.103	0.242	0.112
$n = 200$	0.5	0.540	0.133	0.557	0.159
$n = 200$	1.0	1.077	0.260	1.111	0.313

The generalized odds ratio is

$$\begin{aligned} R(y_1, x_1, y_2, x_2) &= \exp[(y_2 - y_1)\{\log(\alpha + x_1\beta) - \log(\alpha + x_2\beta)\}] \\ &= \exp[(y_2 - y_1)\{\log(1 + x_1\gamma) - \log(1 + x_2\gamma)\}]. \end{aligned}$$

We can estimate the ratio $\gamma = \beta/\alpha$ based on the pairwise conditional likelihood. The corresponding score equation is

$$\frac{\partial \ell_p}{\partial \gamma} = \sum_{i < j} (y_j - y_i) \left\{ \frac{x_i}{1 + x_i\gamma} - \frac{x_j}{1 + x_j\gamma} \right\} \frac{\exp[(y_j - y_i)\{\log(1 + x_i\gamma) - \log(1 + x_j\gamma)\}]}{1 + \exp[(y_j - y_i)\{\log(1 + x_i\gamma) - \log(1 + x_j\gamma)\}]}.$$

The simulation results are reported in Table 21.3. Since both the full likelihood and pairwise conditional likelihood based scores have highly nonlinear forms, we increased the sample size in this setting to $n = 200$. From Table 21.3, the relative efficiency of the maximum pairwise conditional likelihood estimate to the maximum full likelihood estimate ranges from 83 to 91%.

In summary, the relative efficiency of the maximum pairwise conditional likelihood estimator to the MLE depends on the underlying models and values of the parameters. For example, the relative efficiency is relatively high in Examples 2 and 21.3, compared with that in Example 1. The simulation results suggest that the maximum pairwise conditional likelihood indeed produces consistent estimators. Hence it holds promise in situations where the model can only be specified partially.

Under partially specified regression models (21.1.1) and (21.1.2), the odds ratio $R(y_1, x_1, y_2, x_2) = [f(y_1|x_1)f(y_2|x_2)]/[f(y_2|x_1)f(y_1|x_2)]$ plays a key role. The pairwise conditional likelihood obtained by eliminating some nuisance parameters amounts to a binary regression likelihood.

In Kalbfleisch's full conditional likelihood (3.4) of Chap. 3, we can choose a random subset of P , say, P' and then consider the pseudo log-likelihood

$$l_P = a(\phi) \sum_{i=1}^n y_i U(\beta^T x_i) - \log \left[\sum_{j \in P'} \exp \{a(\phi) \sum_{l=1}^n y_{jl} U(\beta^T x_{jl})\} \right].$$

Estimation of β can be based on solving the pseudo-score equation $\partial l_P / \partial \beta = 0$, where

$$\frac{\partial l_P}{\partial \beta} = a(\phi) \sum_{i=1}^n y_i U'(\beta^T x_i) x_i - \frac{\sum_{j \in P'} \sum_{l=1}^n a(\phi) y_{jl} U'(\beta^T x_{jl}) x_{jl} \exp\{a(\phi) \sum_{l=1}^n y_{jl} U(\beta^T x_{jl})\}}{\sum_{j \in P'} \exp\{a(\phi) \sum_{l=1}^n y_{jl} U(\beta^T x_{jl})\}}.$$

This method can be implemented as long as the size P' is not too big. Although such estimators would be computationally more complex, it is reasonable to believe that they would produce more efficient estimators than the pairwise conditional likelihood. As an alternative, we may also construct triple or quadruple wise conditional likelihood.

21.2 Generalized Kendall's Tau for Testing Conditional Independence

Recently, the conditional independence assumption has been used widely in causal inference, genetics, economics, medical studies, machine learning graphical models as well as psychometrics. Let Y, X, S be random variables. Following Dawid (1979), we write

$$Y \perp S | X$$

to denote that Y is independent of S given X . This assumption is different from the familiar unconditional independent assumption $Y \perp S$. Let Y be the response variable and X be a covariate. Also denote the surrogate of X as S . A parametric model is assumed for Y given X

$$f(y|x) = f(y|x, \beta).$$

The conditional independence assumption is equivalent to

$$f(y|s, x) = f(y|x).$$

There are many test statistics such as Pearson's correlation, Spearman's rank correlation and Kendall's tau available for testing the unconditional independence between two random variables. For testing the conditional independence, a commonly used method is the partial correlation coefficient test. Let $g(x) = E(Y|X = x)$ and $h(x) = E(S|X = x)$. Define

$$\rho(Y, S|X) = \frac{\rho\{g(X), h(X)\} - \rho\{X, g(X)\}\rho(X, h(X))}{\sqrt{[1 - \rho^2\{X, g(X)\}][1 - \rho^2\{X, h(X)\}]}}$$

where $\rho(X, Y)$ denotes the Pearson correlation between X and Y . Alternatively

$$\rho(Y, S|X) = \frac{\rho(Y, S) - \rho(Y, X)\rho(S, X)}{\sqrt{[1 - \rho^2(Y, X)][1 - \rho^2(S, X)]}}.$$

In fact $\rho(Y, Z|X)$ equals the correlation between the errors in the regressions $y = g(x) + \epsilon_1$ and $s = h(x) + \epsilon_2$. Evaluation of this test requires estimation of the regression curves g and h . In practice, the popular choices of g and h are identity functions. Kendall's partial tau can be defined analogously, where the Pearson's ρ is replaced by Kendall's tau. Similarly it applies to the Spearman's correlation coefficient. However, as Korn (1984) pointed out, these partial correlations do not need to be consistent estimators of zero under the conditional independence. As a result, type I errors may be inflated.

Next we study a new test statistic. Define the generalized “odds ratio” as

$$R = \frac{f(y_j|x_i, \beta)f(y_i|x_j, \beta)}{f(y_i|x_i, \beta)f(y_j|x_j, \beta)}.$$

Under the conditional independent assumption between Y and S given X , using identity (21.1.3) we can show that

$$E[I\{(y_i - y_j)(s_i - s_j) > 0\} - I\{(y_i - y_j)(s_i - s_j) \leq 0\}R(y_i, x_i, y_j, x_j; \beta)] = 0.$$

The generalized Kendall's tau statistic is defined as

$$T(\beta) = \sqrt{n} \frac{1}{n(n-1)} \sum_{i < j} \xi(y_i, s_i, x_i, y_j, s_j, x_j; \beta), \quad (21.2.8)$$

where

$$\xi(y_i, s_i, x_i, y_j, s_j, x_j; \beta) = [I\{(y_i - y_j)(s_i - s_j) > 0\} - I\{(y_i - y_j)(s_i - s_j) \leq 0\}R(y_i, x_i, y_j, x_j; \beta)].$$

Note that if $x_i = x_j$, then $R = 1$. This becomes the conventional Kendall's tau.

In general β is an unknown parameter. Under H_0 , it can be estimated by using the conditional log-likelihood

$$\ell_c = \sum_{i=1}^n \log f(y_i|x_i, \beta).$$

Denote the maximum likelihood as $\hat{\beta}$.

Theorem 21.5 *Under some regularity conditions, in distribution*

$$T(\hat{\beta}) \rightarrow N(0, \sigma^2),$$

where σ^2 is defined in Eq. (21.2.9).

Proof Note that $\hat{\beta}$ satisfies

$$\frac{\partial \ell_c(\hat{\beta})}{\partial \beta} = 0.$$

Using Taylor's expansion, we have

$$\hat{\beta} - \beta_0 = \left[\frac{\partial^2 \ell_c(\beta_0)}{\partial \beta \partial \beta^T} \right]^{-1} \frac{\partial \ell_c(\beta_0)}{\partial \beta} + o_p(1),$$

or

$$\hat{\beta} - \beta_0 = J^{-1} \sum_{i=1}^n \frac{\partial \log f(y_i | x_i, \beta_0)}{\partial \beta} + o_p(1).$$

$$\begin{aligned} & \frac{2}{n(n-1)} \sum_{i < j} I\{(y_i - y_j)(s_i - s_j) \leq 0\} \partial R(y_i, x_i, y_j, x_j, \beta_0) / \partial \beta \\ & \rightarrow A = E[I\{(y_i - y_j)(s_i - s_j) \leq 0\} \partial R(y_i, x_i, y_j, x_j, \beta_0) / \partial \beta]. \end{aligned}$$

Therefore

$$\begin{aligned} T(\hat{\beta}) &= \sum_{i < j} [1 - I\{(y_i - y_j)(s_i - s_j) \leq 0\} \{1 + R_{ij}(\hat{\beta})\}] \\ &= \sum_{i < j} [1 - I\{(y_i - y_j)(s_i - s_j) \leq 0\} \{1 + R_{ij}(\beta_0)\}] - AJ^{-1} \sum_{i=1}^n \frac{\partial \log f(y_i | x_i, \beta_0)}{\partial \beta} + o_p(1). \end{aligned}$$

Let

$$\psi(y_i, y_j, x_i, x_j, s_i, s_j) = 1 - I\{(y_i - y_j)(s_i - s_j) \leq 0\} \{1 + R_{ij}(\beta_0)\}$$

and

$$\zeta(y_i, x_i, s_i) = E[\psi(y_i, y_j, x_i, x_j, s_i, s_j) | y_i, x_i, s_i].$$

By using Hoeffding's projection, we can show that

$$T(\hat{\beta}) = \sum_{i=1}^n \zeta(y_i, x_i, s_i) - AJ^{-1} \left[\frac{\partial \log f(y_i | x_i, \beta_0)}{\partial \beta} \right] + o_p(1).$$

Therefore

$$T(\hat{\beta}) \rightarrow N(0, \sigma^2),$$

$$\sigma^2 = E[\zeta_i - AJ^{-1} \partial \log f(y_i | x_i, \beta) / \partial \beta]^2. \quad (21.2.9)$$

Remark 1 If there are no covariates, or if $x_i = x_j$ for any i, j , then $R = 1$. The generalized Kendall's tau becomes Kendall's tau.

The p -value can be calculated by using the limiting distribution in this theorem or using a bootstrap method. Below is an outline of the bootstrap method.

Firstly, resample (s_i^*, x_i^*) randomly with replacement from $(s_1, x_1), \dots, (s_n, x_n)$ with equal probability $1/n$. Secondly, resample y_i^* from $f(y|x_i^*, \hat{\beta})$. Denote the bootstrap samples as $(y_i^{*b}, x_i^{*b}, s_i^{*b})$, $b = 1, 2, \dots, B$. Then calculate the generalized Kendall's tau statistic $T^{*b}(\hat{\beta}^{*b})$, $b = 1, 2, \dots, B$. Finally by comparing $|T(\hat{\beta})|$ with $|T^{*b}(\hat{\beta}^{*b})|$ we may find the p -value. Numerical study is given by Ji et al. (2017).

21.3 Applications in Graphical Models and High Dimensional Parameter Problems

Tan et al. (2016) and Ning et al. (2017) have applied the density ratio model (21.1.1) to handle high dimensional parameter problems. If the dimension of θ is p (and p is large) in the density ratio model (21.1.1). They have added a penalty function in Qin and Liang's (1999) and Liang and Qin's (2000) log pairwise conditional likelihood (21.1.5)

$$\ell_P(\theta) = - \sum_{i < j} \log\{1 + R(z_i, z_j; \theta)\} + \sum_{j=1}^p g_\lambda(\theta_j), \quad (21.3.10)$$

where $\lambda \geq 0$ is the turning parameter, and $g_\lambda(\cdot)$ is the penalty function. The popular choice, for example, is the lasso penalty $g_\lambda(\theta_j) = \lambda|\theta_j|$

In statistical physics, computer vision, data mining, and computational biology, graphical probability models have been used extensively recently. A graphical model is a probabilistic model in which a graph expresses the conditional dependence structure between random variables. It is mainly used to explore the interrelationship among a large number of random variables. The pairwise exponential graphical model is given by

$$f(x_1, \dots, x_p) = \left\{ \prod_{j=1}^p f_j(x_j) \right\} \exp \left\{ \sum_{i=1}^p \sum_{j \neq i} \theta_{ij} x_i x_j + C(\Theta) \right\},$$

where $f_j(x_j)$, $j = 1, 2, \dots, p$ may be treated as the (unknown) baseline densities, and $C(\Theta)$, depending on θ_{ij} , $1 \leq i \neq j \leq p$ and f_j , $j = 1, 2, \dots, p$, is the normalizing constant. Since $\exp(\theta_{ii} x_i^2)$ can be absorbed by $f_i(x_i)$, without loss of generality we assume $\theta_{ii} = 0$, $i = 1, 2, \dots, p$. In order to identify Θ we have to impose the constraint $\theta_{ij} = \theta_{ji}$. Each i ($i = 1, 2, \dots, p$) is called a node. Note that $\theta_{ij} = \theta_{ji} = 0$ implies that X_i and X_j or i -th and j -th nodes are conditionally independent given all the other nodes. If $\theta_{ij} = \theta_{ji} \neq 0$, then nodes i and j are connected, or there is an edge between i and j . We are interested in finding all edges, i.e., $\theta_{ij} \neq 0$, $i, j = 1, 2, \dots, p$.

Denote

$$x_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p).$$

The marginal density of X_{-i} is

$$\begin{aligned} & \left\{ \prod_{j \neq i} f_j(x_j) \right\} \exp \left\{ \sum_{j \neq i, k \neq i} \theta_{jk} \theta_{jk} x_j x_k \right\} \int f_i(x_i) \exp \left\{ x_i \left(\sum_{j \neq i} \theta_{ij} x_j + \sum_{j \neq i} \theta_{ji} x_i x_j \right) \right\} \exp(C(\Theta)) dx_i \\ &= \left\{ \prod_{j \neq i} f_j(x_j) \right\} \exp \left\{ \sum_{j \neq i, k \neq i} \theta_{jk} \theta_{jk} x_j x_k \right\} \prod_{j \neq i} a_j(x_j) \exp(C(\Theta)), \end{aligned}$$

where

$$a_j(x_j) = \int \exp\{x_i(\theta_{ij} + \theta_{ji})x_j\} f_i(x_i) dx_i, \quad j \neq i, \quad j = 1, 2, \dots, p.$$

Therefore the conditional density of $X_i | X_{-i} = x_{-i}$ is

$$f_i(x_i) \exp \left\{ \sum_{j \neq i} \theta_{ij} x_i x_j + \theta_{ji} x_i x_j \right\} \prod_{j \neq i} a_j(x_j).$$

This is a density ratio model (21.1.1). The pairwise likelihood (21.1.5) can be applied to eliminate $f_i(x_i) \prod_{j \neq i} a_j(x_j)$. Denote the pairwise likelihood based on $X_i | X_{-i}$ as $L_i(\Theta)$.

Finally, the overall pairwise likelihood is the product of all pairwise likelihoods, i.e.,

$$L = \prod_{i=1}^p L_i(\Theta).$$

To handle the high-dimension parameter problem, we can add a penalty function $g_\lambda(\Theta)$ and study the penalized log-likelihood

$$\ell = \sum_{i=1}^p \log L_i(\Theta) + \sum_{j=1}^p g_\lambda(\theta_j).$$

We refer readers to the works by Tan et al. (2016) and Ning et al. (2017) for details. Chen et al. (2015a,b) used spline method to model the baseline functions $f_i(x_i)$, $i = 1, 2, \dots, p$.

21.4 Profile Likelihood Approach for Continuous Covariate Density Ratio Model

Let us return to the density ratio model with continuous covariate discussed in the last section,

$$f(y|x) = \frac{f(y) \exp(yx\beta)}{\int f(y) \exp(yx\beta) dy},$$

where the baseline density $f(y)$ is not specified. Even though the pairwise conditional likelihood can be used to estimate β , it does not tell us any information on the baseline “carrier density” $f(y)$. Now we derive a nonparametric MLE (NPMLE) for $F(y)$.

Let $(y_i, x_i), i = 1, 2, \dots, n$ be the observed data. The likelihood is

$$L(\beta, F) = \prod_{i=1}^n \frac{\exp(y_i x_i \beta) dF(y_i)}{\int \exp(y x_i \beta) dF(y)}.$$

It can be shown that the nonparametric MLE for $F(y)$, denote as \hat{F} , has jumps only at each of the observed sample data points.

In fact if F does not jump at any one of y_i 's ($i = 1, 2, \dots, n$), then $dF(y_i) = 0$. This leads to a likelihood with zero value. On the other hand if F has an additional mass outside y_i 's, say, y^* , then the likelihood can be written

$$\begin{aligned} L^* &= \prod_{i=1}^n \frac{\exp(y_i x_i \beta) dF(y_i)}{\sum_{j=1}^n \exp(y_j x_i \beta) dF(y_j) + \exp(y^* x_i \beta) dF(y^*)} \\ &\leq \prod_{i=1}^n \frac{\exp(y_i x_i \beta) dF(y_i)}{\sum_{j=1}^n \exp(y_j x_i \beta) dF(y_j)} \\ &= \prod_{i=1}^n \frac{\exp(y_i x_i \beta) \{dF(y_i)/c\}}{\sum_{j=1}^n \exp(y_j x_i \beta) \{dF(y_j)/c\}}, \end{aligned}$$

where

$$c = \sum_{i=1}^n dF(y_i).$$

Therefore we have a larger likelihood by using a new distribution function with jumps $dF(y_i)/c$ at $y_i, i = 1, 2, \dots, n$. This shows that the nonparametric MLE of F has jumps only at the observed data points.

Let $dF(y_i) = p_i, i = 1, 2, \dots, n$. The nonparametric likelihood function is

$$L_n(\beta, p) = \prod_{i=1}^n \frac{p_i \exp(Y_i \beta^T X_i)}{\sum_{k=1}^n p_k \exp(Y_k \beta^T X_i)}. \quad (21.4.11)$$

We maximize $l_n(\beta, p) \equiv \log L_n(\beta, p)$ subject to the constraint such that $\sum_{k=1}^n p_k = 1$ to obtain the NPMLEs of (β, p) , denoted by $(\hat{\beta}_n, \hat{p})$. Consequently $F(y)$ can be estimated by $\hat{F}_n(y) = \sum_{k=1}^n I(Y_k \leq y) p_k$ and $I(\cdot)$ is an indicator function. It is easy to see that $(\hat{\beta}_n, \hat{p})$ exists since the nonparametric likelihood is bounded from above by one.

To maximize $l_n(\beta, p)$ subject to the constraint $\sum_{k=1}^n p_k = 1$, we consider the Lagrangian

$$H_n(\beta, p, \lambda) = l_n(\beta, p) - \lambda \left(\sum_{k=1}^n p_k - 1 \right),$$

where λ is the Lagrange multiplier. We take the derivative of H_n with respect to p_i and set it equal to 0,

$$\frac{\partial H_n(\beta, p, \lambda)}{\partial p_i} = \frac{1}{p_i} - \sum_{j=1}^n \frac{\exp(Y_i \beta^T X_j)}{\sum_{k=1}^n p_k \exp(Y_k \beta^T X_j)} - \lambda = 0.$$

Multiplying both sides by p_i , summing over i , and taking the constraint into account, we obtain

$$\sum_{i=1}^n \sum_{j=1}^n \frac{p_i \exp(Y_i \beta^T X_j)}{\sum_{k=1}^n p_k \exp(Y_k \beta^T X_j)} = n - \lambda.$$

By exchanging the summation indices i and j on the left hand side, it is easy to verify that $\lambda = 0$. This is not surprising since the conditional density $f(y|x)$ is invariant by replacing $f(y)$ with $f(y)/c$. It is not necessary to impose the constraint $\sum_{k=1}^n p_k = 1$.

Therefore

$$\hat{p}_i = \left\{ \sum_{j=1}^n \frac{\exp(Y_i \hat{\beta}_n^T X_j)}{\sum_{k=1}^n \hat{p}_k \exp(Y_k \hat{\beta}_n^T X_j)} \right\}^{-1}. \quad (21.4.12)$$

Based on this result, we use an iterative algorithm to compute the NPMLEs.

Step 1. Start with initial estimates $\beta^{(0)}$ and $p^{(0)}$.

Step 2. Insert $\beta^{(0)}$ and $p^{(0)}$ into the right-hand side of Eq. (21.4.12) to obtain $p^{(1)}$.

Step 3. Insert $p^{(1)}$ into $l_n(\beta, p)$ and maximize the parametric likelihood by solving the score equation

$$\frac{\partial l_n(\beta, p^{(1)})}{\partial \beta} = \mathbf{0}.$$

Step 4. Repeat steps 2 and 3 until the algorithm converges.

The convergence of the proposed algorithm was proved by Davidov and Iliopoulos (2013), which was subsequently improved by Chen (2015).

Alternatively, we use the following reparameterization to reduce the constrained optimization problem to an unconstrained optimization problem

$$p_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^n \exp(\alpha_j)},$$

where $\alpha_n = 0$. Note that this reparameterization has a nice connection with multinomial logit model where α_i can be treated as the intercept for the i th category. We then use the quasi-Newton algorithm to maximize the nonparametric likelihood over $(\beta, \alpha_1, \dots, \alpha_{n-1})$ simultaneously. This quasi-Newton algorithm requires only the input of the nonparametric log-likelihood function and the first partial derivatives of the nonparametric log-likelihood with respect to $(\beta, \alpha_1, \dots, \alpha_{n-1})$. It has been used successfully in maximizing the nonparametric likelihood in semiparametric transformation models with or without censoring. Furthermore, when there is only one binary covariate it can be shown that the NPMLE of β in the density ratio model is exactly the same as the MLE of the log-odds ratio in the logistic regression model fitting the probability of X given Y . In this case, we observe that the NPMLE of β obtained from the quasi-Newton algorithm is the same as the MLE of the log-odds ratio in the logistic regression model obtained from standard statistical software such as SAS and R, which provide an empirical validation of the quasi-Newton algorithm. The quasi-Newton algorithm is computationally efficient and converges very fast. The iterative algorithm also works well and yield the same parameter estimates as the quasi-Newton algorithm in general. However, the iterative algorithm may converge slowly or fail to converge especially when the true regression parameters are large. Moreover, the quasi-Newton algorithm is computationally more efficient than the iterative algorithm.

Luo and Tsai (2012) and Diao et al. (2012) presented some large sample results. Diao et al. (2012) carried out numerical comparisons between the profile likelihood method with the pairwise likelihood and triple likelihood methods. In general they found the loss of efficiency by using conditional approach is small. However, the advantage of the profile likelihood approach is that it can estimate the baseline distribution $F(y)$ as well, though it may be computationally intensive.

Luo and Tsai (2015) also used the estimating function approach by calculating

$$E(Y|x) = \frac{\int y \exp(yx\beta) dF(y)}{\int \exp(yx\beta) dF(y)},$$

where F is estimated by the maximum semiparametric likelihood estimate \hat{F} . Subsequently they estimated β by solving the estimating equation

$$\sum_{i=1}^n x_i \{y_i - \hat{E}(Y|x_i)\} = 0, \quad \hat{E}(Y|x_i) = \frac{\int y \exp(yx_i\beta) d\hat{F}(y)}{\int \exp(yx\beta) d\hat{F}(y)}.$$

In numerical simulations, they found the estimating function approach has similar finite sample performance as the pairwise approach. The estimating equation approach does not show any advantage over the pairwise or triple-wise based conditional likelihood method.

Chen (2007) and Chen et al. (2015a) studied related works by directly modelling the generalized odds ratio. Interested readers may find details from their papers.

Chapter 22

Non-ignorable Missing Data Problems

Biased sampling and non-ignorable missing data are the most difficult missing data problems. In contrast to missing at random, where the missing probability and underlying response model can be separately factored out in the likelihood function, in a non-ignorable missing data problem, they cannot be separated and must be handled simultaneously. Furthermore the underlying model may not be identifiable even in the full parametric setup. First we discuss the full parametric model case.

22.1 Model Identifiability Problem

In this section we discuss two cases: (1) One sample case, and (2) Regression model with covariates.

1. One sample case

Consider an one sample parametric biased sampling problem, where the density is given by

$$g(x) = \frac{w(x)f(x)}{\int w(x)f(x)dx}.$$

Even if both $w(x)$ and $f(x)$ are fully parametric models, the underlying parameters may not be identifiable without imposing strong restrictions.

Example 1 Suppose

$$w(x) = \exp(-x\theta), \quad \theta > 0, \quad f(x) = \lambda \exp(-x\lambda), \quad \lambda > 0, \quad x > 0.$$

Then

$$g(x) = \frac{\exp(-x\theta)\lambda \exp(-x\lambda)}{\int_0^\infty \exp(-x\theta)\lambda \exp(-x\lambda)dx} = (\theta + \lambda) \exp\{-x(\theta + \lambda)\}, \quad x > 0.$$

Based on the observable data from $g(x)$, it is impossible to identify θ and λ .

Example 2 Next we study a missing not at random or nonignorable missing data problem, where X is observable if and only if $D = 1$. Consider the following models:

Model 1.

$$X \sim N(1, 1), \quad P(D = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \quad (\alpha, \beta) = (-3/2, 1).$$

Model 2.

$$X \sim N(2, 1), \quad P(D = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \quad (\alpha, \beta) = (3/2, -1).$$

Conditional on $D = 1$, both models produce likelihoods proportional to

$$\frac{1}{\sqrt{2\pi}} \exp\{-(x-1)^2/2\} \frac{\exp(-3/2+x)}{1 + \exp(-3/2+x)} = \frac{1}{\sqrt{2\pi}} \exp\{-(x-2)^2/2\} \frac{\exp(3/2-x)}{1 + \exp(3/2-x)}.$$

The two examples show that in general the one sample biased sampling problem may have identifiable problem even if both the bias function and the underlying distribution are assumed to have full parametric models. Miao et al. (2016) discussed thoroughly the identification conditions when $f(x)$ is a normal density and $P(D = 1|x) = w(\alpha + x\beta)$, where $w(\cdot)$ is a known and strictly monotonic distribution function with support on $(-\infty, \infty)$.

A regularity condition. For any $\delta > 0$, $\lim_{x \rightarrow -\infty} w(x)/\exp(\delta x) = 0$ or ∞ .

This condition requires the left tail decay rate of the response probability to be not exponential. This condition is satisfied for example for a Probit missingness probability, i.e., $w(x) = \Phi(\alpha + x\beta)$. Unfortunately, the logistic missing probability does not satisfy this condition since

$$\lim_{x \rightarrow -\infty} \frac{\exp(x)}{1 + \exp(x)} \exp(-x) = 1.$$

The following results are due to Miao et al. (2016).

If $X \sim N(\mu, \sigma^2)$ and $P(D = 1|x) = w(\alpha + x\beta)$. Then

(a) σ^2 and $|\beta|$ are identifiable. (b) $\mu, \sigma^2, \alpha, \beta$ are identifiable if the sign of β is known. (c) $\mu, \sigma^2, \alpha, \beta$ are identifiable if the regularity condition holds.

The situation gets better if some x 's are missing, but the total sample size is available. Suppose the observed data are

$$(X_1 D_1, D_1), \dots, (X_N D_N, D_N),$$

where X_i is available if $D_i = 1$. However the information of N is available. The likelihood contribution is

$$\begin{aligned} L &= \prod_{i=1}^N [w(x_i) f(x_i)]^{I(D_i=1)} \left[1 - \int w(x) f(x) dx \right]^{I(D_i=0)} \\ &= \prod_{i=1}^n \frac{w(x_i) f(x_i)}{\int w(x) f(x) dx} \left[\int w(x) f(x) dx \right]^n \left[1 - \int w(x) f(x) dx \right]^{N-n}. \end{aligned}$$

Without loss of generality we assume the first n observations with $D_i = 1$ and the remaining $N - n$ observations with $D_i = 0$. Clearly from the binomial likelihood, n/N is an unbiased estimator of $P(D = 1) = \int w(x) f(x) dx$. This information can be utilized to identify β and λ in Example 1. Since there are two unknown parameters α and β in the weight function $w(x)$ in the second example, the information on N is still not good enough to identify both α and β .

2. Regression case

Consider a regression model with non-ignorable missing data. Let

$$\pi(x, y, \beta) = P(D = 1|x, y, \beta)$$

be the probability of observing Y . Moreover the conditional density of Y given x has a parametric model $f(y|x) = f(y|x\gamma)$. Based on the generic data (d, y, x) , the likelihood is

$$L = [\pi(x, y, \beta) f(y|x\gamma)]^d \left[\int \{1 - \pi(x, y, \beta)\} f(y|x\gamma) dy \right]^{1-d}.$$

Suppose there exists (β^*, γ^*) and (β, γ) such that

(1)

$$\pi(x, y, \beta^*) f(y|x\gamma^*) = \pi(x, y, \beta) f(y|x\gamma)$$

and

(2)

$$\int \{1 - \pi(x, y, \beta^*)\} f(y|x\gamma^*) dy = \int \{1 - \pi(x, y, \beta)\} f(y|x\gamma) dy.$$

In order to identify the underlying parameters from this likelihood, we need to find conditions such that (1) and (2) imply

$$(\beta^*, \gamma^*) = (\beta, \gamma).$$

Note that (2) is equivalent to

$$1 - \int \pi(x, y, \beta^*) f(y|x\gamma^*) dy = 1 - \int \pi(x, y, \beta) f(y|x\gamma) dy,$$

i.e.,

$$\int \pi(x, y, \beta^*) f(y|x\gamma^*) dy = \int \pi(x, y, \beta) f(y|x\gamma) dy.$$

Therefore (2) is determined by (1).

We only need to discuss the identifiability conditions based on data $(Y, x, D = 1)$. If

$$\pi(x, y, \beta^*) f(y|x\gamma^*) = \pi(x, y, \beta) f(y|x\gamma)$$

implies $\beta^* = \beta$ and $\gamma^* = \gamma$, then the model is identifiable.

Miao et al. (2016) gave a counter example for the logistic propensity and normal regression likelihood where the underlying intercept is not identifiable. Note that

$$\frac{1}{1 + \exp(-2 - x + 2y)} \frac{1}{\sqrt{2\pi}} \exp(-(y - 0.5x)^2/2) = \frac{1}{1 + \exp(2 + x - 2y)} \frac{1}{\sqrt{2\pi}} \exp(-(y - 2 - 0.5x)^2/2).$$

More generally, if

$$\pi(x, y, \beta) = \frac{1}{1 + \exp(\beta_0 + x\beta_1 + y\beta_2)}, \quad f(y|x\gamma, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y - \gamma_0 - x\gamma_1)^2/(2\sigma^2)\},$$

$$\pi(x, y, \beta^*) f(y|x\gamma^*, \sigma^*) = \pi(x, y, \beta) f(y|x\gamma, \sigma)$$

then

$$\frac{\pi(x, y, \beta)}{\pi(x, y, \beta^*)} = \frac{f(y|x\gamma^*, \sigma^*)}{f(y|x\gamma, \sigma)}$$

Equivalently

$$\begin{aligned} & \frac{1 + \exp(\beta_0^* + x\beta_1^* + y\beta_2^*)}{1 + \exp(\beta_0 + x\beta_1 + y\beta_2)} \\ &= \frac{\sigma}{\sigma^*} \exp[0.5\{\sigma^{-2} - (\sigma^*)^{-2}\}y^2 + y\{(\gamma_0^* + \gamma_1^*x)/(\sigma^*)^2 - (\gamma_0 + \gamma_1x)/(\sigma)^2\}] \\ & \quad \exp\{0.5(\gamma_0 + \gamma_1x)^2/\sigma^2 - 0.5(\gamma_0^* + \gamma_1^*x)^2/(\sigma^*)^2\}. \end{aligned}$$

(1) Since this is true for any (x, y) , clearly the coefficient of y^2 in the right hand side should be 0. This implies $\sigma = \sigma^*$, i.e. σ is identifiable.

(2) If x takes at least two different values, then coefficient of xy in the right hand side should be 0 since no xy appears in the left hand side. This implies $\gamma_1^*/(\sigma^*)^2 = \gamma/\sigma^2$. Together with (1), we have showed that both σ and γ_1 are identifiable.

Replacing σ^* and γ_1^* by σ and γ , respectively, we have a simplified version

$$\frac{1 + \exp(\beta_0^* + x\beta_1^* + y\beta_2^*)}{1 + \exp(\beta_0 + x\beta_1 + y\beta_2)} = \exp[y(\gamma_0^* - \gamma_0)/\sigma^2 + 0.5\{\gamma_0^2 - (\gamma_0^*)^2\}/\sigma^2 + x(\gamma_0 - \gamma_0^*)\gamma_1/\sigma^2].$$

Since this identity is true for any (x, y) , this is only possible if

$$\beta_0 = -0.5\{\gamma_0^2 - (\gamma_0^*)^2\}/\sigma^2, \quad \beta_1 = -(\gamma_0 - \gamma_0^*)\gamma_1/\sigma^2, \quad \beta_2 = -(\gamma_0^* - \gamma_0)/\sigma^2.$$

Finally we end up to sufficient and necessary conditions for the identifiability for the combination of logistic propensity score and normal regression models based on missing data.

Necessary and Sufficient Conditions

Let $\tau_1 = -2\beta_2(\beta_0 + \beta_2\gamma_0) + \beta_2^3\sigma^2$ and $\tau_2 = \beta_1 + \beta_2\gamma_1$. If either $\tau_1 \neq 0$ or $\tau_2 \neq 0$, then the underlying model is identifiable for $(\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \sigma^2)$.

Exercise 1 Again $f(y|x\gamma, \sigma)$ is given by the normal model but the propensity score is given by

$$\pi(x, y) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 y)},$$

where $\beta_2 \neq 0$. Discuss model identification conditions.

Exercise 2 Discuss model identifiability problems if the propensity score is given by a Probit model

22.2 Semiparametric Approaches for Non-ignorable Missing Data Problems

In this section we discuss some inference methods for missing not at random or non-ignorable missing data.

Let $D = 1$ or 0 be the indicator function on whether Y is observed or not. First we consider a parametric model for the underlying density $f(y|x) = f(y|x\beta)$. However the missing probability

$$P(D = 1|x, y) = P(D = 1|y) = \pi(y)$$

is left arbitrary except for the assumption that it is independent of x . The full likelihood is

$$L = \prod_{i=1}^n [\pi(y_i) f(y_i|x_i\beta)]^{d_i} [\int \{1 - \pi(y)\} f(y|x_i\beta) dy]^{1-d_i}.$$

It would be difficult to maximize this likelihood without specifying the form of $\pi(y)$.

1. Conditional approach

Qin and Liang (1999) and Liang and Qin (2000) discussed a pairwise conditional likelihood approach to eliminate the nuisance function $\pi(y)$.

Conditioning on the complete data and covariate x , we have

$$f(y|x, D=1) = \frac{P(D=1|x, y)f(y|x)}{\int P(D=1|x, y)f(y|x)dy} = \frac{\pi(y)f(y|x\beta)}{\int \pi(y)f(y|x\beta)dy} =: g(y|x).$$

Given the order statistics $Y_{(i)}$ and $Y_{(j)}$ for Y_i and Y_j ($i \neq j$), the pairwise conditional likelihood discussed in Sect. 21.1 implies

$$\frac{g(y_i|x_i\beta)g(y_j|x_j\beta)}{g(y_i|x_i\beta)g(y_j|x_j\beta) + g(y_j|x_i\beta)g(y_i|x_j\beta)} = \frac{f(y_i|x_i\beta)f(y_j|x_j\beta)}{f(y_i|x_i\beta)f(y_j|x_j\beta) + f(y_j|x_i\beta)f(y_i|x_j\beta)}.$$

It successfully eliminates the nuisance function $\pi(y)$. However this does not mean that β is estimable.

Example Suppose

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-0.5(y_i - \alpha - x_i\beta)/\sigma^2\}.$$

We can show that

$$\frac{f(y_i|x_i\beta)f(y_j|x_j\beta)}{f(y_i|x_i\beta)f(y_j|x_j\beta) + f(y_j|x_i\beta)f(y_i|x_j\beta)} = \exp\{(y_i - y_j)(x_i - x_j)\beta/\sigma^2\}.$$

In this example, only β/σ^2 is identifiable.

2. Profile likelihood approach

In Sect. 15.2, we discussed Chen (2001) profile likelihood approach for general outcome dependent sampling problems. Suppose

$$Y|x \sim f(y|x, \beta), \quad X \sim g(x).$$

In outcome dependent sampling, Y is sampled with a specific (known) density $h(y)$, followed by sampling X using the conditional density

$$X|Y \sim \frac{f(y|x, \beta)g(x)}{\int f(y|x, \beta)g(x)}.$$

Therefore the overall likelihood is

$$L = \left\{ \prod_{i=1}^n \frac{f(y_i|x_i, \beta)dG(x_i)}{\int f(y_i|x, \beta)dG(x)} \right\} \left\{ \prod_{i=1}^n h(y_i) \right\}.$$

We can use the first factor to estimate β and $G(x)$.

Returning to the missing data problem discussed in previous section. Conditioning on $D = 1, Y$,

$$X|D=1, Y=y \sim \frac{\pi(y)f(y|x, \beta)dG(x)}{\int \pi(y)f(y|x, \beta)dG(x)} = \frac{f(y|x, \beta)dG(x)}{\int f(y|x, \beta)dG(x)}.$$

Without loss of generality we assume $x_i, i = 1, \dots, n_1$ are available. We need to maximize

$$\prod_{i=1}^{n_1} \frac{f(y_i|x_i, \beta)p_i}{\sum_{j=1}^{n_1} p_j f(y_j|x_j, \beta)}$$

subject to the constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad p_i \geq 0, \quad dG(x_i) = p_i.$$

Essentially this is equivalent to the density ratio model in the continuous covariate case, where $g(x)$ is the baseline ‘‘carrier density’’ and $f(y|x, \beta)$ is the parametric link function between Y and X . Details were given in Chen (2001).

3. Pseudo likelihood approach

If the covariate X is not available for those $D_i = 0$, then the pairwise conditional likelihood is an ideal approach to eliminate the nuisance function $\pi(y)$. The profile likelihood method is also an efficient method. In practical applications, quite frequently X is available for every individual. Neither the pairwise conditional likelihood method nor the profile likelihood method uses x_i when $D_i = 0$. Consequently, both methods may lead to the loss of information.

In an elegant paper, Tang et al. (2003) proposed a conditional likelihood approach for those observed response variables. Observe

$$f(x|y, D=1) = \frac{P(D=1|y)f(y|x)dG(x)}{\int P(D=1|y)f(y|x)dG(x)} = \frac{f(y|x)dG(x)}{\int f(y|x)dG(x)},$$

which is irrelevant to the selection function $\pi(y) = P(D=1|y)$. Since in general all $x_i, i = 1, 2, \dots, n$ are available, we may use the marginal empirical distribution function $G_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$ to estimate $G(x)$. Therefore, the pseudo likelihood is

$$L_E = \prod_{i=1}^{n_1} \frac{f(y_i|x_i, \beta)}{\sum_{j=1}^n f(y_j|x_j, \beta)}.$$

It is expected that the pseudo-likelihood approach to be more informative than the pairwise conditional approach.

Example Consider the case where $Y|x$ has a normal density

$$f(y|x) \propto \exp\{-(y - \beta_0 - x\beta_1)^2/\sigma^2\} = \exp\{-x\beta_0\beta_1/\sigma^2 + (\beta_1/\sigma^2)yx - x^2\beta_1^2/(2\sigma^2)\} \exp(-y^2/\sigma^2).$$

Clearly $\exp(-y^2/\sigma^2)$ is cancelled out in the pseudo likelihood. As a consequence

$$(\beta_0\beta_1/\sigma^2, \beta_1/\sigma^2, \beta_1^2/\sigma^2)$$

are identifiable from the pseudo likelihood. If $\beta_1 \neq 0$, then the original parameters $(\beta_0, \beta_1, \sigma^2)$ are identifiable if x has a continuous density or takes at least three different values.

Tang et al. (2003) also extended their approach to longitudinal data problems. Zhao and Shao (2015) further generalized their approach by allowing $\pi(y)$ to depend on an instrumental variable.

4. Empirical likelihood method for the case without covariate

We consider a case without covariate, in which the missing probability is assumed to be completely known or known up to a set of parameters. However the distribution of Y is left arbitrary.

Leigh (1988) discussed a semiparametric inference method for the natural mortality of ocean fish from tagging experiments. N fish are tagged and released into the ocean. The fish are subject to natural mortality with a constant instantaneous rate θ , as well as to mortality due to fishing, with the fishing recapture times following a distribution function $F(t)$. The probability that a tagged fish will be recaptured within a given time t is then $\int_0^t \exp(-\theta u) dF(u)$, $t > 0$. Let t_1, \dots, t_n be the recapture times. The likelihood for the experiment is

$$L = \left\{ \prod_{i=1}^n \exp(-\theta t_i) dF(t_i) \right\} \left\{ 1 - \int_0^\infty \exp(-\theta t) dF(t) \right\}^{N-n}.$$

It can be written as

$$L = \prod_{i=1}^n \left\{ \frac{\exp(-\theta t_i) dF(t_i)}{\int \exp(-\theta t) dF(t)} \right\} \left\{ \int \exp(-\theta t) dF(t) \right\}^n \left\{ 1 - \int_0^\infty \exp(-\theta t) dF(t) \right\}^{N-n} =: L_1 L_2,$$

where the first term is the biased sampling likelihood, and the second term is the binomial likelihood. Based on L_1 alone, it is impossible to identify both θ and F . The binomial likelihood L_2 contributes an estimating equation $E[n] = N \int \exp(-\theta t) dF(t)$. The combination of L_1 and L_2 is sufficient to identify θ and F .

Leigh (1988) and Li and Qin (1998) discretized F at each of the observed time $t_i, i = 1, 2, \dots, n$. For simplicity denote $p_i = dF(t_i)$, $i = 1, 2, \dots, n$, and $\pi = \sum_{i=1}^n p_i \exp(-\theta t_i)$. The log-likelihood is

$$\ell = \sum_{i=1}^n \{-\theta t_i + \log p_i\} + (N - n) \log(1 - \pi)$$

subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n p_i \exp(-\theta t_i) = \pi.$$

After profiling out the p_i 's, the log-likelihood is

$$\ell = -n\bar{t} - \sum_{i=1}^n \log\{1 + \lambda(\exp(-\theta t_i) - \pi)\} + (N - n) \log(1 - \pi),$$

where the Lagrange multiplier λ is determined by

$$\sum_{i=1}^n \frac{\exp(-\theta t_i) - \pi}{1 + \lambda(\exp(-\theta t_i) - \pi)} = 0.$$

Next we can maximize ℓ with respect to θ . Li and Qin (1998) discussed this problem in details. Moreover, they extended this problem to a more general case where the total number of subjects is known in a left truncation problem.

22.3 Empirical Likelihood Method and Instrument Variable Approach

1. Approach I

Next we assume the missing probability is a given parametric model, say, for example,

$$P(D = 1|y, x) = P(D = 1|y) = \frac{1}{1 + \exp(\beta_1 + \beta_2 y)} =: \pi(y).$$

However we do not make any parametric assumption for the conditional density $f(y|x)$. We must assume Y and X are correlated to each other, otherwise the underlying model is not identifiable. We assume that there exist at least two x_1 and x_2 such that $f(y|x_1) \neq f(y|x_2)$. Let $g(x)$ be the marginal density of X . Based on the observed data the likelihood is

$$L = \prod_{i=1}^n \{\pi(y_i|\beta) f(y_i|x_i) g(x_i)\}^{D_i} \{[1 - \pi(y_i|\beta)] f(y|x_i) g(x_i) dy\}^{1-D_i}.$$

Let \mathcal{A}_i , $i = 1, 2, \dots, I$ be a partition of the covariate space, and

$$\rho_i = P(x \in \mathcal{A}_i).$$

Note that if the observed x_i 's are discrete with finitely many values, then each unique value of x forms a natural partition. The likelihood can be written as

$$L = \prod_{i=1}^I \prod_{j=1}^n \{\pi(y_j|\beta) f(y_j|x_j \in \mathcal{A}_i) \rho_i\}^{d_j I(x_j \in \mathcal{A}_i)} \left[\prod_{i=1}^I \{1 - \int \pi(y|\beta) f(y|x_j \in \mathcal{A}_i) dy\} \rho_i \right]^{(1-d_j) I(x_j \in \mathcal{A}_i)}.$$

Denote

$$\theta_i = \int \pi(y\beta) f(y|x \in \mathcal{A}_i) dy, \quad i = 1, 2, \dots, I$$

and

$$f_i(y) = f(y|x \in \mathcal{A}_i), \quad i = 1, 2, \dots, I.$$

Let

$$y_{ij}, j = 1, 2, \dots, r_i$$

be those observed y_j 's with $d_j = 1, x_j \in \mathcal{A}_i, i = 1, 2, \dots, I$. Also let

$$s_i = \sum_{j=1}^n (1 - d_j) I(x_j \in \mathcal{A}_i), \quad r_i = \sum_{j=1}^n d_j I(x_j \in \mathcal{A}_i), \quad i = 1, 2, \dots, I.$$

The likelihood can be written as

$$L = \left[\prod_{j=1}^n \{\pi(y_j\beta)\}^{d_j} \right] \left\{ \prod_{i=1}^I \prod_{j=1}^{n_i} f_i(y_{ij})(1 - \theta_i)^{s_i} \right\} \left\{ \prod_{i=1}^I \rho_i^{n_i} \right\}.$$

The log-likelihood is

$$\sum_{i=1}^n d_i \log \pi(y_i\beta) + \sum_{i=1}^I \sum_{j=1}^{n_i} \log dF_i(y_{ij}) + \sum_{i=1}^I s_i \log(1 - \theta_i) + \sum_{i=1}^I n_i \log \rho_i.$$

Denote

$$p_{ij} = dF(y_{ij}).$$

We need to maximize the above log-likelihood subject to the constraints

$$\sum_{j=1}^{n_i} p_{ij} \{\pi(y_{ij}) - \theta_i\} = 0, \quad \sum_{j=1}^{n_i} p_{ij} = 1.$$

After profiling out p_{ij} 's we have the log profile likelihood

$$\ell = \sum_{i=1}^n d_i \log \pi(y_i\beta) - \sum_{i=1}^I \sum_{j=1}^{r_i} \log [1 + \lambda_i \{\pi(y_{ij}) - \theta_i\}] + \sum_{i=1}^I s_i \log(1 - \theta_i) + \sum_{i=1}^I n_i \log \rho_i,$$

where the Lagrange multipliers λ_i 's are determined by

$$\sum_{j=1}^{n_i} \frac{\pi(y_{ij}) - \theta_i}{1 + \lambda_i \{\pi(y_{ij}) - \theta_i\}} = 0.$$

We can maximize this empirical likelihood with respect to the underlying parameters. This approach may need a large sample size to guarantee enough samples in each \mathcal{A}_i .

2. Approach II

The following approach was discussed by Guan and Qin (2016).

Let

$$h(y, x) = P(Y = y, X = x | D = 0) = \frac{\{1 - \pi(y|\beta)\}dF(y, x)}{1 - \eta}$$

be the joint density of Y, X for $D = 0$, where $\eta = P(D = 1) = \int \pi(y|\beta)dF(y, x)$. The corresponding cumulative distribution of (Y, X) is denoted by $H(y, x)$. Since Y is not available when $D = 0$, the marginal density is

$$h(x|\beta) = P(X = x | D = 0) = \frac{\int \{1 - \pi(y|\beta)\}f(y|x)dyg(x)}{1 - \eta}.$$

Without loss of generality we assume that the first n_1 observations have complete values of Y, X and the remaining n_0 observations have X value alone ($n = n_0 + n_1$). The likelihood can be written as

$$L = \eta^{n_1}(1 - \eta)^{n_0} \prod_{i=1}^{n_1} \frac{\pi(y_i|\beta)f(y_i, x_i)}{\eta} \prod_{i=n_1+1}^n h(x_i, \beta),$$

where the first term is the binomial likelihood, the second term is the conditional likelihood of (Y, X) conditioning on $D_i = 1$, and the third term is the conditional likelihood of X conditioning on $D_i = 0$.

Note that for any measurable function of X , $\phi(X, \beta)$,

$$E_H[\phi(X, \beta)] = \frac{\int \phi(x, \beta)\{1 - \pi(y|\beta)\}f(y, x)dydx}{1 - \eta} = (1 - \eta)^{-1} E_F[\phi(X, \beta)\{1 - \pi(Y|\beta)\}]. \quad (22.3.1)$$

Since this estimating equation is true for any function of X and β , we expect it to provide information for the underlying parameter β . Moreover some X_i 's from H are available directly, and some (Y_i, X_i) 's from F are available indirectly (a biased version of F). Naturally we can construct two empirical likelihoods to utilize the constraint estimating Eq. (22.3.1) effectively,

$$\ell = \sum_{i=1}^n \{(1-d_i)\log(1-\eta) + d_i \log \pi(y_i|\beta)\} + \sum_{i=1}^{n_1} \log dF(y_i, x_i) + \sum_{j=n_1+1}^n \log dH(x_j).$$

Let

$$p_j = dF(y_j, x_j), \quad j = 1, 2, \dots, n_1, \quad q_j = dH(x_j), \quad j = n_1 + 1, \dots, n,$$

$$\ell = \sum_{i=1}^n \{(1 - d_i) \log(1 - \eta) + d_i \log \pi(y_i|\beta)\} + \sum_{i=1}^{n_1} \log p_i + \sum_{j=n_1+1}^n \log q_j,$$

where the complete observations $(Y_i, X_i, D_i = 1), i = 1, 2, \dots, n_1$ and incomplete observations $X_i, D_i = 0, i = n_1 + 1, \dots, n$ are linked by the constraints $E_F[\psi(y, x)] = 0$, where

$$\psi(y, x) = \begin{pmatrix} \pi(y|\beta) - \eta \\ (1 - \pi(y|\beta))\phi(x) - \theta(1 - \eta) \\ \mu(x) - \hat{\mu} \end{pmatrix}$$

and $\theta = E_H\phi(X)$. Note the last constraint is based on the consideration of the mean estimation of $\mu = E(Y) = E(E(Y|X))$, where $\mu(x)$ is a “working” regression function of $Y|x$ and $\hat{\mu} = n^{-1} \sum_{i=1}^n \mu(x_i)$. The log profile empirical likelihood is

$$\ell = n_0 \log(1 - \eta) + \sum_{i=1}^{n_1} \log \pi(y_i|\beta) - \sum_{j=1}^{n_1} \log[1 + \lambda^T \psi(y_j, x_j)] - \sum_{j=n_1+1}^n \log[1 + \nu^T (\phi(x_j) - \theta)],$$

where the Lagrange multipliers λ and ν are determined by

$$\sum_{j=1}^{n_1} \frac{\psi(y_j, x_j)}{1 + \lambda^T \psi(y_j, x_j)} = 0, \quad \sum_{j=n_1+1}^n \frac{\phi(x_j) - \theta}{1 + \nu^T (\phi(x_j) - \theta)} = 0.$$

We can maximize ℓ with respect to β, θ, η ,

$$\begin{aligned} \frac{\partial \ell}{\partial \eta} &= -\frac{n_0}{1 - \eta} - \sum_{j=1}^{n_1} \frac{\lambda^T \partial \psi(y_j, x_j)/\partial \eta}{1 + \lambda^T \psi(y_j, x_j)} = 0, \\ \frac{\partial \ell}{\partial \theta} &= -\sum_{j=1}^{n_1} \frac{\lambda^T \partial \psi(y_j, x_j)/\partial \theta}{1 + \lambda^T \psi(y_j, x_j)} + \sum_{j=n_1+1}^n \frac{\nu}{1 + \nu^T (\phi(x_j) - \theta)} = 0, \\ \frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{n_1} \frac{\partial \log \pi(y_i|\beta)}{\partial \beta} - \sum_{j=1}^{n_1} \frac{\lambda^T \partial \psi(y_j, x_j)/\partial \beta}{1 + \lambda^T \psi(y_j, x_j)} - \sum_{j=n_1+1}^n \frac{\nu^T \partial \phi(x_j)/\partial \beta}{1 + \nu^T (\phi(x_j) - \theta)} = 0. \end{aligned}$$

We now consider a special case for the choice of ϕ . Let $\mathcal{A}_i, i = 1, 2, \dots, I$ be a partition of the X space and

$$\theta_i = E_H[I(X \in \mathcal{A}_i)].$$

Let

$$m_i = \sum_{j=1}^n (1 - D_j) I(X_j \in \mathcal{A}_i), \quad i = 1, 2, \dots, I.$$

In this special case

$$\psi(y) = \begin{pmatrix} (1 - \pi(y\beta))I(X \in \mathcal{A}_1) - \theta_1(1 - \eta) \\ \dots \\ (1 - \pi(y\beta))I(X \in \mathcal{A}_{I-1}) - \theta_{I-1}(1 - \eta) \\ \pi(y\beta) - \eta \\ \mu(x) - \hat{\mu} \end{pmatrix}.$$

After profiling out $p_i = dF(y_i, x_i)$, $i = 1, 2, \dots, n_1$, the semiparametric log-likelihood is

$$\ell = n_0 \log(1 - \eta) + \sum_{i=1}^{n_1} \log \pi(y_i \beta) + \sum_{i=1}^I m_i \log(1 - \theta_i) - \sum_{j=1}^{n_1} \log\{1 + \lambda^T \psi(y_i, x_i)\},$$

where λ is the Lagrange multiplier determined by

$$\frac{1}{n_1} \sum_{j=1}^{n_1} \frac{\psi(y_j, x_j)}{1 + \lambda^T \psi(y_j, x_j)} = 0,$$

$$p_i = \frac{1}{n_1} \frac{1}{1 + \lambda^T \psi(y_i, x_i)}, \quad i = 1, 2, \dots, n_1,$$

$$\frac{\partial \ell}{\partial \theta_k} = \frac{n_k}{\theta_k} - \frac{n_I}{1 - \theta_+} + n_1 \lambda_i = 0, \quad i = 1, 2, \dots, I - 1.$$

It is expected that if $I = I(n)$, the number of groups, depends on the sample size n and $I(n) \rightarrow \infty$ as $n \rightarrow \infty$, then the proposed estimator utilizes the auxiliary information (22.3.1) effectively. In practical applications, however, a compromise with numerical stability requires limiting the number of constraints, because the finite sample properties of the algorithm will be adversely affected by a large number of constraints.

Exercise Derive the large sample results.

3. Instrumental variable approach

In order to identify the underlying model in the missing not at random case or non-ignorable missing data case, a popular method in economic literature is the instrumental variable approach. An instrumental variable is a variable that is independent of the missing indicator but correlated with the response variable. More specifically, suppose the covariate X can be written as $X = (X_1, X_2)$, then X_2 is called an instrumental variable if

$$P(D = 1|x, y) = \pi(x_1, y), \quad f(y|x) = f(y|x_1, x_2).$$

In statistical literature, we can either specify a model (1) $f(y|x) = f(y|x\beta)$ and leave $\pi(x_1, y)$ arbitrary or (2) specify a model $\pi(x_1, y) = \pi(x_1, y, \theta)$ and leave $f(y|x)$ unknown. In the later case, a popular approach is based on

$$E \left[\phi(x) \left(\frac{D}{\pi(x_1, y, \theta)} - 1 \right) \right] = 0$$

for a user specified function $\phi(x)$ to estimate θ . Then the method of moments can be applied. Interested readers may refer the work by Zhao and Shao (2015) for case (2). In applications, however, it may not be so easy to identify an instrumental variable. Moreover in general this approach has low efficiency.

22.4 Maximum Likelihood Estimation in Call-Back Problem

In marketing researches, social sciences and epidemiological studies, call-back of nonrespondents is standard. If respondents and nonrespondents tend to give different answers, the missing data are called nonignorable, and using them alone may produce biased results. Consider the following call back model originally proposed by Alho (1990a) and discussed by Wood et al. (2006) and Qin and Follmann (2014).

Denote X as the variable of interested. Let p_1 be the probability that an individual responds at the first mailing (or call). For $j = 2, 3, \dots, m$, let p_j be the probability that an individual responds at the j -th mailing (or call), given that he or she has not previously responded. Assume a logistic regression for each response probability

$$p_j = \frac{\exp(\alpha_j + x^T \beta)}{1 + \exp(\alpha_j + x^T \beta)}, \quad j = 1, 2, \dots, m,$$

where β is the common slope and α_j is the time dependent intercept for $j = 1, 2, \dots, m$.

The unconditional response probabilities are

$$\mu_1(x) = p_1(x), \quad \mu_2(x) = p_2(x)\{1 - p_1(x)\}, \quad \dots, \quad \mu_m(x) = p_m(x) \left[\prod_{j=1}^{m-1} \{1 - p_j(x)\} \right].$$

Let

$$\mu_{m+1}(x) = 1 - \sum_{j=1}^m \mu_j(x)$$

be the probability of not responding at all.

Also let

$$v_{ij} = \mu_j(x_i) / \{1 - \mu_{m+1}(x_i)\}, \quad j = 1, 2, \dots, m; i = 1, 2, \dots, n$$

be the conditional probability that he or she responds at the j -th mailing given he or she is a responder. Let u_{ij} be an indicator, being 1 if subject i responded at the j -th mailing and 0 otherwise. Let $R_i = 1$ if subject i is a responder at the m mailing and 0 otherwise. Then conditional on the responders, the conditional likelihood is

$$L_c(\alpha, \beta) = \prod_{R_i=1} v_{i1}^{u_{i1}} \cdots v_{im}^{u_{im}}.$$

Thus the conditional log-likelihood is

$$\begin{aligned} \ell_c(\alpha, \beta) &= \sum_{R_i=1}^m \sum_{j=1}^m u_{ij} \log(v_{ij}) \\ &= \sum_{R_i=1} \left[\sum_{j=1}^m u_{ij} \left\{ \alpha_j - \sum_{k=1}^j \log(1 + \exp(\alpha_k + x_i^T \beta)) \right\} \right. \\ &\quad \left. - \log \sum_{h=1}^m \left\{ \exp(\alpha_h) / \prod_{l=1}^h (1 + \exp(\alpha_l + x_i^T \beta)) \right\} \right]. \end{aligned}$$

Since the conditional log-likelihood does not use the number of non-responders, it does not have a unique maximum. By using this information, Alho (1990a) used m additional estimating equations to estimate α , given β .

$$\alpha_j = -\log \left\{ (n - n_1 - \dots - n_j) / \sum_{i \in I_j} \exp(-x_i^T \beta) \right\}, \quad j = 1, 2, \dots, m,$$

where I_j is the set of individuals responding at the j -th attempt. Based on those responders $R_i = 1$, Alho (1990a) proposed the following Horwitz and Thompson estimator of $v = E(X)$

$$\tilde{v} = \sum_{i=1}^n \frac{R_i x_i}{1 - \mu_{m+1}(x_i, \tilde{\alpha}, \tilde{\beta})}.$$

Next we discuss the maximum semiparametric likelihood estimates of $E(X)$, $F(x)$ and regression parameters β and $\alpha_1, \dots, \alpha_m$.

Denote

$$X_{jk}, \quad j = 1, 2, \dots, m; k = 1, 2, \dots, n_j$$

as the responses at the j -th mailing ($j = 1, 2, \dots, m$) and let n_{m+1} be the number of individuals who never responds. Therefore $n = n_1 + \dots + n_m + n_{m+1}$. The full likelihood is

$$L = \left[\prod_{j=1}^m \prod_{k=1}^{n_j} \mu_j(x_{jk}) dF(x_{jk}) \right] \left[\int \mu_{m+1}(x) dF(x) \right]^{n_{m+1}}.$$

By conditioning on $R_i = 1$, the likelihood can be decomposed as

$$L = \left\{ \prod_{j=1}^m \prod_{k=1}^{n_j} \frac{\mu_j(x_{jk})}{1 - \mu_{m+1}(x_{jk})} \right\} \left\{ \prod_{j=1}^m \prod_{k=1}^{n_j} \frac{\{1 - \mu_{m+1}(x_{jk})\} f(x_{jk})}{\int \{1 - \mu_{m+1}(x)\} f(x) dx} \right\} \\ \left[\left\{ 1 - \int \mu_{m+1}(x) f(x) dx \right\}^{n-n_{m+1}} \left\{ \int \mu_{m+1}(x) f(x) dx \right\}^{n_{m+1}} \right],$$

where the first term is the conditional probability of response at the j -th mailing given X_i and $R_i = 1$, the second is the marginal likelihood of X_i , $i = 1, 2, \dots, n$ for responders, and the third term is the binomial probability of responders or non-responders within the m mailings.

The first factor may be used to make inference for the underlying parameters α_j, β , $j = 1, 2, \dots, m$. However, as Alho (1990a) pointed out the conditional log-likelihood does not have a unique maximum since the parameters α_j, β , $j = 1, 2, \dots, m$ cannot be identified from this likelihood.

Next we study the maximum semiparametric likelihood estimate.

Let

$$\theta_j = \int \mu_j(x) f(x) dx, \quad j = 1, 2, \dots, m,$$

be the probability of response at the i -th mailing. Let

$$\mu(x) = (\mu_1(x), \dots, \mu_m(x))^T, \quad \theta = (\theta_1, \dots, \theta_m)^T.$$

Denote

$$q_{jk} = dF(x_{jk}), \quad j = 1, 2, \dots, m; k = 1, 2, \dots, n_j.$$

We need to maximize the log-likelihood

$$\ell = \sum_{j=1}^m \sum_{k=1}^{n_j} \log \mu_j(x_{jk}, \alpha, \beta) + n_{m+1} \log(1 - \theta_+) + \sum_{j=1}^m \sum_{k=1}^{n_j} \log q_{jk}$$

subject to the constraints

$$\sum_{j=1}^m \sum_{k=1}^{n_j} q_{jk} = 1, \quad q_{jk} \geq 0$$

and

$$\sum_{j=1}^m \sum_{k=1}^{n_j} q_{jk} \{\mu_i(x_{jk}) - \theta_i\} = 0, \quad i = 1, 2, \dots, m,$$

where $\theta_+ = \theta_1 + \dots + \theta_m$. After profiling q_{jk} 's,

$$q_{jk} = \frac{1}{n_+} \frac{1}{1 + \lambda^T \{\mu(x_{jk}) - \theta\}}, \quad j = 1, 2, \dots, m; k = 1, 2, \dots, n_j,$$

where $n_+ = n_1 + \dots + n_m$ and λ is the Lagrange multiplier determined by

$$\sum_{j=1}^m \sum_{k=1}^{n_j} \frac{\mu(x_{jk}) - \theta}{1 + \lambda^T \{\mu(x_{jk}) - \theta\}} = 0.$$

Then the profiled log-likelihood is

$$\begin{aligned} \ell &= \sum_{j=1}^m \sum_{k=1}^{n_j} \log \mu_j(x_{jk}, \alpha, \beta) \\ &\quad - \sum_{j=1}^m \sum_{k=1}^{n_j} \log [1 + \lambda^T \{\mu(x_{jk}, \alpha, \beta) - \theta\}] - n_{m+1} \log \{1 - \theta_+\}. \end{aligned}$$

$$\frac{\partial \ell}{\partial \theta} = \sum_{j=1}^m \sum_{k=1}^{n_j} \frac{\lambda}{1 + \lambda^T \{\mu(x_{jk}, \alpha, \beta) - \theta\}} - \frac{n_{m+1}}{1 - \theta_+} I_m = 0, \quad I_m = (1, 1, \dots, 1)^T$$

or

$$n_+ \lambda = \frac{n_{m+1}}{1 - \theta_+} I_m,$$

or

$$\lambda = \frac{n_{m+1}}{n_+} \frac{1}{1 - \theta_+} I_m.$$

Let $\eta = (\alpha, \beta)^T$.

$$\frac{\partial \ell}{\partial \eta} = \sum_{j=1}^m \sum_{k=1}^{n_j} \frac{\partial \log \mu_j(x_{jk})}{\partial \eta} - \sum_{j=1}^m \sum_{k=1}^{n_j} \frac{\lambda^T \partial \mu / \partial \eta}{1 + \lambda^T \{\mu(x_{jk}, \alpha, \beta) - \theta\}} = 0.$$

The distribution $F(x)$ can be estimated by

$$\hat{F}(x) = \sum_{j=1}^m \sum_{k=1}^{n_j} \hat{q}_{jk} I(x_{jk} \leq x) = \frac{1}{n_+} \sum_{j=1}^m \sum_{k=1}^{n_j} \frac{I(x_{jk} \leq x)}{1 + \hat{\lambda}^T \{\mu(x_{jk}, \hat{\eta}) - \hat{\theta}\}}.$$

Finally $E(X)$ can be estimated by

$$\frac{1}{n_+} \sum_{j=1}^m \sum_{k=1}^{n_j} \frac{x_{jk}}{1 + \hat{\lambda}^T \{\mu(x_{jk}, \hat{\eta}) - \hat{\theta}\}}.$$

Detailed derivations and numerical results can be found in Qin and Follmann (2014).

22.5 Heckman's Sample Selection Model

In his seminal paper, Heckman (1979) studied the sample selection bias as a specification error. It is widely believed that among his many other outstanding contributions in econometrics, he got the Nobel prize for this paper.

Consider two latent dependent variables wage model

$$Y_{1i} = X_{1i}\beta + \epsilon_{1i}, \quad Y_{2i} = X_{2i}\gamma + \epsilon_{2i},$$

where Y_{1i} is the i -th individual's wage and X_{1i} is the corresponding productivity covariate. Y_{1i} is observed only for workers, i.e., only people in work receive a wage.

Y_{2i} is the difference between the wage and the reservation wage W_i for the i -th individual. The reservation wage is the minimum wage at which the i -th individual is prepared to work. If the wage is below that he/she chooses not to work. We observe only an indicator variable for employment defined as $D_i = 1$ if $Y_{2i} > 0$ and 0 otherwise. X_{2i} is the baseline covariate which is always observed. Note that X_{1i} is observable only if $D_i = 1$. It is possible that X_{1i} and X_{2i} may share some common variables.

Heckman (1979) assumed $(\epsilon_{1i}, \epsilon_{2i})$ s are an i.i.d. sample from a bivariate normal distribution with zero mean and covariate matrix Σ .

The likelihood function is

$$L = \prod_{i=1}^n \{P(Y_{2i} < 0 | x_{2i})\}^{I(d_i=0)} [P(Y_{2i} > 0 | x_{2i}) P(Y_{1i} = y_{1i} | Y_{2i} > 0, x_{1i}, x_{2i})]^{I(d_i=1)}.$$

Note that

$$P(Y_{2i} < 0 | x_{2i}) = P(\epsilon_{2i} < -x_{2i}\gamma) = \Phi(-x_{2i}\gamma/\sigma_2) = 1 - \Phi(x_{2i}\gamma/\sigma_2).$$

$$\begin{aligned}
& P(Y_{1i} = y_{1i} | Y_{2i} > 0, x_{1i}, x_{2i}) \\
&= \frac{P(Y_{1i} = y_{1i}, Y_{2i} > 0 | x_{1i}, x_{2i})}{P(Y_{2i} > 0 | x_{2i})} \\
&= \frac{P(Y_{1i} = y_{1i} | x_{1i}) P(Y_{2i} > 0 | Y_{1i} = y_{1i}, x_{1i}, x_{2i})}{\Phi(X_{2i}\gamma/\sigma_2)} \\
&= \frac{1}{\sigma_1} \phi((y_{1i} - x_{1i}\beta)/\sigma_1) \Phi\left(\frac{x_{2i}\gamma + \sigma_2\sigma_1^{-1}\rho(y_{1i} - x_{1i}\beta)}{\sqrt{(1-\rho^2)\sigma_2^2}}\right) \frac{1}{\Phi(X_{2i}\gamma/\sigma_2)},
\end{aligned}$$

where $\phi(z) = d\Phi(z)/dz$ is the standard normal density. Since we only observe $Y_{2i} > 0$ or < 0 , only γ/σ_2 is estimable. Without loss of generality we assume $\sigma_2 = 1$. On the other hand if β and γ have at least one common element, then σ_2 is identifiable.

Using the joint normality of Y_{1i} and Y_{2i} , we can show that

$$E[Y_{1i} | Y_{2i} > 0, x_{1i}, x_{2i}] = x_{1i}\beta + \sigma_1\rho\lambda(x_{2i}\gamma), \quad (22.5.2)$$

where λ is the inverse of Mill's ratio function defined as

$$\lambda(z_i) = \frac{\phi(z_i)}{1 - \Phi(Z_i)}, \quad z_i = -x_{2i}\gamma/\sigma_2.$$

It is a monotonic decreasing function.

Some commonly used approaches are given below.

1. Maximum likelihood method

Directly maximizing the likelihood can be implemented numerically by carefully programming. As an alternative, an EM algorithm can be applied.

2. Two-stage estimation method

Using the likelihood based on the observations $I(Y_{2i} > 0)$'s,

$$L_M = \prod_{i=1}^n [\Phi(x_{2i}\gamma)]^{I(Y_{2i}>0)} [1 - \Phi(x_{2i}\gamma)]^{I(Y_{2i}<0)},$$

we may estimate γ first. Then the method of least squares can be utilized to estimate β in the model $E[Y_{1i} | Y_{2i} > 0] = x_{1i}\beta + \sigma_1\rho\lambda(x_{2i}\gamma)$. Since

$$Var(Y_1 | Y_2 > 0, x_1, x_2) = \sigma_1^2 - \rho\sigma_1[x_2\gamma\lambda(x_2\gamma) + \lambda^2(x_2\gamma)]$$

depends on x_1 and x_2 , to estimate β , we may use Godambe's optimal estimating function theory discussed in Chap. 4.

3. Error distribution generalization

Heckman's selection bias sampling model depends on the joint normal distribution assumption. Marchenko and Genton (2012) generalized the normal error distribution to the t -distribution. Computation, however, is complicated. Nevertheless the EM algorithm can be employed.

4. Semiparametric generalization

Motivated by (22.5.2), Newey (2009) discussed a partial linear model

$$E[Y_{1i}|Y_{2i} > 0, x_i] = x_{1i}\beta + \lambda(x_{2i}\gamma),$$

where the form of $\lambda(\cdot)$ is not specified. A consistent estimator of β was derived by using series expansion of λ ,

5. Puhani's approach

Puhani (2000) noted that the inverse Mill ratio is monotonic decreasing and almost linear except for the right tail. Consequently, there exists a collinearity problem in Heckman model if $x_1 = x_2$. Instead he recommended using

$$E[Y_1|Y_2 > 0, x_1] = x_1\beta$$

directly. A drawback of this approach is that no information on Y_1 is available when $Y_2 < 0$.

6. Generalization of Heckman's model to call back problems

We generalize Heckman's model to the case where there is a call back follow up study,

$$Y_{1i} = X_{1i}\beta + \epsilon_{1i}, \quad Y_{2i} = X_{2i}\gamma + \epsilon_{2i}, \quad Y_{3i} = X_{2i}\xi + \epsilon_{3i},$$

where $(\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i})$ follows a trivariate normal distribution with mean 0 and covariance matrix Σ . We assume Y_{1i} is the response of interest and X_{1i} is the associated covariate. Y_{2i} and Y_{3i} are potential variables governing whether Y_{1i} is observed. If $Y_{2i} > 0$ then Y_{1i} is observable, and the survey is accomplished, i.e., there is no information on Y_{3i} at all. In the presence of one call back, Y_{1i} is also available if $Y_{2i} < 0$ but $Y_{3i} > 0$. Conditioning on the covariates, the likelihood contribution is

$$L = \prod_{i=1}^n \{P(Y_{2i} > 0)P(Y_{1i} = y_{1i}|Y_{2i} > 0)\}^{I(Y_{2i} > 0)} \{P(Y_{2i} < 0, Y_{3i} > 0, Y_{1i} = y_{1i})\}^{I(Y_{2i} < 0, Y_{3i} > 0)} \\ \{P(Y_{2i} < 0, Y_{3i} < 0)\}^{I(Y_{2i} < 0, Y_{3i} < 0)}.$$

With parametric model assumptions for the underlying distributions, maximum likelihood estimation is possible. Further details of this model can be found in Chen et al. (2017).

Chapter 23

Maximum Likelihood Estimation in Capture-Recapture Models

In many fields, such as biology, ecology, demography, epidemiology and reliability study, it is important to know the abundance of a species, the size of a closed population (Borchers et al. 2002), or the number of failed devices in a system. Mark-recapture, or sometimes called capture-recapture, experiments are widely used to collect the necessary data. In such experiments, individuals or animals from a population of interest are captured, marked, and then released. At a later time after the captured and not captured subjects have been mixed, another sample is taken from this population. The mark-recapture experiment is extensively used when it is not practical to count all individuals in the population. Even though this method was originally developed for the estimation of animal abundance, its application in the estimation of population parameters for demographic events has also been growing recently. For example, the U.S. Census Bureau uses a dual system estimation method to estimate the population in the United States (Hogan 1993). This method produces valid population estimates as long as certain assumptions based on the chosen model hold. In epidemiological studies, capture-recapture method is used to estimate the completeness of ascertainment of disease registers. In software inspections, it is also used to estimate the number of defects in an inspected artifact. It is another source of information for deciding whether the artifact requires a reinspection to improve the phase containment of defects. Many other examples on applications of this method can be found through google web search.

23.1 Estimating the Number of Species in the Absence of Covariate

Suppose there are N distinct classes labeled by $i = 1, 2, \dots, N$ and there are (possibly) infinitely many individuals belonging to each class. A typical application of this problem is estimation of species richness in microbial ecology. Suppose

a sample is collected and let X_i be the number of individuals in class i , where

$$P(X_i = j) = p_j, \quad j = 0, 1, 2, \dots,$$

Let n_1, n_2, \dots denote the numbers of species observed once, twice, ..., and so on, in the sample, and n_0 denote the number of unobserved species. More concretely

$$n_j = \sum_{i=1}^N I(X_i = j), \quad j = 0, 1, 2, \dots.$$

The total number of observed individuals is $A = \sum_{j \geq 1} j n_j$. The total number of observed species is $n = n_1 + n_2 + \dots$ and

$$N = n_0 + n_1 + n_2 + \dots = n_0 + n.$$

To estimate N is equivalent to estimating n_0 .

Denote $t = \max(X_1, \dots, X_N)$. Under the i.i.d. assumption for X_i 's, the likelihood function is

$$L = \binom{N}{n_1, \dots, n_t} \prod_{j=0}^t p_j^{n_j} = L_0 L_1$$

where

$$L_0 = \binom{N}{n} p_0^{N-n} (1-p_0)^n, \quad L_1 = \frac{n!}{n_1! \dots n_t!} \prod_{j>0} \left(\frac{p_j}{1-p_0} \right)^{n_j}.$$

The first factor L_0 is a binomial likelihood. Each species i is captured if and only if $X_i \geq 1$ and the corresponding probability is $1 - p_0$. Out of N species, the number of captured species should have a binomial distribution. The second factor L_1 is the conditional likelihood, or zero truncated likelihood.

Note that

$$n = \sum_{j \geq 1} n_j = \sum_{j \geq 1} \sum_{i=1}^N I(X_i = j) = \sum_{i=1}^N I(X_i \geq 1), \quad n_0 = \sum_{i=1}^N I(X_i = 0),$$

$$E(n) = NP(X_i \geq 1) = N(1 - p_0), \quad E(n_0) = Np_0$$

and

$$N = \frac{n}{1 - p_0}.$$

For a given parametric model for $p_j = P(j, \theta)$, $j = 0, 1, 2, \dots$, Sanathanan (1972, 1977) discussed maximum full likelihood estimation of N by treating N as a

parameter in the conditional likelihood. Under some regularity conditions, she showed that both methods produce asymptotically equivalent distribution. In the following we consider two commonly used distributions for p_i .

1. Poisson mixture model

The most popular choice of the probability model for $p_j = P(j, \theta)$ is the Poisson mixture model given by

$$p_j = P(j, \theta) = \int \frac{\theta^j \exp(-\theta)}{j!} dG(\theta), \quad j = 0, 1, 2, \dots$$

where $G(\theta)$ is a cumulative distribution function. A common choice for $G(\theta)$ is the Beta distribution, which leads to the negative binomial distribution.

Using Cauchy-Schwartz's inequality

$$\left\{ \int \theta \exp(-\theta) dG(\theta) \right\}^2 \leq \int \exp(-\theta) dG(\theta) \int \theta^2 \exp(-\theta) dG(\theta),$$

we have

$$p_1^2 \leq 2p_0 p_2, \quad p_0 \geq \frac{p_1^2}{2p_2}.$$

Chao (1987, 1989) estimated the lower bound of N by

$$\hat{N}_{Cao} = n + \frac{n_1^2}{2n_2}.$$

More generally, the following set of inequalities hold

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq \dots$$

In the homogeneity case (no mixture), the inequalities become equalities.

To check whether there is a monotonic trend, a useful strategy is to plot n_i/n_{i-1} against i for $i = 1, 2, \dots$.

2. Binomial mixture model

The binomial mixture model is given by

$$p_j = P(j, \theta) = \int_0^1 \binom{T}{j} \theta^j (1-\theta)^{T-j} dG(\theta), \quad j = 0, 1, 2, \dots, T.$$

where $G(\theta)$ is a cumulative distribution function. If a parametric assumption is made on $G(\theta)$, say, the beta distribution, then it is straightforward to find the maximum likelihood estimation of N .

Bohning et al. (2013) found that for any density $g(\theta) = dG(\theta)/d\theta$, p_j has the following monotonic property

$$a_0 \frac{p_1}{p_0} \leq a_1 \frac{p_2}{p_1} \leq a_2 \frac{p_3}{p_2} \dots, \quad a_j = (j+1)/(T-j), \quad j = 0, 1, 2, \dots, (T-1).$$

Define

$$\eta_j = a_j p_{j+1}/p_j, \quad j = 0, 1, 2, \dots, (T-1).$$

In terms of η_j , $p_{j+1} = \eta_j p_j/a_j = (\eta_j/a_j)(\eta_{j-1}/a_{j-1})p_{j-1} = \dots$ and

$$p_0 = \left\{ 1 + \eta_0/a_0 + (\eta_0/a_0)(\eta_1/a_1) + \dots + \prod_{j=0}^{T-1} (\eta_j/a_j) \right\}^{-1}.$$

Bohning et al. (2016) assumed a model

$$\eta_j = g^{-1}(\beta_0 + z(j)\beta)$$

for some monotone link function, where the common choices of $z(j)$ and η_j are, respectively, $z(j) = j$ and

$$\eta_j = \exp(\beta_0 + \beta_1 j).$$

Wills and Bunge (2015) also used the ratios n_{j+1}/n_j as a function of j to analyze the observed data. Mao and Lindsay (2007) studied Poisson mixture model where the mixture distribution is left unspecified. They showed that it is almost impossible to find a consistent estimator of p_0 . However, they were able to find a lower bound.

23.2 Binomial Detection Model

Next we consider the population size estimation under a binomial detection model when there is no recapture.

Suppose that X_1, \dots, X_N are samples from a super-population with a density $f(x, \theta)$, where the form of $f(x, \theta)$ is known but with an unknown parameter θ . Denote the detection probability (known completely) as

$$P(D = 1|x) = w(x),$$

where X is observed if and only if $D = 1$. Without loss of generality denote the observed data as x_1, \dots, x_n . We are interested in estimating the population size N . The full likelihood is

$$L = \binom{N}{n} \prod_{i=1}^n \{w(x_i) f(x_i, \theta)\}^{d_i} \left[\int \{1 - w(x)\} f(x, \theta) dx \right]^{1-d_i}.$$

By conditioning on $D = 1$, the conditional likelihood is

$$L_C = \prod_{i=1}^n f(X_i | D_i = 1) = \prod_{i=1}^n \frac{w(x_i) f(x_i)}{\Delta}, \quad \Delta = P(D = 1) = \int w(x) f(x, \theta) dx.$$

The full likelihood can be decomposed as

$$L = L_C L_B, \quad L_B = \binom{N}{n} \Delta^n (1 - \Delta)^{N-n}.$$

Up to a constant the log-likelihood is

$$\ell = \log \Gamma(N + 1) - \log \Gamma(N - n + 1) + \ell_B + \ell_c,$$

where

$$\ell_B = n \log \Delta(\theta) + (N - n) \log\{1 - \Delta(\theta)\}, \quad \ell_c = \log L_C.$$

Fewster and Jupp (2009) pointed out that it is reasonable to treat N as a real parameter. Differentiating the log-likelihood with respect to N gives

$$\frac{\partial \ell}{\partial N} = \frac{\partial \log \Gamma(N + 1)}{\partial N} - \frac{\partial \log \Gamma(N - n + 1)}{\partial N} + \log\{1 - \Delta(\theta)\}.$$

Denote the digamma function as

$$\psi(N + 1) = \frac{\partial \log \Gamma(N + 1)}{\partial N}, \quad S_1(N, n) = \psi(N + 1) - \psi(N - n + 1),$$

then the score is

$$S_1(N, n) + \log\{1 - \Delta(\theta)\} = 0.$$

Moreover

$$\frac{\partial \ell}{\partial \theta} = \left[\frac{n}{\Delta(\theta)\{1 - \Delta(\theta)\}} - \frac{N}{1 - \Delta(\theta)} \right] \frac{\partial \Delta(\theta)}{\partial \theta} + \frac{\partial \ell_c}{\partial \theta} = 0.$$

Let \hat{N} be the maximum full likelihood estimate of N .

Using the conditional likelihood approach, as an alternative we solve the conditional score

$$\sum_{i=1}^n \frac{\partial \ell_{ci}}{\partial \theta} = 0.$$

Denote the conditional MLE as $\hat{\theta}_c$. Since $E(D_i/\Delta) = 1$ and $n/\Delta = \sum_{i=1}^N D_i/\Delta$, we may use

$$\hat{N}_c = \frac{n}{\Delta(\hat{\theta}_c)}$$

to estimate N in the conditional likelihood approach.

Empirical studies show that the small-sample distribution of the maximum conditional likelihood estimator is strongly skewed to the right, which may produce Wald-type confidence intervals with lower limits that are less than the number of captured individuals or even negative. An effective transformation suggested by Burnham and then implemented by Chao (1987) is $\{\log(\hat{N}_c - n) - \log(N - n)\}$. It can be shown that

$$\frac{\log(\hat{N}_c - n) - \log(N - n)}{\sqrt{\log\{1 + \hat{N}_c \hat{\sigma}^2 / (\hat{N}_c - n)^2\}}} \rightarrow N(0, 1)$$

in distribution, where $\hat{\sigma}^2$ is an estimator of the asymptotic variance of $\sqrt{N}(\hat{N}_c - N)$. A 95% confidence interval is

$$[n + (\hat{N}_c - n)/a, n + (\hat{N}_c - n)a], \quad a = \exp[1.96\{\log(1 + \hat{\sigma}^2 / (\hat{N}_c - n)^2)\}^{1/2}].$$

It would be interesting to study the relationship between the full maximum likelihood estimator \hat{N} and the maximum conditional likelihood estimator \hat{N}_c .

Fewster and Jupp (2009) showed that

$$\sqrt{N}\{\log(\hat{N}/N), (\hat{\theta} - \theta)\} \rightarrow N(0, \Sigma),$$

$$\sqrt{N}\{\log(\hat{N}_c/N), (\hat{\theta} - \theta)\} \rightarrow N(0, \Sigma).$$

Moreover the difference between the maximum full likelihood estimator and the maximum conditional likelihood estimator satisfies

$$\hat{N}_c - \hat{N} = O_p(1).$$

A natural question concerns the asymptotic behavior of the likelihood ratio statistic

$$R(N) = 2\{\max_{N,\theta} \ell(N, \theta) - \max_{\theta} \ell(N, \theta)\}.$$

It can be shown that its limiting distribution is $\chi^2(1)$ under the null $N = N_0$. We leave this as an exercise to the readers.

Next we consider the situation where the detection function is completely known but the underlying distribution F is not specified.

The maximum conditional likelihood estimator of F is

$$d\hat{F}(x_i) = \frac{1/w(x_i)}{\sum_{j=1}^n 1/w(x_j)}.$$

To maximize the full likelihood L , we let $p_i = dF(x_i)$, $i = 1, 2, \dots, n$. The log profile full likelihood is

$$\ell = - \sum_{i=1}^n \log[1 + \lambda\{w(x_i) - \Delta\}] - n \log \Delta + \log L_B,$$

where

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda\{w(x_i) - \Delta\}}, \quad i = 1, 2, \dots, n$$

and the Lagrange multiplier is determined by

$$\sum_{i=1}^n \frac{w(x_i) - \Delta}{1 + \lambda\{w(x_i) - \Delta\}} = 0.$$

We can maximize ℓ with respect to N ,

$$\frac{\partial \ell}{\partial N} = S_1(N, n) + \log\{1 - \Delta\} = 0,$$

where

$$S_1(N, n) = \frac{\partial \log \Gamma(N+1)}{\partial N} - \frac{\partial \log \Gamma(N-n+1)}{\partial N} + \log\{1 - \Delta(\theta)\}.$$

$$\frac{\partial \ell}{\partial \Delta} = n\lambda - (N-n)\frac{1}{1-\Delta} = 0,$$

or

$$\lambda = \frac{N-n}{n} \frac{1}{1-\Delta}.$$

Note that this is different from the missing not at random problem discussed in Chap. 22, where the population total N is known, but in the population size estimation problem it is the parameter of interest. Again we can construct a full likelihood based confidence interval for N . Detailed derivations can be found in a working paper by Liu et al. (2016).

23.3 Inference in Capture and Re-capture Models

If the sampled individuals are put back in a closed population and one or more samples are recollected, then the situation mimics a capture-recapture experiment. We may allow the detection probability to depend on some finite dimensional parameters. Suppose i.i.d. data

$$X_1, \dots, X_N \sim f(x).$$

Denote p_1 as the probability of capture on the first occasion and p_2 the probability of capture on the second occasion, respectively. The most popular model is

$$p_1(x) = P(D_1 = 1|x) = \frac{\exp(x\beta_1)}{1 + \exp(x\beta_1)}, \quad p_2(x) = P(D_2 = 1|x) = \frac{\exp(x\beta_2)}{1 + \exp(x\beta_2)}.$$

We assume the first capture and second capture are independent of each other for each individual, i.e.,

$$p_{12}(x) = P(D_1 = 1, D_2 = 1|x) = P(D_1 = 1|x)P(D_2 = 1|x) = p_1(x)p_2(x).$$

Of course this assumption can be relaxed. Let $n_1 - m$, $n_2 - m$, m and $N - M$ ($M = n_1 + n_2 - m$) be respectively, the numbers of subjects captured in the first time only, second time only, both times and neither time. The full likelihood is

$$L = \binom{N}{n_1 - m \ n_2 - m \ m \ N - M} \left[\prod_{i=1}^{n_1-m} p_1(x_{1i})\{1 - p_2(x_{1i})\}dF(x_{1i}) \right] \left[\prod_{i=1}^{n_2-m} \{1 - p_1(x_{2i})\}p_2(x_{2i})dF(x_{2i}) \right] \left[\prod_{i=1}^m p_1(x_{3i})p_2(x_{3i})dF(x_{3i}) \right] \left[1 - \int \phi(x)dF(x) \right]^{N-M},$$

where $\phi(x) = p_1(x) + p_2(x) - p_1(x)p_2(x)$ is the probability of at least one capture, x_{1i} 's, x_{2i} 's and x_{3i} 's are, respectively, those caught in first time only, second time only and both times.

Our goal is to estimate N . This can be accomplished by a two-stage approach. First fixing N we profile out F, β_1, β_2 . In the second stage we can search for N such that the likelihood achieves the maximum.

Let

$$\phi_1(x) = p_1(x)\{1 - p_2(x)\}, \quad \phi_2(x) = \{1 - p_1(x)\}p_2(x), \quad \phi_{12} = p_1(x)p_2(x)$$

$$\theta_1 = \int p_1(x)\{1 - p_2(x)\}dF(x), \quad \theta_2 = \int \{1 - p_1(x)\}p_2(x)dF(x), \quad \theta_{12} = \int p_1(x)p_2(x)dF(x).$$

The full likelihood can be decomposed as

$$L = \left(\begin{matrix} N & - \\ n_1 - m & n_2 - m & m & N - M \end{matrix} \right) \theta_1^{n_1-m} \theta_2^{n_2-m} \theta_{12}^m (1 - \theta_1 - \theta_2 - \theta_{12})^{N-M}$$

$$\left[\prod_{i=1}^{n_1-m} \frac{p_1(x_{1i})\{1-p_2(x_{1i})\}dF(x_{1i})}{\theta_1} \right] \left[\prod_{i=1}^{n_2-m} \frac{\{1-p_1(x_{2i})\}p_2(x_{2i})dF(x_{2i})}{\theta_2} \right]$$

$$\left[\prod_{i=1}^m \frac{p_1(x_{3i})p_2(x_{3i})dF(x_{3i})}{\theta_{12}} \right].$$

Let $\eta = \theta_1 + \theta_2 - \theta_{12}$ and $M = n_1 + n_2 - m$. Note that $\eta = \int \phi(x)dF(x)$ is the overall capturing probability. As an alternative

$$L \propto \binom{N}{M} (1 - \eta)^{N-M} P(D_1 = i, D_2 = j | D_1 + D_2 > 0) P(D_1 + D_2 > 0)$$

$$\propto \binom{N}{M} \eta^M (1 - \eta)^{N-M} P(D_1 = i, D_2 = j | D_1 + D_2 > 0, x) P(x | D_1 + D_2 > 0)$$

$$\propto \binom{N}{M} \eta^M (1 - \eta)^{N-M} P(D_1 = i, D_2 = j | D_1 + D_2 > 0, x) \frac{\phi(x)dF(x)}{\int \phi(x)dF(x)}.$$

Alho (1990b) used a conditional approach to estimate β_1 and β_2 .

$$P(D_1 + D_2 > 0 | x) = \frac{\exp(x\beta_1) + \exp(x\beta_2) + \exp(x\beta_1 + x\beta_2)}{\{1 + \exp(x\beta_1)\}\{1 + \exp(x\beta_2)\}} := \Delta(x).$$

$$P(D_1 = 1, D_2 = 0 | D_1 + D_2 > 0, x) = \frac{\exp(x\beta_1)}{\Delta(x)},$$

$$P(D_1 = 0, D_2 = 1 | D_1 + D_2 > 0, x) = \frac{\exp(x\beta_2)}{\Delta(x)},$$

$$P(D_1 = 1, D_2 = 1 | D_1 + D_2 > 0, x) = \frac{\exp(x\beta_1 + x\beta_2)}{\Delta(x)}.$$

Finally N can be estimated by

$$\sum_{D_{1i} + D_{2i} > 0} \frac{1}{\phi(x_i)}.$$

Conditioning on at least one catch, we can find the joint density of X and catching status D_1 and D_2 through

$$f(D_1 = i, D_2 = j, x | D_1 + D_2 > 0) = \frac{P(D_1 = i, D_2 = j | x) dF(x)}{1 - \int P(D_1 = 0, D_2 = 0 | x) dF(x)}$$

$$= \frac{\exp(ix\beta_1 + jx\beta_2)}{\{1 + \exp(x\beta_1)\}\{1 + \exp(x\beta_2)\}} dF(x) \frac{1}{1 - \int P(D_1 = 0, D_2 = 0 | x) dF(x)}.$$

We call this is a “truncation approach”. The truncation likelihood is

$$L = \left[\prod_{i=1}^{n_1-m} \frac{p_1(x_{1i})\{1 - p_2(x_{1i})\}dF(x_{1i})}{\theta_1} \right] \left[\prod_{i=1}^{n_2-m} \frac{\{1 - p_1(x_{2i})\}p_2(x_{2i})dF(x_{2i})}{\theta_2} \right]$$

$$\left[\prod_{i=1}^m \frac{p_1(x_{3i})p_2(x_{3i})dF(x_{3i})}{\theta_{12}} \right]$$

$$\left(\frac{\theta_1}{\theta_1 + \theta_2 + \theta_{12}} \right)^{n_1-m} \left(\frac{\theta_2}{\theta_1 + \theta_2 + \theta_{12}} \right)^{n_2-m} \left(\frac{\theta_{12}}{\theta_1 + \theta_2 + \theta_{12}} \right)^m.$$

Note that the first three terms are biased sampling problems. The profile likelihood methods discussed in Chap. 11 can be applied. Liu et al. (2016) showed that the maximum full likelihood estimate is superior to the conditional likelihood or truncation likelihood approach. They also studied multiple capture and recapture problems.

Without making any parametric assumption on the detection functions $p_i(x)$, $i = 1, 2$, Chen and Lloyd (2000) used a nonparametric kernel method to estimate them.

Chapter 24

A Review of Survival Analysis

Survival analyses have been developed extensively over the past 50 years. The fundamental challenge lies in the incomplete observation of failure times due to censoring. Another layer of complication which arises frequently in epidemiological studies or econometric researches is the selection bias reduced by left truncation in addition to right censoring. In this chapter, in the absence of covariates, we briefly derive the Kaplan–Meier estimator for right-censored data and Lynden-Bell's (1971) product-limit estimate for left-truncated data. Then we review different inference methods for the Cox proportional hazards model, accelerated failure time model (AFT), and quantile regression model in the presence of covariates. This chapter mainly focuses on heuristic and exploratory arguments rather than rigorous mathematical proofs. We provide different ways of interpreting a statistical problem. Readers interested in mathematical details are referred to those theoretical survival books such as Andersen et al. (1993), Kalbfleisch and Prentice (2002), Van der Vaart and Wellner (1992), Tsiatis (2006) and Kosorok (2008a). We hope materials presented in this chapter will shed lights on this topic, especially for readers who are new to this area of research. The latest developments on right-censored length-biased sampling problems will be discussed in the next chapter.

24.1 Kaplan–Meier Estimator

Denote by T the survival time of interest. Due to loss to follow-up or end of study, T may not be observed. Let C be the censoring time. The observed survival data are $\{Y_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i)\}, i = 1, 2, \dots, n$, where the lifetime T_i and C_i have distribution functions $F(t)$ and $G(c)$, respectively. Moreover, we assume that T_i and C_i are mutually independent of each other. We do not make any parametric assumption on the forms of F and G . The log-likelihood can be written as

$$\ell = \sum_{i=1}^n \delta_i \log dF(y_i) + (1 - \delta_i) \log \bar{F}(y_i) = \sum_{i=1}^n \{\delta_i \log \lambda(y_i) + \log \bar{F}(y_i)\}, \quad (24.1.1)$$

where $\lambda(t) = dF(t)/\bar{F}(t)$ is the hazard function.

Denote by

$$t_1 < t_2 < \dots < t_{h_1}, \quad h_1 \leq n$$

the ordered and distinct **observed data points** (both censored and uncensored event times). Define

$$\xi_j = \sum_{i=1}^n \delta_i I(y_i = t_j), \quad \eta_j = \sum_{i=1}^n I(y_i = t_j), \quad j = 1, 2, \dots, h,$$

so that ξ_j and η_j are the numbers of uncensored failures events and censored events at t_j , respectively. The log-likelihood is

$$\ell = \sum_{j=1}^{h_1} \xi_j \log \lambda(t_j) + \sum_{j=1}^{h_1} \eta_j \log \bar{F}(t_j).$$

Let $\bar{F}(t_1) = 1$, and denote

$$\lambda(t_i) = \frac{\bar{F}(t_i) - \bar{F}(t_{i+1})}{\bar{F}(t_i)} = \lambda_i.$$

$$\bar{F}(t_j) = \prod_{i=1}^{j-1} \frac{\bar{F}(t_{i+1})}{\bar{F}(t_i)} = \prod_{i=1}^{j-1} \left(1 - \frac{\bar{F}(t_i) - \bar{F}(t_{i+1})}{\bar{F}(t_i)}\right) = \prod_{i=1}^{j-1} (1 - \lambda_i), \quad j \geq 2.$$

The log-likelihood can be reexpressed as

$$\begin{aligned} \ell &= \sum_{i=1}^{h_1} \xi_i \log \lambda_i + \sum_{i=1}^{h_1-1} \left(\sum_{j=i+1}^{h_1} \eta_j \right) \log(1 - \lambda_i) \\ &= \xi_{h_1} \log \lambda_{h_1} + \sum_{i=1}^{h_1-1} [\xi_i \log \lambda_i + \left(\sum_{j=i+1}^{h_1} \eta_j \right) \log(1 - \lambda_i)]. \end{aligned}$$

Given that $0 \leq \lambda_i \leq 1$, it is easy to see that ℓ achieves the maximum when $\lambda_i = 0$ if $\xi_i = 0$. When $\xi_i > 0$ maximizing ℓ with respect to λ_i 's gives

$$\hat{\lambda}_i = \xi_i / (\xi_i + \sum_{j=i+1}^{h_1} \eta_j), \quad i = 1, 2, \dots, h_1 - 1,$$

and $\lambda_{h_1} = 1$ if $\xi_{h_1} > 0$. Therefore the nonparametric MLE of F has a jump size of zero at the censored data points, and if the largest observation is censored, then the survival function is not defined beyond that point.

Intuitively, we can use a different argument to show that F has jumps only at the failure time points. In fact in order to maximize this log-likelihood, F must have jumps at each of the observed failure time point $\delta_i = 1$. Otherwise we have $\log 0 = -\infty$ in the log-likelihood. Denote

$$t_1 < t_2 < \dots < t_h$$

as ordered **unique failure times** for $\{y_1, \dots, y_n\}$, $h \leq h_1 \leq n$. Note that

$$\bar{F}(y_i) = P(T > y_i) \leq P(T > t_j) = \bar{F}(t_{ij}), \quad t_{ij} = \max\{t_j : t_j \leq y_i\},$$

and the log-likelihood satisfies

$$\ell = \sum_{j=1}^h \delta_j dF(t_j) + \sum_{i=1}^n (1 - \delta_i) \log \bar{F}(y_i) \leq \sum_{j=1}^h \delta_j dF(t_j) + \sum_{i=1}^n (1 - \delta_i) \log \bar{F}(t_{ij}).$$

To maximize this log-likelihood, we only need to consider distributions with jumps at all failure time points.

An approximate formula for the Kaplan–Meier is defined as

$$1 - \hat{F}_{KM}(t) = \exp\{-\hat{\Lambda}(t)\} \approx \prod_{t_i \leq t} (1 - \hat{\lambda}_i),$$

where $\hat{\Lambda}(t) = \sum_{t_i \leq t} \hat{\lambda}_i$.

Alternatively we can use an imputation method to derive the self-consistent estimator. Let $t_1 < t_2 < \dots < t_h$ be the unique failure times. In the absence of right censoring, the log-likelihood can be written as

$$\ell = \sum_{j=1}^h \sum_{i=1}^n \{I(T_i = t_j)\} \log p_j, \quad p_j = dF(t_j), \quad j = 1, 2, \dots, h.$$

In the presence of right censoring, we can impute the indicator function with

$$P(T_i = t_j | T_i > y_i, \delta_i = 0) = \frac{f(t_j)}{\bar{F}(y_i)}.$$

By taking those uncensored individuals, the overall weight at t_j is

$$w_j = \sum_{i=1}^n \left\{ \delta_i I(T_i = t_j) + (1 - \delta_i) \frac{p_j}{\sum_{k=1}^h p_k I(T_k \geq y_i)} \right\}.$$

As a result, the imputed log-likelihood is

$$\ell_I = \sum_{j=1}^h w_j \log p_j.$$

The maximizers satisfy

$$p_j = w_j/n, j = 1, 2, \dots, h.$$

We may also derive the self-consistent algorithm starting from the empirical survival function.

$$1 - \hat{F}(t_j) = n^{-1} \sum_{i=1}^n I(T_i \geq t_j), \quad j = 1, 2, \dots, h.$$

Since not all $I(T_i > t_j)$'s are observable, we need to impute them by using the observed quantities. Let $1 - \hat{F}(t_j), j = 1, 2, \dots, h$ be the solution of the following equation

$$1 - \hat{F}(t_j) = n^{-1} \sum_{i=1}^n E[I(T_i \geq t_j) | \delta_i, y_i] = n^{-1} \sum_{i=1}^n \left[\delta_i I(T_i \geq t_j) + (1 - \delta_i) \frac{1 - \hat{F}(\max(t_j, y_i))}{1 - \hat{F}(y_i)} \right].$$

Efron (1967) showed that $1 - \hat{F}(t_j), j = 1, 2, \dots, h$ is the Kaplan–Meier estimator (**exercise**).

Exercise Consider the log-likelihood (24.1.1). Without loss of generality we assume there is no ties and the first h observations are failures. Let

$$t_1 < t_2 < \dots < t_h$$

be the ordered failure times. Define $p_i = dF(t_i), i = 1, 2, \dots, h$. Then (24.1.1) can be written as

$$\ell = \sum_{i=1}^h \log p_i + \sum_{i=1}^n (1 - \delta_i) \log \left\{ \sum_{j=1}^h I(t_j \geq y_i) p_j \right\}.$$

Derive Kaplan–Meier estimate by directly maximizing ℓ subject to the constraints

$$p_i \geq 0, \quad i = 1, 2, \dots, h, \quad \sum_{i=1}^h p_i = 1.$$

Inverse Probability Weighted Considerations

Consider the conditional likelihood given being uncensored:

$$L_{c1} = \prod_{\delta_i=1} \frac{\bar{G}(y_i)dF(y_i)}{\int \bar{G}(y)dF(y)}.$$

This is a biased sampling problem with sampling weight function $\bar{G}(y)$. When \bar{G} is known, clearly $F(y)$ can be estimated by the maximization of L_{c1} ,

$$\hat{F}_{IW}(y) = n^{-1} \sum_{i=1}^n \delta_i I(y_i \leq y) / \bar{G}(y_i).$$

On the other hand, if $\bar{G}(y)$ is unknown, we can maximize the log-likelihood

$$\ell^* = \sum_{i=1}^n \delta_i \log \bar{G}(y_i) + (1 - \delta_i) \log dG(c_i)$$

to get the Kaplan–Meier estimator $\hat{G}_{KM}(c)$.

In fact if G is replaced by the Kaplan–Meier estimator \hat{G}_{KM} , then the resulting inverse probability weighted estimator

$$\tilde{F}_{IW}(y) = n^{-1} \sum_{i=1}^n \frac{\delta_i I(Y_i \leq y)}{1 - \hat{G}_{KM}(Y_i)},$$

is identical to the Kaplan–Meier estimator \hat{F}_{KM} (Satten and Datta 2001).

We end up with a paradox associated with survival function estimation in the presence of nuisance parameters or functions. In a statistical model with unknown nuisance parameters, the efficiency of an estimator of a parameter usually increases when the nuisance parameters are known. However the opposite result occurs in this example since \tilde{F}_{IW} is the nonparametric MLE or Kaplan–Meier estimator. Note that we also observed this counterintuitive phenomenon in Chap. 19 where the inverse weighted or augmented inverse weighted estimator of the population mean with estimated propensity score is more efficient than its counterpart with the true propensity score.

Exercise Similarly if we condition on $\delta_i = 0$, then the likelihood is

$$L_{c2} = \prod_{\delta_i=0} \frac{\bar{F}(y_i)dG(y_i)}{\int \bar{F}(y)dG(y)}.$$

This is also a biased sampling problem. For fixed F , we can estimate G by

$$\hat{G}(y) = n^{-1} \sum_{i=1}^n (1 - \delta_i) I(y_i \leq y) / \bar{F}(y_i).$$

If we iteratively maximize L_{c1} and L_{c2} , what type of estimator do we end up to? Is it the Kaplan–Meier estimator?

24.2 Nonparametric MLE for Left Truncated Data

Since Lynden-Bell's (1971) work in astronomy, the product-limit estimator for truncation problem has become very popular in the statistical literature. Among others, Woodroffe (1985) and Wang et al. (1986) discussed this problem thoroughly from a theoretical perspective. Moreover, Vardi (1982a,b, 1985) pointed out the close connection between truncation problems and the general biased sampling problems. Some generalizations from left-truncated and right-censored data were studied by Tsai et al. (1987) and Lai and Ying (1994).

We begin by considering the left truncation problem. We assume that the observed lifetime Y is subject to the left truncation time A . Suppose $Y \sim F(y)$ and $A \sim H(a)$. The observed data have a density

$$(Y, A)|Y > A \sim \frac{h(a)f(y)}{P(Y > A)}, \quad y > a.$$

Denote by $[Y, A]$ the joint density of Y and A . Let $[Y|A]$ and $[A]$ be the conditional density and marginal density, respectively. The joint density can be decomposed as

$$[Y, A] \cong [Y|A][A] \sim \left\{ \frac{f(y)}{\bar{F}(a)} \right\} \left\{ \frac{\bar{F}(a)h(a)}{\int \bar{F}(a)h(a)da} \right\}, \quad (24.2.2)$$

or

$$[Y, A] \cong [A|Y][Y] \sim \left\{ \frac{h(a)}{H(y)} \right\} \left\{ \frac{H(y)dF(y)}{\int H(y)dF(y)} \right\},$$

respectively. Let (y_i, a_i) , $i = 1, 2, \dots, n$ be the observed truncation data. Denote the likelihood as

$$L(F, H) = \prod_{i=1}^n \left\{ \frac{dF(y_i)}{\bar{F}(a_i)} \right\} \left\{ \frac{\bar{F}(a_i)dH(a_i)}{\int \bar{F}(a)dH(a)} \right\} = L_1(F)L_2(H|F), \quad (24.2.3)$$

or equivalently

$$L(F, H) = \prod_{i=1}^n \left\{ \frac{dH(a_i)}{H(y_i)} \right\} \left\{ \frac{H(y_i)dF(y_i)}{\int H(y)dF(y)} \right\} = L_1^*(H)L_2^*(F|H). \quad (24.2.4)$$

Lynden-Bell's (1971) Product-Limit Estimator

For convenience, we assume the observed survival data do not have ties. By rearranging the observed data, without loss of generality, we write

$$y_1 < y_2 < \cdots < y_n.$$

First we maximize the conditional likelihood

$$L_1(F) = \prod_{i=1}^n \frac{dF(y_i)}{\bar{F}(a_i)}.$$

Let

$$\lambda_i = dF(y_i)/\bar{F}(y_i), i = 1, 2, \dots, n$$

be the hazard at time y_i . Using the relationship between \bar{F} and λ ,

$$\bar{F}(y) = \exp\{-\Lambda(y)\},$$

where $\Lambda(y) = \sum_{i=1}^n \lambda_i I(y_i \leq y)$ is the cumulative hazard. Then the conditional likelihood $L_1(F)$ can be written as

$$L_1(F) = \prod_{i=1}^n \lambda_i \exp[-\{\Lambda(y_i) - \Lambda(a_i)\}].$$

The log conditional likelihood is

$$\ell = \sum_{i=1}^n \log \lambda_i - \sum_{j=1}^n \sum_{i=1}^n I(a_i < y_j \leq y_i) \lambda_j.$$

Differentiating ℓ with respect to λ_k , we have

$$\frac{\partial \ell}{\partial \lambda_k} = \frac{1}{\lambda_k} - \sum_{i=1}^n I(a_i < y_k \leq y_i) = 0.$$

The solution is

$$\hat{\lambda}_k = \frac{1}{\sum_{i=1}^n I(a_i < y_k \leq y_i)}.$$

The Lynden-Bell's (1971) product-limit estimator is given by

$$1 - \hat{F}(y_k) = \prod_{i=1}^{k-1} (1 - \hat{\lambda}_i).$$

This estimator is similar to the Kaplan–Meier estimator except that the indicator for the “risk set” at y_k has changed from $\sum_{j=1}^n I(y_k \leq y_j)$ to $\sum_{j=1}^n I(a_j < y_k \leq y_j)$. This modification can be justified by the fact that, due to left truncation, we have to take the “delayed entry” a_j into account.

Note that the product-limit estimator $\hat{F}(y)$ maximizes the first factor in (24.2.3), i.e., the conditional likelihood $L_1(F)$. A natural question is whether this estimator is efficient. In theory the most efficient estimator should maximize the joint likelihood $L(F, H)$.

For the given product limit estimator $\hat{F}(y)$, the marginal likelihood of A ,

$$L_2(H|\hat{F}) = \prod_{i=1}^n \frac{\{1 - \hat{F}(a_i)\}dH(a_i)}{\int \{1 - \hat{F}(a)\}dH(a)}$$

is a biased sampling likelihood. The nonparametric MLE for dH is

$$d\hat{H}(a_i) = \frac{\{1 - \hat{F}(a_i)\}^{-1}}{\sum_{j=1}^n \{1 - \hat{F}(a_j)\}^{-1}}.$$

Consequently, the maximum of the likelihood L_2 is $L_2(\hat{H}|\hat{F}) = c = (1/n)^n$, which does not depend on \hat{F} . Therefore indeed (\hat{F}, \hat{H}) maximize $L(F, H)$ indeed.

By symmetry, we can show that $\max_F L_2^*(F|H) = (1/n)^n$. If we directly maximize $L_1^*(H)$ in (24.2.4), then we obtain a similar product-limit estimator for H , denoted as \tilde{H} . A natural question is whether $\tilde{H} = \hat{H}$? Wang (1987) showed analytically that the two product-limit estimators for F and H indeed maximize the joint likelihood. We can use an alternative argument.

Based on the two different decompositions (24.2.3) and (24.2.4), we have

$$\max_{F,H} L_1(F)L_2(H|F) = \max_F L_1(F)c = L_1(\hat{F})c$$

and

$$\max_{F,H} L_1^*(H)L_2^*(F|H) = \max_H L_1^*(H)c = L_1^*(\tilde{H})c.$$

Therefore

$$L_1(\hat{F}) = L_1^*(\tilde{H}).$$

Note that

$$\begin{aligned} L_1(\hat{F})c &= L_1(\hat{F}) \max_H L_2(H|\hat{F}) = L_1(\hat{F})L_2(\hat{H}|\hat{F}) = L_1^*(\hat{H})L_2^*(\hat{F}|\hat{H}) \\ &\leq L_1^*(\hat{H}) \max_F L_2^*(F|\hat{H}) = L_1^*(\hat{H})c \leq \max_H L_1^*(H)c = L_1(\tilde{H})c. \end{aligned}$$

Since $L_1^*(\tilde{H}) = L_1(\hat{F})$, this implies that \hat{H} maximizes $L_1^*(H)$, i.e., \hat{H} must be the product limit estimator for H ! In other words, both product limit estimators \hat{F} and \tilde{H} maximize the joint likelihood.

Exercise 1 What happens if we iteratively maximize the marginal likelihoods $L_2(H|F)$ and $L_2^*(F|H)$ with respect to H and F , respectively? Do we obtain the same MLE?

Let F_k and H_k be the solutions in the k -th iteration. If F_k and H_k do converge to some functions, say, F and H , then they satisfy the following equations

$$H(a) = \left[\sum_{i=1}^n \bar{F}^{-1}(a_i) I(a_i \leq a) \right] \left[\sum_{i=1}^n \bar{F}^{-1}(a_i) \right]^{-1}$$

and

$$F(y) = \left[\sum_{i=1}^n H^{-1}(y_i) I(y_i \leq y) \right] \left[\sum_{i=1}^n H^{-1}(y_i) \right]^{-1}.$$

Equivalently

$$\sum_{i=1}^n \bar{F}^{-1}(a_i) \{I(a_i \leq a) - H(a)\} = 0, \quad \sum_{i=1}^n H^{-1}(y_i) \{I(y_i \leq y) - F(y)\} = 0,$$

$$\int \bar{F}^{-1}(t) \{I(t \leq a) - H(a)\} d\hat{G}_1(t) = 0, \quad \hat{G}_1(t) = n^{-1} \sum_{i=1}^n I(a_i \leq t),$$

$$\int H^{-1}(t) \{I(t \leq y) - F(y)\} d\hat{G}_2(t) = 0, \quad \hat{G}_2(t) = n^{-1} \sum_{i=1}^n I(y_i \leq t).$$

If the above equations have an unique solution, then the iterative maximization method produces the same MLE, why?

Exercise 2 Find the product-limit estimator when the lifetime is subject to left truncation and right censoring.

Exercise 3 Consider the three dimensional truncation problem discussed in Kalbfleisch and Lawless (1989) and Wang (1992). Let X, Y, Z be three positive and independent random variables with densities $f(x), g(y), h(z)$, respectively. Furthermore, (X, Y, Z) are observed if and only if $X + Y + Z \leq \tau$, where τ is a fixed number. A typical example is that X, Y, Z are, respectively the HIV infection time, AIDS onset time, and reporting lag time, and τ is the case report time. The joint likelihood is

$$(X, Y, Z) | X + Y + Z \leq \tau \sim \frac{f(x)g(y)h(z)}{P(X + Y + Z \leq \tau)}, \quad 0 \leq x + y + z \leq \tau.$$

Note that

$$X|X \leq \tau - Y - Z, Y = y, Z = z \sim \frac{f(x)}{F(\tau - y - z)}, \quad x \leq \tau - y - z.$$

We can use the product-limit estimator to estimate F . Similarly symmetry among X, Y, Z , the product limiting estimators can also be applied to estimating G and H . Note

$$Y|X \leq \tau - Y - Z, Z = z \sim \frac{g(y)F(\tau - y - z)}{\int F(\tau - y - z)g(y)dy}$$

and

$$Z|X \leq \tau - Y - Z \sim \frac{h(z)\int F(\tau - y - z)g(y)dy}{\{\int F(\tau - y - z)g(y)dy\}h(z)dz}.$$

If F and G are known, we can treat the marginal density of $Z|X \leq \tau - Y - Z$ as a biased sampling problem with weight function $\int F(\tau - y - z)g(y)dy$. Then H can be estimated by the inverse probability weighted estimator as discussed in Chap. 10. Similarly for fixed G, H , we can estimate F and for fixed F, H we can estimate G .

Use the iterative maximum marginal likelihood method discussed above to find the nonparametric MLE for the distributions of X, Y, Z . Discuss their efficiency compared with the product limiting estimators. Details can be found in Wang (1992).

Semiparametric Approach

Wang (1989) considered a semiparametric estimation of H by postulating a parametric model

$$f(y) = f(y, \theta).$$

Then the truncation likelihood (24.2.2) is

$$L = \left(\prod_{i=1}^n \frac{f(y_i, \theta)}{\bar{F}(a_i, \theta)} \right) \prod_{i=1}^n \left(\frac{\bar{F}(a_i, \theta)dH(a_i)}{\int \bar{F}(a, \theta)dH(a)} \right).$$

We can estimate θ by maximizing the first term, that is, the conditional likelihood. Then maximizing the second term, that is the marginal likelihood, Wang (1989) found the biased sampling estimator

$$\left[\sum_{i=1}^n I(y_i \leq y) / \bar{F}(y_i, \hat{\theta}) \right] \left[\sum_{i=1}^n 1 / \bar{F}(y_i, \hat{\theta}) \right]^{-1}$$

is more efficient than the product-limit estimator $\hat{H}(y)$.

Qin and Wang (2001) generalized this result to a two-sample problem, where

$$f_i(y) = f_i(y, \theta_i), i = 1, 2,$$

and the two groups are subject to left truncation with the same distribution function H . The likelihood is

$$\prod_{j=1}^{n_1} \left\{ \frac{f_1(y_{1j}, \theta_1)}{\bar{F}_1(a_{1j}, \theta_1)} \frac{\bar{F}_1(a_{1j}, \theta_1) dH(a_{1j})}{\int \bar{F}_1(a, \theta_1) dH(a)} \right\} \left\{ \prod_{j=1}^{n_2} \frac{f_1(y_{2j}, \theta_2)}{\bar{F}_2(a_{2j}, \theta_2)} \frac{\bar{F}_2(a_{2j}, \theta_2) dH(a_{2j})}{\int \bar{F}_2(a, \theta_2) dH(a)} \right\}.$$

For fixed θ_1, θ_2 , this is a two sample bias sampling problem concerning about the maximum likelihood estimator of H . Methods discussed in Chaps. 10 and 11 can be applied directly.

Moreover, Li and Qin (2006) considered the following two sample semiparametric truncation model

$$(X, T)|X > T \sim \frac{dF_1(x)dG_1(t)}{\int F_1(t)dG_1(t)},$$

$$(Y, Z)|Y > Z \sim \frac{dF_2(y)dG_2(z)}{\int F_2(y)dG_2(y)},$$

where $dF_1(x)$ and $dF_2(x)$ are linked via the exponential tilting model

$$dF_2(x) = \exp\{\alpha + \phi(x, \beta)\}dF_1(x)$$

and the truncation distributions G_1 and G_2 are unspecified. They developed an iterative algorithm to obtain the maximum likelihood estimation of β , F_1 , G_1 and G_2 . Again their algorithm is developed by using the two decompositions (24.2.3) and (24.2.4) for the joint likelihood function.

First, the likelihood is decomposed as

$$L^* = \left\{ \prod_{i=1}^{n_1} \frac{dF_1(x_i)}{F_1(t_i)} \prod_{j=1}^{n_2} \frac{dF_2(y_j)}{F_2(z_j)} \right\} \left\{ \prod_{i=1}^{n_1} \frac{F_1(t_i)dG_1(t_i)}{\int F_1(t)dG_1(t)} \prod_{j=1}^{n_2} \frac{F_2(z_j)dG_2(z_j)}{\int F_2(z)dG_2(z)} \right\} = L_1^* L_2^*,$$

where L_1^* is the conditional likelihood of x_i and y_j given t_i and z_j , $i = 1, 2, \dots, n_1$; $j = 1, 2, \dots, n_2$, and L_2^* is the marginal likelihood of t_i , $i = 1, 2, \dots, n_1$ and z_j , $j = 1, 2, \dots, n_2$.

For fixed F_1 and F_2 , the marginal likelihood L_2^* is maximized at

$$G_1(t) = \sum_{i=1}^{n_1} \frac{I(t_i \leq t)}{F_1(t_i)} \Bigg/ \sum_{i=1}^{n_1} \frac{1}{F_1(t_i)}, \quad G_2(z) = \sum_{j=1}^{n_2} \frac{I(z_j \leq z)}{F_2(z_j)} \Bigg/ \sum_{j=1}^{n_2} \frac{1}{F_2(z_j)}.$$

Substituting these G_1 and G_2 in L^* , we have

$$\begin{aligned} L^* &= cL_1^* = c \prod_{i=1}^{n_1} \frac{dF_1(x_i)}{F_1(t_i)} \prod_{j=1}^{n_2} \frac{dF_2(y_j)}{F_2(z_j)} \\ &= c \prod_{i=1}^{n_1} \frac{dF_1(x_i)}{F_1(t_i)} \prod_{j=1}^{n_2} \frac{\exp(\alpha + \phi(y_j; \beta)) dF_1(y_j)}{\int_0^{z_j} \exp(\alpha + \phi(y; \beta)) dF_1(y)}, \end{aligned} \quad (24.2.5)$$

where c is a constant. However, it is not clear how to maximize L_1^* in (24.2.5) with respect to F_1 and (α, β) .

Instead of maximizing (24.2.5), we may consider a different decomposition of the likelihood function:

$$L = \left\{ \prod_{i=1}^{n_1} \frac{dG_1(t_i)}{\bar{G}_1(x_i)} \prod_{j=1}^{n_2} \frac{dG_2(z_j)}{\bar{G}_2(y_j)} \right\} \left\{ \prod_{i=1}^{n_1} \frac{\bar{G}_1(x_i) dF_1(x_i)}{\int \bar{G}_1(x) dF_1(x)} \prod_{j=1}^{n_2} \frac{\bar{G}_2(y_j) dF_2(y_j)}{\int \bar{G}_2(y) dF_2(y)} \right\} = L_1 L_2,$$

where $\bar{G}_1 = 1 - G_1$, $\bar{G}_2 = 1 - G_2$, L_1 is the conditional likelihood of t_i and z_j given x_i and y_j , $i = 1, 2, \dots, n_1$; $j = 1, 2, \dots, n_2$, and L_2 is the marginal likelihood of x_i , $i = 1, 2, \dots, n_1$ and y_j , $j = 1, 2, \dots, n_2$.

For fixed (G_1, G_2) and (α, β) , we can maximize L_2 . Let

$$dH_1(x) = \frac{\bar{G}_1(x) dF_1(x)}{\int \bar{G}_1(x) dF_1(x)} \quad \text{and} \quad dH_2(x) = \frac{\bar{G}_2(x) \exp(\alpha + \phi(x; \beta)) dF_1(x)}{\int \bar{G}_2(y) \exp(\alpha + \phi(y; \beta)) dF_1(y)},$$

then

$$dH_2(x) = \frac{\bar{G}_2(x)}{\bar{G}_1(x)} \exp(\alpha^* + \phi(x; \beta)) dH_1(x),$$

where $\alpha^* = \log\{\int \bar{G}_1(x) dF_1(x) / \int \exp(\phi(y; \beta)) \bar{G}_2(y) dF_1(y)\}$. Again, let

$$\{w_1, w_2, \dots, w_n\} = \{x_1, \dots, x_{n_1}; y_1, \dots, y_{n_2}\}$$

be the pooled lifetimes from the two samples, where $n = n_1 + n_2$. Also let $dH_1(w_i) = p_i$, $i = 1, 2, \dots, n$. Note that

$$\begin{aligned} L_2 &= \prod_{i=1}^{n_1} dH_1(x_i) \prod_{j=1}^{n_2} \left\{ \frac{\bar{G}_2(y_j)}{\bar{G}_1(y_j)} \exp(\alpha^* + \phi(y_j; \beta)) dH_1(y_j) \right\} \\ &= \left\{ \prod_{i=1}^n p_i \right\} \prod_{j=1}^{n_2} \left\{ \frac{\bar{G}_2(y_j)}{\bar{G}_1(y_j)} \exp(\alpha^* + \phi(y_j; \beta)) \right\}. \end{aligned}$$

Maximizing L_2 subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0, \quad \sum_{i=1}^n \frac{\bar{G}_2(w_i)}{\bar{G}_1(w_i)} \exp(\alpha^* + \phi(w_i; \beta)) p_i = 1,$$

we have

$$p_i = \frac{1}{n_1} \frac{1}{1 + \rho \bar{G}_2(w_i) \exp(\alpha^* + \phi(w_i; \beta)) / \bar{G}_1(w_i)}, \quad \rho = n_2/n_1.$$

Therefore

$$l_2 = \log L_2 = - \sum_{i=1}^n \log \{1 + \rho \bar{G}_2(w_i) \exp(\alpha^* + \phi(w_i; \beta)) / \bar{G}_1(w_i)\} + \sum_{j=1}^{n_2} \left\{ \alpha^* + \phi(y_j; \beta) + \log \frac{\bar{G}_2(y_j)}{\bar{G}_1(y_j)} \right\}. \quad (24.2.6)$$

We can maximize l_2 to obtain point estimators of (α^*, β) which depend on G_1 and G_2 . Note that $H_1(t)$ and $H_2(t)$ can be estimated by

$$H_1(t) = \sum_{i=1}^n p_i I(w_i \leq t), \quad H_2(t) = \sum_{i=1}^n \frac{\bar{G}_2(w_i)}{\bar{G}_1(w_i)} \exp(\alpha^* + \phi(w_i; \beta)) p_i I(w_i \leq t),$$

respectively. Hence F_1 and F_2 can be estimated by

$$F_1(t) = \sum_{i=1}^n \frac{1}{\bar{G}_1(w_i)} p_i I(w_i \leq t) / \sum_{i=1}^n \frac{1}{\bar{G}_1(w_i)} p_i$$

and

$$F_2(t) = \sum_{i=1}^n \frac{\exp(\alpha^* + \phi(w_i; \beta))}{\bar{G}_1(w_i)} p_i I(w_i \leq t) / \sum_{i=1}^n \frac{\exp(\alpha^* + \phi(w_i; \beta))}{\bar{G}_1(w_i)} p_i,$$

respectively. Denote

$$L_1^* = L_1^*(F_1, F_2), L_2^* = L_2^*(G_1, G_2 | F_1, F_2)$$

and

$$L_1 = L_1(G_1, G_2), L_2 = L_2(F_1, F_2 | G_1, G_2).$$

Then

$$L^* = L_1^*(F_1, F_2) L_2^*(G_1, G_2 | F_1, F_2) = L_1(G_1, G_2) L_2(F_1, F_2 | G_1, G_2) = L.$$

In general, if $G_1 \neq G_2$, for given $(G_1^{(k)}, G_2^{(k)})$, let $(F_1^{(k)}, F_2^{(k)})$ maximize $L_2(F_1, F_2 | G_1^{(k)}, G_2^{(k)})$. Also, for given $(F_1^{(k)}, F_2^{(k)})$, let $(G_1^{(k+1)}, G_2^{(k+1)})$ maximize $L_2^*(G_1, G_2 | F_1^{(k)}, F_2^{(k)})$. Therefore

$$\begin{aligned}
L_1(G_1^{(1)}, G_2^{(1)})L_2(F_1, F_2|G_1^{(1)}, G_2^{(1)}) &\leq L_1(G_1^{(1)}, G_2^{(1)})L_2(F_1^{(1)}, F_2^{(1)}|G_1^{(1)}, G_2^{(1)}) \\
&= L_1^*(F_1^{(1)}, F_2^{(1)})L_2^*(G_1^{(1)}, G_2^{(1)}|F_1^{(1)}, F_2^{(1)}) \\
&\leq L_1^*(F_1^{(1)}, F_2^{(1)})L_2^*(G_1^{(2)}, G_2^{(2)}|F_1^{(1)}, F_2^{(1)}) \\
&= L_1(G_1^{(2)}, G_2^{(2)})L_2(F_1^{(1)}, F_2^{(1)}|G_1^{(2)}, G_2^{(2)}) \\
&\dots
\end{aligned}$$

This process can be iterated until the likelihood $L = L^*$ no longer increases. The limits of $(G_1^{(k)}, G_2^{(k)})$ and $(F_1^{(k)}, F_2^{(k)})$ as $k \rightarrow \infty$ are the estimators of (G_1, G_2) and (F_1, F_2) .

In the statistical literature this is called the alternating divergence maximization (or minimization) algorithm discussed in Sect. 4.6. More details can be found in Csiszar and Shield (2004).

24.3 Comparable Set Approach in Truncation Problems

Next we discuss a special feature of truncation or double truncation problems.

1. Comparable set approach in left truncation problems

We discuss a test problem between a lifetime and a truncation variable or a covariate based on truncation data. Due to truncation, not all pairs can be used to evaluate the Kendall's tau statistic. To generalize Kendall's tau based on truncated data, Bhattacharya et al. (1983) and Tsai (1990) introduced the concept of comparable pairs. Suppose $(y_i, a_i), i = 1, 2, \dots, n$ are the observed pairs from left truncated population, where $y_i > a_i, i = 1, 2, \dots, n$. A pair of indices (i, j) with $i \neq j$ is said to be comparable if $y_i > a_j$ and $y_j > a_i$. In other words a_i, a_j both can be candidates to truncate y_i and y_j . Define

$$\mathcal{C} = \text{set of comparable pairs.}$$

Note that

$$P(Y_i = y_i, Y_j = y_j | Y_i > a_i, Y_j > a_j) = \frac{f(y_i)}{\bar{F}(a_i)} \frac{f(y_j)}{\bar{F}(a_j)}, \quad y_i > a_i, \quad y_j > a_j.$$

If the (i, j) -th pair is comparable, there are two possible combinations: (1) a_i and a_j truncate y_i and y_j , respectively, i.e., the pairs are (y_i, a_i) , (y_j, a_j) , and (2) a_i and a_j truncate y_j and y_i , respectively, i.e., the pairs are (y_i, a_j) and (y_j, a_i) . Given the two possible combinations, the observed case (1) has a probability

$$\frac{\{f(y_i)/\bar{F}(a_i)\}f(y_j)/\bar{F}(a_j)}{\{f(y_i)/\bar{F}(a_i)\}f(y_j)/\bar{F}(a_j) + \{f(y_i)/\bar{F}(a_j)\}f(y_j)/\bar{F}(a_i)} = 1/2.$$

In other words Y_i and Y_j are exchangeable given that the pairs are comparable. More specifically,

$$P\{Y_i = y_i, Y_j = y_j | Y_i \geq \max(a_i, a_j), Y_j \geq \max(a_i, a_j)\} = \frac{f(y_i)f(y_j)}{\bar{F}^2\{\max(a_i, a_j)\}}, \quad y_i, y_j \geq \max(a_i, a_j).$$

If we are interested in testing the lifetime Y and covariate Z are independent each other, then the generalized Kendall's tau is defined by

$$\hat{\tau} = \sum_{(i,j) \in \mathcal{C}} \text{sign}[(y_i - y_j)(z_i - z_j)] / \#\mathcal{C},$$

where $\#\mathcal{C}$ is the number of all possible comparable pairs. The large sample results and some generalizations to left truncation and right censored problems can be found in Tsai (1990). In general the number of comparable pairs drops quickly when the proportion of truncation gets large. The generalized Kendall's tau statistic based on truncation data may have low efficiency.

2. Double truncation problems with astronomy applications

Efron and Petrosian (1999) found an application of doubly truncated data in astronomy. The observed data consist of n pairs (z_i, y_i) , with y_i a real valued response and z_i a covariate, with observation y_i restricted to a known region $R_i = [u_i, v_i]$.

$$\text{Data} = \{(z_i, y_i), \text{ with } y_i \in R_i = [u_i, v_i] \text{ for } i = 1, 2, \dots, n\}.$$

The n quadruplets (z_i, y_i, u_i, v_i) are observed independently of one another. Furthermore the regions R_i can depend on z_i .

First we consider the no covariate case.

$$Y|u < Y < v \sim \frac{f(y)}{F(v) - F(u)}, \quad u \leq y \leq v.$$

The likelihood is

$$L = \prod_{i=1}^n \frac{f(y_i)}{F(v_i) - F(u_i)}.$$

The nonparametric MLE puts all its mass on the observed responses y_1, \dots, y_n . The general results by Turnbull (1976) on the nonparametric EM algorithm can be applied here. However the underlying distribution F can be estimated only in the interval (a, b) , where $a = \min(y_1, \dots, y_n)$, $b = \max(y_1, \dots, y_n)$. In other words, only the truncated version $dF(y)/\{F(b) - F(a)\}$ is estimable.

The concept of comparable sets can be generalized naturally in the double truncation problems.

A permutation of $y = (y_1, y_2, \dots, y_n)$, say $y^* = (y_1^*, \dots, y_n^*)$, is observable if the permuted values all fall within their truncated regions; that is, if $y_i^* \in R_i$ for $i = 1, 2, \dots, n$. Denote

$$\mathcal{Y} = \text{set of observable permutations.}$$

Intuitively, for given covariates z_1, \dots, z_n and truncation times $[u_i, v_i]$, $i = 1, 2, \dots, n$, comparable set contains all permutations of y_1^*, \dots, y_n^* such that (y_i^*, u_i, v_i, z_i) , $i = 1, 2, \dots, n$ are the possible observations, in other words, $y_i^* \in [u_i, v_i]$, $i = 1, 2, \dots, n$.

Under H_0 : Y and Z are independent, the conditional distribution of y^* given the order statistics $\{(y_{(1)}, \dots, y_{(n)}, z_1, \dots, z_n) \text{ and } (u_1, v_1), \dots, (u_n, v_n)\}$ is uniform over \mathcal{Y} . In fact the observed vector $y = (y_1, \dots, y_n)$ has probability

$$\prod_{i=1}^n [dF(y_i)/(F(v_i) - F(u_i))].$$

For any comparable permutation set (y_1^*, \dots, y_n^*) , the probability density is

$$\prod_{i=1}^n [dF(y_i^*)/(F(v_i) - F(u_i))].$$

Given the comparable sets, due to cancellation the probability of (y_1, \dots, y_n) is $1/k$, where $k = \#\mathcal{Y}$ is the number of all possible comparable permutations. Again the number of comparable pairs decreases when the truncation proportion gets larger. Chen and Liu (2007) proposed an efficient sequential importance sampling strategy for generating permutations under truncation constraint which leads to an accurate calculation of P-value in the testing problems.

Next we consider a semiparametric approach based on doubly truncated data. Consider a density ratio model

$$f(y|z) = \frac{f(y) \exp(yz\beta)}{\mu(z\beta)}, \quad \mu(z\beta) = \int f(y) \exp(yz\beta),$$

where the baseline ‘‘carrier density’’ $f(y)$ is not specified. The simple conditional likelihood approach discussed in Chap. 21 does not work since not all pairs are comparable. However we can use the pairwise conditional likelihood approach for all possible comparable pairs, the resulting likelihood is

$$L_P = \prod_{y_i \in R_j, y_j \in R_i} \frac{\exp\{(y_i z_i + y_j z_j)\beta\}}{\exp\{(y_i z_i + y_j z_j)\beta\} + \exp\{(y_j z_i + y_i z_j)\beta\}},$$

where $R_i = [u_i, v_i]$, $i = 1, 2, \dots, n$. We can maximize L_P for β estimation.

More details on the conditional pairwise likelihood method with applications in recurrent event data can be found in Huang et al. (2010).

24.4 A Review of Cox Regression Model in Survival Analysis

The most popular regression model in survival analysis is the Cox proportional hazards model. This model assumes a multiplicative covariate effect on the risk of the failure event. Specifically the hazard function for individuals with a covariate variable x is given by multiplying an unknown baseline hazard with a parametric function through

$$\lambda(t|x) = \lambda(t) \exp(x\beta), \quad t > 0.$$

Let $\Lambda(t) = \int_0^t \lambda(u)du$ be the cumulative baseline hazard. Then the survival function is

$$\bar{F}(t|z) = \bar{F}^{\exp(z\beta)}(t) = \exp\{-\Lambda(t) \exp(z\beta)\}.$$

We can understand this model through the cumulative hazard transformation of the survival time T

$$P(\Lambda(T) > t) = P(T > \Lambda^{-1}(t)) = \exp\{-\Lambda(\Lambda^{-1}(t)) \exp(x\beta)\} = \exp\{-t \exp(x\beta)\}.$$

In other words, $\Lambda(T)$ has an exponential distribution with rate $\exp(x\beta)$. In this section we review some most popular inference methods for the Cox regression model.

1. Rank-based likelihood approach

Since $\Lambda(t)$ is a monotonic increasing function, the ranks of T_1, \dots, T_n and the ranks of $\Lambda(T_1), \dots, \Lambda(T_n)$ are identical. Let $T_{(1)} \cdots T_{(n)}$ be the order statistics and $R = [(1), \dots, (n)]$ be the rank statistics. Let $x_{(i)}, i = 1, 2, \dots, n$ be the covariates associated with $T_{(i)}, i = 1, 2, \dots, n$. Without loss of generality we can assume that T_1, \dots, T_n come from exponential distributions with rates $\theta_i = \exp(x_i\beta), i = 1, 2, \dots, n$ in the rank likelihood calculation below.

Directly working out the integration, we have (**exercise**)

$$\begin{aligned} L_M(\beta) &= P\{R = [(1), \dots, (n)] | x_1, \dots, x_n\} \\ &= \int_{t_{(1)} < \dots < t_{(n)}} \prod_{i=1}^n \theta_{(i)} \exp(-\theta_{(i)} t_{(i)}) dt_{(1)} \cdots t_{(n)} \\ &= \prod_{i=1}^n \frac{\exp(x_{(i)}\beta)}{\sum_{\ell \in R(T_{(i)})} \exp(x_\ell\beta)}, \end{aligned}$$

where $R(T_{(i)}) = \{T_j : T_j \geq T_{(i)}\}$ is the risk set at $T_{(i)}$, i.e., all individuals who survived up to time point $T_{(i)}$. Kalbfleisch and Prentice (1973) derived this rank based marginal likelihood. Clearly, the baseline cumulative hazard function $\Lambda(t)$ does not play a role in the rank likelihood calculation.

Since this is the marginal rank likelihood, some nice properties of the likelihood function are inherited. For example, the likelihood ratio statistic has a chi-squared

limiting distribution, and

$$E \left[\frac{\partial^2 L_M(\beta)}{\partial \beta \partial \beta^T} \right] = -E \left[\left(\frac{\partial L_M(\beta)}{\partial \beta} \right) \left(\frac{\partial L_M(\beta)}{\partial \beta} \right)^T \right].$$

Under the general transformation model, in principle the rank likelihood method, which does not depend on the unknown transformation function, can be applied. However, there is no closed form in the numerical integration. For example, the proportional odds ratio model is given by

$$P(T > t|x) = \frac{\exp(\Lambda(t) + x\beta)}{1 + \exp(\Lambda(t) + x\beta)},$$

or

$$P(\Lambda(T) > t|x) = P(T > \Lambda^{-1}(t)|x) = \frac{\exp(t + x\beta)}{1 + \exp(t + x\beta)}.$$

The transformation normal model is given by

$$P(T > t|x) = 1 - \Phi\{\Lambda(t) + x\beta\},$$

where $\Lambda(t)$ is a monotonic non-decreasing function and $\Phi(\cdot)$ is the normal cumulative distribution function.

We can denote the conditional density of $\Lambda(T)$ given x as $f(t|x\beta)$, then the likelihood based on rank statistics is

$$n! \int \cdots \int_{t_{(1)} < t_{(2)} < \cdots < t_{(n)}} \prod_{i=1}^n f(t_{(i)}|x_{(i)}\beta) dt_{(1)} \cdots dt_{(n)}.$$

Let $n! \prod_{i=1}^n f(t_{(i)}|0)$ be the density for the order statistics $t_{(1)}, \dots, t_{(n)}$ under $\beta = 0$. Then using Hoeffding's (1951) formula, we have

$$n! \int \cdots \int_{t_{(1)} < t_{(2)} < \cdots < t_{(n)}} \prod_{i=1}^n f(t_{(i)}|x_{(i)}\beta) dt_{(1)} \cdots dt_{(n)} = E \left[\frac{\prod_{i=1}^n f(T_{(i)}|x_{(i)}\beta)}{\prod_{i=1}^n f(T_{(i)}|0)} \right],$$

where $T_{(1)}, \dots, T_{(n)}$ are order statistics generated from $f(t|0)$. Doksum (1987) proposed to evaluate this integral by using Monte Carlo methods. However, this is a cumbersome task as it involves $n!$ terms. When n is large, the computation burden grows exponentially. A compromise is to construct pairwise or triple wise likelihoods. More general cases for dealing with right censoring in the Cox proportional hazards model can be found in Kalbfleisch and Prentice (2002).

2. Partial likelihood method

Next, we use the conditional likelihood method to eliminate the nuisance baseline function $\lambda(t)$. Denote the observed data as (Y, x, δ) , where

$$Y = \min(T, C), \quad \delta = I(T \leq C).$$

Denote the conditional densities of T and C given x , respectively, as

$$T \sim f(t|x), \quad C \sim g_c(c|x),$$

the corresponding survival functions are denoted as $\bar{F}(t|x)$ and $\bar{G}_c(c|x)$. The marginal density of X is $\eta(x)$.

The partial likelihood method was proposed by Cox (1972, 1975). Essentially at each failure time point it is a conditional likelihood approach for a “case and control” problem. It is implemented as follows.

At each failure time point t_i , $\delta_i = 1$, we can define artificial cases and controls as follows:

Cases are defined as deaths at t_i ($T = t_i$) and controls are defined as those individuals who survive and remain under observation at t_i ($Y \geq t_i$). Given a case, the observed covariate has a probability density

$$h_{1t}(x) = P(X = x|T = t, \delta = 1) = \frac{\bar{G}_c(t|x)f(t|x)\eta(x)}{P(Y = t, \delta = 1)}.$$

Similarly the density for a control covariate is given by

$$h_{0t}(x) = P(X = x|Y \geq t) = \frac{\bar{G}_c(t|x)\bar{F}(t|x)\eta(x)}{P(Y > t)}.$$

Note that the density ratio between cases and controls is

$$h_{1t}(x)/h_{0t}(x) = \frac{P(Y > t)}{P(Y = t, \delta = 1)} \frac{f(t|x)}{\bar{F}(t|x)} = \frac{P(Y > t)}{P(Y = t, \delta = 1)} \lambda(t) \exp(x\beta) = \exp\{\alpha(t) + x\beta\},$$

where $\alpha(t) = \log \lambda(t) + \log\{P(Y > t)/P(Y = t, \delta = 1)\}$. In other words, at time t the density of “cases” and that of “controls” satisfy an exponential tilting model. Note that t can be treated as a stratification variable.

We can use the conditional approach discussed in Chap. 12 to eliminate the nuisance function $\alpha(t)$. Given covariates x_{11}, \dots, x_{1k_1} of “cases” and covariates x_{01}, \dots, x_{0k_0} of “controls” at time t , the probability of observing “cases” and “controls” covariates is

$$\begin{aligned} & \frac{h_{1t}(x_{11}) \dots h_{1t}(x_{1k_1}) h_{0t}(x_{01}) \dots, h_{0t}(x_{0k_0})}{\sum_{\tau} h_{1t}\{\tau(x_{11})\} \dots h_{1t}\{\tau(x_{1k_1})\} h_{0t}\{\tau(x_{01})\} \dots, h_{0t}\{\tau(x_{0k_0})\}} \\ &= \frac{\exp(x_{11}\beta) \dots \exp(x_{1k_1}\beta)}{\sum_{\tau} \exp\{\tau(x_{11})\beta\} \dots \exp\{\tau(x_{1k_1})\beta\}}, \end{aligned}$$

where τ are all possible k_1 “cases” and k_0 “controls” combinations. In this case the nuisance function $\alpha(t)$ is eliminated. In the absence of ties, it reduces to the Cox partial likelihood up to a constant. Finally the overall conditional likelihood is the product of conditional likelihoods at all failure time points.

3. Godambe's justification

Essentially Cox's partial log-likelihood is the summation of all log conditional likelihoods constructed at all observed failure times. A natural question is whether the simple summation of log conditional likelihoods is optimal. Godambe (1985) found that, indeed, the optimal estimating equation approach matches the Cox partial likelihood score. Godambe (1985) used the following arguments.

If the joint density of (y_1, \dots, y_n) can be decomposed as

$$(y_1, \dots, y_n) \sim f(y_1, \dots, y_n, \theta, \phi) = \prod_{i=1}^n f_{i-1}(y_i, \theta, \phi),$$

where

$$f_{i-1}(y_i, \theta, \phi) = f(y_i | y_1, \dots, y_{i-1}, \theta, \phi),$$

and θ is the parameter of interest and ϕ is the nuisance parameter. If $t_i(y_1, \dots, y_i)$, $i = 1, 2, \dots, n$, is a sufficient statistic for ϕ , then

$$f_{i-1}(y_i | t_i) = f(y_i | y_1, \dots, y_{i-1}, t_i, \theta, \phi) = f(y_i | y_1, \dots, y_{i-1}, t_i, \theta),$$

which is independent of the nuisance parameter ϕ . Godambe (1985) showed that the partial likelihood score for θ

$$\sum_{i=1}^n \partial \log f_{i-1}(y_i | t_i, \theta) / \partial \theta$$

is optimal in the class \mathcal{L} , where

$$\mathcal{L} = \{g : |g = \sum_{i=1}^n h_i a_{i-1}\},$$

$E_{i-1}[h_i(y_1, \dots, y_i; \theta)] = 0$, E_{i-1} denotes the expectation holding the first $i-1$ values, namely y_1, \dots, y_{i-1} , fixed, and a_{i-1} is a function of y_1, \dots, y_{i-1} and θ .

The partial likelihood score for θ (Cox 1975) is

$$\sum_{i=1}^n \frac{\partial \log f_{i-1}(y_i | t_i, \theta)}{\partial \theta} = \sum_{i=1}^n h_i(\theta),$$

where t_i is the pooled covariate information of x_{11}, \dots, x_{1k_1} and x_{01}, \dots, x_{0k_0} in "cases" and "controls" at time t_i , and y_i is the observed covariate information for "cases" and "controls" in the conditional logistic approach in justification 3 above.

We can check $E_{i-1}(h_i) = 0$ and $E(h_i h_j) = 0$, $i \neq j = 1, 2, \dots, n$. In the class of estimating equations $\sum_{i=1}^n h_i a_{i-1}$, the optimal one is given by (5.1.1) in Chap. 5, i.e., $\sum_{i=1}^n h_i a_{i-1}^*$, where

$$a_{i-1}^* = \frac{E_{i-1}\{\partial^2 \log f_{i-1}(y_i|t_i, \theta)/\partial\theta^2\}}{E_{i-1}\{\partial \log f_{i-1}(y_i|t_i, \theta)/\partial\theta\}^2} = -1$$

since $f_{i-1}(y_i|t, \beta)$ is the genuine conditional density which satisfies the information identity. Therefore Cox's partial likelihood score is optimal.

Since Cox's (1972; 1975) landmark partial likelihood work, it has become the standard method in regression analysis of survival data. Unfortunately this method does not work if the conditional hazard model departs from the proportional hazards model, for example, if the model follows the proportional odds ratio model or general transformation model. As an alternative, we discuss a powerful profile likelihood method proposed by Breslow (1972). Some variations were discussed by Bailey (1984) and Ren and Zhou (2011) more recently. Zeng and Lin (2007) discussed this method for general transformation models such as those with a possibly random frailty for the baseline hazard.

4. Two versions of the profile likelihood approach

Even though the partial likelihood approach is efficient for estimating the log hazard ratio parameter β , it does not provide any information for the baseline hazard $\lambda(t)$. In many applications, for example for the purpose of prediction of survival probability for a given covariate, however, $\lambda(t)$ would also be of interest. Another very powerful tool in handling the nuisance parameter problem is the profile likelihood method. For the Cox model, Breslow (1972) used a piecewise constant hazard function approach. We will study two different versions. The first one is based on the maximization of baseline hazard function and the second one is based on the maximization of baseline distribution function. Even though large sample results are equivalent for the two methods, their behavior may be different in small sample size problems.

Version I

We can discretize the baseline hazard such that it has jumps only at each of the observed failure time. Without loss of generality, we assume

$$t_1 < t_2 < \dots < t_n.$$

Similar to previous discussion in no covariate case, the baseline hazard function $\lambda(t)$ has jumps only at those observed failure time points. The log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^n [\delta_i \{\log \lambda(t_i) + x_i \beta\} - \Lambda(t_i) \exp(x_i \beta)] \\ &= \sum_{i=1}^n [\delta_i \{\log \lambda_i + x_i \beta\} - (\sum_{j=1}^i \lambda_j) \exp(x_i \beta)] \\ &= \sum_{i=1}^n [\delta_i \{\log \lambda_i + x_i \beta\} - \lambda_i \sum_{j=i}^n \exp(x_i \beta)], \end{aligned}$$

where we have used the fact that the cumulative hazard $\Lambda(t_i) = \sum_{j=1}^i \lambda_j$, $\lambda_j = \lambda(t_j)$. If $\delta_i > 0$, differentiating ℓ with respect to λ_i , we have

$$\frac{\partial \ell}{\partial \lambda_i} = \frac{\delta_i}{\lambda_i} - \sum_{j=i}^n \exp(x_j \beta) = 0.$$

Easily we have

$$\hat{\lambda}_i = \frac{\delta_i}{\sum_{j=i}^n \exp(x_j \beta)}.$$

Replacing λ_i with $\hat{\lambda}_i$ in ℓ yields the profile likelihood for β

$$L(\beta) = \prod_{i=1}^n \left(\frac{\exp(x_i \beta)}{\sum_{j \geq i} \exp(x_j \beta)} \right)^{\delta_i}.$$

Version II

The second approach is more like empirical likelihood method by discretizing the baseline distribution function. It was studied by Bailey (1984) and Kalbfleisch and Prentice (2002) (Sect. 4.3). Later Ren and Zhou (2011) gave details for the large sample results.

Without loss of generality we assume there are no ties

$$t_1 < t_2 < \cdots < t_n.$$

The log-likelihood is

$$L = \prod_{i=1}^n [dF(t_i|x_i)]^{\delta_i} [\bar{F}(t_i|x_i)]^{1-\delta_i}.$$

Denote

$$p_i = dF(t_i) = P(Y \leq t_i) - P(Y < t_i), \quad i = 1, 2, \dots, n, \quad p_{n+1} = P(Y > t_n),$$

then the p_i 's satisfy the constraints

$$\sum_{i=1}^{n+1} p_i = 1, \quad p_i \geq 0.$$

Moreover

$$dF(t_i|x_i) = \beta_i \bar{F}^{\beta_i-1}(t_i)p_i, \quad \beta_i = \exp(x_i \beta) > 0.$$

The likelihood can be written as

$$L = \prod_{i=1}^n \{dF(t_i)\beta_i\}^{\delta_i} \{\bar{F}(t_i)\}^{\beta_i - \delta_i} = \prod_{i=1}^n (\beta_i p_i)^{\delta_i} \left(\sum_{j=i+1}^{n+1} p_j \right)^{\beta_i - \delta_i}.$$

Define

$$\lambda_i = p_i/b_i, \quad i = 1, 2, \dots, n, \quad b_i = \sum_{j=i}^{n+1} p_j,$$

then

$$b_1 = 1, \quad b_{n+1} = p_{n+1}, \quad 1 - \lambda_i = b_{i+1}/b_i,$$

$$b_{i+1} = b_i(1 - \lambda_i) = b_i \prod_{j=1}^{i-1} (1 - \lambda_j),$$

$$p_i = \lambda_i b_i = \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j).$$

As a consequence

$$L = \prod_{i=1}^n (\beta_i \lambda_i)^{\delta_i} (1 - \lambda_i)^{d_i - \delta_i},$$

where

$$d_i = \beta_i + \dots + \beta_n.$$

Taking derivative with respect to λ_i yields

$$\frac{\partial \ell}{\partial \lambda_i} = \frac{\delta_i}{\lambda_i} - \frac{d_i - \delta_i}{1 - \lambda_i} = 0.$$

$$\hat{\lambda}_i = \delta_i/d_i, \quad i = 1, 2, \dots, n.$$

The profile likelihood is

$$L = \prod_{i=1}^n \left(\frac{\beta_i}{d_i} \right)^{\delta_i} \left(\frac{d_i - \delta_i}{d_i} \right)^{d_i - \delta_i} =: L_c(\beta) \prod_{i=1}^n \left(\frac{d_i - \delta_i}{d_i} \right)^{d_i - \delta_i},$$

where $L_c(\beta)$ is the Cox partial likelihood derived before. Define

$$\psi_n(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}, \quad \phi_n(\beta) = \frac{\partial \log L_c(\beta)}{\partial \beta}.$$

Ren and Zhou (2011) showed that

$$\psi_n(\beta) = \phi_n(\beta) + O_p\left(\frac{\log n}{n}\right).$$

As a result, the full profile likelihood estimation and the Cox partial likelihood estimation of β have the same asymptotic distribution.

Exercise Under the regularity conditions specified in Andersen et al. (1993), in distribution the profile likelihood ratio statistic satisfies

$$R(\beta) = 2[\max_{\beta} \log L(\beta) - \log L(\beta_0)] \rightarrow \chi^2(p),$$

where p is the dimension of β .

Remark The profile likelihood method is a very powerful tool in statistical inference. In a comprehensive discussion paper by Zeng and Lin (2007), the profile likelihood method was applied to find the semiparametric MLE under general transformation models. Unfortunately there is no closed form in the maximization stage. Some iteration approaches are needed.

5. Projection method

Now we discuss the projection method by treating the baseline hazard $\lambda(t)$ as a nuisance function. For any given sub-model $\lambda(t) = \lambda(t, \eta)$, the log-likelihood based on a generic data set is

$$\ell = \delta(\log \lambda(Y, \eta) + x\beta) - \int_0^Y \lambda(u, \eta) \exp(x\beta) du.$$

The score estimating function for β is

$$S_{\beta} = \frac{\partial \ell}{\partial \beta} = \delta x - \int_0^Y \lambda(u, \eta) \exp(x\beta) x du = \int_0^{\infty} x dM(u),$$

where $dM(t) = dN(t) - \lambda(t, \eta) \exp(x\beta) I(Y \geq t) dt$, $N(u) = \delta I(Y \geq u)$. The score for the nuisance parameter η is

$$\frac{\partial \ell}{\partial \eta} = \delta \frac{\partial \lambda(Y, \eta)/\partial \eta}{\lambda(Y, \eta)} - \int_0^Y \partial \lambda(u, \eta)/\partial \eta \exp(x\beta) du.$$

Let

$$h(u) = \frac{\partial \lambda(u, \eta)/\partial \eta}{\lambda(u, \eta)},$$

then the nuisance parameter space is given by

$$\begin{aligned}
S_\eta &= \delta h(Y) - \int_0^Y h(u)\lambda(u) \exp(x\beta)du \\
&= \delta h(Y) - \int_0^\infty h(u)\lambda(u)I(Y \geq u) \exp(x\beta)du \\
&= \int h(u)[dN(u) - \lambda(u)I(Y \geq u) \exp(x\beta)du] := \int h(u)dM(u).
\end{aligned}$$

The nuisance parameter space is spanned by $\{\int h(u)dM(u) : h \text{ is a measurable function}\}$.

Define

$$S_A = \int \{x - \bar{x}(u)\}dM(u), \quad \bar{x}(u) = \frac{\sum_{i=1}^n x_i \exp(x_i\beta)I(Y_i \geq u)}{\sum_{i=1}^n \exp(x_i\beta)I(Y_i \geq u)}.$$

Then S_β can be decomposed as

$$S_\beta = S_A + \int \bar{x}(u)dM(u).$$

Exercise Show

$$E[\int \{x - \bar{x}(u)\}dM(u) \int h(u)dM(u)] = 0.$$

Therefore S_A is orthogonal to the nuisance parameter space. Moreover S_A does not depend on the nuisance function $\lambda(t)$, we can use it as an estimating function for β . In fact the partial likelihood score for β is equivalent to $S_A(\beta) = 0$. Begun et al. (1983) calculated the information lower bound for the Cox proportional hazards model. Small and McLeish (1989) also discussed this projection method.

6. Extension to left truncation and right censored case for Cox regression model

Methods discussed above can be easily adapted to estimate the underlying parameters in the Cox model based on left truncation and right-censored data. Let A be a left truncation variable and T be the lifetime. An individual is observable if and only if $T > A$. Moreover T can be right censored by C . Define $Y = \min(T, C)$, $\delta = I(T \leq C)$. The conditional likelihood is

$$(A, Y, \delta)|x, T > A = a \sim \frac{f^\delta(y|x)\bar{F}^{1-\delta}(y|x)}{\bar{F}(a|x)}.$$

Again at each of the failure time t , we can construct “case and control” data, where cases satisfy

$$X|Y = t, \delta = 1, T > A = a \sim \frac{\bar{G}_c(t|x)f(t|x)g(x)/\bar{F}(a)}{P(Y = t, \delta = 1)/\bar{F}(a)} = \frac{\bar{G}_c(t|x)f(t|x)g(x)}{P(Y = t, \delta = 1)}$$

and controls satisfy

$$X|Y > t > A, T > A = a \sim \frac{\bar{G}_c(t|x)\bar{F}(t|x)g(x)/\bar{F}(a)}{P(Y > t)/\bar{F}(a)} = \frac{\bar{G}_c(t|x)\bar{F}(t|x)g(x)}{P(Y > t)}.$$

Note that the ratio of the covariate densities between case and control is

$$\left\{ \frac{\bar{G}_c(t|x)f(t|x)g(x)}{P(Y = t, \delta = 1)} \right\} \left\{ \frac{\bar{G}_c(t|x)\bar{F}(t|x)g(x)}{P(Y > t)} \right\}^{-1} = \frac{P(Y > t)f(t|x)}{\bar{F}(t|x)P(Y = t, \delta = 1)} = \exp\{\alpha(t) + x\beta\},$$

where $\alpha(t) = \log\{\lambda(t)P(Y > t)/P(Y = 1, \delta = 1)\}$. Again this is the density ratio model discussed in Sect. 11.1. We can use the conditional method discussed in Chap. 12 to eliminate the nuisance parameter $\alpha(t)$ at each death time t . Kalbfleisch and Lawless (1991) found that the partial likelihood approach for truncation data has exactly the same form as that in the absence of truncation except for replacing risk set by $\sum_{j=1}^n I(y_j \geq y_i > a_j)$ to account for the left truncation.

Exercise Derive the maximum likelihood estimate by profiling out either the baseline hazard function or the baseline distribution function based on left truncation and right censored data.

24.5 Inferences for AFT Model and Quantile Regression Model

Other than the proportional hazards model, the accelerated failure time model (AFT), which relates covariates linearly to the logarithm of the survival time, has been one of the most commonly used regression models for analyzing right censored survival data (Kalbfleisch and Prentice 2002). The linear regression structure after the log-transformation of failure time and the straightforward interpretation of the regression coefficients are especially appealing to biomedical and industrial investigators. We present some commonly used statistical inference methods for the AFT model, followed by the method proposed by Peng and Huang (2008) for the quantile regression models.

1. Buckley and James imputation estimation

Using imputation idea, Buckley and James (1979) proposed an imputation method for AFT model based on right-censored data. For convenience we denote T as the log transformed version of the lifetime. The AFT model is given by

$$T = x\beta + \epsilon.$$

The observed data are $Y_i = \min(T_i, C_i)$, $\delta_i = I(Y_i \leq C_i)$, $X_i, i = 1, 2, \dots, n$. In the absence of right censoring, the ordinary least squares method solves

$$\sum_{i=1}^n x_i(t_i - x_i\beta) = 0.$$

Since not all T_i 's are available due to right censoring, we need to impute those censored T_i . Note that

$$E[T|T > c, x] = \frac{E\{TI(T > c)|x\}}{P(T > c|x)} = x\beta + \frac{E[\epsilon I(\epsilon > c - x\beta)]}{P(T > c|x)} = x\beta + \frac{\int_{c-x\beta}^{\infty} \epsilon dF(\epsilon)}{\bar{F}(c - x\beta)}.$$

Finally we can replace T by

$$\tilde{T} = \delta T + (1 - \delta) \left\{ x\beta + \frac{\int_{c-x\beta}^{\infty} \epsilon dF(\epsilon)}{\bar{F}(c - x\beta)} \right\}.$$

An iterative algorithm can be used to solve the above equation.

- (1) For fixed β , use the transformed data $\min(T_i - x_i\beta, c_i - x_i\beta)$, $\delta_i, i = 1, 2, \dots, n$ to estimate F by the Kaplan–Meier estimator.
- (2) Solve the least squares by using the imputed \tilde{T} .
- (3) Repeat Steps (1) and (2) until convergence.

Ritov (1990) gave some theoretical justifications on Buckley and James' estimator.

2. The log rank based method

Tsiatis (1990) and Wei et al. (1990) identified a class of estimators for regression parameters in a linear regression model with right censored data. They used linear rank tests for right censored data as estimating equations. They showed that their estimators are consistent and have asymptotical normal distributions.

Recall the score estimating equation for Cox model $\lambda(t|x) = \lambda(t) \exp(x\gamma)$ is

$$\sum_{i=1}^n \delta_i \left(x_i - \frac{\sum_{j=1}^n x_j \exp(x_j\gamma) I(y_j \geq y_i)}{\sum_{j=1}^n \exp(x_j\gamma) I(y_j \geq y_i)} \right) = 0.$$

If $\gamma = 0$, it becomes

$$\sum_{i=1}^n \delta_i \left(x_i - \frac{\sum_{j=1}^n x_j I(y_j \geq y_i)}{\sum_{j=1}^n I(y_j \geq y_i)} \right) = 0.$$

In the AFT model,

$$\log Y_i = x_i\beta + \epsilon_i, i = 1, 2, \dots, n$$

We can treat $Z_i = \min(\log Y_i - x_i\beta, \log C_i - x_i\beta)$, $\delta_i = I(Y_i \leq C_i)$, $i = 1, 2, \dots, n$ as i.i.d. data. Therefore we can use

$$\sum_{i=1}^n \delta_i \left(x_i - \frac{\sum_{j=1}^n x_j I(\log y_j - x_j\beta \geq \log y_i - x_i\beta)}{\sum_{j=1}^n I(\log y_j - x_j\beta \geq \log y_i - x_i\beta)} \right) = 0$$

as the estimating equation for β .

3. Projection method in AFT model

Next we study the projection method by treating β as the parameter of interest and the underlying density f as a nuisance function. Again we assume

$$Y = x\beta + \epsilon, \quad \epsilon \sim f(\epsilon).$$

The corresponding hazard function is denoted as $\lambda(t)$. With censored data the log-likelihood is

$$\ell = \sum_{i=1}^n \delta_i \log \lambda(y_i - x_i\beta) - \int_0^\infty \lambda(u - x_i\beta) I(y_i \geq u) du.$$

The score for β is

$$\begin{aligned} S_\beta &= \sum_{i=1}^n \left[-x_i \delta_i \frac{\lambda'(y_i - x_i\beta)}{\lambda(y_i - x_i\beta)} + x_i \int \frac{\lambda'(u - x_i\beta)}{\lambda(u - x_i\beta)} \lambda(u - x_i\beta) I(y_i \geq u) du \right] \\ &= - \sum_{i=1}^n x_i \int \frac{\lambda'(u - x_i\beta)}{\lambda(u - x_i\beta)} dM_i(u), \end{aligned}$$

where $\lambda'(u) = d\lambda(u)/du$, and

$$M_i(u) = \delta_i I(y_i - x_i\beta \leq u) - \lambda(u - x_i\beta) I(y_i - x_i\beta \geq u).$$

The score for the nuisance parameter η is

$$\begin{aligned} S_\eta &= \sum_{i=1}^n \delta_i \frac{\partial \lambda(y_i - x_i\beta, \eta)}{\partial \eta} - \int \frac{\partial \lambda(u - x_i\beta, \eta)}{\partial \eta} \lambda(u - x_i\beta) I(y_i \geq u) du \\ &=: \int h(u - x_i\beta) dM_i(u). \end{aligned}$$

Note that

$$S_\beta = - \sum_{i=1}^n \int \frac{\lambda'(u - x_i\beta)}{\lambda(u - x_i\beta)} \{x_i - \bar{x}(u)\} dM_i(u) - \sum_{i=1}^n \int \bar{x}(u) dM_i(u), \quad \bar{x}(u) = \frac{\sum_{i=1}^n x_i I(y_i - x_i\beta \geq u)}{\sum_{i=1}^n I(y_i - x_i\beta \geq u)}.$$

It can be shown that the two terms are orthogonal to each other. Therefore we can use

$$\sum_{i=1}^n \int \frac{\lambda'(u - x_i\beta)}{\lambda(u - x_i\beta)} \{x_i - \bar{x}(u)\} dM_i(u) = 0$$

to estimate β . Unfortunately this estimating equation depends on $\lambda'(u)$ and $\lambda(u)$ which in general are unknown. This is a typical problem in general semiparametric models, where even though it is possible to find the efficient score by using the projection method, the score function cannot be implemented due to the presence of some unknown baseline function.

Lin and Chen (2013) used a kernel method to estimate $\lambda'(t)/\lambda(t)$. As an alternative we may postulate a “working model”

$$\lambda(u) = \lambda(u, \eta).$$

Since $\lambda'(u - x_i\beta)/\lambda(u - x_i\beta)$ plays a role of weighting function in above estimating equation, even a misspecified form of $\lambda(u, \eta)$ would retain the consistency. For fixed η , we can find the asymptotic variance of β , denoted it as $V(\beta_0, \eta)$. Then we can minimize $\text{trace}\{V(\eta_0, \eta)\}$ with respect to η to find an optimal estimate of β . A different approach is to maximize the parametric likelihood with respect to η .

It is worth mentioning that Zeng and Lin (2007) developed an approximate non-parametric maximum likelihood method for the accelerated failure time model with possible time-dependent covariates. The regression parameters are estimated by maximizing a kernel-smoothed profile likelihood function. The maximization can be achieved through conventional gradient-based search algorithms. The resulting estimators are consistent and asymptotically normal. However it is not easy to implement their method in practical applications due to the complication and difficult on the choice of window size in the kernel smoothing stage.

4. Quantile regression for right censored data

The classical work on quantile regression by Koenker and Bassett (1978) is influential in econometric and statistical literature. A good reference is Koenker's monograph (Koenker 2005). Quantile regression model is a generalization of the median regression. The linear median regression model assumes

$$T_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where the median of ϵ given x is 0. In general, $\epsilon_i, i = 1, 2, \dots, n$ do not need to be identically distributed. This is in contrast to the homogeneous linear regression model where the error distribution is identical. As a consequence

$$P(T_i < x_i\beta | x_i) = 1/2.$$

A natural generalization is

$$P\{Y_i < x_i \beta(\tau) | x_i\} = \tau, \quad 0 < \tau < 1 \quad (24.5.7)$$

for the τ -th quantile. If this is true for all $\tau \in (0, 1)$, then we have a quantile regression process. This is a very rich semiparametric model. Mathematically it can be shown that $\beta(\tau)$ can be estimated by minimizing $\sum_{i=1}^n \rho_\tau(y_i - x_i \beta)$, where $\rho_\tau(\epsilon) = \{\tau - I(\epsilon < 0)\}$ is called the check function.

As discussed before the general transformation model takes the form

$$h(T) = x\beta + \epsilon,$$

where $h(\cdot)$ is an unknown monotonic non-decreasing function. The Cox regression model can be treated as a special case. In fact the Cox proportional hazards model is equivalent to

$$\log \Lambda_0(T) = x\beta + \epsilon$$

where $\Lambda_0(t)$ is the underlying baseline cumulative hazard and ϵ has an extreme distribution. On the other hand if ϵ has a logistic distribution, it reduces to Bennett's (1983) proportional odds model. Moreover, if the transformation function $h(t)$ is known but the error distribution is unknown, for example, $h(t) = \log t$, then the transformation model becomes the accelerated failure time model. A common feature of these three models is the i.i.d. error assumption, which implies that, for some appropriate choice of $h(\cdot)$, the transformed survival times can be expressed as a pure location shift model. If we formulate a family of conditional quantile models for $h(T)$, then

$$Q_{h(T)}(\tau|x) = x\beta + F^{-1}(\tau), \quad 0 < \tau < 1$$

where $Q_{h(T)}(\tau|x)$ is the conditional quantile of $h(T)$. Clearly the general quantile process (24.5.7) is a wider class than the transformation models.

Next we briefly review the results by Peng and Huang (2008) on the regression quantile problem for right censored survival data.

Let

$$Y_i = \min(T_i, C_i), \quad \delta_i = I(T_i \leq C_i), \quad X_i \quad i = 1, 2, \dots, n$$

be the observed lifetime (possible right censored), censoring indicator and covariate. For the underlying lifetime T , we assume a log quantile regression process

$$P(\log T \leq x\beta(\tau)|x) = \tau, \quad \tau \in (0, 1)$$

Equivalently,

$$P\{T \leq \exp(x\beta(\tau))|x\} = \tau.$$

In terms of the conditional distribution function $F(\cdot|x)$ and quantile function $F^{-1}(\cdot|x)$,

$$F\{\exp(x\beta(\tau))|x\} = \tau, \quad F^{-1}(\tau|x) = \exp(x\beta(\tau)).$$

Denote the survival function of C as $\bar{G}_c(c|x) = P(C > c|x)$. For right censored data, we need to modify above estimating equation by taking the right censoring into account. Note that

$$\begin{aligned} E[\Delta I\{Y < \exp(x\beta(\tau))\}|x] &= E[\bar{G}_C(T|x)I\{T < \exp(x\beta(\tau))\}|x] \\ &= \int_0^{\exp(x\beta(\tau))} \bar{G}_C(y|x)dF(y|x) = \int_0^\tau \bar{G}_C\{F^{-1}(t|x)|x\}dt \\ &= \int_0^\tau \bar{G}_C\{\exp(x\beta(t))\}dt = E\left[\int_0^\tau I(Y > \exp(x\beta(t)))/(1-t)dt\right]. \end{aligned}$$

Therefore

$$E\left[\Delta I\{Y < \exp(x\beta(\tau))\} - \int_0^\tau I(Y > \exp(x\beta(t)))/(1-t)dt\right] = 0.$$

Peng and Huang (2008) used this estimating equation to make inference for β . To implement it, they chose

$$0 < \tau_1 < \tau_2 < \dots < \tau_k < 1$$

and then solved the estimating equations

$$\sum_{i=1}^n x_i \left[\Delta_i I\{Y < \exp(x_i\beta(\tau_k))\} - \int_0^{\tau_k} I(Y_i > \exp(x_i\beta(t)))/(1-t)dt \right] = 0$$

iteratively. Fortunately they found that the linear programming can be adapted. An R package is available to implement their method.

<http://svitsrv25.epfl.ch/R-doc/library/quantreg/html/crq.html>.

24.6 Double Empirical Likelihoods Utilizing Auxiliary Information

Information sharing has been becoming very popular in modern researches. It was demonstrated in Sect. 14.4, auxiliary information such as covariate specific survival probability can be utilized to enhance estimation efficiency in case-control studies. The Gail model (1983) used the marginal hazard function information extracted from the surveillance, epidemiology, and end results (SEER) program of the National Cancer Institute to find a more efficient estimator for some rare cancer problems. Liu et al. (2014) proposed some new methods to estimate the risk of disease with time-to-event data and applied their methods to data sets from the Women's Health Initiative.

Below we illustrate a survival analysis problem with the use of auxiliary information on the covariate specific survival. We consider the Cox proportional hazards model

$$\lambda(t|x, z) = \lambda(t) \exp(x\beta + \gamma z),$$

where the baseline $\lambda(t)$ is not specified. The cumulative hazard is denoted as $\Lambda(t) = \int_0^t \lambda(s)ds$.

The full conditional (conditioning on X and Z) log-likelihood is

$$\ell = \sum_{i=1}^n [\delta_i \{\log \lambda(t_i) + x_i\beta + z_i\gamma\} - \exp(x_i\beta + z_i\gamma)\Lambda(t_i)].$$

To maximize this likelihood with respect to $\lambda(\cdot)$, Breslow (1972) showed that $\lambda(t)$ may be discretized at all those failure data points. Rewrite the log-likelihood as

$$\ell = \sum_{i=1}^n [\delta_i \{\log \lambda_i + x_i\beta + z_i\gamma\} - \exp(x_i\beta + z_i\gamma) \sum_{j=1}^n \delta_j \lambda_j I(t_j \leq t_i)]. \quad (24.6.8)$$

By differentiating ℓ with respect to λ_i and setting it to 0, we have

$$\hat{\lambda}_i = \frac{\delta_i}{\sum_{j=1}^n \exp(x_j\beta + z_j\gamma) I(t_j \geq t_i)}.$$

The cumulative hazard can be estimated by the so called Breslow estimator

$$\hat{\Lambda}(t) = \sum_{i=1}^n \hat{\lambda}_i I(t_i \leq t).$$

Zhou (2015) explored utilization of the baseline cumulative hazard information at a specified time point, for example, when $\Lambda(t_0)$ is known. Then we need to maximize the log-likelihood ℓ subject to the constraint

$$\sum_{i=1}^n \delta_i I(t_i \leq t_0) \lambda_i - \Lambda(t_0) = 0, \quad \Lambda(t_0) = \Lambda_0.$$

More generally, Zhou (2015) considered the hazard constrained inference based on

$$\int g(s)d\Lambda(s) = \theta.$$

In particular if $g(s) = I(s \leq t_0)$ then $\theta = \Lambda_0 = \Lambda(t_0)$. Let

$$H = \sum_{i=1}^n [\delta_i \{\log \lambda_i + x_i \beta + z_i \gamma\} - \exp(x_i \beta + z_i \gamma) \sum_{j=1}^n \delta_j \lambda_j I(t_j \leq t_i)] - n\nu [\sum_{i=1}^n \delta_i \lambda_i I(t_i \leq t_0) - \Lambda_0].$$

Differentiating H with respect to λ_i 's and some algebra, we can show that

$$\frac{\partial H}{\partial \lambda_i} = \frac{\delta_i}{\lambda_i} - \sum_{j=1}^n \exp(x_j \beta) I(T_j \geq T_i) - n\nu \delta_i I(t_i \leq t_0) = 0.$$

The constrained partial likelihood score is

$$S_P = \sum_{i=1}^n \delta_i X_i - \sum_{i=1}^n \delta_i \frac{\sum_{j=1}^n x_j \exp(x_j \beta) I(T_j \geq T_i)}{\sum_{j=1}^n \exp(x_j \beta) I(T_j \geq T_i) + n\nu I(T_i \leq t_0) \delta_i},$$

where ν is the Lagrange multiplier determined by

$$\sum_{i=1}^n \frac{\delta_i I(T_i \leq t_0)}{\sum_{j=1}^n \exp(x_j \beta) I(T_j \geq T_i) + n\nu I(T_i \leq t_0) \delta_i} - \Lambda_0 = 0.$$

Zhou (2015) showed that the likelihood ratio statistic for the underlying parameter converges to a standard chi-squared distribution.

Next we consider a situation where the auxiliary information is given by the covariate specific disease prevalence

$$P(Y > t_0 | a < X < b) = \phi(t_0, a, b).$$

This type information is commonly available from SEER data base. For example, if X denotes disease onset age, then different survival rates can be obtained for those younger or older than 65 ovarian cancer patients.

Equivalently, the auxiliary information is

$$P(Y > t_0, a < X < b) - \phi(t_0, a, b) P(a < X < b) = 0,$$

or

$$E[I(Y > t_0, a < X < b) - \phi(t_0, a, b) I(a < X < b)] = 0.$$

Using the Cox proportional hazard model, we have

$$\int \int \bar{F}^{\exp(x\beta_1 + z\beta_2)}(t_0) I(a < z < b) g(x, z) dx dz - \phi(t_0, a, b) E[I(a < X < b)] = 0,$$

or

$$E \left[\exp\{-\Lambda(t_0) \exp(X\beta_1 + Z\beta_2)\} I(a < X < b) - \phi(t_0, a, b) I(a < X < b) \right] = 0,$$

where $g(x, z)$ is the marginal density of X and Z . If there are different $a_i, b_i, i = 1, 2, \dots, I$ such that the corresponding $\phi(t_0, a_i, b_i)$'s are available, we can denote

$$\psi_i(t_0, x, z) = \exp\{-\Lambda(t_0) \exp(X\beta_1 + Z\beta_2)\}I(a_i < Z < b_i) - \phi(t_0, a_i, b_i)I(a_i < Z < b_i),$$

$$i = 1, 2, \dots, I.$$

Let

$$\psi(t_0, x, z, \beta) = (\psi_1(t_0, x, z), \dots, \psi_I(t_0, x, z, \beta)).$$

Using the marginal constraints, we can write out the profile marginal empirical likelihood as

$$\ell_M = - \sum_{i=1}^n \log[1 + \gamma^T \psi(t_0, x_i\beta_1 + z_i\beta_2, \Lambda)], \quad (24.6.9)$$

where γ is the Lagrange multiplier determined by the constraint equation

$$\sum_{i=1}^n \frac{\psi(t_0, x_i\beta_1 + z_i\beta_2, \Lambda)}{1 + \gamma^T \psi(t_0, x_i\beta_1 + z_i\beta_2, \Lambda)} = 0.$$

Finally the full log-likelihood is the summation of the two empirical likelihoods

$$\ell_F = \ell_c + \ell_M. \quad (24.6.10)$$

Huang et al. (2016) derived the large sample results. Their numerical results showed that the double empirical likelihood method may improve the estimation efficiency for β_1 substantially.

24.7 Case-Cohort Study

Case-cohort study is cost effective in data collection stage. In a large cohort with infrequent failures, ascertainment of covariate histories on all cohort members may be too expensive. Furthermore early censored individuals may not be as informative as failed individuals. Therefore it seems reasonable to obtain history information for all failed individuals (cases) and those randomly selected matched controls. In other words we end up with a biased sampling problem by design. This should be in contrast to the biased sampling problems discussed in earlier chapters where the selection bias occurs naturally and is out of our control. There are many publications addressing the regression analysis of cohort data with the above designs; see Thomas (1977), Oakes (1981), Prentice (1986), Self and Prentice (1988), Kalbfleisch and Lawless (1988), Langholz and Thomas (1990, 1991), Goldstein and Langholz (1992), Borgan and Langholz (1993), Langholz and Goldstein (1996), Breslow (1996), Samuelsen (1997), Chen and Lo (1999) and Chen (2001); Chen et al. (2004), among many others.

Most papers discussed the adjustment of risk sets in the Cox regression analysis. Recently Zhou et al. (2017) have discussed case-cohort studies with interval-censored data.

1. Parametric approach

First we present the work by Kalbfleisch and Lawless (1988) using a full parametric modelling approach. Field data provide more realistic information concerning the life distribution of equipment in actual use than do laboratory life test data. But, generally the (field) data are seriously incomplete. For example, only failure (but not nonfailure) information is recorded, hence data on failures may be incomplete. Therefore failure-record data may not be very informative. The utility of failure-record data can be increased greatly by collecting a supplementary sample on items that do not fail. This general approach is widely used in retrospective or case-control studies. A variety of supplementary sampling schemes have been proposed in the literature, see Kalbfleisch and Lawless (1988). Various pseudo likelihoods have been constructed for making inferences about the underlying parameters.

Suppose N items are in field use. Denote the failure time as T and the associated covariate as X . A parametric density $f(y|x, \theta)$ for T given X is assumed. Let $g(x)$ be the marginal density of covariate X . Field record data arise when the i -th item is sampled if and only if $T < \tau$ for some pre-specified time τ . Typically the time interval $(0, \tau]$ is the warrant period. For those failed individuals before τ , the failure time t_i and covariate x_i are recorded. For those individuals not failed before τ , the only available information is $I(T_i > \tau)$ but the covariate information x_i is not observable. The likelihood function is

$$L = \left\{ \prod_{t_i \leq \tau} f(t_i|x_i, \theta) g(x_i) \right\} \left\{ \prod_{t_i > \tau} \int \bar{F}(\tau|x, \theta) g(x) dx \right\}.$$

Clearly this likelihood involves the density of $g(x)$ which in general is unknown. As an alternative we may consider the conditional likelihood

$$L_c = \prod_{t_i \leq \tau} \frac{f(t_i|x_i, \theta)}{F(\tau|x_i, \theta)}.$$

In this approach the information on the total number N of items is not used. Furthermore, Kalbfleisch and Lawless (1988) pointed out the conditional likelihood or even the full likelihood approach if $g(x)$ is known or known up to some parameters is not very efficient due to lack of covariate information x_i 's for those not failed items. The utility of failure record data can be increased greatly by collecting a supplementary sample on items that do not fail. This general approach is widely used in retrospective or case-control studies. For the present, we consider the following scheme: the failure record data are supplemented by selecting a sample of those items that do not fail by time τ and, for each sampled item, determining the corresponding x . This can be implemented under various sampling schemes. If all x_i 's were available, then the log-likelihood is

$$\ell_F = \sum_{i=1}^N I(t_i \leq \tau) \log f(t_i|x_i, \theta) + I(t_i > \tau) \log \bar{F}(\tau|x_i, \theta).$$

Without loss of generality we assume $t_i \leq \tau, i = 1, 2, \dots, n$. Let $V_i = 1, i = n+1, \dots, N$ if the i -th item is selected in the supplementary samples. Let $p_i = P(V_i = 1) > 0, i = n+1, \dots, N$ be the sampling probability. Since

$$\sum_{i=n+1}^N \frac{V_i}{p_i} \log \bar{F}(\tau|x_i, \theta)$$

is an unbiased estimator of $\sum_{i=n+1}^N \log \bar{F}(\tau|x_i, \theta)$, naturally we can use pseudo log-likelihood

$$\ell_P = \sum_{i=1}^n I(t_i \leq \tau) \log f(t_i|x_i, \theta) + \sum_{i=n+1}^N \frac{V_i}{p_i} \log \bar{F}(\tau|x_i, \theta)$$

to make inference on θ . Large sample results and simulation results on the efficiency gain by comparing the conditional likelihood approach are discussed in Kalbfleisch and Lawless (1988).

Next we consider the maximum full likelihood approach. We only consider the non-supplementary sample case. Some modifications are needed for dealing with supplementary data. Denote

$$\Delta = P(T > \tau) = \int \bar{F}(\tau|x, \theta) dG(x).$$

Let $n_0 = \sum_{i=1}^n I(t_i > \tau)$. Without loss of generality we assume the first n_1 observations with $t_i \leq \tau$. The full log-likelihood is

$$\ell = \sum_{i=1}^n I(t_i \leq \tau) \{\log f(t_i|x_i, \theta) + \log p_i\} + n_0 \log \Delta,$$

where $p_i = dG(x_i), i = 1, 2, \dots, n_1$ with constraint

$$\sum_{i=1}^{n_1} p_i \{\bar{F}(\tau|x_i, \theta) - \Delta\} = 0.$$

We can show that the profile empirical likelihood is

$$\ell = \sum_{i=1}^{n_1} [\log f(t_i|x_i, \theta) - \log \{1 + \lambda(\bar{F}(\tau|x_i, \theta) - \Delta)\}] + n_0 \log \Delta$$

subject to the constraint

$$\sum_{i=1}^{n_1} \frac{\bar{F}(\tau|x_i, \theta) - \Delta}{1 + \lambda(\bar{F}(\tau|x_i, \theta) - \Delta)} = 0.$$

Finally we can maximize ℓ with respect to θ and Δ .

Exercise Derive the large sample results.

2. Semiparametric model approach

In the general case cohort study, τ is a random variable, i.e., it serves as a censoring variable.

Consider a cohort of size N . Let $(Y_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i), X_i, i = 1, 2, \dots, N)$ be the full cohort data. Denote $N_1 = \sum_{i=1}^N \delta_i$ and $N_0 = \sum_{i=1}^N (1 - \delta_i)$ as the total numbers of failures and non failures, respectively. The information on N_0 might be known or unknown. The classic case and control design takes n_1 cases (failures) and n_0 controls (non failures) randomly respectively. The case cohort design takes all N_1 cases and m subjects from the entire cohort without replacement. A nested case and control design takes all cases and m subjects without replacement from the risk sets at each failure time t_i .

Now we discuss case cohort design in detail.

- (1) Take a random sample of all cohort members to form a subcohort, denoted by \tilde{C} .
- (2) Add failures that are not sampled: non-subcohort failures.

Under this sampling design, an analysis problem arises since even though subcohort is representative of the entire full cohort, the non-subcohort cases are not; person-time by cases is over represented, or, equivalently cases are over represented in the risk sets.

Prentice (1986) used the following approach. Form a risk set at every failure time, but only use subcohort members as controls in each risk set t_i . Then the case-cohort data can be analysed using Cox regression with conditional logistic likelihood contributions from each case-cohort risk set in which the risk set includes all those survived subcohort individuals at each failure time t_i- .

If all covariate are available then the Cox partial likelihood score is

$$\sum_i \delta_i \left\{ x_i - \frac{\sum_{j=1}^n I(Y_j \geq t_i) x_j \exp(x_j \beta)}{\sum_{j=1}^n I(Y_j \geq t_i) \exp(x_j \beta)} \right\} = 0.$$

Based on the case cohort data, the pseudo score is

$$\sum_i \delta_i \left\{ x_i - \frac{\sum_{j \in \tilde{C}} I(Y_j \geq t_i) x_j \exp(x_j \beta)}{\sum_{j \in \tilde{C}} I(Y_j \geq t_i) \exp(x_j \beta)} \right\} = 0.$$

Since $\sum_{j \in \tilde{C}} Y_j(t_i) x_j^k \exp(x_j \beta)$ is an unbiased estimator of $\sum_{j=1}^n Y_j(t_i) x_j^k \exp(x_j \beta)$, $k = 0, 1$. In Prentice (1986) paper, the pseudo score is given by

$$\sum_i \delta_i \left\{ x_i - \frac{\sum_{j \in \tilde{C} \cup \{i\}} I(Y_j \geq t_i) x_j \exp(x_j \beta)}{\sum_{j \in \tilde{C} \cup \{i\}} I(Y_j \geq t_i) \exp(x_j \beta)} \right\} = 0.$$

Note that if $i \in \tilde{C}$, then there is no difference, otherwise i is included in the risk set for a failure if not it is in the selected subcohort \tilde{C} . Basically the two methods are asymptotically equivalent. Large sample theory was derived by Self and Prentice (1988).

Next we present the results by Chen and Lo (1999) and Chen (2001) and Chen et al. (2012) in the Cox regression model set up. Chen and Lo (1999) discussed classical case and control sampling. Define

$$p = P(\delta = 1), \quad m(t) = E[X|Y = t, \delta = 1].$$

Let $\bar{G}(c|x)$ be the survival function of censoring variable C given X . Note that the observed survival time Y satisfies

$$E[I(Y \geq t)|x] = \bar{G}(t|x)\bar{F}(t|x).$$

Under the proportional hazards model,

$$m(t) = \frac{\int xf(t|x)\bar{G}(t|x)g(x)dx}{\int f(t|x)\bar{G}(t|x)g(x)dx} = \frac{\int x\lambda(t) \exp(X\beta)\bar{G}(t|x)\bar{F}(t|x)g(x)dx}{\int \lambda(t) \exp(X\beta)\bar{G}(t|x)\bar{F}(t|x)g(x)dx} = \frac{E[X \exp(X\beta)I(Y \geq t)]}{E[\exp(X\beta)I(Y \geq t)]}.$$

Therefore $\delta[X - m(Y)]$ has mean zero.

Moreover $m(t)$ can be written as

$$m(t) = \frac{P(\delta = 1)E[X \exp(X\beta)I(Y \geq t)|\delta = 1] + P(\delta = 0)E[X \exp(X\beta)I(Y \geq t)|\delta = 0]}{P(\delta = 1)E[\exp(X\beta)I(Y \geq t)|\delta = 1] + P(\delta = 0)E[\exp(X\beta)I(Y \geq t)|\delta = 0]}.$$

In classical case and control sampling, denote C_1 and C_0 as the randomly selected case subcohort and control subcohort, respectively. Assume there are k_1 and k_0 cases and controls in C_1 and C_0 , respectively. Replacing the unknown quantities on the right hand side of $m(t)$, we have an estimating equation

$$\sum_{i \in C_1} \delta_i \left\{ x_i - \frac{(\hat{p}/k_1) \sum_{j \in C_1} I(Y_j \geq t_i) x_j \exp(x_j \beta) + \{(1 - \hat{p})/k_0\} \sum_{j \in C_0} I(Y_j \geq t_i) x_j \exp(x_j \beta)}{\sum_{j \in C_1} (\hat{p}/k_1) I(Y_j \geq t_i) \exp(x_j \beta) + \{(1 - \hat{p})/k_0\} \sum_{j \in C_0} I(Y_j \geq t_i) \exp(x_j \beta)} \right\} = 0.$$

Note that \hat{p} can be N_1/N if $N_1 = \sum_{i=1}^N \delta_i$ is available.

If all $(Y_i, X_i, \delta_i = 1)$, $i = 1, 2, \dots, N_1$ are known (we have assumed the first N_1 individuals are cases), we can estimate $pE[\exp(X\beta)I(Y \geq t)|\delta = 1]$ by

$$(N_1/N) \frac{1}{N_1} \sum_{i=1}^{N_1} \delta_i \exp(x_i \beta) I(y_i \geq t).$$

Similarly $(1-p)E[\exp(X\beta)I(Y \geq t)|\delta = 0]$ can be estimated by

$$\{N_0/(Nk_0)\} \sum_{i \in C_0} \exp(x_i \beta) I(y_i \geq t).$$

The score estimating equation is

$$\sum_{i=1}^N \delta_i \left\{ x_i - \frac{N^{-1} \sum_{j=1}^{N_1} I(Y_j \geq t_i) x_j \exp(x_j \beta) + \{N_0/(Nk_0)\} \sum_{j \in C_0} I(Y_j \geq t_i) x_j \exp(x_j \beta)}{N^{-1} \sum_{j=1}^{N_1} I(Y_j \geq t_i) \exp(x_j \beta) + \{N_0/(Nk_0)\} \sum_{j \in C_0} I(Y_j \geq t_i) \exp(x_j \beta)} \right\} = 0.$$

Next we consider a typical nested case and control design where all cases are included. For each observed case failed at t_j , m controls are randomly sampled from his/her risk set excluding the candidate cases, which is of size

$$n^-(t_j) = \sum_{i=1}^n I(y_i \geq t_j) - 1.$$

The m controls are sampled without replacement from the finite population. In the Bernoulli sampling, each eligible subject in the risk set of t_j is sampled independently with probability $m/n^-(t_j)$ as a control for t_j . Denote V_{i0} as whether subject i ever sampled as a control and

$$V_i = \delta_i + (1 - \delta_i)V_{0i}$$

indicates being sampled into the nested case and control design.

If i is a failure, then, with probability one, the individual will be selected. On the other hand if i is a non-failure, the individual will not appear in the risk set $R(t_k) = \{y_l : y_l \geq t_k\}$ if $t_k > y_i$. y_i can only appear in those risk sets $R(t_k)$, $t_k \leq y_i$. y_i is not selected in $R(t_k)$, $t_k \leq y_i$ with probability $1 - m/n^-(t_k)$. Samuelsen (1997 [Biometrika]) derived the inclusion probability as

$$\pi_i = P(V_i = 1 | \mathcal{D}) = \delta_i + (1 - \delta_i)p_{0i}, \quad p_{0i} = 1 - \prod_{y_j \leq y_i} \left(1 - \frac{m\delta_j}{n^-(y_j)}\right).$$

For the Bernoulli sampling, let B_{jk} denote an indicator that takes value 1 if subject k is sampled as a control for subject j and 0 otherwise.

$$\pi_i = 1 - \prod_{j: x_j \leq t} \left\{ 1 - \frac{\delta_j \sum_{l=1}^N B_{jl}}{n^-(x_j)} \right\}.$$

Define $V_{0i}(t)$ as the indicator that individual i is selected as a control at time t . Then the indicator for i is not selected as control is

$$1 - V_{0i} = \prod_{t \leq t_i} \{1 - V_{0i}(t)\},$$

where t ranges for all death points. The conditional inclusion probabilities p_{0ij} that neither i nor j is selected as control is given by

$$\begin{aligned} p_{0ij} &= E[(1 - V_{0i})(1 - V_{0j})|\mathcal{F}] = E \left[\prod_{t,s} \{1 - V_{0i}(t)\}\{1 - V_{0j}(s)\} |\mathcal{F} \right] \\ &= \prod_{t \neq s} E[\{1 - V_{0i}(t)\}|\mathcal{F}] E\{1 - V_{0j}(s)\} |\mathcal{F} \prod_t E[\{1 - V_{0i}(t)\}\{1 - V_{0j}(t)\}|\mathcal{F}]. \end{aligned}$$

Note that the control samples at different times are independent, i.e., $V_{0i}(t)$ and $V_{0j}(s)$ are independent if $t \neq s$.

$$\begin{aligned} E[\{1 - V_{0i}(x_k)\}\{1 - V_{0j}(x_k)\}|\mathcal{F}] &= 1 - E[V_{0i}(x_k)|\mathcal{F}] - E[V_{0j}(x_k)|\mathcal{F}] + E[V_{0i}V_{0j}|\mathcal{F}] \\ &= 1 - 2 \frac{m}{Y_k - 1} + P(V_{0i} = 1|\mathcal{F})P(V_{0j} = 1|V_{0i} = 1, \mathcal{F}) \\ &= 1 - 2 \frac{m}{Y_k - 1} + \frac{m}{Y_k - 1} \frac{m - 1}{Y_k - 2} \end{aligned}$$

for k with $\delta_k = 1$. Define

$$\rho_{ij} = \frac{q_{0ij}}{(1 - p_{0i})(1 - p_{0j})} - 1,$$

then

$$\rho_{ij} = \prod \left\{ 1 - 2 \frac{m}{Y_k - 1} + \frac{m(m - 1)\delta_k}{(Y_k - 1)(Y_k - 2)} \right\} / (1 - m\delta_k/(Y_k - 1))^2 - 1.$$

Suppose that k individuals had events. The full log-likelihood is

$$\ell = \sum_{i=1}^k \{\log \lambda(t_i) - x_i \beta - \Lambda(t_i) \exp(x_i \beta)\} - \sum_{i=1}^k \sum_{j=1}^{n_i} \Lambda(t_i) \exp(x_{ij} \beta).$$

We can use inverse probability weighted method to estimate this log-likelihood

$$\hat{\ell} = \sum_{i=1}^k \{\log \lambda(t_i) - x_i \beta - \Lambda(t_i) \exp(x_i \beta)\} - \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{\delta_{ij}}{\pi_{ij}} \Lambda(t_i) \exp(x_{ij} \beta),$$

where

$$P(\delta_{ij} = 1 | y_i, \delta_i, i = 1, 2, \dots, n) = \pi_{ij}.$$

The general formula for π_{ij} is given, respectively, for classical case and control design, case cohort design, nested case cohort design as follows:

(1) Classical case and control design

$$\pi_{ij} = \begin{cases} \frac{m_1(m_1-1)}{n_1(n_1-1)}, & \text{if both } i \text{ and } j \text{ are failures;} \\ \frac{m_0(m_0-1)}{n_0(n_0-1)}, & \text{if both } i \text{ and } j \text{ are non failures;} \\ \frac{m_0 m_1}{n_0 n_1}, & \text{else.} \end{cases}$$

(2) Case and control design

$$\pi_{ij} = \begin{cases} 1, & \text{if both } i \text{ and } j \text{ are failures;} \\ \frac{m(m-1)}{n(n-1)}, & \text{if both } i \text{ and } j \text{ are non failures;} \\ \frac{m}{n} & \text{else.} \end{cases}$$

(3) Nested case cohort design.

$$\pi_{ij} = \begin{cases} 1, & \text{if both } i \text{ and } j \text{ are failures;} \\ \pi_j, & \text{if } i \text{ is a failure } j \text{ is not;} \\ \pi_i, & \text{if } j \text{ is a failure } i \text{ is not.} \end{cases}$$

In the case that both i and j are non-failures (Exercise)

$$\pi_{ij} = \pi_i + \pi_j - 1 + \prod_{t < \min(y_i, y_j)} \left(\frac{n_t^- - m - 1}{n_t^- - 1} \right)^{dN(t)} \prod_{t < \max(y_i, y_j)} \left(\frac{n_t^- - m}{n_t^-} \right)^{dN(t)}.$$

Samuelsen (1997) discussed a slightly more general weighted estimation method with left truncation data in addition to the right censoring. The weighted log-likelihood is

$$\ell = \sum_i \frac{V_i \ell_i}{p_i},$$

where

$$\ell_i = \log \left(\frac{f^{\delta_i}(y_i|x_i) \bar{F}^{1-\delta_i}(y_i|x_i)}{\bar{F}(a_i|x_i)} \right) = \delta_i \log \lambda(y_i|x_i) - \{\Lambda(y_i) - \Lambda(a_i)\},$$

where a_i is the left truncation time.

The baseline cumulative hazard can be estimated by modifying a Breslow (1972) type estimator, see, for example, Borgan and Langholz (1993). For more discussions, we refer readers to the work by Borgan and Langholz (1993).

As an alternative method to the inverse probability weighted estimation, Yao (2015) proposed profile likelihood approach in the case-cohort design. In particular a covariate X is selected with the probability

$$P(V = 1|Y, \delta) = \pi(y, \delta), \quad X \sim h(x).$$

This is a covariate missing at random problem. The likelihood function is

$$L = \prod_{i=1}^n P(Y = y_i, \delta_i) [\prod_{i=1}^n P(X = x_i|Y_i, \delta_i, V_i = 1) dH(x_i)]^{V_i}.$$

The log-likelihood is

$$\ell = \sum_{i=1}^n I(v_i = 1) [\delta_i \{ \log f(y_i|x_i) + \log \bar{G}_C(y_i|x_i) + \log dH(x_i) \} \\ + \{1 - I(v_i = 1)\} [\delta_i \log \int f(y_i|x) \bar{G}_C(y_i|x) dH(x) + (1 - \delta_i) \log \bar{F}(y_i|x) g_c(y|x) dH(x)]].$$

For $G_c(y|x) = G_c(y)$, Yao (2015) studied semiparametric maximum likelihood estimation of Λ and H by discretizing Λ at each of the observed data points, and discretizing H at each of the observed covariates X_i , $V_i = 1$. Even though this approach can produce the most efficient estimator, unfortunately the assumption of censoring distribution of C being independent of covariate may not be satisfied in practical applications. A related previous work in covariate missing data problem with Cox regression model can be found in Chen and Little (1999).

As an alternative approach, Chen et al. (2012) proposed to maximize

$$\sum_{i=1}^n \frac{V_i}{\pi_i} \ell_i,$$

where ℓ_i is the likelihood contribution from the i -th individual, and $P(V_i = 1|y_i, \delta_i) = \pi(y_i, \delta_i)$. The support points can be chosen to be all observed failure time points. It is possible to obtain a closed form under the Cox regression model. However iteration algorithms are needed for inference under semiparametric transformation models.

The large sample theory in general is difficult. Nan and Wellner (2013) developed a so-called Z-estimator theory for case-cohort studies. We refer readers to their paper.

Chapter 25

Length Biased Sampling, Multiplicative Censoring and Survival Analysis

With the preparation of the previous chapter on survival analysis, we are ready to move to the latest development on the survival analysis based on length-biased and right-censored data. We start from Vardi's (1989) seminal *Biometrika* paper.

25.1 Vardi's Four Equivalent Problems

Vardi (1982a,b) discussed nonparametric maximum likelihood estimation based on length-biased sampling and renewal process with incomplete renewal data. Furthermore Vardi (1989) discussed four important and closely related statistical problems. (1) Estimation in the multiplicative censoring problem. (2) Nonparametric estimation in renewal processes. (3) Nonparametric deconvolution, and (4) Estimation in a monotonic decreasing density model.

We start from the nonparametric maximum likelihood estimation based on the multiplicative censoring data problem.

Problem 1 Estimating a survival function with multiplicative censoring.

Suppose we have two independent data sets

$$X_1, \dots, X_m \sim i.i.d.F(x), \quad Z_1, \dots, Z_n \sim i.i.d.F(x).$$

In addition we denote another set of independent and uniformly distributed r.v.'s as

$$U_1, \dots, U_n \sim U(0, 1).$$

Instead of observing Z_1, \dots, Z_n and U_1, \dots, U_n , we can only observe

$$Y_1 = Z_1 U_1, \dots, Y_n = Z_n U_n.$$

Vardi (1989) termed U_i a multiplicatively censoring for Z_i . It follows from $Y = ZU \leq Z$ that

$$\begin{aligned} P(Y \leq y) &= P(ZU \leq y, Z \leq y) + P(ZU \leq y, Z > y) \\ &= P(Z \leq y) + P(U \leq y/z, Z > y) \\ &= F(y) + \int_y^\infty \frac{y}{z} dF(z). \end{aligned}$$

As a result, the density function of Y is $\int_y^\infty z^{-1} dF(z)$, which is a monotonic non-increasing function of y .

Based on the x_i 's and y_j 's, we can estimate F nonparametrically. Suppose

$$t_1 < t_2 < \dots < t_h$$

are the distinct values of $x_1, \dots, x_m; y_1, \dots, y_n$ in an increasing order, where $h \leq m+n$. Let ξ_j and η_j be the multiplicity of x 's and y 's at t_j , i.e.,

$$\xi_j = \sum_{i=1}^m I(x_i = t_j), \quad \eta_j = \sum_{i=1}^n I(y_i = t_j).$$

The likelihood is

$$L = \prod_{j=1}^h p_j^{\xi_j} \left(\sum_{k=j}^h t_k^{-1} p_k \right)^{\eta_j}$$

subject to

$$p_j \geq 0, \quad j = 1, 2, \dots, h, \quad \sum_{j=1}^h p_j = 1.$$

Exercise In addition to the jumps at the observed data points $\{t_1, \dots, t_h\}$ if F has a jump at point $t_0 \neq t_i, i = 1, 2, \dots, h$, then show that one can always find another probability distribution F^* with jumps only at $\{t_1, \dots, t_h\}$ such that likelihood $L(F^*) \geq L(F)$.

Therefore we only need to consider those probabilities with mass at the observed data points. To take the missing data problem into account, naturally we can use an EM algorithm to accomplish the maximization.

Vardi's EM Algorithm

Note that

$$Z \sim dF(z), \quad Y|Z = z \sim U(0, z),$$

$$Z|Y = y \sim \frac{z^{-1} dF(z)}{\int_y^\infty z^{-1} dF(z)}, \quad z \geq y > 0.$$

If x_i 's and z_j 's are available, clearly one may estimate $p_j = P(X = t_j)$ by the empirical frequency

$$(m+n)^{-1} \left\{ \sum_{i=1}^m I(X_i = t_j) + \sum_{i=1}^n I(Z_i = t_j) \right\}.$$

Since z_i 's are not observable, we can replace them by conditional expectations

$$\begin{aligned} p_j^{new} &= (m+n)^{-1} \left[\sum_{i=1}^m I(x_i = t_j) + \sum_{i=1}^n I(z_i = t_j) | (x_1, \dots, x_m; y_1, \dots, y_n) \right] \\ &= (m+n)^{-1} \left[\xi_j + t_j^{-1} p_j^{old} \sum_{k=1}^j \eta_k \left(\sum_{i=k}^h t_i^{-1} p_i^{old} \right)^{-1} \right]. \end{aligned} \quad (25.1.1)$$

Let $\rho = m/(m+n)$. Note that the empirical distribution of $m^{-1} \sum_{i=1}^m I(x_i \leq t) \rightarrow F(t)$, and the empirical distribution $n^{-1} \sum_{i=1}^n I(y_i \leq t) \rightarrow F_Y(t) = P(Y \leq t)$, then we end up with an integral equation

$$d\hat{F}(t) = \rho dF(t) + (1 - \rho) \frac{\hat{F}(t)}{t} \int_0^t du \left\{ \int_{v \geq u} v^{-1} dF(v) / \int_{v \geq u} v^{-1} d\hat{F}(v) \right\}.$$

Vardi (1989) and Vardi and Zhang (1992) showed that this integral equation has a consistent solution for F . Moreover Asgharian and Wolfson (2005) studied large sample results for prospectively (m and n are random, see Problem 2 below) and retrospectively (m and n are fixed) collected data.

Question arises as to whether the EM algorithm converges. Based on the result by Csiszar and Tusnády (1984), the EM algorithm (or alternating maximizing algorithm) will converge to the globe MLE if the likelihood function is a strict log-concave and the domain is a convex region. To prove this, we consider the transformation

$$q_j = p_j/t_j, \quad Q_j = \sum_{k=j}^h q_k.$$

Then the likelihood is proportional to

$$L \propto \prod_{j=1}^h q_j^{\xi_j} Q_j^{\eta_j}$$

subject to the constraint

$$\sum_{j=1}^h t_j q_j \leq 1, \quad q_j \geq 0.$$

The log-likelihood is

$$\ell = C + \sum_{j=1}^h [\xi_j \log q_j + \eta_j \log Q_j],$$

where C is a constant. Straightforward algebra shows that

$$\alpha^T \left(\frac{\partial^2 \ell}{\partial q \partial q^T} \right) \alpha = - \sum_{j=1}^h \alpha_j^2 \frac{\xi_j}{q_j^2} - \sum_{j=1}^h (\alpha_j + \dots + \alpha_h)^2 \frac{\eta_j}{Q_j^2},$$

which is strictly positive unless $\alpha = 0$. Therefore we have shown that the log likelihood is strict concave. As a consequence, the EM algorithm converges to the unique maximum.

Problem 2 Nonparametric maximum likelihood estimation for length-biased and right-censored data

Consider a prevalent cohort study in which subjects are diagnosed with a disease and are at risk for a failure event. Let \tilde{T} be the duration from the disease onset to failure with the unbiased density function $f(t) = df(t)$ and survival function $\bar{F}(t)$. The observed data include backward recurrence time A (from disease onset to study entry), forward recurrence time V (from study entry to failure), and total “renewal time” $T = A + V$. Based on the renewal theory discussed in Chap. 2, the joint distribution of (A, V) is

$$\frac{f(a+v)}{\mu}, a, v > 0, \text{ where } \mu = \int t f(t) dt.$$

The marginal density of T is $dG(t) = tdf(t)/\mu$. When the prevalent cohort is followed prospectively, V is subject to right censoring. The censoring time, denoted by C , is measured from the study entry. Let $\delta = I(V < C) = I(A + V < C + A)$ be the censoring indicator. Assume that (A, V) is independent of C . Let $X = \min(A + V, A + C)$ be the possibly right censored “renewal time”. Denote the observed data as (X_i, A_i, δ_i) , $i = 1, 2, \dots, n$.

Note that the survival function $\bar{F}(t)$ of \tilde{T} can be written as

$$\bar{F}(t) = \int_t^\infty dF(u) = \mu \int_t^\infty u^{-1} dG(u).$$

The likelihood for the observed data (X_i, A_i, δ_i) is proportional to

$$\prod_{i=1}^n \frac{f^{\delta_i}(X_i) \bar{F}^{1-\delta_i}(X_i)}{\mu} \propto \prod_{i=1}^n [dG(X_i)]^{\delta_i} \prod_{i=1}^n \left\{ \int_{x \geq X_i} x^{-1} dG(x) \right\}^{1-\delta_i}. \quad (25.1.2)$$

Therefore this likelihood has exactly the same form as the one discussed before for the multiplicative censoring likelihood. The only difference is that $\sum_{i=1}^n \delta_i$ is a random variable here while it is fixed in the multiplicative censoring problem. Fortunately the same algorithm can be applied to derive the nonparametric MLE. However, the large sample properties are different. Using the relationship between G and F , $dF(t) = t^{-1}dG(t)/\int t^{-1}dG(t)$, one can derive the NPMLE for F . Asgharian et al. (2002) discussed this problem in details.

Problem 3 Nonparametric deconvolution.

Suppose we observe a sample from the distribution of the sum of two independent random variables, X and Y , where $X \sim F$ (unknown), and Y has a known distribution with density given by

$$g(y) = \exp(y), \quad -\infty < y < 0.$$

Based on the observed $Z_i = X_i + Y_i$, $i = 1, 2, \dots, n$, we are interested in estimating the distribution of F .

Note that $-Y$ has a standard exponential random variable, or equivalently $U = \exp(Y)$ is a standard uniform random variable. Using the exponential transformation, we have

$$\exp(Z) = \exp(X) \exp(Y) = \exp(X)U.$$

Therefore the distribution function of the random variable $\exp(X)$ can be estimated using the multiplicative censoring method discussed before. Unfortunately in this case there is no training sample from $\exp(X)$.

Problem 4 Estimating a monotone decreasing density.

Suppose Y_1, \dots, Y_n are independent and identically distributed nonnegative random variable with an unknown distribution G . Based on the prior information we know that G has a monotone non-increasing density g . This is a classical problem discussed by Grenander (1956a,b). Now we use a different approach to examine this problem.

Note that backward time A has a non-increasing density given by

$$A \sim \frac{\bar{F}(a)}{\mu}, \quad \mu = \int_0^\infty adF(a).$$

Let

$$dG(a) = \frac{adF(a)}{\mu},$$

then

$$A \sim \int_a^\infty \frac{dF(s)}{s}.$$

Again methods discussed in multiplicative censoring problems can be used here. In this case there is no training samples directly from F .

25.2 A New EM Algorithm

Even though Vardi's (1989) EM algorithm is elegant for nonparametric estimation of F , Qin et al. (2011) observed that Vardi's (1989) method is often difficult to impose constraints on F directly when F is estimated from the NPMLE of G (length biased version of F), because the constraints on F may not be easily translated to the constraints on the NPMLE of G . To overcome this difficult, Qin et al. (2011) proposed a direct maximization method for F .

As demonstrated in Vardi (1989), to maximize (25.1.2), it suffices to consider the discrete version of distribution F , i.e., $p(\tilde{T} = t_i) = p_i$, non-parametrically on the point masses at

$$t_1 < t_2 < \cdots < t_k,$$

where t_1, \dots, t_k are the ordered unique failure and censoring times for $\{X_1, \dots, X_n\}$, $k \leq n$.

Let $\tau = t_k = \max\{X_1, \dots, X_n\}$, i.e., the largest observation. In principle, the length-biased observations (A, X) can be equivalently generated from a truncation model with

$$A \sim U(0, \hat{\tau}), \quad T \sim F, \quad \text{on } (0, \hat{\tau}), \quad (25.2.3)$$

where A and T are independent, $P(T = t_i) = dF(t_i) = p_i$ and $\sum_{i=1}^k p_i = 1$, and $(A, T = X)$ is observed if and only if $T > A$. The probability of observing a length-biased observation under this setting is

$$\pi = P(T > A) = E(T)/\tau = \sum_{i=1}^k p_i t_i / \tau.$$

Qin et al. (2011) proposed a new EM algorithm with a different missing mechanism to directly estimate the target distribution, F . For a cohort subject to left truncation, a biased sample on n subjects denoted by $O = \{(X_1, \delta_1, A_1), \dots, (X_n, \delta_n, A_n)\}$, $A_i \leq X_i, i = 1, \dots, n\}$, is observed, whereas the data on m subjects are left truncated. Here the latent left-truncated data are denoted by $O^* = \{(T_1^*, A_1^*), \dots, (T_m^*, A_m^*), A_i^* > T_i^*, i = 1, 2, \dots, m\}$. The random integer m then follows a negative binomial distribution with parameter π . The probability mass function of m is

$$\binom{m+n-1}{m} (1-\pi)^m \pi^n, \quad m = 0, 1, 2, \dots \quad \text{and} \quad E(m|O) = n(1-\pi)/\pi.$$

Following the principle of the EM algorithm, we treat $\{O, O^*\}$ as the “complete data”, and interpret the pseudo missing data, which are also referred as “ghosts” data in Turnbull (1976), as $O^* = \{(T_1^*, A_1^*), \dots, (T_m^*, A_m^*)\}$ and the observed ‘incomplete data’ as O . We derive the full likelihood including the component of the truncated observations. The log-likelihood based on the complete data $\{O, O^*\}$ is

$$\sum_{j=1}^k \left[\sum_{i=1}^n I(T_i = t_j) + \sum_{i=1}^m I(T_i^* = t_j) \right] \log p_j, \quad (25.2.4)$$

where $T_i \geq A_i$, $i = 1, 2, \dots, n$ and $T_k^* < A_k^*$, $k = 1, \dots, m$. Then conditioned on the observed data,

$$\begin{aligned} E \left[\sum_{i=1}^n I(T_i = t_j) \middle| O \right] &= \sum_{i=1}^n \left\{ \delta_i I(T_i = t_j) + (1 - \delta_i) P(T_i = t_j | T_i \geq A_i, T_i > x_i) \right\} \\ &= \sum_{i=1}^n \left[\delta_i I(X_i = t_j) + (1 - \delta_i) \frac{I(x_i < t_j) p_j}{\int_{x_i}^{\infty} f(s) ds} \right], \end{aligned}$$

because $P(T_i = t_j, T_i > x_i) = I(x_i < t_j) p_j$ and $P(T_i > x_i, T_i \geq a_i) = P(T > x_i) = \int_{x_i}^{\infty} f(s) ds$. Here $\int_{x_i}^{\infty} f(s) ds = \sum_{j=1}^k p_j I(t_j > x_i)$ with the discretized function of F . Conditional on the observed data O , the expectation for the missing left-truncated data can be expressed as

$$E \left\{ E \left[\sum_{i=1}^m I(T_i^* = t_j) \middle| m \right] \middle| O, T^* < A^* \right\}.$$

Under the truncation model specified in (25.2.3),

$$EI(T^* = t_j | O, T^* < A^*) = Pr(T^* = t_j, A^* > t_j) / Pr(T^* < A^*) = \frac{p_j(1 - t_j/\tau)}{1 - \pi}.$$

This together with $E(m|O) = n(1 - \pi)/\pi$,

$$E \left\{ E \left[\sum_{i=1}^m I(T_i^* = t_j) \middle| m \right] \middle| O, T^* < A^* \right\} = \frac{n(1 - \pi)}{\pi} \frac{(1 - t_j/\tau)p_j}{1 - \pi} = \frac{n}{\pi}(1 - t_j/\tau)p_j.$$

Subject to $\sum_{i=1}^k p_i = 1$ and $p_i \geq 0$, we maximize the expected complete-data log-likelihood conditional on the observed data via the EM algorithm,

$$\ell_E(p) = \sum_{j=1}^k w_j \log p_j, \quad (25.2.5)$$

where $p = (p_1, \dots, p_k)$, and

$$w_j = \sum_{i=1}^n \left[\delta_i I(X_i = t_j) + (1 - \delta_i) \frac{p_j I(X_i < t_j)}{\sum_{j=1}^k p_j I(X_i < t_j)} \right] + \frac{n}{\pi}(1 - t_j/\tau)p_j.$$

By simple algebra, $\sum_{j=1}^k w_j = n + n(1 - \pi)/\pi = n/\pi$. The following iterative EM algorithm can be used to solve \hat{p}_j for $j = 1, \dots, k$.

Step 1 Select an arbitrary $p_j^{(0)}$ satisfying $\sum_{j=1}^k p_j^{(0)} = 1$, $p_j^{(0)} \geq 0$.

Step 2 Solve $p_j^{(1)}$ by maximizing (25.2.5), so that we replace $p_j^{(0)}$ with

$$\hat{p}_j^{(1)} = \frac{\hat{\pi}^{(0)}}{n} \left\{ \sum_{i=1}^n \left[\delta_i I(X_i = t_j) + (1 - \delta_i) \frac{\hat{p}_j^{(0)} I(X_i < t_j)}{\sum_{j=1}^k \hat{p}_j^{(0)} I(X_i < t_j)} \right] + \frac{n}{\hat{\pi}^{(0)}} (1 - \frac{t_j}{\tau}) \hat{p}_j^{(0)} \right\}, \quad (25.2.6)$$

$$\text{where } \hat{\pi}^{(0)} = \sum_{j=1}^k t_j \hat{p}_j^{(0)} / \tau.$$

With a given convergence criterion, we can solve p_j iteratively. Let \hat{p}_j denote the MLE of $p_j, j = 1, \dots, k$, the NPMLE $\hat{F}(t) = \sum_{j=1}^k \hat{p}_j I(t_j \leq t)$, $\hat{\pi} = \int t d\hat{F}(t)/\tau$, and

$$Q_1^n(t) = \frac{1}{n_1} \sum_{i=1}^n \delta_i I(X_i \leq t), \quad Q_0^n(t) = \frac{1}{n_0} \sum_{i=1}^n (1 - \delta_i) I(X_i \leq t),$$

where $n_1 = \sum_{i=1}^n \delta_i$ and $n_0 = \sum_{i=1}^n (1 - \delta_i)$. Thus, the limiting form of (25.2.6) is

$$d\hat{F}(t) = \hat{\pi} n_1 n^{-1} dQ_1^n(t) + \hat{\pi} n_0 n^{-1} d\hat{F}(t) \int_0^t \frac{dQ_0^n(s)}{1 - \hat{F}(s)} + (1 - t/\tau) d\hat{F}(t). \quad (25.2.7)$$

Remark 1 In contrast to the NPMLE for traditional survival analyses, where the baseline survival or hazard function has jumps only at the observed failure time points, the NPMLE for length-biased data has jumps at all observed but unique points including censored times.

Remark 2 One interesting observation is that the newly proposed EM algorithm with the unbiased distribution function F is essentially equivalent to that for Vardi's EM algorithm based on a 'multiplicative-censorship' model with the biased distribution function G . Denoting $d\hat{G}(t) = td\hat{F}(t)/\hat{\mu}$, where $\hat{\mu} = \hat{\pi}\tau$, we can re-express the equation in the new EM algorithm as an equation of \hat{G} ,

$$d\hat{G}(t) = n_1 n^{-1} dQ_1^n(t) + n_0 n^{-1} \frac{d\hat{G}(t)}{t} \int_0^t \left[\int_s^\infty r^{-1} d\hat{G}(r) \right]^{-1} dQ_0^n(s),$$

which is the same equation derived by Vardi (1989). The advantage of the new EM algorithm is that it directly estimates the target distribution function of the unbiased data, which allows one to directly impose constraints on F . This advantage will be further elucidated in the next section on the maximum semiparametric likelihood estimation based on the Cox regression model.

Remark 3 The “missing” data (i.e., the left-truncated failure times), $\{T_1^*, \dots, T_m^*\}$ are assumed not subject to right censoring. It is clear that whether T^* is subject to right censoring or not is irrelevant in the derivation of the above EM algorithm.

Remark 4 In practice, we can estimate τ by the maximum of observation times, $X_{(n)} = \max_{1 \leq i \leq n} X_i$. As n tends to infinity, we can prove that $X_{(n)} \rightarrow \tau$ in probability.

Tail Problems for Length Biased Sampling Data

Suppose the support τ of F is finite.

Note that the backward time satisfies

$$A \sim \frac{\bar{F}(a)}{\mu}, \quad \mu = \int_0^\tau \bar{F}(a)da.$$

(1) Suppose there is no follow up at all. Observe

$$\begin{aligned} P(\max_{1 \leq i \leq n} A_i < \tau - \epsilon) &= \left[\frac{\int_0^{\tau-\epsilon} \bar{F}(a)da}{\mu} \right]^n \\ &= \left[1 - \frac{\int_{\tau-\epsilon}^\tau \bar{F}(a)da}{\mu} \right]^n. \end{aligned}$$

As long as $\int_{\tau-\epsilon}^\tau \bar{F}(a)da > 0$ for any $\epsilon > 0$, then $\max_{1 \leq i \leq n} A_i \rightarrow \tau$.

If we use the Mean Value Theorem in the integral, then

$$P(\max_{1 \leq i \leq n} A_i < \tau - \epsilon) = [1 - \bar{F}(\xi)\epsilon/\mu]^n,$$

where $\tau - \epsilon \leq \xi \leq \tau$. The problem is that $\bar{F}(\xi) \rightarrow 0$ as $\epsilon \rightarrow 0$. We are not sure whether the convergence rate of $\max_{1 \leq i \leq n} A_i \rightarrow \tau$ is $n^{-1/2}$.

(2) Suppose there are follow up times C_i , $i = 1, 2, \dots, n$.

The observed data are

$$\tilde{Y}_i = \min(A_i + V_i, A_i + C_i), \quad Y_i = A_i + V_i, \quad i = 1, 2, \dots, n.$$

For any small $\epsilon > 0$,

$$P(\max_{1 \leq i \leq n} \tilde{Y}_i < \tau - \epsilon) = [P(\tilde{Y}_1 < \tau - \epsilon)]^n.$$

Note that the joint density of A and Y is

$$(A, Y) \sim \frac{f(y)}{\mu}, \quad y > a.$$

$$\begin{aligned}
P(\tilde{Y}_1 < \tau - \epsilon) &= 1 - E[I(Y > \tau - \epsilon)I(A + C > \tau - \epsilon)] \\
&= 1 - E[I(Y > \tau - \epsilon)\bar{G}_c(\tau - \epsilon - A)] \\
&= \int_{\tau-\epsilon}^{\tau} f(y)dy \int_0^y \bar{G}_c(\tau - \epsilon - a)da/\mu \\
&= 1 - \int_{\tau-\epsilon}^{\tau} w(y)f(y)dy/\mu \\
&= 1 - w(\xi)f(\xi) \int_{\tau-\epsilon}^{\tau} dy/\mu = 1 - w(\xi)f(\xi)\epsilon/\mu,
\end{aligned}$$

where $w(y) = \int_0^y \bar{G}(\tau - \epsilon - a)da$, $\tau - \epsilon \leq \xi \leq \tau$, and the Mean Value Theorem is used in the last step. Therefore

$$P(\max_{1 \leq i \leq n} \tilde{Y}_i < \tau - \epsilon) = [1 - w(\xi)f(\xi)\epsilon/\mu]^n.$$

If $\epsilon = n^{-2/3}$, then

$$P(\max_{1 \leq i \leq n} \tilde{Y}_i < \tau - n^{-2/3}) = [1 - w(\xi)f(\xi)n^{-2/3}/\mu]^n = \exp[-nw(\xi)f(\xi)n^{-2/3}/\mu + \dots].$$

As $n \rightarrow \infty$, we have

$$w(\xi) = \int_0^{\xi} \bar{G}(\tau - n^{-2/3} - a)da = \int_{\tau - \xi - n^{-2/3}}^{\tau - n^{-2/3}} \bar{G}(s)ds \rightarrow \int_0^{\tau} \bar{G}(s)ds > 0.$$

Also

$$\int_0^{\tau} \bar{G}(s)ds = E\left[\int_0^{\tau} I(C > s)ds\right] = E\left[\int_0^{\min(C, \tau)} ds\right] = \min(E(C), \tau).$$

If $E(C) > 0$, we have $w(\tau) > 0$. Moreover since τ is the largest support of F ,

$$f(\xi) \rightarrow f(\tau) > 0.$$

The conclusion is that, as long as subjects are followed for an additional period of time C , no matter how short it is,

$$\max_{1 \leq i \leq n} \tilde{Y}_i \rightarrow \tau$$

in a rate faster than $n^{-1/2}$ but slower than n^{-1} . Therefore we have proved the following theorem.

Theorem 25.1 Suppose $E(C) > 0$ and τ is the upper bound of the distribution function. Then for $1 > q > 1/2$

$$n^q(X_{(n)} - \tau) = o_p(1).$$

Exercise

(1) Find the limiting distribution of

$$n\{\max(y_i) - \tau\}.$$

(2) Use bootstrap method to approximate the limiting distribution.

25.3 Cox Model with Length Biased Sampling Data

In this section we study the semiparametric maximum likelihood estimation based on length biased data for the Cox regression model. Wang (1996) discussed this problem in the absence of right censoring by using an inverse weighted method for the risk sets.

There is a subtle difference between Vardi's definition of survival function with the conventional one. Vardi (1989) defined

$$S(t) = P(Y \geq t),$$

where the point mass $P(Y = t)$ is included in the definition. In order to maximize the length biased sampling likelihood we can show the observed data points are the supports of $S(t)$. In other words, for any estimator of the cumulative hazard function $\Lambda(t)$ with jumps outside of the event times, we can find a greater likelihood with jumps on the observed data points only. To better understand this issue, for simplicity, we use an example with four data points to illustrate.

Let $t_1 < t_2 < t_3 < t_4$ be the observed times, where t_1, t_4 are failures and t_2, t_3 are censored times, respectively. Under the proportional hazards model $S(t|z) = [S_0(t)]^{\exp(z\beta)}$, where z is the covariate. For simplicity we denote $S_i(t) = S(t|z_i) = [S_0(t)]^{\beta_i}$, $i = 1, 2, 3, 4$, $\beta_i = \exp(z_i\beta) > 0$, where $S_0(t)$ is the baseline survival function. We consider two cases.

Case 1 In addition to the masses at t_i , $i = 1, 2, 3, 4$, F has an extra mass at t^* , where $t_2 < t^* < t_3$.

Case 2 In addition to the masses at t_i , $i = 1, 2, 3, 4$, F has an extra mass at t^* , where $t^* < t_1$.

We consider case 1 first.

Let $p_i = dF(t_i)$, $i = 1, 2, 3, 4$, where $F(t) = 1 - S_0(t)$. The likelihood is

$$\begin{aligned} L_n(t_1, t_2, t_3, t_4) &= \beta_1 p_1 [S_0(t_1)]^{\beta_1-1} \beta_4 p_4 [S_0(t_4)]^{\beta_4-1} [S_0(t_2)]^{\beta_2} [S_0(t_3)]^{\beta_3} / (\mu_1 \mu_2 \mu_3 \mu_4) \\ &= \beta_1 \beta_4 p_1^{\beta_1} p_4^{\beta_4} (p_1 + p_2 + p_3 + p_4)^{\beta_1-1} (p_2 + p_3 + p_4)^{\beta_2} \\ &\quad \times (p_3 + p_4)^{\beta_3} / (\mu_1 \mu_2 \mu_3 \mu_4), \end{aligned}$$

where

$$\begin{aligned}\mu_i &= \int_0^{t_4} [S_0(t)]^{\beta_i} dt \\ &= \int_0^{t_1} 1 dt + \int_{t_1}^{t_2} (p_2 + p_3 + p_4)^{\beta_i} dt + \int_{t_2}^{t_3} (p_3 + p_4)^{\beta_i} dt + \int_{t_3}^{t_4} p_4^{\beta_i} dt \\ &= t_1 + (t_2 - t_1)(p_2 + p_3 + p_4)^{\beta_i} + (t_3 - t_2)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i}, \quad i = 1, 2, 3, 4.\end{aligned}$$

If an additional positive mass p^* is added at t^* , where $t_2 < t^* < t_3$, the corresponding likelihood is then

$$\begin{aligned}L_n(t_1, t_2, t^*, t_3, t_4) &= \beta_1 p_1 [S_0(t_1)]^{\beta_1-1} \beta_4 p_4 [S_0(t_4)]^{\beta_4-1} [S_0(t_2)]^{\beta_2} [S_0(t_3)]^{\beta_3} / (\mu_1^* \mu_2^* \mu_3^* \mu_4^*) \\ &= \beta_1 \beta_4 p_1 p_4^{\beta_4} (p_1 + p_2 + p^* + p_3 + p_4)^{\beta_1-1} (p_2 + p^* + p_3 + p_4)^{\beta_2} \\ &\quad \times (p_3 + p_4)^{\beta_3} / (\mu_1^* \mu_2^* \mu_3^* \mu_4^*),\end{aligned}$$

where

$$\begin{aligned}\mu_i^* &= \int_0^{t_4} [S_0^*(t)]^{\beta_i} dt \\ &= \int_0^{t_1} 1 dt + \int_{t_1}^{t_2} (p_2 + p^* + p_3 + p_4)^{\beta_i} dt + \int_{t_2}^{t^*} (p^* + p_3 + p_4)^{\beta_i} dt \\ &\quad + \int_{t^*}^{t_3} (p_3 + p_4)^{\beta_i} dt + \int_{t_3}^{t_4} p_4^{\beta_i} dt \\ &= t_1 + (t_2 - t_1)(p_2 + p^* + p_3 + p_4)^{\beta_i} + (t^* - t_2)(p^* + p_3 + p_4)^{\beta_i} \\ &\quad + (t_3 - t^*)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i} \\ &> t_1 + (t_2 - t_1)(p_2 + p^* + p_3 + p_4)^{\beta_i} + (t^* - t_2)(p_3 + p_4)^{\beta_i} \\ &\quad + (t_3 - t^*)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i} \\ &= t_1 + (t_2 - t_1)\{(p_2 + p^*) + p_3 + p_4\}^{\beta_i} + (t_3 - t_2)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i} \\ &=: \mu_i^{**}.\end{aligned}$$

Therefore the likelihood

$$L_n(t_1, t_2, t^*, t_3, t_4) < L_n^*(t_1, t_2, t_3, t_4),$$

where

$$\begin{aligned}L_n^*(t_1, t_2, t_3, t_4) &= \beta_1 \beta_4 p_1 p_4^{\beta_4} \{p_1 + (p_2 + p^*) + p_3 + p_4\}^{\beta_1-1} \{(p_2 + p^*) + p_3 + p_4\}^{\beta_2} \\ &\quad \times (p_3 + p_4)^{\beta_3} / (\mu_1^{**} \mu_2^{**} \mu_3^{**} \mu_4^{**}).\end{aligned}$$

is the likelihood with jumps only on t_1, t_2, t_3, t_4 , with the re-defined mass of $p_2^* = p_2 + p^*$ at t_2 . Therefore, shifting the mass p^* at t^* to t_2 will increase the likelihood.

Next we consider case 2.

If an extra mass p^* at t^* , where $t^* < t_1$, is added, then the likelihood is

$$\begin{aligned} L_n(t^*, t_1, t_2, t_3, t_4) \\ = \beta_1 p_1 [S_0(t_1)]^{\beta_1-1} \beta_4 p_4 [S_0(t_4)]^{\beta_4-1} [S_0(t_2)]^{\beta_2} [S_0(t_3)]^{\beta_3} / (\mu_1^* \mu_2^* \mu_3^* \mu_4^*) \\ = \beta_1 \beta_4 p_1 p_4^{\beta_4} (p_1 + p_2 + p_3 + p_4)^{\beta_1-1} (p_2 + p_3 + p_4)^{\beta_2} (p_3 + p_4)^{\beta_3} / (\mu_1^* \mu_2^* \mu_3^* \mu_4^*), \end{aligned}$$

where

$$\begin{aligned} \mu_i^* &= \int_0^{t_4} [S_0^*(t)]^\beta dt \\ &= \int_0^{t^*} 1 dt + \int_{t^*}^{t_1} (p_1 + p_2 + p_3 + p_4)^{\beta_i} dt + \int_{t_1}^{t_2} (p_2 + p_3 + p_4)^{\beta_i} dt \\ &\quad + \int_{t_2}^{t_3} (p_3 + p_4)^{\beta_i} dt + \int_{t_3}^{t_4} p_4^{\beta_i} dt \\ &= t^* + (t_1 - t^*)(p_1 + p_2 + p_3 + p_4)^{\beta_i} + (t_2 - t_1)(p_2 + p_3 + p_4)^{\beta_i} \\ &\quad + (t_3 - t_2)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i} \\ &\geq t^* (p_1 + p_2 + p_3 + p_4)^{\beta_i} + (t_1 - t^*)(p_1 + p_2 + p_3 + p_4)^{\beta_i} + (t_2 - t_1)(p_2 + p_3 + p_4)^{\beta_i} \\ &\quad + (t_3 - t_2)(p_3 + p_4)^{\beta_i} + (t_4 - t_3)p_4^{\beta_i} \\ &= \Delta^{\beta_i} \left[t_1 + (t_2 - t_1) \left(\frac{p_2 + p_3 + p_4}{\Delta} \right)^{\beta_i} + (t_3 - t_2) \left(\frac{p_3 + p_4}{\Delta} \right)^{\beta_i} + (t_4 - t_3) \left(\frac{p_4}{\Delta} \right)^{\beta_i} \right] \\ &=: \Delta^{\beta_i} \mu_i^{**}, \end{aligned}$$

where $\Delta = p_1 + p_2 + p_3 + p_4$. Therefore

$$\begin{aligned} L_n(t^*, t_1, t_2, t_3, t_4) \\ = \beta_1 \beta_4 p_1 p_4^{\beta_4} (p_1 + p_2 + p_3 + p_4)^{\beta_1-1} (p_2 + p_3 + p_4)^{\beta_2} (p_3 + p_4)^{\beta_3} / (\mu_1^* \mu_2^* \mu_3^* \mu_4^*) \\ \leq \beta_1 \beta_4 (p_1/\Delta) (p_4/\Delta)^{\beta_4} \{(p_2 + p_3 + p_4)/\Delta\}^{\beta_2} \{(p_3 + p_4)/\Delta\}^{\beta_3} / (\mu_1^{**} \mu_2^{**} \mu_3^{**} \mu_4^{**}). \end{aligned}$$

In other words, dropping the point mass p^* at t^* ($t^* < t_1$) and redistributing it to other points will increase the likelihood.

Therefore we only need to consider those discrete baseline distributions or hazard functions with supports at each of the observed data points. Motivated by the EM algorithm discussed above in the absence of covariates, Qin et al. (2011) proposed the following EM algorithm to deal with the covariate case.

Recall that $0 = t_0 < t_1 < t_2 < \dots < t_k < \infty$ denotes distinct failure and censored time points. For $i = 1, \dots, n$, let $T_{ij}^*, j = 1, 2, \dots, m_i$ be the truncated latent data corresponding to covariate Z_i . We consider the discretized version of $\Lambda(u) = \sum_{t_j < u} \lambda_j$, where λ_j is the positive jump at time t_j for $j = 1, \dots, k$, and $\lambda = (\lambda_1, \dots, \lambda_k)$. For notational convenience, denote $f_i(t) = dF(t|Z_i)$. The log-likelihood based on the complete data is then

$$\sum_{j=1}^k \sum_{i=1}^n \left[I(T_i = t_j) + \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \right] \log f_i(t_j).$$

Conditional on the observed data relative to the i th subject, $\mathcal{O}_i = \{X_i, A_i, \delta_i, Z_i\}$, we obtain the expectation that

$$\begin{aligned} w_{ij} &= E \left[I(T_i = t_j) + \sum_{l=1}^{m_i} I(T_{il}^* = t_j) \middle| \mathcal{O}_i \right] \\ &= \delta_i I(X_i = t_j) + (1 - \delta_i) \frac{p_{ij} I(X_i < t_j)}{\sum_{j=1}^k p_{ij} I(X_i < t_j)} + \frac{1}{\mu_i} (1 - t_j/\tau) p_{ij}, \end{aligned}$$

where

$$p_{ij} = \lambda_j \exp(\beta' Z_i) \exp \left\{ - \sum_{l=1}^j \lambda_l \exp(\beta' Z_i) \right\}, \quad \text{and } \mu_i = \sum_{j=1}^k t_j p_{ij} / \tau.$$

Thus, the expected complete-data log-likelihood function conditional on the observed data is as follows:

$$\ell_E(\beta, \lambda) = \sum_{i=1}^n \sum_{j=1}^k w_{ij} \log f_i(t_j) \tag{25.3.8}$$

$$= \sum_{j=1}^k w_{+j} \log \lambda_j + \sum_{i=1}^n w_{i+} \beta' Z_i - \sum_{l=1}^k \sum_{j=l}^k \sum_{i=1}^n w_{ij} \exp(\beta' Z_i) \lambda_l, \tag{25.3.9}$$

where $w_{+j} = \sum_{i=1}^n w_{ij}$, and $w_{i+} = \sum_{j=1}^k w_{ij}$. In the M-step, we maximize the expected complete-data log-likelihood function conditional on the observed data with respect to the baseline hazard function at t_j , for $j = 1, \dots, k$,

$$\frac{\partial \ell_E(\beta, \lambda)}{\partial \lambda_j} = \frac{w_{+j}}{\lambda_j} - \sum_{l=j}^k \sum_{i=1}^n w_{il} \exp(\beta' Z_i) = 0,$$

which leads to a closed form for λ_j , denoted by

$$\lambda_j(\beta) = \frac{w_{+j}}{\sum_{l=j}^k \sum_{i=1}^n w_{il} \exp(\beta' Z_i)}. \tag{25.3.10}$$

Here, λ_j is the maximizer of the M-step, which is neatly expressed as a function of β . Next, we maximize the expected complete-data log-likelihood function with respect to β

$$\frac{\partial \ell_E(\beta, \lambda)}{\partial \beta} = \sum_{i=1}^n w_{i+} Z_i - \sum_{l=1}^k \sum_{j=l}^k \sum_{i=1}^n w_{ij} Z_i \exp(\beta' Z_i) \lambda_l. \quad (25.3.11)$$

By inserting $\lambda_j(\beta)$ of (25.3.10) into the Eq.(25.3.11), β can be solved from the following equation,

$$\sum_{i=1}^n w_{i+} Z_i - \sum_{l=1}^k w_{+l} \left\{ \frac{\sum_{i=1}^n \sum_{j=l}^k w_{ij} Z_i \exp(\beta' Z_i)}{\sum_{i=1}^n \sum_{j=l}^k w_{ij} \exp(\beta' Z_i)} \right\} = 0, \quad (25.3.12)$$

which is equivalent to maximizing the complete-data profile likelihood function for Λ . With the estimated λ_j ($j = 1, \dots, k$) and β , we can update the expectation of the likelihood via w_{ij} in (25.3.8) and repeat the M-step until the estimators of β and λ_j ($j = 1, \dots, k$) converge.

At the M-step, the estimating Eq. (25.3.12) reveals that we may use existing software for conventional right-censored data to estimate the covariate coefficient β under the Cox proportional hazards model. To simplify description, consider a model with one covariate Z . First we need to create a vector with a length of nk for the weight function defined by $W_{nk} = (w_{11}, \dots, w_{1k}, w_{21}, \dots, w_{2k}, \dots, w_{n1}, \dots, w_{nk})$, which is estimated at the E-step. The corresponding failure time data and covariate vectors are constructed with the same length as W_{nk} , $T_{nk} = (t_1, \dots, t_k, \dots, t_1, \dots, t_k)$ and $Z_{nk} = (Z_1, \dots, Z_1, \dots, Z_n, \dots, Z_n)$, respectively. By using the function “coxph” in S-PLUS (or R) with the “weights” option, we obtain the estimator of β at the M-step from

> coxph(Surv(T_{nk} , Δ) ~ Z_{nk} , weights = W_{nk}),

where the censoring indicator, $\Delta = (1, \dots, 1)$, is an identity vector of length nk .

Finally the maximum semiparametric likelihood can be found by iterating the above steps until convergence. Numerical results can be found in Qin et al. (2011).

25.4 Composite Partial Likelihood Approach

As discussed in Chap. 4 that the main motivations for constructing composite likelihood are the computational simplicity and less model dependency. In this section we will construct a composite partial likelihood using the special structure of length biased survival data. First we consider the ideal case where there is no right censoring.

Given covariate X , the joint density of backward time A and forward time V is given by (Theorem 2.4 in Chap. 2)

$$(A, V)|x \sim \frac{f(a + v|x)}{\mu(x)}, \quad a, v > 0,$$

where $\mu(x) = \int y f(y|x) dy$ is a normalizing constant. The marginal densities are, respectively,

$$A|x \sim \frac{\bar{F}(a|x)}{\mu(x)}, \quad a > 0, \quad V|x \sim \frac{\bar{F}(v|x)}{\mu(x)}, \quad v > 0.$$

As a consequence the product of the two conditional likelihoods of $A_i|(v_i, x_i)$ and $V_i|(a_i, x_i)$ is

$$\left\{ \prod_{i=1}^n \frac{f(y_i|x_i)}{\bar{F}(a_i|x_i)} \right\} \left\{ \prod_{i=1}^n \frac{f(y_i|x_i)}{\bar{F}(v_i|x_i)} \right\}, \quad (25.4.13)$$

where $y_i = a_i + v_i$, $i = 1, 2, \dots, n$. Under the proportional hazards model assumption

$$\lambda(t|x) = \lambda(t) \exp(x\beta),$$

we may apply Cox's partial likelihood to the "augmented data" (y_i, a_i, x_i) , (y_i, v_i, x_i) , $i = 1, 2, \dots, n$ by pretending independence between observations. Therefore the augmented partial likelihood score is (Sect. 24.4 approach 6)

$$\sum_{i=1}^n \left[x_i - \frac{\sum_{j=1}^n x_j \exp(x_j\beta) \{I(y_j \geq y_i \geq a_j) + I(y_j > y_i > v_j)\}}{\sum_{j=1}^n \exp(x_j\beta) \{I(y_j \geq y_i \geq a_j) + I(y_j > y_i > v_j)\}} \right] = 0. \quad (25.4.14)$$

Note that V_i and A_i have identical distribution, $E[I(Y > t > A)] = E[I(Y > t > V)]$, essentially the risk sets are augmented. Therefore we may expect a more accurate estimator of β .

Generalization to the Censored Case

Next we discuss the right censoring case. Note that the joint density of $(Y = A + V, V)$ is

$$(Y, V)|x \sim \frac{f(y|x)}{\mu(x)}, \quad y > v,$$

Conditioning on the failed individual ($\delta = 1$), we have

$$(Y, V)|\delta = 1 \sim \frac{\tilde{G}_c(v|x)f(y|x)}{P(\delta = 1)\mu(x)}, \quad y > v,$$

$$(v|\delta = 1) \sim \frac{\tilde{G}_c(v|x)}{P(\delta = 1)\mu(x)} \int_v^\infty f(y|x)dy = \frac{\tilde{G}_c(v|x)}{P(\delta = 1)\mu(x)} \bar{F}(v|x),$$

where $\tilde{G}_c(c|x)$ is the survival function of the censoring variable C , which may depends on x . Therefore

$$(Y|\delta = 1, v) \sim \frac{f(y|x)}{\bar{F}(v|x)}, \quad y > v,$$

i.e., conditional on the uncensored forward time, the survival time Y has a truncation density $f(y|x)/\bar{F}(y|x)$. As a result we can augment the observed data from

$$(\delta_i, y_i, a_i, x_i), i = 1, 2, \dots, n$$

to

$$(\delta_i, y_i, a_i, x_i), i = 1, 2, \dots, n, \quad (\delta_i = 1)(y_i, v_i, x_i), i = 1, 2, \dots, n.$$

Then the augmented log-likelihood is

$$\begin{aligned} \ell_{Aug} = & \sum_{i=1}^n [\delta_i \log f(y_i|x_i) + (1 - \delta_i) \log \bar{F}(y_i|x_i) - \log \bar{F}(a_i|x_i)] \\ & + \sum_{i=1}^n \delta_i \{\log f(y_i|x_i) - \log \bar{F}(v_i|x_i)\}. \end{aligned} \quad (25.4.15)$$

Using Cox's partial likelihood or the profile likelihood method in Chap. 24, we have score estimating equation

$$\sum_{i=1}^n \delta_i \left[x_i - \frac{\sum_{j=1}^n x_j \exp(x_j \beta) \{I(y_j \geq y_i \geq a_j) + \delta_j I(y_j > y_i > v_j)\}}{\sum_{j=1}^n \exp(x_j \beta) \{I(y_j \geq y_i \geq a_j) + \delta_j I(y_j > y_i > v_j)\}} \right] = 0. \quad (25.4.16)$$

Huang and Qin (2012) discussed large sample results based on the composite partial likelihood argument.

The composite partial likelihood method is very easy to implement in R program. Without loss of generality we assume the first n_1 observations are failures. We can create a new data set

$$(\delta_i^*, y_i^*, a_i^*, x_i^*), i = 1, 2, \dots, n + n_1,$$

where

$$\delta_i^* = \delta_i, y_i^* = y_i, a_i^* = a_i, x_i^* = x_i, i = 1, 2, \dots, n,$$

and

$$\delta_{n+i}^* = 1, y_{n+i}^* = y_i, a_{n+i}^* = v_i, x_{n+i}^* = x_i, i = 1, 2, \dots, n_1.$$

Then the log hazard ratio parameter β can be estimated by using `coxph(Surv(a*, y*, delta*) ~ x)`.

We have to estimate the variance of the maximum composite partial likelihood estimator $\hat{\beta}$ by using a bootstrap method due to the dependence between the original data and the augmented data. To do so, we can simply re-sample the original data $(\delta_i, y_i, a_i, x_i), i = 1, 2, \dots, n$.

Exercise 1 Study the augmented partial likelihood ratio statistic.

$$R(\beta_0) = 2\{\ell_{Aug}(\hat{\beta}) - \ell_{Aug}(\beta_0)\},$$

where

$$\exp\{\ell_{Aug}(\beta)\} = \prod_{i=1}^n \left[\frac{\exp(x_i\beta)}{\sum_{j=1}^n \exp(x_j\beta)\{I(y_j > y_i > a_i) + \delta_j I(y_j > y_i > v_i)\}} \right]^{2\delta_i}.$$

Exercise 2 In the absence of covariates, use the composite partial likelihood to construct a augmented product-limiting estimator of the underlying survival function. Compare this new estimator with Vardi's (1989) estimator.

25.5 Linear Rank Statistics with Cross-Sectional Data

In this section we discuss inference problems for the accelerated failure time model (AFT). We start from the hypothesis test and then move on to estimation. Our aim is to develop useful methods for analyzing length-biased sampling data and backward time data.

Unlike the full parametric approach which depends on the correctness of the underlying distribution assumption, the rank test is robust to this assumption. If the distributional assumption is correct, then the rank test is fully efficient. On the other hand, the rank test can produce a valid test even if the distributional assumption is misspecified. In conventional survival analyses, Prentice (1978) introduced a general class of linear rank tests for the accelerated failure model in the presence of right censoring. The basic idea of his approach is to specify a score function for a classical linear rank test in the absence of censoring, then to construct all possible ranking of the (unobserved) uncensored values which are consistent with censored sample, and finally to assign to each observation the average of all possible scores it could have received in the absence of censoring, giving equal weighting to each possible uncensored ranking. In this approach there is no ordering of censored values between adjacent uncensored values. Ying (1990) generalized Prentice's (1978) linear rank statistics to the left truncation and right censored case. Below we construct the linear rank statistics for both backward times and length biased sampling data.

1. A Brief Review of the Conventional Censored Data Linear Rank Statistics

We begin by reviewing the linear rank statistics in conventional right censoring problems. Suppose that observations $Y_i, i = 1, 2, \dots, n$ follow an accelerated life model

$$T_i = \log Y_i = x_i\beta + \epsilon_i, \quad (25.5.17)$$

where $\epsilon_1, \dots, \epsilon_n$ are independently and identically distributed with density $f(\epsilon)$ and survival function $\bar{F}(\epsilon)$, and x_1, \dots, x_n are the associated covariates. Let $\tau_{(1)} < \dots < \tau_{(k)}$ denote the ordered observed log failure times ($\tau_{(0)} = 0, \tau_{(k+1)} = \infty$). Let $\tau_{i1}, \dots, \tau_{im_i}$ be the set of m_i subjects censored between $\tau_{(i)}$ and $\tau_{(i+1)}$. The corresponding covariates are denoted by x_{i1}, \dots, x_{im_i} . In the calculation of rank likelihood, Prentice (1978) followed closely those calculations in the marginal likelihood for β in the proportional hazards model (Kalbfleisch and Prentice 1973). In evaluating the total probability that the uncensored rank vector should be one of those possible on the sample, one first calculates the probability of the event $\tau_{ij} \geq \tau_{(i)}, j = 1, 2, \dots, m_i$, given the uncensored observations $\tau_{(1)} < \dots < \tau_{(k)}$. This gives $\prod_{i=1}^k \left[f(\tau_{(i)} - x_{(i)}\beta) \prod_{j=1}^{m_i} \bar{F}(\tau_{(i)} - x_{(ij)}\beta) \right]$. Then the rank likelihood is

$$P(\{r\}) = \int_{\tau_{(1)} < \dots < \tau_{(k)}} \prod_{i=1}^k \left[f(\tau_{(i)} - x_{(i)}\beta) \prod_{j=1}^{m_i} \bar{F}(\tau_{(i)} - x_{(ij)}\beta) d\tau_{(i)} \right].$$

In other words there is no ordering of censored Y (or T) values between adjacent uncensored values. Then linear rank statistics can be derived from $\partial \log P(\{r\}) / \partial \beta|_{\beta=0}$. More details can be found in the book by Kalbfleisch and Prentice (2002).

2. Linear Rank Statistics with Length Biased Data

As discussed in the introduction section the length biased sampling data and backward times occur frequently in the cross sectional studies. Denote the conditional density function for the lifetime given covariate x as $f(y|x)$. In the presence of length biased sampling, let A and V be the backward time and forward time respectively. Then the observed lifetime is $Y = A + V$. The joint density of Y, A for a given covariate X is (Theorem 2.4 in Chap. 2)

$$\frac{f(y|x)}{\mu(x)}, \quad y \geq a \geq 0, \quad \mu(x) = \int y f(y|x) dy.$$

Consequently the densities of Y and A for a given covariate x are, respectively, $yf(y|x)/\mu(x)$ and $\bar{F}(a|x)/\mu(x)$.

Under the accelerated life model, $f(y|x) = f(\log y - x\beta)/y$. With length biased sampling, the observed lifetime Y has a density

$$Y|x \sim \frac{f(\log(y) - x\beta)}{\int f(\log(y) - x\beta) dy}, \quad y > 0.$$

Therefore $T = \log Y$ has a density

$$T \sim \frac{\exp(t)f(t - x\beta)}{\int f(t - x\beta) \exp(t) dt} := \frac{\exp(t)f(t - x\beta)}{\mu(x\beta)}, \quad -\infty < t < \infty.$$

Note that

$$\mu(x\beta) = \int f(\log y - x\beta) dy = \int \exp(x\beta + t)f(t)dt = \mu \exp(x\beta), \quad (25.5.18)$$

where $\mu = \int \exp(t)f(t)dt$. Denote the observed order statistic as

$$t_{(1)} < \dots < t_{(n)}.$$

For simplicity the corresponding covariates are denoted as $x_{(1)}, \dots, x_{(n)}$. Then the rank likelihood is

$$P(\beta) = \int_{t_{(1)} < \dots < t_{(n)}} \exp(n\bar{t})f(t_{(1)} - x_{(1)}\beta) \dots f(t_{(n)} - x_{(n)}\beta) dt_{(1)} \dots dt_{(n)} / \{\mu(x_1) \dots \mu(x_n)\},$$

where $\bar{t} = n^{-1} \sum_{i=1}^n t_i$. The log rank likelihood is

$$\begin{aligned} \log P(\beta) &= \log \left[\int_{t_{(1)} < \dots < t_{(n)}} \exp(n\bar{t})f(t_{(1)} - x_{(1)}\beta) \dots f(t_{(n)} - x_{(n)}\beta) dt_{(1)} \dots dt_{(n)} \right] \\ &\quad - \sum_{i=1}^n x_i \beta - n \log \mu. \end{aligned}$$

Differentiating $\log P(\beta)$ with respect to β , we have

$$\begin{aligned} \frac{\partial \log P(\beta)}{\partial \beta} &= - \left[\sum_i \int_{t_{(1)} < \dots < t_{(n)}} \exp(n\bar{t})x_{(i)} \frac{\partial \log f}{\partial t_{(i)}} f(t_{(1)} - x_{(1)}\beta) \dots f(t_{(n)} - x_{(n)}\beta) dt_{(1)} \dots dt_{(n)} \right] \\ &\quad \left[\int_{t_{(1)} < \dots < t_{(n)}} \exp(n\bar{t})f(t_{(1)} - x_{(1)}\beta) \dots f(t_{(n)} - x_{(n)}\beta) dt_{(1)} \dots dt_{(n)} \right]^{-1} - \sum_{i=1}^n x_i. \end{aligned}$$

Evaluating $\partial \log P(\beta)/\partial \beta$ at $\beta = 0$, we end up to

$$\frac{\partial \log P(0)}{\partial \beta} = - \sum_i x_{(i)} E[\partial \log f / \partial t_{(i)}] - \sum_{i=1}^n x_i,$$

where the expectation is respect to the i -th order statistic from a distribution with density $\exp(t)f(t)/\mu$. Note that

$$E[\partial \log f / \partial t] = \int \partial \log f / \partial t \exp(t)f(t)dt / \mu = \int \partial f / \partial t \exp(t)dt / \mu = -1,$$

$$\frac{\partial \log P(0)}{\partial \beta} = - \sum_i x_{(i)} E[\partial \log f / \partial t_{(i)}] + n\bar{x}E[\partial \log f / \partial t], \quad \bar{x} = n^{-1} \sum_{i=1}^n x_i.$$

Observe

$$E[x_{(i)}] = \bar{x}$$

and

$$\sum_i E[\partial \log f / \partial t_{(i)}] = \sum_i E[\partial \log f / \partial t_i] = nE[\partial \log f / \partial t]$$

under $H_0 : \beta = 0$, $\partial \log P(0) / \partial \beta$ has zero mean. The score statistic based on the rank likelihood can be rewritten as

$$\sum_i [x_{(i)} - \bar{x}] E[-\partial \log f / \partial t_{(i)}].$$

Equivalently let

$$c_i = E[\phi(u_{(i)})], \quad \phi(u) = -f'(G^{-1}(u))/f(G^{-1}(u)), \quad G(u) = \int_0^u \exp(t)f(t)dt/\mu$$

and denote

$$u_{(1)} < \dots < u_{(n)}$$

as order statistics from $U(0, 1)$. Then the linear rank statistic is

$$LR = \sum_{i=1}^n (x_{(i)} - \bar{x}) c_i.$$

One key feature of the linear rank statistic is it is unbiased under null $\beta = 0$ no matter what the choice of f is, even though a correct choice can produce the most powerful test.

3. Linear Rank Statistics for Backward Times

In a cross-sectional study without follow up time, the backward time has density

$$A|x \sim \frac{\bar{F}(a|x)}{\mu(x)}, \quad a > 0, \quad \mu(x) = \int \bar{F}(a|x) da.$$

In other words every individual is right censored. Denote $A_{(1)} < \dots < A_{(n)}$ as order statistics and $R = [(1), \dots, (n)]$ as rank statistics, respectively. For simplicity the corresponding covariates of $A_{(i)}$ are denoted as $x_{(i)}$, $i = 1, 2, \dots, n$. The rank likelihood is

$$P_A(\beta) = P\{R = [(1), \dots, (n)]|x_1, \dots, x_n\} = \int_{a_{(1)} < \dots < a_{(n)}} \prod_{i=1}^n \frac{\bar{F}(a_{(i)}|x_{(i)}\beta)}{\mu(x_{(i)}\beta)} da_{(i)}.$$

Under the AFT model (25.5.17) and the Eq. (25.5.18), the log rank likelihood is

$$\begin{aligned} \log P_A &= \log \left[\int_{a_{(1)} < \dots < a_{(n)}} \prod_{i=1}^n \bar{F}(\log a_{(i)} - x_{(i)}\beta) \right] - \sum_{i=1}^n x_i\beta - n \log \mu, \\ \frac{\partial \log P_A}{\partial \beta} |_{\beta=0} &= \sum_{i=1}^n \left[\int_{a_{(1)} < \dots < a_{(n)}} \frac{\partial \bar{F}(\log a_{(i)} - x_{(i)}\beta)}{\partial \beta} \prod_{i=1}^n \bar{F}(\log a_{(i)} - x_{(i)}\beta) da_{(i)} \right]_{\beta=0} \\ &\quad \left[\int_{a_{(1)} < \dots < a_{(n)}} \prod_{i=1}^n \bar{F}(\log a_{(i)} - x_{(i)}\beta) \right]_{\beta=0}^{-1} - \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n x_{(i)} E \left[\frac{f(\log a_{(i)})}{\bar{F}(\log a_{(i)})} \right] - \sum_{i=1}^n x_i. \end{aligned}$$

The expectation is respect to the i -th order statistic from a distribution generated from the density $\bar{F}(\log a)/\mu$.

Also note that

$$\sum_{i=1}^n E \left[\frac{f(\log a_{(i)})}{\bar{F}(\log a_{(i)})} \right] = \sum_{i=1}^n E \left[\frac{f(\log a_i)}{\bar{F}(\log a_i)} \right] = n \int f(\log a) da / \mu = n.$$

Finally the linear rank statistic can be written as

$$LR_A = \sum_{i=1}^n (x_{(i)} - \bar{x}) c_i, \quad c_i = E \left[\frac{f(\log a_{(i)})}{\bar{F}(\log a_{(i)})} \right].$$

Chan and Qin (2015) implemented this statistic numerically. Readers may find details in their paper.

Exercise Generalize the linear rank statistics to the length biased and right censored data problems.

25.6 Estimating Equations Derived from an Embedded Likelihood

Ning et al. (2014) discussed a general method for constructing score estimating equations in the AFT model by embedding it into a larger semiparametric family.

This method is applicable to many other types of incomplete data under the AFT model assumption. For illustration, we discuss length biased sampling data only.

As shown before, the length biased sampling likelihood is

$$(A, T, \delta)|x \sim \frac{f^\delta(t|x)\bar{F}^{1-\delta}(t|x)}{\mu(x)}, \quad t > a, \quad \mu(x) = \int_0^\infty tf(t|x)dt.$$

Suppose in the absence of length biased sampling, T satisfies an AFT model, then

$$T|x \sim f(t|x) = \exp(-x\beta)f(t \exp(-x\beta)), \quad t > 0.$$

Note that

$$Y = \log T|x \sim f\{\exp(y - x\beta)\}\exp(y - x\beta) := g(y - x\beta), \quad -\infty < y < \infty,$$

is a linear model, where $g(s) = f(\exp(s))\exp(s)$. Therefore the length biased sampling version is

$$(A, T, \delta)|x \sim \frac{\{\exp(-x\beta)f(t \exp(-x\beta))\}^\delta \bar{F}^{1-\delta}(t \exp(-x\beta))}{\mu(x)},$$

where

$$\mu(x) = \int_0^\infty t \exp(-x\beta)f(t \exp(-x\beta))dt.$$

Let $T^* = T \exp(-x\beta)$, $A^* = A \exp(-x\beta)$, $\delta = I(Y < A + C) = I(Y^* < A^* + C^*)$, $C^* = C \exp(-x\beta)$, then

$$(A^*, T^*, \delta)|x \sim \frac{f^\delta(t^*)\bar{F}^{1-\delta}(t^*)}{\mu_0} \exp\{(1 - \delta)x\beta\}, \quad t^* > a^*, \quad \mu_0 = \int_0^\infty tf(t)dt.$$

Denote the corresponding hazard of f as $\lambda(t)$. We can enlarge it through a proportional hazards model

$$\lambda(t) = \lambda(t) \exp(x\alpha).$$

Replacing $\lambda(t)$ by this enlarged hazard, we have the log-likelihood

$$\begin{aligned} \ell = & \sum_{i=1}^n [\delta_i \{\log \lambda(t_i^*) + x_i \alpha\} - \Lambda(t_i^*) \exp(x_i \alpha)] \\ & - \log \int \exp\{-\Lambda(s) \exp(x_i \alpha)\} ds] + \sum_{i=1}^n (1 - \delta_i) x_i \beta. \end{aligned}$$

Taking derivative with respect to α , we arrive at

$$\sum_{i=1}^n x_i \left\{ \delta_i - \Lambda(t_i^*) + \frac{\int \Lambda(s) \exp(-\Lambda(s))ds}{\int \exp(-\Lambda(s))ds} \right\} = 0.$$

Based on (A_i^*, T_i^*, δ_i) , $i = 1, 2, \dots, n$, finally we can plug in Vardi's (1989) baseline cumulative hazard estimator $\hat{\Lambda}$ in the estimating equation. Of course iterations are needed since A_i^* and T_i^* involve the unknown parameter β . Simulations and large sample results are given in Ning et al. (2014).

Exercise Derive the score estimating equations by embedding the underlying density in the proportional odds ratio model.

25.7 Generalized Multiplicative Censoring and Truncation

So far we have assumed that the truncation time has a uniform distribution. We now relax this assumption.

Due to the fact that the censoring distribution function can be factored out in the likelihood for right censored data, the information on censoring distribution would not provide any help for the estimation of lifetime distribution. However, the information for the truncation distribution in left truncation data problems would be very informative for the estimation of lifetime distribution (Sect. 24.2). Wang (1989, 1992) studied a semiparametric model when the truncation distribution is assumed to have a parametric form but the lifetime distribution is left arbitrary. She found that the maximum semiparametric likelihood estimation is much more efficient than the nonparametric product limit estimation. Unfortunately, her method can only deal with left truncation but not right censoring. In practice, left truncation and right censored data occur frequently, for example in AIDS studies by Lagakos et al. (1988), Kalbfleisch and Lawless (1991) and Wang (1992).

Next we study the semi-parametric maximum likelihood estimate for a lifetime distribution G under the following "generalized multiplicative-censorship" model.

Let

$$X_i \sim dG(x), \quad Z_j \sim dG(z), \quad U_i \sim U(0, 1), \quad V_j \sim U(0, 1), \quad i = 1, 2, \dots, m; \\ j = 1, 2, \dots, n,$$

where random variables X_i, Z_j, U_i, V_j are independent. Suppose two groups of data are available

$$(S_i, X_i), i = 1, 2, \dots, m; \quad Y_i, i = 1, 2, \dots, n,$$

where (S_i, X_i) , $i = 1, 2, \dots, m$ and (Y_j, Z_j) , $j = 1, 2, \dots, n$ are linked by

$$S_i = H^{-1}(U_i H(X_i)), \quad Y_j = H^{-1}(V_j H(Z_j)),$$

where H is a specified parametric monotonic increasing function or distribution function depending on a parameter θ and H^{-1} is the inverse function of H . Note that $S_i \leq H^{-1}(H(X_i)) = X_i$ and $Y_j \leq H^{-1}(H(Z_j)) = Z_j$. In other words, S_i and Y_j are stochastically smaller than X_i and Z_j , respectively (Chap. 1). We call this as a “generalized multiplicative censoring model”. If $H(t) = t$, it becomes the multiplicative censoring problem discussed by Vardi (1989). Our main interest is the estimation of G and θ .

Note that

$$P(Y \leq t | Z = z) = P(VH(z) \leq H(t) | Z = z) = \frac{H(t)}{H(z)}, \quad t \leq z.$$

The marginal density of Y is

$$Y \sim \int_{z>t} \frac{h(t)}{H(z)} dG(z).$$

The conditional density of Z given Y is

$$Z|Y = t \sim \frac{h(t)}{H(z)} dG(z) \left(\int_{z>t} \frac{h(t)}{H(z)} dG(z) \right)^{-1} = \frac{1}{H(z)} dG(z) \left(\int_{z>t} \frac{1}{H(z)} dG(z) \right)^{-1}, \quad t \leq z.$$

The overall likelihood is

$$L = \prod_{i=1}^m \frac{h(s_i)}{H(x_i)} dG(x_i) \prod_{i=1}^n h(y_i) \int_{z \geq y_i} \frac{1}{H(z)} dG(z). \quad (25.7.19)$$

Similar to Vardi’s (1989) approach, we can show that the nonparametric MLE for G has jumps at each of the observed data points.

Before deriving the nonparametric MLE, we first establish a connection between the generalized “multiplicative censoring model” and the left truncation and right censoring data problems.

Connection with Truncation Data Problem When the Truncation Distribution is Known up to a Parameter

Let $\tilde{Y} = \min(Y, C)$, $\delta = I(Y \leq C)$. Denote the left truncation time as A . Assume $A \sim h(a)$, $Y \sim f(y)$. The observed left truncation and right censored data have a density

$$(A, \tilde{Y}, \delta) | Y > A \sim \frac{h(a)f^\delta(y)\bar{F}^{1-\delta}(y)}{P(Y > A)}.$$

Note that

$$P(Y > A) = \int \bar{F}(a)h(a)da = \int H(y)f(y)dy,$$

$$\frac{h(a)f^\delta(y)\bar{F}^{1-\delta}(y)}{P(Y > A)} = \frac{f^\delta(y)\bar{F}^{1-\delta}(y)}{\bar{F}(a)} \frac{\bar{F}(a)h(a)}{\int \bar{F}(a)h(a)da}.$$

If h is unknown, then the maximum likelihood estimation of H is

$$\left[\sum_i \frac{I(a_i \leq a)}{\bar{F}(a_i)} \right] \left[\sum_i \frac{1}{\bar{F}(a_i)} \right]^{-1}.$$

On the other hand, if the form of h is known, let

$$dG(y) = \frac{H(y)f(y)dy}{\int H(y)f(y)dy}, \quad C = P(Y > A),$$

then

$$\bar{F}(y_i)/C = \int_{y_i}^{\infty} dF(y)/C = \int_{y_i}^{\infty} \frac{1}{H(y)} dG(y)$$

and

$$\frac{h(a_i)f^\delta(y_i)\bar{F}^{1-\delta}(y_i)}{P(Y > A)} = h(a_i) \left(\frac{1}{H(y_i)} dG(y_i) \right)^{\delta_i} \left(\int_{y_i}^{\infty} \frac{1}{H(y)} dG(y) \right)^{1-\delta_i}.$$

Therefore the likelihood based on the observed data ($\tilde{y}_i = \min(y_i, c_i)$, $\delta_i = I(y_i \leq c_i)$, $i = 1, 2, \dots, n$) is

$$\left[\prod_{i=1}^n h(a_i) \left(\frac{1}{H(y_i)} \right)^{\delta_i} \right] \left[\prod_{i=1}^n (dG(y_i))^{\delta_i} \prod_{i=1}^n \left(\int_{y_i}^{\infty} \frac{1}{H(y)} dG(y) \right)^{1-\delta_i} \right].$$

It can be written as

$$\left[\prod_{i=1}^n \left(\frac{h(a_i)}{H(y_i)} dG(y_i) \right)^{\delta_i} \right] \left[\prod_{i=1}^n \left(h(y_i) \int_{y_i}^{\infty} \frac{1}{H(y)} dG(y) \right)^{1-\delta_i} \right] \left[\prod_{i=1}^n \left(\frac{h(a_i)}{h(y_i)} \right)^{1-\delta_i} \right]. \quad (25.7.20)$$

Comparing (25.7.19) and (25.7.20), we need to augment a term

$$\left[\prod_{i=1}^n \left(\frac{h(a_i)}{h(y_i)} \right)^{1-\delta_i} \right]$$

for the truncation likelihood problem. However in terms of estimating G we may use likelihood (25.7.19).

Next we study the nonparametric MLE for G based on likelihood (25.7.19).

Let

$$0 < t_1 < t_2 < \dots < t_\nu$$

be the distinct values of $x_1, \dots, x_m; y_1, \dots, y_n$, where $\nu \leq n + m$. Let

$$\xi_j = \sum_{i=1}^m I(x_i = t_j), \quad \eta_j = \sum_{i=1}^n I(y_i = t_j), \quad j = 1, 2, \dots, \nu.$$

Let $p_j = dG(t_j)$, $j = 1, 2, \dots, \nu$. The likelihood (25.7.19) is

$$L = \prod_{i=1}^m \frac{h(s_i)}{H(x_i)} \prod_{j=1}^\nu h^{\eta_j}(t_j) \prod_{j=1}^\nu p_j^{\xi_j} \left(\sum_{k=j}^\nu \frac{1}{H(t_k)} p_k \right)^{\eta_j}$$

subject to the constraint

$$p_j \geq 0, \quad (j = 1, 2, \dots, \nu), \quad \sum_{j=1}^\nu p_j = 1.$$

As Vardi (1989) did, we may use transformation

$$q_j = p_j/H(t_j), \quad Q_j = \sum_{k=j}^\nu q_k, \quad j = 1, 2, \dots, \nu.$$

For fixed θ , the log-likelihood becomes

$$\begin{aligned} \ell &= \sum_{i=1}^m \log h(s_i) + \sum_{j=1}^\nu \eta_j [\log h(t_j) - \log H(t_j)] + \sum_{j=1}^\nu \xi_j \log H(t_j) \\ &\quad + \sum_{j=1}^\nu [\xi_j \log q_j + \eta_j \log Q_j]. \end{aligned}$$

subject to the constraint

$$\sum_{j=1}^\nu H(t_j) q_j \leq 1.$$

It can be shown that for fixed θ the log-likelihood is a concave function. Therefore Csiszar and Tusnády (1984) result implies the convergence of EM algorithm.

An EM Algorithm

In order to maximize the likelihood, we must discretize F at each $0 < t_1 < \dots < t_\nu$. The log-likelihood based on the full data is

$$\begin{aligned}
\ell &= \sum_{j=1}^{\nu} \sum_{i=1}^m I(x_i = t_j) \log p_j + \sum_{j=1}^{\nu} \sum_{i=1}^n I(z_i = t_j) \log p_j \\
&\quad + \sum_{i=1}^m \{\log h(s_i) - \log H(x_i)\} + \sum_{i=1}^n \{\log h(y_i) - \log H(z_i)\} \\
&= \sum_{j=1}^{\nu} \sum_{i=1}^m I(x_i = t_j) \log p_j + \sum_{j=1}^{\nu} \sum_{i=1}^n I(z_i = t_j) \log p_j \\
&\quad + \sum_{i=1}^m \{\log h(s_i) - \log H(x_i)\} + \sum_{j=1}^{\nu} \left\{ \sum_{i=1}^n I(y_i = t_j) \right\} \log h(t_j) \\
&\quad - \sum_{j=1}^{\nu} \left\{ \sum_{i=1}^n I(z_i = t_j) \right\} \log H(t_j).
\end{aligned}$$

Let

$$w_j = E[\sum_{i=1}^m I(x_i = t_j) + \sum_{i=1}^n I(z_i = t_j) | (s_1, x_1, \dots, s_m, x_m), (y_1, \dots, y_n)] = \xi_j + v_j,$$

where

$$v_j = \frac{p_j}{H(t_j)} \sum_{k=1}^j \eta_k \left(\sum_{i=k}^{\nu} \frac{p_i}{H(t_i)} \right)^{-1}.$$

The log-likelihood is

$$\begin{aligned}
\ell &= \sum_{j=1}^{\nu} w_j \log p_j + \sum_{i=1}^m \log h(s_i) - \sum_{j=1}^{\nu} \xi_j \log H(t_j) \\
&\quad + \sum_{i=1}^n \log h(y_i) - \sum_{j=1}^{\nu} v_j \log H(t_j).
\end{aligned}$$

Maximizing ℓ with respect to p_j and θ ,

$$p_j = (n+m)^{-1} w_j, \quad \sum_{j=1}^{\nu} w_j = n+m$$

$$\sum_{i=1}^m \frac{\partial \log h(s_i, \theta)}{\partial \theta} - \sum_{j=1}^{\nu} \xi_j \frac{\partial \log H(t_j, \theta)}{\partial \theta} + \sum_{j=1}^n \frac{\partial \log h(y_j, \theta)}{\partial \theta} - \sum_{j=1}^{\nu} v_j \frac{\partial \log H(t_j, \theta)}{\partial \theta} = 0.$$

Finally

$$dF(t_j) = \frac{p_j}{H(t_j)} \left\{ \sum_{i=1}^{\nu} \frac{p_i}{H(t_i)} \right\}^{-1}.$$

Huang et al. (2015) studied this problem thoroughly. Shen (2007, 2009) proposed a closely related approach. Mandel (2007) studied the special case when the truncation distribution is completely known.

25.8 The Multi-sample Wicksell Corpuscle Problem

Based on the discussions on different length biased sampling and multiplicative censoring problems, now we are ready to tackle the maximum likelihood estimation based on multiple-sample data for Wicksell corpuscle problems (Chan and Qin 2016).

Suppose spherical particles of different radii are randomly distributed in a three dimensional space, where the center of the sphere is distributed according to a stationary spatial Poisson process. Denote the density of the radii of the particles as $f(R)$, $R > 0$. It would be straightforward to estimate it if R can be directly sampled. Due to technical difficulties, a two-dimensional planar sampling and one dimensional linear probe sampling are used in practice.

1. Two Dimensional Planar Sample

Let r be the radii of the two-dimensional circular profiles. Two layers of bias occur in the planar sample.

(1) Spheres with larger radii are more likely to be sampled. In other words this is a length biased sampling problem. The sampling distribution of R is

$$f^S(R) = \frac{Rf(R)}{\int_0^\infty Rf(R)dR}, \quad R > 0.$$

(2) The radii r of the two-dimensional circular profiles are indirect measurements and are always smaller than the radii of the sphere being sampled. Given a sphere with radius R is being sampled, Wicksell (1925) showed that r has density

$$g(r|R) = \frac{r}{R\sqrt{R^2 - r^2}}, \quad 0 < r \leq R.$$

By taking the two layer biased sampling into consideration, we can show that the sampled r has density

$$g(r) = \int g(r|R)f^S(R)dR = \frac{r}{\int_0^\infty Rf(R)dR} \int_r^\infty \frac{f(R)dR}{\sqrt{R^2 - r^2}}, \quad r > 0.$$

2. One-Dimensional Linear Probe Sampling

Let l be the half length of the traces where the linear probe intersects the spheres. Similar to the two-dimensional case, two layer bias sampling occurs in the linear probe sampling. (1) The spheres are being sampled proportional to their squared radii.

$$h^S(R) = \frac{R^2 f(R)}{\int_0^\infty R^2 f(R) dR}, \quad R > 0.$$

(2) The half-lengths of the measurements are always smaller than the radii of the spheres. Given a sphere with radius R is sampled, Watson (1971) showed that the sampling distribution of l is given by

$$h(l|R) = \frac{2l}{R^2}, \quad 0 < l \leq R.$$

As a result the sampling distribution of l is

$$h(l) = \int h^S(R) h(l|R) dR = \frac{2l \{1 - F(l)\}}{\int_0^\infty R^2 f(R) dR}, \quad l > 0.$$

Let

$$X_i \sim H_i(x) = \frac{x^i dF(x)}{\mu_i}, \quad \mu_i = \int_0^\infty x^i dF(x), \quad i = 1, 2.$$

Then mathematically it can be shown (exercise) that the planar probe sampling is equivalent to observing $Y_1 = X_1 \sqrt{1 - U^2}$. Similarly the linear probe is equivalent to observing $Y_2 = X_2 \sqrt{U}$, where $U \sim U(0, 1)$. This is also related to Vardi's (1989) multiplicative censoring problem. However instead of uniform censoring, the censoring variables are $\sqrt{1 - U^2}$ and \sqrt{U} , respectively. In statistical literature the Wicksell corpuscle problem based on two dimensional planar sample alone has been discussed extensively. A comprehensive introduction of this problem is given in a recent monograph by Groeneboom and Jongbloed (2014). In particular they considered shape constrained estimation problems. However in multiple sample problems such as when both planar sample and linear probe sample are available, it is not clear how to generalise their method to find the nonparametric maximum likelihood of F .

3. Nonparametric Maximum Likelihood for Wicksell Corpuscle Problem Based on Multiple Samples

Using the connection with multiplicative censoring, we can estimate F by combining the EM algorithm and biased sampling techniques. We can treat X_1 and X_2 as complete data. Since this is a biased sampling problem with weights x and x^2 , respectively, in Chap. 10 we already discussed how to estimate F . However the observed data are Y_1 and Y_2 , we can use EM algorithm to impute X_1 and X_2 . We leave this as an exercise. Readers may find details in Chan and Qin (2016).

25.9 Missing Information Principle

In this section we discuss the general missing data information principle for handling survival data. Basically it is a method to impute incomplete data in the estimating equation setup. Sun et al. (2016) are studying this principle for many types of survival data, including right censoring, left truncation or both, and length biased sampling data. A previous work on constructing M-estimators of regression parameters in the presence of left truncation and right censoring on the observed responses can be found in Lai and Ying (1994).

We start from the basic survival estimating function. Let T be the survival time. Denote $N(t) = I(T \leq t)$ as the counting function. The basic estimating function is

$$dM(t) = dN(t) - \lambda(t)I(T \geq t)dt, \quad M(t) = N(t) - \int_0^t \lambda(u)I(Y > u)du.$$

It is well known that $M(t)$ is a martingale with mean 0 (for example, Chap. 5 in Kalbfleisch and Prentice (2002), Chap. 1 in Fleming and Harrington (1994)). Solving

$$\sum_{i=1}^n dM_i(t) = 0,$$

we can estimate the baseline hazard

$$\hat{\lambda}(t)dt = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n I(T_i \geq t)}.$$

This is the well known Nelson estimator.

Next we consider three cases. (1) There is right censoring but no left truncation. (2) In addition to right censoring there exists left truncation also. (3) The collected are length biased and right censored.

1. Right Censoring Only

Let C be the censoring time. The observed data are $Y = \min(T, C)$, $\delta = I(T \leq C)$. In this case, we cannot use $M(t)$ as an estimating function since T is not observable for the censored individuals. To utilize $M(t)$, we can employ the imputation method. Note that

$$M(t) = \delta M(t) + (1 - \delta)M(t).$$

Replacing $M(t)$ with $E[M(t)|\delta = 0, Y = y]$ for those censored individuals, we find that

$$P(T = t|C < T, C = y) = \frac{f(t)}{\bar{F}(y)}I(t > y),$$

$$P(T > t|C < T, C = y) = I(t < y) + I(t \geq y)\frac{\bar{F}(t)}{\bar{F}(y)}.$$

Therefore

$$\begin{aligned}
& \delta dM(t) + (1 - \delta)E[dM(t)] \\
&= \delta dN(t) - \delta \lambda(t)I(T > t)dt + (1 - \delta)\frac{f(t)}{\bar{F}(y)}I(t > y)dt \\
&\quad - (1 - \delta)\lambda(t) \left[I(t < y) + I(t > y)\frac{\bar{F}(t)}{\bar{F}(y)} \right] dt \\
&= \delta dN(t) - \lambda(t)[\delta I(T > t) + (1 - \delta)I(y > t)]dt \\
&= \delta dN(t) - \lambda(t)I(Y > t)dt.
\end{aligned}$$

This is the basic estimating equation for right censored survival data.

2. Left Truncation and Right Censoring

Next we study the imputation method for the left truncation and right censoring case. We consider two cases.

(1) The censoring time starts from the beginning of a study. (2) The censoring time starts from the truncation time. In general the truncation distribution function is unknown. In the approaches below we condition on the truncation time. As a consequence the truncation distribution function does not play a role in the estimation stage.

Version 1 The censoring time starts from the beginning of a study, i.e., the available data are $(Y = \min(T, C), A, \delta = I(T < C))$, where $T > A$. We can replace $M(t)$ by

$$M(t) = \delta M(t) + (1 - \delta)E[M(t)|O] + E[M(t)|O^*],$$

where O^* denotes those truncated individuals, i.e., $T^* < A^* = a$. We then calculate the conditional expectations.

$$\begin{aligned}
& P(T = t|T > a, T > c) - \lambda(t)P(T > t|T > a, T > c) \\
&= \frac{f(t)}{\bar{F}(c)}I(t > c) - \lambda(t)\frac{\bar{F}(\max(c, t))}{\bar{F}(c)} \\
&= -\lambda(t)I(t < c),
\end{aligned}$$

$$\begin{aligned}
& P(T = t|T < a) - \lambda(t)P(T > t|T < a) = \frac{f(t)}{F(a)}I(t < a) - \lambda(t)\frac{\bar{F}(t) - \bar{F}(a)}{F(a)}I(t < a) \\
&= \lambda(t)I(t < a)\frac{\bar{F}(a)}{F(a)}.
\end{aligned}$$

On average, the $F(a)/\bar{F}(a)$ terms are truncated. Therefore

$$E[M(t)|O^*] = \frac{F(a)}{\bar{F}(a)}\lambda(t)\frac{\bar{F}(a)}{F(a)}I(t < a) = \lambda(t)I(t < a).$$

We can show that the basic imputed estimating equation is

$$\delta M(t) - (1 - \delta)\lambda(t)I(t < c) + \lambda(t)I(t < a) = \delta dN(t) - \lambda(t)I(\min(c, T) > t > a).$$

This is precisely the estimating equation for the truncation data problem discussed in the last chapter.

An Alternative Imputation Method

We consider an alternative imputation method. First we consider the case with no right censoring. The basic estimating equation

$$dM(t) = dN(t) - I(T > t)\lambda(t)dt$$

can be written as

$$dM(t) = I(T < a)dM(t) + I(T > a)dM(t).$$

Under truncation $dM(t)$ is observable if and only if $T > A$. Therefore we need to replace the first term by its expectation. From $0 = E[dM(t)]$ we have

$$E[I(T < a)dM(t)] = -E[I(T > a)dM(t)].$$

As a consequence, $dM(t)$ can be replaced by

$$-E[I(T > a)dM(t)] + I(T > a)dM(t).$$

Note that

$$E[dN(t)|T > a] = P(T = t|T > a) = \frac{f(t)}{\bar{F}(a)}I(t > a),$$

$$\lambda(t)P(T > t|T > a) = \lambda(t)\frac{\bar{F}(\max(t, a))}{\bar{F}(a)},$$

$$E[dM(t)|T > a] = -\lambda(t)I(t < a).$$

Finally we can replace $dM(t)$ by

$$\begin{aligned} & -I(T > a)E[dM(t)|T > a] + I(T > a)dM(t) \\ &= I(T > a)I(t < a)\lambda(t) + I(T > a)dM(t), \end{aligned}$$

or

$$I(T > a)[dN(t) - \lambda(t)I(a < t)] = I(T > a)dN(t) - I(T > t > a)\lambda(t)dt.$$

This is exactly the estimating function for the truncation data problem.

The censoring case can be derived similarly. In fact, let $Y = \min(T, C)$, $\delta = I(T \leq C)$.

$$\begin{aligned} & E[\{dN(t) - \lambda(t)I(T > t > a)\}|T > A, A = a, \delta, Y = y > a] \\ &= \delta\{dN(t) - \lambda(t)I(T > t > a)\} + (1 - \delta)\left[\frac{f(t)}{\bar{F}(y)}I(t > y) - \lambda(t)\frac{\bar{F}(\min(t, y))}{\bar{F}(y)}\right] \\ &= \delta\{dN(t) - \lambda(t)I(T > t > a)\} - (1 - \delta)\lambda(t)I(Y > t > a) \\ &= dN(t) - I(Y > t > a)\lambda(t)dt. \end{aligned}$$

Version 2 Suppose A and T are independent. The observable data satisfy $T > A$. Denote the residual lifetime as $V = T - A$. It may be censored by C . The observed data are

$$\delta = I(V \leq C), \quad A + \min(V, C).$$

In this setup censoring starts from the truncation time A . Assume that

$$T \sim f(t), \quad A \sim g(a).$$

Again the basic estimating equation is given by

$$N(t) = I(T \leq t), \quad dM(t) = dN(t) - \lambda(t)I(T > t)dt.$$

We need to impute T if $T > A = a$, $V > C = c$ or $T < A = a$. Conditioning on $T > A$, $V > c$, and the observed truncation time $A = a$,

$$\begin{aligned} P(T = t|A = a, V > c, C = c) &= \frac{P(T = t, A = a, V > c)}{P(A = a, V > c)} \\ &= \frac{f(t)g(a)}{g(a)\bar{F}(a+c)}I(t > a+c), \end{aligned}$$

or

$$\begin{aligned} P(T = t|A = a, V > c, C = c) &= \frac{f(t)}{\bar{F}(a+c)}I(t > a+c) \\ &= \lambda(t)\frac{\bar{F}(t)}{\bar{F}(a+c)}I(t > a+c), \end{aligned}$$

$$\begin{aligned} \lambda(t)P(T > t|A = a, V > c) &= \lambda(t)\frac{P(T > t, A = a, V > c)}{P(A = a, V > c)} \\ &= \lambda(t)\frac{\bar{F}(\max(t, a+c))}{\bar{F}(a+c)}. \end{aligned}$$

Moreover

$$\begin{aligned} P(T = t|T < a, A = a) - \lambda(t)P(T > t|T < a, A = a) &= \frac{f(t)}{F(a)}I(t < a) \\ &\quad - \lambda(t)\frac{F(a) - F(t)}{F(a)}I(t < a), \end{aligned}$$

or

$$P(T = t|T < a, A = a) - \lambda(t)P(T > t|T < a, A = a) = \lambda(t)\frac{\bar{F}(a)}{F(a)}I(t < a).$$

On average, the $F(a)/\bar{F}(a)$ terms are truncated, therefore

$$E[dM(t)|O^*] = \lambda(t)I(t < a).$$

Finally

$$\delta M(t) + (1 - \delta)E[dM(t)|O] + E[dM(t)|O^*] = \delta dM(t) - \lambda(t)I(\min(T, A + C) > t > A).$$

We conclude that, whether the censoring time starts from the beginning of a study or from the truncation time, we end up with the same imputation estimating equations.

So far we have not used the distribution information on the truncation variable A . If this information indeed is available such as in the length biased sampling case, then we can incorporate it in the imputation stage.

3. Length Biased Sampling with Right Censoring

Next we utilize the uniform distribution assumption for the truncation variable A in the imputation stage.

Let τ be the largest observation among all observed lifetime data. Mathematically the length biased sample problem is equivalent to a truncation model with

$$A \sim U(0, \tau), \quad T \sim F(y), \quad 0 \leq T \leq \tau.$$

We assume that before observing n sample points $O = \{(Y_1, \delta_1, A_1), \dots, (Y_n, \delta_n, A_n), A_i \leq Y_i, i = 1, 2, \dots, n\}$ there are m sample points $O^* = \{(T_1^*, A_1^*), \dots, (T_m^*, A_m^*), A_i^* > T_i^*, i = 1, 2, \dots, m\}$. Then m has a negative binomial distribution with parameter μ ; that is the probability mass function of m is

$$\binom{m+n-1}{m} (1-\pi)^m \pi^n, \quad m = 0, 1, 2, \dots$$

where $\pi = \mu/\tau = \int_0^\tau y dF(y)/\tau$. Thus the missing data are m and $(Y_1^*, A_1^*), \dots, (A_m^*, A_m^*)$. Note that

$$En(n+m)^{-1} = \pi, \quad m = n(1-\pi)/\pi.$$

Consider

$$E[M(t)|T > A, T > c, A = a]$$

and

$$E[M(t)|T < A] = \frac{f(t)(1 - t/\tau)}{1 - \mu} - \lambda(t) \frac{\int_t^\tau f(s)(1 - s/\tau)ds}{1 - \mu}, \mu = \int_0^\tau \bar{F}(s)ds/\tau.$$

The difference in the imputation stage for the length biased case and for the general truncation case is that the former one is conditional on the event $T < A$ and the latter one is conditional on $T < A, A = a$. Therefore

$$\begin{aligned} \delta M(t) - (1 - \delta)\lambda(t)I(t < y)dt + \lambda(t)dt\bar{F}(t)(1 - t/\tau)/\mu - \lambda(t)dt \frac{\int_t^\tau f(s)(1 - s/\tau)ds}{\mu} \\ = \delta M(t) - (1 - \delta)\lambda(t)dtI(t < y) + \lambda(t)dt \int_t^\tau \bar{F}(s)ds/(\mu\tau) \\ = \delta dN(t) - \lambda(t)I(y \geq t) + \lambda(t) \int_t^\tau \bar{F}(s)ds / \int_0^\tau \bar{F}(s)ds, \end{aligned}$$

where $y = \min(T, a + c) = \min(a + v, a + c)$. Therefore

$$\sum_{i=1}^n \left[\delta_i dN_i(t) - \lambda(t) \left\{ I(y_i \geq t) - \frac{\int_t^\tau \bar{F}(s)ds}{\int_0^\tau \bar{F}(s)ds} \right\} \right] = 0.$$

As a consequence

$$\hat{\lambda}(t)dt = \left[\sum_{i=1}^n \delta_i dN_i(t) \right] \left[\sum_{i=1}^n \left\{ I(y_i \geq t \geq a_i) + I(a_i \geq t) - \frac{\int_t^\tau \bar{F}(s)ds}{\int_0^\tau \bar{F}(s)ds} \right\} \right]^{-1}.$$

Different Versions of Imputation Methods

We found the basic truncation estimating equation in the absence of knowledge of truncation distribution is

$$E[\delta dN(t) - \lambda(t)I(Y \geq t > A)dt] = 0.$$

The indicator function can be written as

$$I(Y \geq t > A) = I(Y \geq t) - I(A \geq t).$$

In the presence of uniform distribution knowledge for A in the length biased sample problem, Luo and Tsai (2009) used $P(A \geq t|y, \delta)$ to replace $I(A \geq t)$ in the truncation estimating equation. Unfortunately, in general $P(A \geq t|y, \delta)$ depends on the censoring distribution G of C . G needs to be estimated by the Kaplan–Meier method.

If we replace $I(A \geq t)$ by $P(A \geq t) = \int_t^\infty \bar{F}(s)ds/\mu$ in the truncation estimating equation, then we have a self consistent estimating equation

$$E[\delta dN(t) - \lambda(t)\{I(Y \geq t) - \int_t^\infty \bar{F}(s)ds/\mu\}] = 0.$$

This is precisely the same maximum likelihood estimating equation derived before.

All imputation methods discussed above can be applied to regression models, particular for the Cox regression model. Let

$$\begin{aligned} R_i(t, \beta) &= \left\{ I(y_i \geq t \geq a_i) + I(a_i \geq t) - \frac{\int_t^\tau \bar{F}(s|x_i/\beta)ds}{\int_0^\tau \bar{F}(s|x_i/\beta)ds} \right\} \\ &= \left\{ I(y_i \geq t \geq a_i) + I(a_i \geq t) - \frac{\int_t^\tau \exp\{-\Lambda(s) \exp(x_i/\beta)ds\}}{\int_0^\tau \exp\{-\Lambda(s) \exp(x_i/\beta)ds\}} \right\}. \end{aligned}$$

The basic estimating equation for the baseline hazard is

$$\sum_{i=1}^n \{\delta_i dN_i(t) - \lambda(t) \exp(x_i/\beta) R_i(t, \beta) dt\} = 0.$$

The estimating equation for β is

$$\sum_{i=1}^n \delta_i \left[x_i - \frac{\sum_{j=1}^n x_j \exp(x_j/\beta) R_j(t_i, \beta)}{\sum_{j=1}^n \exp(x_j/\beta) R_j(t_i, \beta)} \right] = 0.$$

From above self consistent estimating equations we can solve $\lambda(t)$ and then iteratively update it. Unfortunately this approach may lead to some negative values of $\lambda(t)$ numerically. For an initial value $\lambda_0(t)$, instead we can update it by

$$\lambda(t) = \frac{\sum_{i=1}^n \delta_i dN_i(t) + \lambda_0(t) \exp\{-\Lambda_0(t)\}(1-t/\tau)/\mu}{\sum_{i=1}^n I(y_i > t) + \int_t^\tau f_0(s)(1-s/\tau)ds/\mu},$$

where the integral in the denominator can be written as

$$\begin{aligned} -\int_t^\tau (1-s/\tau) d\bar{F}(s) &= -(1-s/\tau) \bar{F}(s)|_t^\tau + \int_t^\tau \bar{F}(s) ds/\tau \\ &= (1-t/\tau) \bar{F}(t) + \int_t^\tau \bar{F}(s) ds/\tau. \end{aligned}$$

Sun et al. (2016) found this algorithm works well.

Exercise 1 Apply the missing information principle to the distribution function estimation based on doubly truncated data discussed in Chaps. 3 and 24.

Exercise 2 Apply the missing data information principle to the regression quantile process for right censored data.

Exercise 3 Extend the missing information principle to the competing risk models discussed in Kalbfleisch and Prentice (2002) and Lawless (2003).

Exercise 4 Use the missing information principle to derive the nonparametric MLE for the underlying distribution function based on window censored recurrent data discussed in Problem 9 in Chap. 3.

25.10 Case and Control Study with Prevalent Cases

We conclude this chapter with an application in case-control studies with prevalent cases.

It has been a thorny issue on the use of prevalent cases in case-control studies. It is a well known problem that the prevalent cases may cause bias. Below we discuss how to adjust the density ratio model for a valid inference if prevalent cases are used. Moreover, it is possible to combine incident cases, prevalent cases and controls for more efficient inference.

Begg and Gray (1987) discussed the prevalent cases in case and control study. Unfortunately their arguments are not easily understandable. With the preparation of the basic concepts on length biased sampling, we believe the following approach is easier for readers.

Let X be covariate associated with an incident case or a randomly selected control. Let $D = 1$ or 0 denote disease or no disease. As usual, a logistic regression is assumed for disease status

$$P(D = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

In rare diseases, it is difficult to accrue incident cases. On the other hand sampling of prevalent cases, that is subjects who had disease onset times before the sampling time and are still alive, may be an economic and effective method. In order to use the prevalent cases, one has to adjust the selection bias since only healthier cases (live up to sampling time) can be selected. As a result they do not representative of the general population cases.

Denote the density of X for an incident case as $g(x)$. Let A be the time from disease onset to sampling time. In a stationary case, A should have a uniform distribution. For an observed prevalent case

$$X|T > A, A = a \sim \frac{\bar{F}(t|x)g(x)}{\int \bar{F}(t|x)g(x)dx},$$

where $\bar{F}(t|x)$ is the survival function for an individual who had covariate X at the disease onset time. Denote the density of X in the controls as $g_0(x)$. As shown in Chap. 11 the logistic assumption is equivalent to

$$g(x)/g_0(x) = \exp(\alpha + \beta x).$$

The density of X for a prevalent case is

$$X|T > A \sim \frac{\int_0^\tau \bar{F}(t|x)g(x)dt}{\int_0^\tau \int \bar{F}(t|x)g(x)dtdx} = \frac{\mu(x)g(x)}{\int \mu(x)g(x)dx},$$

where

$$\mu(x) = \int_0^\tau \bar{F}(t|x)dt.$$

The joint density of (A, X) for a prevalent case is

$$\begin{aligned} (X, A)|T \geq A &\sim (X|T \geq A)(A|X, T \geq A) \\ &= \frac{\int \bar{F}(t|x)dtg(x)}{\int \int \bar{F}(t|x)dtg(x)dx} \frac{\bar{F}(a|x)}{\int \bar{F}(t|x)dt} \\ &= \frac{\mu(x)g(x)}{\int \mu(x)g(x)dx} \frac{\bar{F}(a|x)}{\mu(x)} \\ &=: \eta(x) \frac{\bar{F}(a|x)}{\mu(x)}. \end{aligned}$$

The density ratio of X for the prevalent cases and controls is linked by

$$\eta(x)/g_0(x) = \mu(x) \exp(\alpha + \beta x).$$

A very common assumption in survival analysis is the accelerated failure time model where

$$\bar{F}(a|x) = \bar{F}(\log a - x\gamma).$$

The corresponding mean is

$$\mu(x) = \int_0^\infty \bar{F}(\log t - x\gamma)dt = \int_{-\infty}^\infty \bar{F}(u) \exp(u + x\gamma)du,$$

or

$$\mu(x) = \exp(\gamma_0 + \gamma x), \quad \int_{-\infty}^\infty \bar{F}(u) \exp(u)du =: \exp(\gamma_0).$$

The density ratio between a prevalent case and a control is

$$\eta(x)/g_0(x) = \exp\{\alpha^* + (\beta + \gamma)x\}.$$

Therefore the densities of covariate X among the incident cases, prevalent cases and controls are linked by density ratio models. Begg and Gray (1987) and Chan (2013) discussed this problem based on controls and prevalent cases only. A more comprehensive discussion of this problem based on incident cases, prevalent cases and controls are under the way by Maziarz, Liu, Qin and Pfeiffer.

Chapter 26

Applications of the Pool Adjacent Violation Algorithm (PAVA) in Statistical Inferences

The isotonic regression solves many order restricted maximum likelihood estimation problems. This method, especially with the combination of the celebrated EM algorithm, is a powerful mathematical tool to tackle many important and difficult statistical problems. In this chapter we give a few examples to illustrate this method. In general the theoretical results are difficult for shape restricted inferences. For fundamental computational aspects, we refer the readers to the excellent books by Barlow et al. (1972) and Robertson et al. (1988). Recent results on theoretical developments can be found, among others, in Groeneboom and Jongbloed (2014). Sun (2006)'s book discusses applications of order restricted inferences in survival analysis with current status data. Our main interest is to convert some seemingly unrelated statistical problems to order restricted inferences. Hopefully, methods discussed in this chapter may shed light on new challenging statistical problems.

26.1 Pool Adjacent Violation Algorithm (PAVA)

First we briefly review the pool adjacent violation algorithm (PAVA).

In animal studies, it is very common to study the dose-toxicity relationship. Let $p(d)$ be the probability of toxicity to dose level d . It is usually reasonable to assume that $p(d)$ is a monotonic function of dose levels. Suppose there are k possible dose levels. Denote $p_i = p(d_i)$, $i = 1, 2, \dots, k$. If n_i experiments are conducted at d_i with r_i successes, the log-likelihood is

$$\ell = \sum_{i=1}^k \{r_i \log p_i + (n_i - r_i) \log(1 - p_i)\}.$$

We need to maximize this likelihood subject to the constraint

$$p_1 \leq p_2 \leq \cdots \leq p_k.$$

Ayer et al. (1955) found a solution for this problem. The general theory was developed by Barlow et al. (1972). The detailed pool adjacent violation algorithm is described as follows.

For any i , first one can perform the unconstrained maximum likelihood estimation (MLE) of p_i , denoted as $\hat{p}_i = r_i/n_i$. It would be desirable if \hat{p}_i satisfy the monotonic property. On the other hand, if any two adjacent \hat{p}_i is in the wrong order, i.e., $r_i/n_i > r_{i+1}/n_{i+1}$, then at the doses d_i and d_{i+1} one can replace each of the unconstrained MLEs \hat{p}_i and \hat{p}_{i+1} by $(r_i + r_{i+1})/(n_i + n_{i+1})$, i.e., pooling doses d_i and d_{i+1} together and recalculating the MLE. These two doses may now be considered as a block, and a number of such blocks may need to be formed. After the process of block formation is completed, it may be that the adjacent block, or block and dose proportions are found to be out of order. In such a case the relevant blocks are pooled to form larger blocks if necessary. Continue this process until the estimated probabilities satisfy the monotonic increasing order constraint.

Mathematically it can be shown that the inequality constrained maximum likelihood estimators are

$$\hat{p}_i = \max_{1 \leq u < i} \min_{i < v \leq k} \left\{ \sum_{j=u}^v r_j / \sum_{j=u}^v n_j \right\}, \quad i = 1, 2, \dots, k.$$

In the special case where $n_i = n$ for all i ,

$$\hat{p}_i = \max_{1 \leq u < i} \min_{i < v \leq k} \frac{\sum_{j=u}^v r_j}{n(v-u+1)}.$$

The algorithm has been implemented in R package, for example, Iso, or isotone.

An excellent introduction on PAVA algorithm in R program can be found from Jan de Leeuw, Kurt Hornik and Patrick Mair.

<http://CRAN.R-project.org/package=isotone>.

26.2 Applications of Pool Adjacent Violation Algorithm in Exponential Families

A fundamental problem in isotonic regression analysis is estimating the mean under a monotonic non-decreasing constraint. Given k normal distributions, $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, k$, let \bar{x}_i be the sample mean from the i -th distribution and n_i the sample size. For simplicity, we assume that σ_i^2 are known. Consider the problem of maximizing the likelihood subject to an order restriction on the means

$$\max_{\mu_i} \prod_{i=1}^n \frac{\sqrt{n_i}}{\sqrt{2\pi}\sigma_i} \exp\{-n_i(\bar{x}_i - \mu_i)^2/(2\sigma_i^2)\},$$

where $\mu_i \leq \mu_j$, $i \leq j$.

A more general problem is the order restricted nonparametric regression, where prior knowledge of monotonic mean response in terms of covariate x is available. Denote

$$Y_i = \mu(x_i) + \varepsilon_i,$$

where $\mu_1 = \mu_1(x_1) \leq \mu_2 = \mu(x_2) \leq \dots \leq \mu_n = \mu(x_n)$, if $x_1 \leq x_2 \leq \dots \leq x_n$.

Solving above two problems is equivalent to minimizing

$$\sum_{i=1}^n (y_i - \mu_i)^2 w_i$$

subject to the constraint $\mu_1 \leq \mu_2 \leq \dots \leq \mu_n$, where $x_1 \leq x_2 \leq \dots \leq x_n$ are given numbers and $w_1, \dots, w_n \geq 0$ are nonnegative weights. In the first example $w_i = n_i/\sigma_i^2$, $i = 1, 2, \dots, n$, and in the second example, $w_i = 1$, $i = 1, 2, \dots, n$. The solution is

$$\hat{\mu}_i = \max_{s \leq t} \min_{t \geq i} Av(\{s, \dots, t\}), \quad (26.2.1)$$

where

$$Av(\{s, \dots, t\}) = \frac{\sum_{i=s}^t y_i w_i}{\sum_{i=s}^t w_i}.$$

Proof can be found, for example, in Barlow et al. (1972) and Robertson et al. (1988).

As an alternative, the inequality constrained Lagrangian multiplier method can be used. Define

$$H = \sum_{i=1}^n (y_i - \mu_i)^2 w_i + \sum_{i=1}^{n-1} \lambda_i (\mu_{i+1} - \mu_i).$$

Using Kuhn–Tucker conditions (for example, Boyd and Vandenberghe 2004), we can show that the solution must be a step function, having blocks of equal μ_i 's.

Exercise Use the Lagrangian multiplier method to show the solution of the monotonic constrained minimization problem,

$$\min_{\mu_1 \leq \mu_2 \leq \dots \leq \mu_n} \sum_{i=1}^n w_i (y_i - \mu_i)^2,$$

is given in (26.2.1).

Order Restricted Inference in Exponential Families

Robertson et al. (1988) discussed the monotonic constrained inference in a very general exponential family.

Suppose Y_{ij} , $j = 1, 2, \dots, n_i$ ($i = 1, 2, \dots, n$) are independent samples from

$$f(y; \theta_i, \phi_i) = \exp\{p_1(\theta_i)p_2(\phi_i)K(y; \phi_i) + S(y; \phi_i) + q(\theta_i, \phi_i)\}, \quad (26.2.2)$$

where it is assumed that

$$\frac{\partial q(\theta_i, \phi_i)}{\partial \theta_i} = -\theta_i \frac{\partial p_1(\theta_i)}{\partial \theta_i} p_2(\phi_i),$$

and ϕ_i is known ($i = 1, 2, \dots, n$). We can show that

$$E[K(Y_i; \phi_i)] \frac{\partial p_1(\theta_i)}{\partial \theta_i} p_2(\phi_i) = \theta_i \frac{\partial p_1(\theta_i)}{\partial \theta_i} p_2(\phi_i).$$

As a consequence

$$E[K(Y_i; \phi_i)] = \theta_i, \quad \text{var}[K(Y_i; \phi_i)] = \left[\frac{\partial p_1(\theta_i)}{\partial \theta_i} p_2(\phi_i) \right]^{-1}.$$

The log-likelihood is

$$\ell = \sum_{i=1}^n p_1(\theta_i) p_2(\phi_i) n_i \hat{\theta}_i + n_i q(\theta_i, \phi_i) + C,$$

where $\hat{\theta}_i = n_i^{-1} \sum_{j=1}^{n_i} K(y_{ij}, \phi_i)$, and C is a constant (independent of θ_i). Clearly the unrestricted MLE of θ_i is $\hat{\theta}_i$.

Theorem 26.1 *The monotonic constrained MLE is a isotonic regression of $\hat{\theta}_i$ with weights $w(x_i) = n_i p_2(\phi_i)$, i.e.,*

$$\operatorname{argmin} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 n_i p_2(\phi_i).$$

Example The gamma extremum problem.

$$f(y, \theta, \tau) = \frac{y^{\tau-1} \exp(-y/\theta)}{\theta^\tau \Gamma(\tau)}, \quad y > 0$$

where θ, τ are shape and scale parameters, respectively. Suppose

$$Y_{ij} \sim f(y, \theta_i, \tau_i), \quad j = 1, 2, \dots, n_i; \quad i = 1, 2, \dots, n,$$

where τ_1, \dots, τ_n are known. The log-likelihood is

$$\ell = - \sum_{i=1}^n n_i \tau_i \left(\frac{\bar{y}_i}{\tau_i} \frac{1}{\theta_i} + \log \theta_i \right) + \text{const.}, \quad \bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}.$$

We are interested in maximizing this log-likelihood subject to the constraint $\theta_1 \leq \theta_2 \leq \dots \leq \theta_n$.

The unrestricted maximum likelihood estimate for θ_i is given by $\hat{\theta}_i = \sum_{j=1}^{n_i} Y_{ij}/\tau_i$. The restricted maximum likelihood estimate is the isotonic regression of $\hat{\theta}_i$ with weights $w_i = n_i \tau_i$. Equivalently, we need to minimize

$$\sum_{i=1}^n (\bar{y}_i/\tau_i - \theta_i)^2 w_i, \quad w_i = n_i \tau_i$$

In this example if the gamma density is parametrized by

$$f(y, \theta, \tau) = \frac{y^{\tau-1} \theta^\tau \exp(-y\theta)}{\Gamma(\tau)}, \quad y > 0,$$

then the isotonic regression of $\hat{\theta}_i = \tau_i/(n_i^{-1} \sum_{j=1}^{n_i} y_{ij})$ with weights $w_i = n_i \tau_i$ does not work. For details, see Robertson et al. (1988).

Exercise Discuss the Poisson extreme problem and geometric extreme problem.

Order Restricted Inference for Multinomial Distribution

Suppose we would like to maximize a product

$$L(p) = p_1^{y_1} \cdots p_n^{y_n}$$

subject to the constraints

$$p_1 \geq p_2 \geq \cdots \geq p_n \geq 0, \quad \sum_{i=1}^n p_i w_i = c, \quad w_i \geq 0, \quad (26.2.3)$$

where c is a given value, typically $c = 1$ for the multinomial parameter problem.

Denote

$$z_i = \frac{y_i}{y_+} \frac{c}{w_i}, \quad y_+ = \sum_{j=1}^n y_j.$$

Theorem 26.2 *The solution for maximizing $L(p)$ subject to the constraints (26.2.3) is the one that minimizes*

$$\sum_{i=1}^n (z_i - p_i)^2 w_i$$

subject to the constraint $p_1 \geq p_2 \geq \dots \geq p_n$.

Note the equality constraint can be written as

$$\sum_{i=1}^n p_i w_i / c = 1.$$

Treating $w_i/c, i = 1, 2, \dots, n$ as new weights, we only need to prove the case for $c = 1$.

Denote

$$\ell(p) = \log L(p) = \sum_{i=1}^n y_i \log p_i.$$

Let

$$p_i = \frac{r_i}{\sum_{j=1}^n r_j w_i}, i = 1, 2, \dots, n,$$

then $\sum_{j=1}^n p_i w_i = 1$.

$$\begin{aligned} \ell(r) &= \sum_{i=1}^n y_i \log(r_i) - \left(\sum_{i=1}^n y_i \right) \log \left(\sum_{j=1}^n r_j w_j \right) \\ &= \sum_{i=1}^n y_i \log(r_i) - y_+ \log \left(\sum_{j=1}^n r_j w_j \right) \\ &= \left[\sum_{i=1}^n y_i \log(r_i) - y_+ \left(\sum_{j=1}^n r_i w_i \right) \right] + \left[y_+ \left(\sum_{j=1}^n r_i w_i \right) - y_+ \log \left(\sum_{j=1}^n r_j w_j \right) \right] \\ &= \ell_1(r) + \ell_2(r). \end{aligned}$$

(a) Subject to $r_1 \geq r_2 \geq \dots \geq r_n$, first we show that $\ell_1(r)$ achieves maximum only if $\sum_{i=1}^n r_i w_i = 1$.

In fact if $\ell_1(r^*) \geq \ell_1(r)$ for any r , then for $\Delta = \sum_{i=1}^n r_i^* w_i$,

$$\begin{aligned} \ell_1(r^*/\Delta) &= \sum_{i=1}^n y_i \log(r_i^*/\Delta) - y_+ \left(\sum_{i=1}^n r_i^* w_i / \Delta \right) \\ &= \sum_{i=1}^n y_i \log(r_i^*) - y_+ \log(\Delta) - y_+ \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n y_i \log(r_i^*) - y_+ \Delta + y_+ \Delta - y_+ \log(\Delta) - y_+ \\
&= \ell_1(r^*) + \ell_2(r^*) - y_+.
\end{aligned}$$

Following the fact that $t > 0$, we have $t - \log t \geq 1$ and

$$\ell_1(r^*/\Delta) \geq \ell_1(r^*) \geq \ell_1(r).$$

Thus we showed (a).

On the other hand if $\sum_{i=1}^n r_i w_i = 1$, then

$$\ell_1(r) = \sum_{i=1}^n y_i \log(r_i) - y_+ (\sum_{i=1}^n r_i w_i) = \sum_{i=1}^n y_i \log(r_i) - y_+ = \ell(r) - y_+.$$

Therefore we have proved that maximizing $\ell(r)$ with respect to r subject to the monotonic constraint and the equality constraint $\sum_{i=1}^n r_i w_i = 1$ is equivalent to maximizing $\ell_1(r)$ with respect to r subject to the monotonic constraint only.

Note that $\ell_1(r)$ can be treated as the log-likelihood from n independent gamma densities

$$f_i(t, y_i, r_i) = \frac{t^{y_i-1} r_i^{y_i} \exp(-tr_i)}{\Gamma(y_i)}, \quad t > 0,$$

with observed values $t_i = y_+ w_i, i = 1, 2, \dots, n$. As shown before PAVA does not work by directly maximizing $\ell_1(r)$ with respect to $r_1 \geq r_2 \geq \dots \geq r_n$. Let $\lambda_i = 1/r_i, i = 1, 2, \dots, n$. We can use PAVA to maximize $\ell_1(\lambda)$ subject to the constraint $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Using Theorem 26.1, equivalently we need to minimize

$$\sum_{i=1}^n y_i (y_+ w_i / y_i - \lambda_i)^2.$$

Denote the solution as $\hat{\lambda}_i, i = 1, 2, \dots, n$. Finally

$$\hat{r}_1 = 1/\hat{\lambda}_1 \geq \hat{r}_2 = 1/\hat{\lambda}_2 \geq \dots \geq \hat{r}_n = 1/\hat{\lambda}_n.$$

It is not hard to show that $\hat{r}_i, i = 1, 2, \dots, n$ also are the solution of

$$\min_r \sum_{i=1}^n \{y_i/(y_+ w_i) - r_i\}^2 w_i,$$

subject to $r_1 \geq r_2 \geq \dots \geq r_n$.

Mixture of Gamma Distributions

For simplicity, consider a gamma mixture model with two components

$$f(y, \theta, \tau) = \lambda \frac{y^{\tau_1-1} \exp(-y/\theta_1)}{\theta_1^{\tau_1} \Gamma(\tau_1)} + (1-\lambda) \frac{y^{\tau_2-1} \exp(-y/\theta_2)}{\theta_2^{\tau_2} \Gamma(\tau_2)}, \quad y > 0,$$

where the proportion parameter λ is known. Clearly this is not an exponential family. Direct use of the pool adjacent violation algorithm does not work. Fortunately, we can use a combination of EM algorithm and PAVA to solve this problem. In fact, we will show later that this combination is a very powerful tool to solve many parametric, semiparametric, and nonparametric problems.

An Application of the Isotonic Regression Problem

To answer the question of whether low-volume hospitals have higher mortality rates than high-volume hospitals have, one way is to model patient mortality outcome as a function of hospital volume, adjusting for some important patient-level predictors. Another approach is to directly test the association between the hospital's standardized mortality ratio (number of deaths divided by number of patients) and hospital volume. Betensky et al. (2006) studied the correlation of the bivariate data, $(Y_i/n_i, n_i)$, where Y_i is a binomial(n_i, p_i). They found that the popular Kendall's tau is an invalid test for the bivariate association. In contrast, Pearson's correlation coefficient test, used mainly for normally distributed data, can be used to detect the correlation in the presence of nonlinear dependence. In this section, we will study Pearson's test thoroughly and give our insights on its rationale. As an alternative, we recommend the use of the monotonic likelihood ratio test.

Denote N independent binomial observations as $Y_i \sim \text{binomial}(n_i, p_i)$, $i = 1, 2, \dots, N$. The null hypothesis is that the probability p_i is independent of n_i , i.e.,

$$H_0 : p_1 = p_2 = \dots = p_N = p.$$

Since Y_i/n_i is an unbiased estimator of p_i , we may test that there is no association between Y_i/n_i and n_i . Let $S_i = Y_i/n_i$ and $T_i = n_i$, $i = 1, 2, \dots, N$. Pearson's correlation coefficient test, Spearman's test and Kendall's tau test are the most popular test statistics for testing the association between random variables. In order to use them, a basic assumption is that (S_i, T_i) , $i = 1, 2, \dots, N$ are independent and identically distributed. However, this basic assumption is violated for S_i, n_i , $i = 1, 2, \dots, N$.

(1) Clearly $(Y_i/n_i, n_i)$'s are not identically distributed since volumes between large and small hospitals are different.

(2) Clearly Y_i/n_i and n_i are not independent of each other since

$$P(Y_i/n_i > 0|n_i) = P(Y_i > 0|n_i) = 1 - (1 - p_i)^{n_i},$$

which depends on n_i .

Pearson's correlation coefficient test is mainly designed to test the association between normal random variables. Spearman's and Kendall's tau tests, on the other hand, are more powerful for detecting the nonlinear dependence. Intuitively, it seems that Spearman's and Kendall's tau tests are more appropriate for the case considered here than Pearson's correlation coefficient test since n_i 's are not normally distributed data. Note that S_i may behave like normal variables since they are the averages of Bernoulli outcomes. However, Betensky et al. (2006) observed that Pearson's test is more appropriate in this case. Kendall's tau cannot be used at all since it has inflated type I error rate. Next we examine the three tests in details.

Pearson's correlation coefficient is defined as

$$\rho_P = \frac{\sum_{i=1}^N (S_i - \bar{S})(T_i - \bar{T})}{\sqrt{\sum_{i=1}^N (S_i - \bar{S})^2 \sum_{i=1}^N (T_i - \bar{T})^2}} = \frac{\sum_{i=1}^N S_i T_i - N \bar{S} \bar{T}}{\sqrt{\sum_{i=1}^N (S_i - \bar{S})^2 \sum_{i=1}^N (T_i - \bar{T})^2}}.$$

Replacing S_i and T_i by Y_i/n_i and n_i , respectively, we may define

$$\eta = \sum_{i=1}^N S_i T_i - N \bar{S} \bar{T} = \sum_{i=1}^N Y_i - N \left[N^{-1} \sum_{i=1}^N Y_i/n_i \right] \left[N^{-1} \sum_{i=1}^N n_i \right].$$

Let $C = N^{-1} \sum_{i=1}^N n_i$. Taking expectation with respect to η , we have

$$E(\eta) = \sum_{i=1}^N n_i p_i - N \left[\frac{1}{N} \sum_{i=1}^N p_i \right] C = \sum_{i=1}^N (n_i - C) p_i.$$

Clearly $E(\eta) = 0$ if $p_1 = p_2 = \dots = p_N$. However, in general the reverse relationship does not necessarily hold. Nevertheless, we have the following Lemma.

Lemma 26.3 Suppose $n_1 \geq n_2 \geq \dots \geq n_N$. The corresponding p_i 's satisfy $p_1 \leq p_2 \leq \dots \leq p_N$. Then $E(\eta) = 0$ if and only if $p_1 = p_2 = \dots = p_N = p_0$.

Proof If $p_1 = p_2 = \dots = p_N = p_0$, then

$$E(\eta) = \sum_{i=1}^N (n_i - C) p_0 = 0$$

since $C = N^{-1} \sum_{i=1}^N n_i$.

On the other hand, if $E(\eta) = 0$, we need to prove that $p_1 = p_2 \dots = p_N$. Without loss of generality, we can assume that there is an integer k between 1 and N such that

$$n_1 \geq n_2 \geq \dots \geq n_{k-1} \geq C \geq n_k \geq \dots \geq n_N.$$

Since $\sum_{i=1}^N (n_i - C) = 0$,

$$\sum_{i=1}^{k-1} (n_i - C) = - \sum_{i=k}^N (n_i - C).$$

Using the non-decreasing assumption on p_i 's,

$$\begin{aligned} 0 &= E(\eta) = \sum_{i=1}^N (n_i - C)p_i \\ &= \sum_{i=1}^{k-1} (n_i - C)p_i + \sum_{i=k}^N (n_i - C)p_i \\ &\leq \sum_{i=1}^{k-1} (n_i - C)p_{k-1} + \sum_{i=k}^N (n_i - C)p_i \\ &= - \sum_{i=k}^N (n_i - C)p_{k-1} + \sum_{i=k}^N (n_i - C)p_i \\ &= \sum_{i=k}^N (n_i - C)(p_i - p_{k-1}) \leq 0. \end{aligned}$$

This implies

$$p_{k-1} = p_k = \cdots = p_N.$$

Again

$$\begin{aligned} 0 &= E(\eta) = \sum_{i=1}^N (n_i - C)p_i \\ &\leq \sum_{i=1}^{k-1} (n_i - C)p_i + \sum_{i=k}^N (n_i - C)p_{k-1} \\ &= \sum_{i=1}^{k-1} (n_i - C)p_i - \sum_{i=1}^{k-1} (n_i - C)p_{k-1} \\ &= \sum_{i=1}^{k-1} (n_i - C)(p_i - p_{k-1}) \leq 0. \end{aligned}$$

This implies

$$p_1 = p_2 = \cdots = p_{k-1}.$$

This completes the proof.

Now we consider Kendall's tau statistic. It is defined as

$$\tau = \tau_c - \tau_d, \quad \tau_c = P\{(S_i - S_j)(T_i - T_j) > 0\}, \quad \tau_d = P\{(S_i - S_j)(T_i - T_j) < 0\}.$$

The basic assumption in Kendall's tau statistic is that (S_i, T_i) 's are independent and identically distributed data. For fixed n_i, n_j , however,

$$P\{(S_i - S_j)(T_i - T_j) > 0\} = P\{(n_i - n_j)(S_i - S_j) > 0\}$$

and

$$\begin{aligned} P(Y_i/n_i > Y_j/n_j) &= P(Y_j < (n_i/n_j)Y_i) = E\{\text{pbinom}([n_i/n_j]Y_i - 1], n_j, p_j)\} \\ &= \sum_{k=0}^{n_i} \text{pbinom}([(n_i/n_j)k - 1], n_j, p_j) \text{dbinom}(k, n_i, p_i), \end{aligned}$$

where pnorm and dnorm are binomial cumulative and density functions, respectively, and $[x]$ is the integer part of x . Clearly the indexes i and j are not exchangeable. As a result this test may not have the correct type one error.

Similar arguments also apply to Spearman's correlation coefficient test.

Monotone Likelihood Ratio Test

As an alternative, we may employ the likelihood ratio statistic to test $H_0 : p_1 = p_2 = \dots = p_N$. The log-likelihood based on $Y_i \sim \text{binomial}(n_i, p_i)$, $i = 1, 2, \dots, N$ is

$$\ell = \sum_{i=1}^N [Y_i \log p_i + (n_i - Y_i) \log(1 - p_i)].$$

Without loss of generality, we can assume that

$$n_1 \geq n_2 \geq \dots \geq n_N.$$

As expected, higher volume hospitals do not have larger mortality than lower volume hospitals have. The alternative hypothesis is

$$H_A : p_1 \leq p_2 \leq \dots \leq p_N.$$

The likelihood ratio statistic is defined as

$$R = 2[\max_{p_1 \leq p_2 \leq \dots \leq p_N} \ell(p_1, p_2, \dots, p_N) - \max_p \ell(p, p, \dots, p)].$$

The maximization can be achieved by using the pool-adjacent-violation algorithm. Under the alternative, we can estimate p_i by

$$\hat{p}_i = \max_{1 \leq u \leq i} \min_{i \leq \nu \leq N} \left\{ \sum_{j=u}^{\nu} Y_j / \sum_{j=u}^{\nu} n_j \right\}, \quad i = 1, 2, \dots, N.$$

Under H_0 , p can be estimated by

$$\tilde{p} = \sum_{i=1}^N Y_i / \sum_{i=1}^N n_i.$$

To find the p -value, we can generate N independent parametric bootstrap samples

$$Y_1^b \sim B(n_1, \tilde{p}), \dots, Y_N^b \sim B(n_N, \tilde{p}).$$

Then R can be re-calculated through the parametric bootstrap data. Denoted bootstrap R values as

$$R^1, \dots, R^B.$$

The p -value can be calculated by comparing R calculated from the original data with R^b , $b = 1, 2, \dots, B$.

We conducted a small simulation study. In this study, $N = 9$,

$$(n_1, \dots, n_9) = (100, 50, 50, 20, 20, 10, 10, 5, 5).$$

Under H_0 : $\mathbf{p} = (0.056, \dots, 0.056)$. Under alternative H_A ,

$$\mathbf{p} = (0.02, 0.03, 0.03, 0.05, 0.05, 0.15, 0.15, 0.3, 0.3).$$

The significant level is set at 0.05. 1000 simulations were repeated. In order to find the p -values for the likelihood ratio test, 1000 binomial replicates were drawn in each simulation. Under H_0 , the likelihood ratio test rejected 42 times, the Pearson's correlation coefficient test rejected 71 times, the Spearman's correlation test rejected 170 times, and the Kendall's τ test rejected 181 times. This result shows that Spearman's correlation test and Kendall's τ do not have a correct type I error rate.

Under the alternative, the likelihood ratio rejected 908 times, Pearson's correlation coefficient test rejected 84 times, Spearman's correlation test rejected 462 times and Kendall's τ test rejected 522 times. This shows that the likelihood ratio test is the most powerful one. The Pearson's correlation coefficient test has the correct type one error but has low power.

26.3 Estimating Monotonic Decreasing Density and Hazard Functions

We start from the maximum likelihood estimation for a monotonic decreasing density, i.e., the Grenander (1956a,b) estimator. Recall that this problem was discussed in last chapter using backward time and EM algorithm. Grenander (1956a,b) directly maximized the nonparametric likelihood subject to the monotonic density constraint. Let the observed data be

$$X_1, \dots, X_n \sim f(x).$$

The order statistics are denoted as

$$0 < X_{(1)} \leq \dots \leq X_{(n)}.$$

We need to maximize the likelihood

$$L = \sum_{i=1}^n \log f(X_{(i)})$$

subject to the constraints

$$f_1 \geq f_2 \geq \dots \geq f_n,$$

and

$$X_{(1)}f_1 + (X_{(2)} - X_{(1)})f_2 + \dots + (X_{(n)} - X_{(n-1)})f_n = 1,$$

where $f_i = f(X_{(i)})$, $i = 1, 2, \dots, n$ and the equality constraint comes from the fact $\int f(x)dx = 1$. Without imposing the monotonic constraint, the naive estimate is

$$f_n(x) = \begin{cases} 0 & \text{for } x < 0 \\ n^{-1}(X_{(i)} - X_{(i-1)})^{-1}, & \text{for } X_{(i-1)} < x \leq X_{(i)} \\ 0 & \text{for } X > X_{(n)}. \end{cases}$$

Denote

$$w_i = (X_{(i)} - X_{(i-1)}), \quad g_i = \frac{1}{nw_i}.$$

This is equivalent to maximizing

$$\sum_{i=1}^n g_i w_i \log f_i$$

subject to the constraints

$$f_1 \geq f_2 \geq \dots \geq f_n,$$

and

$$\sum_{i=1}^n [g_i - f_i]w_i = 0.$$

Using Theorem 26.2, we can use PAVA to find

$$\hat{f}_n(x) = \max_{s \leq i-1} \min_{t \geq i} \left[\frac{t-s}{n(X_{(t)} - X_{(s)})} \right].$$

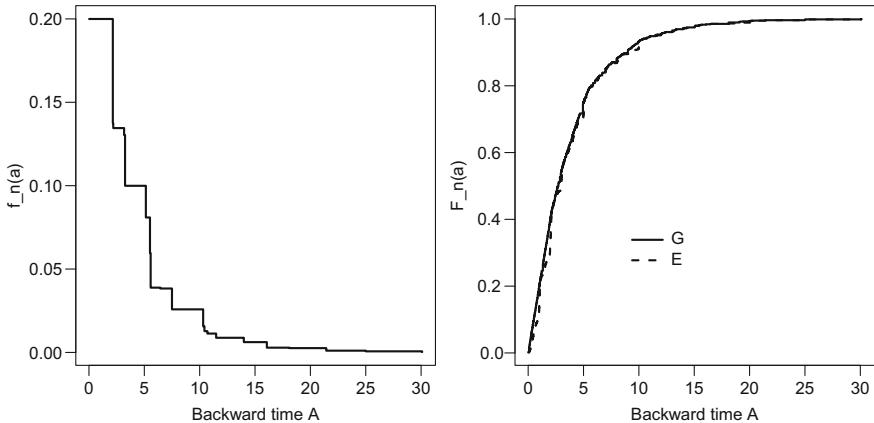


Fig. 26.1 Grenander density estimator (left panel) and cumulative empirical distribution (E) versus shaped restricted distribution (G) (right panel)

Example It is known that the backward time in the renewal process has a monotonic decreasing density, in the graph below, we plot Grenander's estimator based on the backward time from the Canadian dementia data. Vardi (1989) used the EM algorithm to derive this estimator (Fig. 26.1).

Note that the cumulative distribution $F(t) = \int_0^t f(x)dx$ is concave since f is monotonic non-increasing. The conventional empirical distribution $F_n(t) = n^{-1} \sum_{i=1}^n I(x_i \leq t)$ is not necessarily concave. Naturally, the shape restricted nonparametric MLE is $\hat{F}(t)$ with the shape of a tight string tied to the origin and wraps around $F_n(t)$ from above. In fact the shape restricted nonparametric density estimator is the left-continuous slope of $\hat{F}(t)$. More details can be found in Robertson et al. (1988).

Large Sample Property of Grenander's Estimator

In order to show consistency of the nonparametric or semiparametric maximum likelihood estimator, we need the general theory on empirical processes. To appreciate this difficult problem, we discuss as follows briefly:

The log-likelihood is

$$n^{-1}\ell = n^{-1} \sum_{i=1}^n \log f(x_i) = \int \log f(x) dF_n(x),$$

where $F_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$ be the empirical distribution. Let \hat{f}_n be the restricted MLE. Then for any monotonic density, including the true density

$$\ell(f_0) \leq \ell(\hat{f}_n),$$

or equivalently

$$0 \leq \int \log\{\hat{f}_n(x)/f_0(x)\}dF_n(x) = \int \log\{\hat{f}_n(x)/f_0(x)\}\{dF_n(x) - dF_0(x)\} \\ + \int \log\{\hat{f}_n(x)/f_0(x)\}dF_0(x).$$

Using the fact $\log x = 2 \log \sqrt{x} \leq 2(\sqrt{x} - 1)$ for $x > 0$ and $\int f_0(x)dx = \int \hat{f}_n(x)dx = 1$,

$$\int \{\sqrt{f_0(x)} - \sqrt{\hat{f}(x)}\}^2 dx \leq 0.5 \int \log\{\hat{f}_n(x)/f_0(x)\}\{dF_n(x) - dF_0(x)\}.$$

The Hellinger distance between f_0 and \hat{f} converges to 0 if we can show uniform convergence

$$\sup_{\phi \in \Psi} \int \phi(x)\{dF_n(x) - dF_0(x)\},$$

where $\Psi = \{\phi | \phi = \log(f(x)/f_0(x))\}$ for all possible monotonic decreasing density. This involves the entropy bound calculation. We refer readers to advanced books on empirical processes, such as Pollard (1984), van der Vaart and Wellner (1996), Kosorok (2008a) and Groeneboom and Jongbloed (2014).

Another method to show consistency is based on the directional derivative. If \hat{f} is a monotonic density and maximizes the log-likelihood

$$\ell = n^{-1} \sum_{i=1}^n \log f(x_i) = \int \log f(x)dF_n(x), \quad F_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x),$$

then

$$\ell(\hat{f}) - \ell((1 - \varepsilon)\hat{f} + \varepsilon f) \geq 0, \quad \varepsilon \geq 0$$

for any monotonic density f . Note that

$$\lim_{\varepsilon \rightarrow 0+} \frac{\ell(\hat{f}) - \ell((1 - \varepsilon)\hat{f} + \varepsilon f)}{\varepsilon} \geq 0,$$

or

$$\int (f - \hat{f})/\hat{f} dF_n(x) \leq 0.$$

Specifically if we choose f by the true density f_0 , as $n \rightarrow \infty$,

$$\int f_0^2(x)/f^*(x)dx \leq 1,$$

where f^* is the limit of any subsequence of \hat{f} . Finally

$$\begin{aligned} 0 \leq \int \frac{(f_0 - f^*)^2}{f^*} dx &= \int \frac{f_0^2 - 2f_0 f^* + f^* f^*}{f^*} dx \\ &= \int \frac{f_0^2}{f^*} dx - 2 + 1 \leq 0, \end{aligned}$$

where we have used the fact that both f_0 and f^* are densities. Therefore \hat{f} is a consistent estimator of f_0 .

Prakasa Rao (1969) obtained the asymptotic distribution for the Grenander estimator $\hat{f}_n(t)$. In particular, if $f'(t_0) \neq 0$, $0 < t_0 < \infty$, then

$$n^{1/3}\{f_n(t_0) - f(t_0)\} \rightarrow |0.5f(t_0)f'(t_0)|^{1/3}C,$$

where $C = \text{armax}_{s \in R}\{W(s) - s^2\}$ and $W(s)$ is a two-sided standard Brownian motion in R^1 with $W(0) = 0$.

Kosorok (2008a,b) found that the Grenander estimator using bootstrap samples from the empirical distribution function F_n or its least concave majorant \hat{F}_n , does not have any weak limit in probability. The m out of n ($m \ll n$) bootstrap estimation of f , however, is shown to be consistent.

Estimating a Monotonic Density Under Selection Bias Sampling

Suppose

$$X_1, \dots, X_n \text{ i.i.d. } \sim \frac{w(x)f(x)}{\int w(x)f(x)dx}, \quad x \geq 0,$$

where $f(x)$ is a monotonic decreasing density.

If $w(x)$ is not a constant, the likelihood is invariant when f is replaced by f/c . We do not need to impose the constraint $\int f(x)dx = 1$. Denote the order statistics as

$$0 = x_{(0)} \leq x_{(1)} \leq \dots \leq x_{(n)}.$$

Let $f_i = f(x_{(i)})$, $i = 1, 2, \dots, n$. Note that

$$\begin{aligned} \int w(x)f(x)dx &= \int_0^{x_{(1)}} w(x)f(x)dx + \int_{x_{(1)}}^{x_{(2)}} w(x)f(x)dx + \dots + \int_{x_{(n-1)}}^{x_{(n)}} w(x)f(x)dx \\ &\geq f(x_{(1)}) \int_0^{x_{(1)}} w(x)dx + f(x_{(2)}) \int_{x_{(1)}}^{x_{(2)}} w(x)dx + \dots + f(x_{(n)}) \int_{x_{(n-1)}}^{x_{(n)}} w(x)dx \\ &:= \sum_{i=1}^n f_i a_i, \quad a_i = \int_{x_{(i-1)}}^{x_{(i)}} w(x)dx \end{aligned}$$

The log-likelihood satisfies

$$\ell \leq \sum_{i=1}^n \log[f_i / \sum_{i=1}^n f_i a_i] + \text{constant} = \sum_{i=1}^n \log g_i + \text{constant},$$

where $g_i = f_i / \sum_{j=1}^n f_j a_j$, $i = 1, 2, \dots, n$, then

$$\sum_{i=1}^n g_i a_i = 1, \quad g_1 \geq g_2 \geq \dots \geq g_n.$$

Using Theorem 26.2, we can use PAVA to estimate g_i , $i = 1, 2, \dots, n$.

Exercise Consider a two-sample biased sample problem,

$$X_1, \dots, X_m \sim i.i.d. f(x), \quad Y_1, \dots, Y_n \sim i.i.d. g(y) = yf(y) / \int_0^\infty yf(y)dy.$$

Find the maximum likelihood estimation of f under a monotonic constraint.

Maximum Likelihood Estimation of a Monotonic Hazard Model

Suppose X_1, \dots, X_n *i.i.d.* $\sim f(x)$, the corresponding hazard function $\lambda(x)$ is assumed to be monotonic non-decreasing. Without loss of generality, denote $t_1 < t_2 < \dots < t_h$ ($h \leq n$) as the ordered unique observations. Denote $\lambda_i = \lambda(t_i)$, $i = 1, 2, \dots, h$. The log-likelihood is

$$\ell = \sum_{i=1}^h \log dF(x_i) = \sum_{i=1}^h [d_i \log \lambda_i + (n_i - d_i) \log(1 - \lambda_i)],$$

where $d_i = \sum_{j=1}^n I(x_j = t_i)$, $n_i = \sum_{j=1}^n I(x_j \geq t_i)$, $i = 1, 2, \dots, h$. The unrestricted nonparametric MLE of λ_i 's are

$$\hat{\lambda}_i = d_i / n_i, \quad i = 1, 2, \dots, h.$$

Under the monotonic restriction

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_h,$$

we have a series of binomial likelihoods. Therefore the PAVA algorithm can be applied to find the constrained MLE.

Exercise Derive the maximum likelihood estimation of a monotonic non-increasing hazard function for length biased sampling data.

Inference Under a Two-Sample Monotonic Likelihood Ratio Assumption

As demonstrated in Chap. 1, among different stochastic orderings, the likelihood ratio ordering is the strongest one. Consider a two sample problem

$$X_1, \dots, X_m \sim f(x), \quad Y_1, \dots, Y_n \sim g(x).$$

The likelihood ratio ordering assumes that $f(x)/g(x)$ is monotonic non-decreasing, or,

$$f(x) = \frac{\bar{H}(x)dG(x)}{\int \bar{H}(x)dG(x)},$$

where $\bar{H}(x)$ is an unknown survival function. We are interested in estimating G and H non-parametrically. Dykstra et al. (1995) used the following approach. Let

$$t_1 < t_2 < \dots < t_h, \quad h \leq m + n$$

be the distinct pooled data, and

$$p_i = dF(t_i), \quad dG(t_i) = q_i, \quad i = 1, 2, \dots, h, \quad m_i = \sum_{k=1}^m I(X_k = t_i), \quad n_i = \sum_{k=1}^n I(Y_k = t_i).$$

Also denote

$$\theta_i = mp_i / (mp_i + nq_i), \quad \phi_i = mp_i + nq_i.$$

Then

$$p_i = \theta_i \phi_i / m, \quad q_i = \phi_i (1 - \theta_i) / n, \quad i = 1, 2, \dots, h.$$

The likelihood can be written as

$$L = \prod_{i=1}^h p_i^{m_i} q_i^{n_i} = \left(\frac{1}{m}\right)^m \left(\frac{1}{n}\right)^n \prod_{i=1}^h \theta_i^{m_i} (1 - \theta_i)^{n_i} \prod_{i=1}^h \phi_i^{m_i + n_i},$$

where

$$\theta_i = \frac{1}{1 + \rho q_i / p_i}, \quad \rho = n/m.$$

Since q_i/p_i is monotonic non-increasing, θ_i 's are monotonic non-decreasing.

$$\sum_{i=1}^h \phi_i = m + n, \quad \sum_{i=1}^h \theta_i \phi_i = m.$$

After profiling out the ϕ_i 's (subject to the constraint $\sum_{i=1}^h \phi_i = m + n$ only),

$$\hat{\phi}_i = m_i + n_i.$$

PAVA can be applied to maximize $\prod_{i=1}^h \theta_i^{m_i} (1 - \theta_i)^{n_i}$ subject to constraints $\theta_1 \leq \theta_2 \leq \dots \leq \theta_h$. Moreover, the order restricted nonparametric likelihood estimate $\hat{\theta}_i$

satisfies $\sum_{i=1}^h \hat{\theta}_i(m_i + n_i) = \sum_{i=1}^h m_i = m$. (**Exercise**). As a result, the constrained ($\sum_{i=1}^h \theta_i \phi_i = m$) and the unconstrained MLE are exactly the same. This is similar to the result by Prentice and Pyke (1979) discussed in Chap. 11 about an exponential tilting assumption for the underlying density ratio model.

26.4 Cosslett's Maximum Likelihood Estimation and Related Problems

Discrete choice model and binomial regression model are very popular in economics and medical studies. The commonly used models include, the logistic regression model, Probit model or complement log-log model etc. A common feature among these models is that for a given covariate X , the discrete choice response depends on $x\beta$ monotonically. More specifically, assume a binary choice model

$$P(Y = 1|x) = P(\varepsilon < x\beta) = F(x\beta),$$

where $\varepsilon \sim F$ is an unknown distribution function. It is desirable to estimate β without assuming the form of F .

The log-likelihood is

$$\ell = \sum_{i=1}^n [Y_i \log F(x_i\beta) + (1 - Y_i) \log\{1 - F(x_i\beta)\}].$$

If β is known, this is exactly the same problem discussed in Sect. 26.1 and then F can be estimated by using the PAVA. For unknown β we need to maximize this likelihood with respect to F and β . Since the intercept and the scale of β can be absorbed by F , we can assume that there is not intercept term and the first component of β is 1.

Cosslett (1987) used the following two-step approach.

Step 1. For fixed β , sort $x_i\beta, i = 1, 2, \dots, n$

$$(x\beta)_{(1)} \leq (x\beta)_{(2)} \leq \dots \leq (x\beta)_{(n)}.$$

Maximize this likelihood with respect to F subject to the monotonic constraint. This can be accomplished by PAVA. Denote the maximizer as \hat{F} .

Step 2. Maximize the profile likelihood

$$\ell_P = \sum_{i=1}^n [Y_i \log \hat{F}((x\beta)_{(i)}) + (1 - Y_i) \log\{1 - \hat{F}((x\beta)_{(i)})\}]$$

with respect to β .

We can iterate steps 1 and 2 until convergence.

Using the classical result from Kiefer and Wolfowitz (1956), Cosslett (1987) showed consistency of the maximum likelihood estimate of β . However Huang and Wellner (1997) pointed out that the large sample results, such as the central limiting theory etc., are still open problems.

Next we discuss three related problems.

1. Use empirical likelihood to calibrate auxiliary information

Chen and Qin (2013) used a calibration method to extract auxiliary information in a partially specified linear regression model under a monotonic constraint. Suppose the response Y and covariate (x, z) are linked by

$$Y = g(x\beta) + h(z) + \varepsilon,$$

where the form of $g(x)$ is known but h is a nondecreasing unknown function. Denote the observed data as

$$(D_i, Y_i, D_i X_i, Z_i), i = 1, 2, \dots, n,$$

where (Y_i, Z_i) are available for each individual and X_i is available only if $D_i = 1$. Denote the propensity score as

$$\pi(y, z) = P(D = 1|x, y, z) = P(D = 1|y, z, \gamma).$$

Using the binomial likelihood

$$L_B = \prod_{i=1}^n \pi^{D_i}(y_i, z_i, \gamma) \{1 - \pi(y_i, z_i, \gamma)\}^{1-D_i}$$

we can estimate γ . Denote the MLE as $\hat{\gamma}$.

Without loss of generality, we assume the first n_1 observations have x_i values. Conditional on $D = 1$, the likelihood is

$$L_c = \prod_{i=1}^{n_1} P(y_i, x_i, z_i | D_i = 1) = \prod_{i=1}^{n_1} \frac{\pi(y_i, z_i, \gamma) dF(y_i, x_i, z_i)}{\int \pi(y, z, \gamma) dF(y, x, z)}.$$

Denote $p_i = dF(y_i, x_i, z_i)$, $i = 1, 2, \dots, n_1$, the log-conditional likelihood can be written as

$$\ell_c = \sum_{i=1}^{n_1} \log p_i - n_1 \log \Delta,$$

where $\Delta = \sum_{i=1}^{n_1} p_i \pi(y_i, z_i, \gamma)$. We can replace γ by $\hat{\gamma}$. As discussed in Chap. 19, we can find the maximum likelihood estimation of p_i subject to constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad \sum_{i=1}^{n_1} p_i \{\psi(y_i, z_i) - \bar{\psi}\} = 0, \quad \bar{\psi} = n^{-1} \sum_{i=1}^n \psi(y_i, z_i),$$

where the last constraint is based on the efficient consideration. The optimal choice of ψ was discussed in Chap. 19.

Denote the maximum constrained empirical likelihood estimator as $\hat{p}_i, i = 1, 2, \dots, n_1$. For fixed β we need to minimize

$$\sum_{i=1}^{n_1} \hat{p}_i \{y_i - g(x_i \beta) - h(z_i)\}^2$$

subject to the constraints

$$h(z_1) \leq h(z_2) \leq \dots \leq h(z_{n_1}), \quad z_1 \leq z_2 \leq \dots \leq z_{n_1},$$

where without loss of generality we have assumed that $z_i, i = 1, 2, \dots, n_1$ are arranged in ascending order. Finally we can search for β , such that it achieves the minimum value. More details can be found in Chen and Qin (2013).

In this problem, the calibration method is crucial for the weighted least squares estimation under monotonic constraint. However, it is not clear how to perform the order restricted minimization by using the augmented inverse weighted method discussed in Chap. 19.

2. Using PAVA to Stabilize Propensity Score in Casual Inference

As discussed in Chap. 19, the inverse probability weighted and augmented inverse probability weighted estimators may have extremely large variances in simulation studies when the propensity score is too close to 0. Replacing the logistic regression estimation of propensity score by a PAVA based estimator, recently, Qin et al. (2017a) have found that this new inverse probability weighted estimator or augmented inverse probability weighted estimator produces more stable estimators for the population mean.

3. Order Restricted Maximum Likelihood in a General Semiparametric Transformation Model

Next we present an exciting application of the order restricted maximum semiparametric likelihood estimation. This method provides an elegant solution to the well-known, but difficult, semiparametric transformation model used in econometric literature. This model is given by

$$h(y_i) = x_i \beta + \varepsilon_i,$$

where h is an unknown monotonic transformation function, and the error distribution F_ε is also unspecified. This model encompasses many well known semiparametric models. For example, if h is known but F_ε is unknown, this becomes the accelerated

failure time model discussed in Chap. 25. If $h(y, \lambda) = (y^\lambda - 1)/\lambda$, $0 \leq \lambda \leq 2$, and F_ε is a normal distribution, then this is the celebrated Box-Cox transformation model. On the other hand, if F_ε is known but h is unknown, then it becomes the semiparametric transformation model. The Cox regression model is a special case of this model, corresponding to F_ε being an extreme distribution.

Han's (1987) maximum rank correlation discussed in Sect. 15.4 is a valid method to estimate β in this transformation model. However, his method is not a likelihood based approach which may lead to the loss of information. In the following we discuss an order restricted pairwise maximum likelihood method proposed by Yu et al. (2017).

Using the conventional likelihood methods, it would be difficult to deal with two sets of unknown functions h and F_ε . Instead, we consider a pairwise rank likelihood method, which automatically eliminates h . Note that for $1 \leq i < j \leq n$,

$$\begin{aligned} P(Y_i \leq Y_j | x_i, x_j) &= P(h(Y_i) \leq h(Y_j) | x_i, x_j) \\ &= P(\varepsilon_i - \varepsilon_j \leq (x_j - x_i)\beta) = F((x_j - x_i)\beta), \end{aligned}$$

where F is the distribution of $\varepsilon_i - \varepsilon_j$, which is symmetric. The pairwise rank log-likelihood is

$$L_P(\beta, F) = \sum_{i < j} \{I(Y_i \leq Y_j)F((x_j - x_i)\beta) + I(Y_i > Y_j)\log\{1 - F((x_j - x_i)\beta)\}\}. \quad (26.4.4)$$

Yu et al. (2017) proposed a two-step approach.

(1) For fixed β , sort $(x_j - x_i)\beta$ in ascending order. Then the pool adjacent violation algorithm is used to profile out F , denoted as \hat{F} .

(2) Search for β that maximizes $L_P(\beta, \hat{F}(\beta))$.

Repeat above two steps until convergence.

Finally after finding $\hat{\beta}$, we can estimate h by

$$\min_h \{h(y_i) - x_i \hat{\beta}\}^2$$

subject to the constraint that h is monotonic non-decreasing. This can be accomplished by straightforwardly using PAVA. Finally the error distribution of ε can be estimated using the residual distribution based on $\hat{h}(y_i) - x_i \hat{\beta}$, $i = 1, 2, \dots, n$.

Remark It was pointed out earlier that Han's maximum correlation coefficient method is a generalization of Manski's maximum score method from a discrete response to a continuous response problem. Essentially the order restricted maximum pairwise rank likelihood (26.4.4) is a generalization of Cosslett's (1983) method from choice based models to continuous response models. Since this method ranks all possible $(x_j - x_i)\beta$, $i \neq j$, it is expected to be comparable to the quadrupewise rank method (Abrevaya 1999, also Sect. 15.5) and to be more efficient than the maximum rank correlation method (Han's 1987). This conjecture is supported by

numerical results in Yu et al. (2017). When the underlying Cox model is correctly specified, the loss of efficiency by using this method compared with using the Cox partial likelihood method is moderate. Moreover, the PAVA based maximum pairwise rank likelihood method (26.4.4) is much faster to implement than Abrevaya's (1999) quadruple-wise rank method.

26.5 Maximum Binomial Likelihood Estimation in a Genetic Mixture Model

In genetic studies, it is common to use a mixture model with known mixing proportions to fit the phenotype data given different hidden genotypes. Examples of such studies include, among others, kin-cohort studies and quantitative trait locus (QTL) studies. In a kin-cohort study (Struewing et al. 1997 and Wacholder et al. 1998), a volunteer with or without disease (called proband) agrees to be genotyped. The accompanied information on the history of disease in the first-degree relatives of the proband is also ascertained. Due to high cost and other reasons, relatives' genotype information cannot be collected. Compared with traditional cohort or case-control designs, the kin-cohort design is a promising alternative for estimating penetrance of an identified rare autosomal mutation. It has some practical advantages, including its feasibility, efficiency to implement, and ability to study the effects of an autosomal dominant mutation on several disease outcomes. The design is, however, subject to several biases. Other than recall bias, probands may have systematic selection bias since probands' tendency to participate depends on their disease status. Due to the fact that relatives are not genotyped, data collected from relatives are mixture results of "bad gene carrier" or "not bad gene carrier". The kin-cohort design has been used to estimate the probability that Ashkenazi Jewish women with specific mutations of BRCA1 or BRCA2 will develop breast cancer (Wacholder et al. 1998).

A closely related example is the quantitative trait locus (QTL) studies. QTL analysis is an important method to explore the genetic influences on biological traits. The aim of a QTL study is to determine the association between the genetic variability at known locations on chromosomes and the variability in the observed traits or genotypes (Lander and Botstein 1989). In standard interval mapping genotypes of markers are observed at known locations, but the genotypes between the markers are missing. The frequency of the genotypes can be determined by the recombination fractions between the locus and the flanking markers. As a result, the distribution of the observed phenotype is a mixture of two (or more) components, where each component is the phenotype density for the given genotype. More details can be found in Lander and Botstein (1989) and the book by Wu et al. (2007).

For simplicity we only focus on a two components mixture model. Let

$$X_i \sim H_i(x) = \lambda_i F(x) + (1 - \lambda_i)G(x), \quad i = 1, 2, \dots, n,$$

where λ_i 's are known. In order this model to be identifiable, we assume not all λ_i 's are the same.

We are interested in estimating $F(x)$ and $G(x)$. Since $E[I(X_i \leq t)] = H_i(t)$, $i = 1, 2, \dots, n$, a simple approach is to solve $F(t)$ and $G(t)$ from this type of estimating equations. Unfortunately, there is no guarantee on the monotonic property of the estimated distribution functions. Moreover, it is possible that the solution may have negative values for small sample size, especially at the tails.

In contrast to the rich parametric or semiparametric regression models, not many flexible models are available in finite mixture models other than fully parametric mixture models. It is well known that the maximum full parametric likelihood method produces the most efficient estimate if the underlying parametric model is correctly specified. On the other hand, it is not robust to model misspecification. Without using the parametric assumption on the two components in the mixture model, we may directly maximize the nonparametric likelihood

$$\prod_{i=1}^n \{\lambda_i dF(x_i) + (1 - \lambda_i) dG(x_i)\}$$

with respect to F and G subject to the constraints

$$dF(x_i) \geq 0, \quad dG(x_i) \geq 0, \quad \sum_{i=1}^n dF(x_i) = 1, \quad \sum_{i=1}^n dG(x_i) = 1.$$

Unfortunately Ma and Wang (2012) showed that the nonparametric likelihood estimators fail to be consistent. We further point out that the maximum likelihood estimates fail again even if one of the components, say, $G(x)$ is completely known.

Kiefer and Wolfowitz (1956) first introduced the concept of generalized maximum likelihood estimate (GMLE) for non-dominated family of probability measures \mathcal{P} as follows. For P_1, P_2 in \mathcal{P} , let

$$f(X; P_1, P_2) = \frac{dP_1}{d(P_1 + P_2)}(X)$$

be the Radon–Nikodym derivative of P_1 with respect to $P_1 + P_2$. If X represents the observed data vector, \hat{P} is a GMLE if and only if

$$f(X; \hat{P}, P) \geq f(X; P, \hat{P}), \quad \text{for all } P \text{ in } \mathcal{P}.$$

A distribution function is said to be a GMLE if the probability measure, which induces the distribution function, is a GMLE. Johansen (1978) showed that the Kaplan–Meier estimate is a GMLE for right censored data when there are no restrictions on the underlying distribution function.

Proposition.

If G is a distribution for a continuous random variable, then the maximum likelihood estimator of F is inconsistent in the mixture model

$$X_1, \dots, X_n \sim i.i.d. \lambda dF(x) + (1 - \lambda)dG(x),$$

when λ and $G(x)$ are known but F is unknown.

Proof Using the definition of the generalized maximum likelihood estimate (GMLE), in the simple mixture model discussed above we need to find a dominate measure Q such that \hat{F} maximizes

$$\prod_{i=1}^n \frac{\lambda dF(x_i) + (1 - \lambda)dG(x_i)}{dQ} \leq \prod_{i=1}^n \frac{\lambda d\hat{F}(x_i) + (1 - \lambda)dG(x_i)}{dQ},$$

where $Q = \{\lambda F + (1 - \lambda)G\} + \{\lambda \hat{F} + (1 - \lambda)G\}$. Since $dG(x) = g(x)dx$ is a continuous distribution, the increment $dG(x)$ can be infinitesimal. This is equivalent to maximizing $\lambda^n \prod_{i=1}^n dF(x_i) + \max_i O(dx_i)$, or to maximizing $\prod_{i=1}^n dF(x_i)$. The resulting GMLE is the empirical distribution function $F_n(x) = n^{-1} \sum_{i=1}^n I(x_i \leq x)$. Clearly in the presence of mixture model, $F_n(x) \rightarrow \lambda F(x) + (1 - \lambda)G(x) \neq F(x)$, which leads to inconsistent estimator. This completes the proof.

Since the nonparametric maximum likelihood method fails to produce consistent estimates in the mixture model, as an alternative, we can investigate maximizing a binomial likelihood estimation method for F and G discussed by Huang et al. (2007) and Ma and Wang (2012). To find the connection between the maximum nonparametric likelihood and maximum binomial likelihood, we will first consider a non-mixture case.

Suppose $X_1, \dots, X_n \sim i.i.d. F(x)$. Then the nonparametric likelihood is

$$\prod_{i=1}^n dF(x_i).$$

Without loss of generality we assume $x_1 \leq x_2 \leq \dots \leq x_n$. Denote $p_i = dF(x_i)$, $i = 1, 2, \dots, n$. It is well known that if we maximize $\prod_{i=1}^n p_i$ subject to the constraints

$$\sum_{i=1}^n p_i = 1, \quad p_i \geq 0,$$

then $\hat{p}_i = 1/n$. As a result, the nonparametric likelihood estimation of F is the empirical distribution function $F_n(t) = n^{-1} \sum_{i=1}^n I(x_i \leq t)$.

On the other hand, we can construct a binomial likelihood at each observed data point x_i

$$\ell_i = \sum_{j=1}^n I(X_j \leq x_i) \log F(x_i) + \sum_{j=1}^n I(X_j > x_i) \log\{1 - F(x_i)\}.$$

If we maximize $\sum_{i=1}^n \ell_i$ with respect to $F(x_i)$ subject to the monotonic constraint $F(x_1) \leq F(x_2) \leq \dots \leq F(x_n)$, then

$$\hat{F}(x_i) = n^{-1} \sum_{j=1}^n I(X_j \leq x_i)$$

since the empirical distribution function satisfies the monotonic constraint automatically. This implies that in the non-mixture case, maximizing the nonparametric likelihood $\prod_{i=1}^n dF(x_i)$ with respect to dF is equivalent to maximizing the binomial likelihood with respect to $F(x)$ subject to the monotonic constraint $F(x_1) \leq F(x_2) \leq \dots \leq F(x_n)$. As a result, the binomial likelihood is fully efficient. Based on this result, we anticipate that maximizing the binomial likelihood method may produce highly efficient estimates in general and in more complex situations such as the mixture model problems discussed above.

Now we move back to the binomial likelihood for the mixture model discussed before. Let t_1, \dots, t_h ($h \leq n$) be the ordered unique observed data points. At each time point t_j , we form a log binomial likelihood where

$$\ell_j = \sum_{i=1}^n [I(x_i \leq t_j) \log\{\lambda_i F(t_j) + (1 - \lambda_i) G(t_j)\} + I(x_i > t_j) \log\{\lambda_i \bar{F}(t_j) + (1 - \lambda_i) \bar{G}(t_j)\}].$$

The overall log binomial likelihood is

$$\ell = \sum_{j=1}^n \ell_j.$$

The score equation for $F(t_j)$ is

$$\frac{\partial \ell_j}{\partial F(t_j)} = \sum_{i=1}^n \frac{\lambda_i I(x_i \leq t_j)}{\lambda_i F(t_j) + (1 - \lambda_i) G(t_j)} - \frac{\lambda_i I(x_i > t_j)}{\lambda_i \bar{F}(t_j) + (1 - \lambda_i) \bar{G}(t_j)} = 0.$$

Naturally we wondered whether it is necessary to consider a class of weighted estimating equations

$$\sum_{i=1}^n w_i \left(\frac{\lambda_i I(x_i \leq t_j)}{\lambda_i F(t_j) + (1 - \lambda_i) G(t_j)} - \frac{\lambda_i I(x_i > t_j)}{\lambda_i \bar{F}(t_j) + (1 - \lambda_i) \bar{G}(t_j)} \right) = 0.$$

Since $\text{var}\{\partial\ell_j/\partial F(t_j)\} = -E[\partial^2\ell_j/\partial F(t_j)\partial F(t_j)]$, based on Godambe's (1960) optimal estimating function theory it can be shown that the optimal weights are $w_i = 1, i = 1, 2, \dots, n$.

Both Huang et al. (2007) and Ma and Wang (2012) only considered the point-wise estimation for $F(t)$ for $t = t_1, \dots, t_h$. However there is no guarantee that their methods produce a genuine distribution function. In order to achieve the monotonicity, we need to maximize the binomial likelihood subject to the constraints

$$F(t_1) \leq F(t_2) \leq \dots \leq F(t_h),$$

and

$$G(t_1) \leq G(t_2) \leq \dots \leq G(t_h),$$

where $t_1 < t_2 < \dots < t_h$.

Direct use of the isotonic regression method does not work since the mixture model discussed above does not belong to the exponential family. In order to accomplish this maximization, we need to use a combination of the EM and PAVA algorithms.

Let D_i be a latent binary variable

$$P(D_i = 1) = \lambda_i, i = 1, 2, \dots, n.$$

If $D_i = 1$ or 0, then X_i is generated from $F(x)$ or $G(x)$, respectively. The complete data are $D_i, I(X_i \leq t_j), i = 1, 2, \dots, n, j = 1, 2, \dots, h$. The complete data log-likelihood is

$$\begin{aligned} \ell_c = & \sum_{j=1}^n \sum_{i=1}^n [D_i I(X_i \leq t_j) \log\{\lambda_i F(t_j)\} + D_i I(X_i > t_j) \log\{\lambda_i \bar{F}(t_j)\} \\ & + (1 - D_i) I(X_i \leq t_j) \log\{(1 - \lambda_i) G(t_j)\} + (1 - D_i) I(X_i > t_j) \log\{(1 - \lambda_i) \bar{G}(t_j)\}]. \end{aligned}$$

Since D_i is not available in general, we need to impute it for given observable quantities $I(X_i \leq t_j)$. Define

$$u_{ij} = P(D_i = 1 | X_i \leq t_j) = \frac{\lambda_i F(t_j)}{\lambda_i F(t_j) + (1 - \lambda_i) G(t_j)},$$

$$v_{ij} = P(D_i = 1 | X_i > t_j) = \frac{\lambda_i \bar{F}(t_j)}{\lambda_i \bar{F}(t_j) + (1 - \lambda_i) \bar{G}(t_j)}.$$

The imputed log-likelihood is

$$\begin{aligned} E[\ell_c | O] = & \sum_{j=1}^n \sum_{i=1}^n [u_{ij} I(X_i \leq t_j) \log\{\lambda_i F(t_j)\} + v_{ij} I(X_i > t_j) \log\{\lambda_i \bar{F}(t_j)\} \\ & + (1 - u_{ij}) I(X_i \leq t_j) \log\{(1 - \lambda_i) G(t_j)\} + (1 - v_{ij}) I(X_i > t_j) \log\{(1 - \lambda_i) \bar{G}(t_j)\}]. \end{aligned}$$

Therefore the unrestricted maximum likelihood estimates of $F(t_j)$ and $G(t_j)$ are, respectively,

$$\hat{F}(t_j) = \frac{\sum_{i=1}^n u_{ij} I(X_i \leq t_j)}{\sum_{i=1}^n u_{ij} I(X_i \leq t_j) + \sum_{i=1}^n v_{ij} I(X_i > t_j)},$$

and

$$\hat{G}(t_j) = \frac{\sum_{i=1}^n (1 - u_{ij}) I(X_i \leq t_j)}{\sum_{i=1}^n (1 - u_{ij}) I(X_i \leq t_j) + \sum_{i=1}^n (1 - v_{ij}) I(X_i > t_j)}.$$

Then PAVA can be used to achieve the monotonicity:

$$\tilde{F}(t_j) = \max_{s \leq j} \min_{t \geq j} \frac{\sum_{h=s}^t \sum_{i=1}^n u_{ih} I(X_i \leq t_h)}{\sum_{h=s}^t \sum_{i=1}^n u_{ih} I(X_i \leq t_h) + \sum_{i=1}^n v_{ih} I(X_i > t_h)}.$$

Repeat the above procedures until converge.

In the presence of right censoring, the maximum binomial likelihood estimation becomes more complex since we cannot directly observe $I(X_i > t_j)$ for individuals who are lost to follow up before time t_j . A natural approach is to impute this indicator function. We leave this as an exercise for readers.

Suppose $\lambda_i, i = 1, 2, \dots, n$ are random variables with a common distribution function $\eta(\lambda)$. To show consistency of the binomial likelihood, we use the Kullback–Leibler information argument. By the Law of Large Numbers, it can be shown that in probability

$$\begin{aligned} & n^{-2} \ell(\hat{F}_0, \hat{G}_0, F, G) \\ &= \int [1 - \{\lambda \bar{F}_0(t) + (1 - \lambda) G_0(t)\} \bar{H}_C(t) \\ &\quad - (1 - \hat{F}_0)(t) \int_0^t \frac{\lambda \bar{F}_0(u) + (1 - \lambda) \bar{G}_0(u)}{\lambda(1 - \hat{F}_0(u)) + (1 - \lambda)(1 - \hat{G}_0(u))} dH_C(u)] \\ &\quad \times \log[\lambda F(t) + (1 - \lambda) G(t)] d\eta(\lambda) d\xi(t) \\ &+ \int [\{\lambda \bar{F}_0(t) + (1 - \lambda) G_0(t)\} \bar{H}_C(t) + (1 - \lambda) \hat{F}_0(t) \\ &\quad - (1 - \lambda) \hat{G}_0(t) \int_0^t \frac{\lambda \bar{F}_0(u) + (1 - \lambda) \bar{G}_0(u)}{\lambda \bar{F}(u) + (1 - \lambda) \bar{G}(u)} dH_C(u)] \\ &\quad \log[\lambda \bar{F}(t) + (1 - \lambda) \bar{G}(t)] d\eta(\lambda) d\xi(t) + o_p(1) \\ &\rightarrow \int \{\lambda F_0(t) + (1 - \lambda) G_0(t)\} \log[\lambda F(t) + (1 - \lambda) G(t)] d\eta(\lambda) d\xi(t) \\ &+ \int \{\lambda \bar{F}_0(t) + (1 - \lambda) \bar{G}_0(t)\} \log[\lambda \bar{F}(t) + (1 - \lambda) \bar{G}(t)] d\eta(\lambda) d\xi(t). \end{aligned}$$

This integration must achieve maximum at the true $F_0(t)$ and $G_0(t)$ by the Kullback–Leibler information inequality. As a consequence we have shown that the maximum binomial likelihood estimate is consistent.

Qin et al. (2014) applied the maximum binomial likelihood method to analyze a genetic data set.

26.6 Maximum Likelihood Estimation Based on Current Status Data

In this section, we discuss maximum likelihood estimation based on current status data. First we study the nonparametric MLE for current status data. Then, we make a connection between AFT model for current status data and Cosslett's (1983) binary choice model. Third, we develop a new algorithm by the combination of EM and PAVA for analyzing competing risk current status data. Finally, we investigate the maximum likelihood estimation for the transformation normal model with current status data.

(1) Nonparametric MLE

Suppose T is the lifetime and C is the examining time. We only know whether T is shorter or longer than C , where C is observable. A typical example is that T is the mail delivery time each day, and C is the examining time. At time C , one only knows whether or not the mail has been delivered. In animal studies, it is a standard procedure to sacrifice animals to examine tumor onset time. In the absence of tumor, the tumor onset time is later than the examining time. Otherwise, the onset time is earlier than the examining time, but the exact onset time is unknown. Compared with conventional right censored data where the exact time T is observed if $T < C$, the information based on the current status data is much weaker than that provided by the right censored data. As a consequence, the underlying distribution function estimation has a slower than root- n convergence rate. Many examples in medical applications for current status data can be found in the excellent monograph by Sun (2006).

Assume that $T \sim F(t)$ and $C \sim G(c)$. The observed data are denoted as $\delta_i = I(T_i \leq C_i)$, C_i , $i = 1, 2, \dots, n$. The likelihood is

$$L = \prod_{i=1}^n F^{\delta_i}(c_i) \{1 - F(c_i)\}^{1-\delta_i} dG(c_i).$$

The log-likelihood can be written as

$$\ell = \sum_{i=1}^n \delta_i \log F(c_i) + (1 - \delta_i) \log \{1 - F(c_i)\} + \sum_{i=1}^n \log dG(c_i).$$

Note that G is ancillary for estimating F . In principle this is exactly the same problem discussed in Sect. 26.1 on the dose finding example subject to the monotonic constraints

$$c_1 \leq c_2 \leq \cdots \leq c_n, \quad F(c_1) \leq F(c_2) \leq \cdots \leq F(c_n).$$

We can use PAVA to estimate F , denoted as \hat{F} . It was shown that \hat{F} has a cubic root convergence rate. In general, the limiting distribution is complicated and not Gaussian. Details can be found in Groeneboom and Jongbloed (2014).

(2) AFT Model with Current Status Data

Under the accelerated failure time model (AFT)

$$\log T = x\beta + \varepsilon, \quad \varepsilon \sim F(\cdot)$$

with current status data, the log-likelihood is

$$\ell = \sum_{i=1}^n I(T_i < c_i) \log F(\log(c_i) - x_i\beta) + I(T_i > c_i) \log\{1 - F(\log(c_i) - x_i\beta)\}.$$

If let $z\gamma = \gamma_1 z_1 + \gamma_2 z_2$, where $z_1 = \log(c)$, $Z_2 = x$, $\gamma_1 = 1$ and $\gamma_2 = -\beta$, then mathematically this is exactly equivalent to Cosslett's (1987) binary choice model. Note that except for the intercept β is estimable. However, in Cosslett's setting, only the direction of γ is estimable. Again a two-step approach is needed. First fix β , use PAVA to profile out F . Then search for β in the profile log-likelihood in the second step.

Huang and Wellner (1997) established some large sample results. However, the root- n consistency for β and Central Limit theory are still open problems. Some recent work can be found in Groeneboom and Jongbloed (2014).

(3) Maximum Likelihood Estimation Based on Current Status Data with Competing Risks

There are many discussions about estimating cumulative incidence functions based on right censored data or current status competing risks data. A good reference for right censored competing risk data can be found in the books by Kalbfleisch and Prentice (2002) and Lawless (2003). A nice algorithm to compute the nonparametric maximum likelihood estimator based on left truncated and right censored competing risk data was developed by Hudgens et al. (2001). Jewell et al. (2003) discussed two nonparametric estimators of the cumulative incidence functions: the nonparametric maximum likelihood estimator and a simpler naive nonparametric maximum likelihood estimator based on reduced current status data for the cause of interest. To find the nonparametric maximum likelihood estimators, Jewell and Kalbfleisch (2004) generalized the pool adjacent violation algorithm from a series of binomial trials subject to order constraints, to a series of ordered trinomial parameters. Groeneboom et al. (2008a,b) derived large sample properties of the two nonparametric estimators.

They showed that both estimators converge at a cube root rate, but have different limiting distributions. The nonparametric maximum likelihood estimator was shown to be superior to the naive estimator in terms of mean squared error, both empirically and asymptotically.

Below we discuss a more flexible algorithm through a combination of the EM algorithm and PAVA (Huang et al. 2016). This new approach can be easily adapted to the case in which the exact competing failure times is observable for one of them if it occurred prior to the monitoring time.

Define two sub-distribution functions as

$$F_j(t) = P(T \leq t, J = j), \quad j = 1, 2.$$

The overall survival is $S(t) = 1 - F_1(t) - F_2(t)$. For current status data, each individuals' information on survival status is available only at a single time C . The observed data can be represented as $Y = (C, \Delta, \Phi)$, where $\Delta = 1$ if $T \leq C$ with $J = 1$, and $\Phi = 1$ if $T < C$ with $J = 2$. In other words, an individual is known to have failed at the observation time C then the cause of failure is also available.

The likelihood is

$$L_1 = \prod_{i=1}^n \{F_1(c_i)\}^{\delta_i} \{F_2(c_i)\}^{\phi_i} \{1 - F_1(c_i) - F_2(c_i)\}^{1-\delta_i-\phi_i}.$$

Let

$$c_{(1)} \leq c_{(2)} \leq \cdots \leq c_{(n)}.$$

Jewell and Kalbfleisch (2004) studied the constrained nonparametric MLE subject to the constraints

$$F_j(c_{(1)}) \leq F_j(c_{(2)}) \leq \cdots \leq F_j(c_{(n)}), \quad j = 1, 2, \quad F_1(c_{(j)}) + F_2(c_{(j)}) \leq 1.$$

An iterative PAVA for F_1 and F_2 was proposed in their paper. Lim et al. (2009) applied the geometric programming method to the same problem considered by Jewell and Kalbfleisch (2004).

Next we study a new algorithm. Denote

$$G_j(t) = P(T \leq t | J = j), \quad j = 1, 2, \quad \pi_j = P(J = j).$$

The likelihood can be written as

$$\begin{aligned} L_1 &= \prod_{i=1}^n \{\pi_1 G_1(c_i)\}^{\delta_i} \{(1 - \pi_1) G_2(c_i)\}^{\phi_i} \{1 - \pi_1 G_1(c_i) - \pi_2 G_2(c_i)\}^{1-\delta_i-\phi_i} \\ &= \prod_{i=1}^n \{\pi_1 G_1(c_i)\}^{\delta_i} \{(1 - \pi_1) G_2(c_i)\}^{\phi_i} \{\pi_1 \bar{G}_1(c_i) + (1 - \pi_1) \bar{G}_2(c_i)\}^{1-\delta_i-\phi_i}. \end{aligned}$$

This is equivalent to a mixture model by introducing an indicator variable D_i ,

$$P(D_i = 1) = \pi_1, \quad P(T < c_i | D_i = j) = G_j(c_i), \quad j = 1, 2,$$

where D_i is observable if $T < c_i$, however it is not if $T > c_i$. Noting

$$P(T > c_i) = \pi_1 \bar{G}_1(c_i) + (1 - \pi_1) \bar{G}_2(c_i),$$

we can impute D_i for given $T > c_i$ through

$$w_i =: P(D_i = 1 | T > c_i) = \frac{\pi_1 \bar{G}_1(c_i)}{\pi_1 \bar{G}_1(c_i) + (1 - \pi_1) \bar{G}_2(c_i)}.$$

Therefore the imputed log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^n I(D_i = 1) I(T_i < c_i) \log\{\pi_1 G_1(c_i)\} + I(D_i = 2) I(T_i < c_i) \log\{(1 - \pi_1) G_2(c_i)\} \\ &\quad + \sum_{j=1}^n w_j I(T_j > c_j) \log\{\pi_1 \bar{G}_1(c_j)\} + (1 - w_j) I(T_j > c_j) \log\{(1 - \pi_1) \bar{G}_2(c_j)\} \\ &= \sum_{i=1}^n \{I(D_i = 1) I(T_i < c_i) + w_i I(T_i > c_i)\} \log \pi_1 \\ &\quad + \sum_{i=1}^n \{I(D_i = 2) I(T_i < c_i) + (1 - w_i) I(T_i > c_i)\} \log(1 - \pi_1) \\ &\quad + \sum_{i=1}^n \{I(D_i = 1) I(T_i < c_i) \log G_1(c_i) + w_i I(T_i > c_i) \log \bar{G}_1(c_i)\} \\ &\quad + \sum_{i=1}^n \{I(D_i = 2) I(T_i < c_i) \log G_2(c_i) + (1 - w_i) I(T_i > c_i) \log \bar{G}_2(c_i)\}. \end{aligned}$$

The maximum likelihood estimate of π_1 has a closed form

$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^n \{I(D_i = 1) I(T_i < c_i) + w_i I(T_i > c_i)\}.$$

Without loss of generality we assume that

$$c_1 \leq c_2 \leq \dots \leq c_n.$$

Then we need to maximize above likelihood with respect to G_1, G_2 subject to the constraints

$$G_1(c_1) \leq G_1(c_2) \leq \dots \leq G_1(c_n)$$

and

$$G_2(c_1) \leq G_2(c_2) \leq \dots \leq G_2(c_n).$$

Easily PAVA can be applied. In practical implementation, we need give an initial value (G_1 , G_2) for calculating w_i , then iterate this process until it converges.

The advantage of the newly proposed method is that it can be easily adapted to handle the situation when the failure time can be observed exactly for one type of event. Without loss of generality, we assume the first type event time is available if a failure is observed. The likelihood can be written as

$$L_2 = \prod_{i=1}^n \{\pi_1 dG_1(c_i)\}^{\delta_i} \{(1 - \pi_1) dG_2(c_i)\}^{\phi_i} \{\pi_1 \bar{G}_1(c_i) + (1 - \pi_1) \bar{G}_2(c_i)\}^{1-\delta_i-\phi_i}.$$

This can be also treated as a mixture model

$$P(D_i = 1) = \pi_1, \quad P(T < c_i | D_i = j) = G_j(c_i), \quad j = 1, 2,$$

where if $T < c_i$, then we can observe either $D_i = 1$ and $T = t_i$ (c_i is not available) or $D_i = 2$ but $T < c_i$ (c_i is available), and if $T > c_i$ then we cannot observe D_i . The “complete data” log-likelihood is

$$\begin{aligned} \ell_2 = & \sum_{i=1}^m I(D_i = 1) I(T_i \leq C_i) \log\{\pi_1 dG_1(t_i)\} + I(D_i = 2) I(T_i \leq C_i) \log\{(1 - \pi_1) dG_2(c_i)\} \\ & + \sum_{j=1}^n w_j I(T_j > c_j) \log\{\pi_1 \bar{G}_1(c_j)\} + (1 - w_j) I(T_j > c_j) \log\{(1 - \pi_1) \bar{G}_2(c_j)\}. \end{aligned}$$

Denote

$$\bar{G}_1(c_i) = \exp\{-\Lambda_1(c_i)\}.$$

In this case G_1 can be estimated by the Kaplan–Meier or the Nelson method, and G_2 can be estimated by PAVA. Again iterations are needed.

The newly proposed method can also be adapted to the case where the cause of failure is missing for a portion of patients, for example, Dinse (1986). Let η be the indicator for the missingness of δ , where $\eta = 1$ if δ is observed for a failure, and 0 otherwise. For example, in situations where some of the systems or subjects are being studied, the exact failure cause cannot be identified easily but the test procedures can restrict the cause to some subset. Even though this problem has been discussed extensively in right censored data in statistical literature, for example, among others, such as Dinse (1986), Flehinger et al. (1998), it remains an open problem for current status data.

Define

$$P(D_i = 1) = \pi_1, \quad P(T < c_i | D_i = j) = G_j(c_i), \quad j = 1, 2,$$

where if $T < c_i$ then we observe either $D_i = 1$ or $D_i = 2$ and c_i if $\eta_i = 1$, but we observe $T < c_i$ only if $\eta_i = 0$. Moreover if $T > c_i$ then we cannot observe D_i ($\eta_i = 0$).

Without loss of generality we assume the first n_1 individuals with observations $T_i \leq c_i, i = 1, 2, \dots, n_1$. Let

$$v_i = P(D_i = 1, \eta_i = 0 | T_i \leq c_i) = \frac{\pi_1 G_1(c_i)}{\pi_1 G_1(c_i) + \pi_2 G_2(c_i)}, i = 1, 2, \dots, n_1.$$

Again for those censored individuals we never observe the status of D_i , i.e., $\eta_i = 0$. Denote

$$w_i = P(D_i = 1 | T_i > c_i) = \frac{\pi_1 \bar{G}_1(c_i)}{\pi_1 \bar{G}_1(c_i) + \pi_2 \bar{G}_2(c_i)}, i = n_1 + 1, \dots, n.$$

The full log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^{n_1} \eta_i I(D_i = 1) I(T_i < c_i) \log\{\pi_1 G_1(c_i)\} + \eta_i I(D_i = 2) I(T_i < c_i) \log\{(1 - \pi_1) G_2(c_i)\} \\ &\quad + \sum_{i=1}^{n_1} (1 - \eta_i) v_i I(T_i < c_i) \log\{\pi_1 G_1(c_i)\} + (1 - \eta_i)(1 - v_i) I(T_i < c_i) \log\{(1 - \pi_1) G_2(c_i)\} \\ &\quad + \sum_{j=n_1+1}^n w_j I(T_j > c_j) \log\{\pi_1 \bar{G}_1(c_j)\} + (1 - w_j) I(T_j > c_j) \log\{(1 - \pi_1) \bar{G}_2(c_j)\} \\ &= \left[\sum_{i=1}^{n_1} \{\eta_i I(D_i = 1) + (1 - \eta_i) v_i\} I(T_i < c_i) + \sum_{j=n_1+1}^n w_j I(T_j > c_j) \right] \log \pi_1 \\ &\quad + \left[\sum_{i=1}^{n_1} \{\eta_i I(D_i = 2) + (1 - \eta_i)(1 - v_i)\} I(T_i < c_i) + \sum_{j=n_1+1}^n (1 - w_j) I(T_j > c_j) \right] \log(1 - \pi_1) \\ &\quad + \left[\sum_{i=1}^{n_1} \{\eta_i I(D_i = 1) + (1 - \eta_i) v_i\} I(T_i < c_i) \log G_1(c_i) + \sum_{j=n_1+1}^n w_j I(T_j > c_j) \log \bar{G}_1(c_j) \right] \\ &\quad + \left[\sum_{i=1}^{n_1} \{\eta_i I(D_i = 2) + (1 - \eta_i)(1 - v_i)\} I(T_i < c_i) \log G_2(c_i) \right. \\ &\quad \left. + \sum_{j=n_1+1}^n (1 - w_j) I(T_j > c_j) \log \bar{G}_2(c_j) \right]. \end{aligned}$$

We can estimate π_1 by

$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n_1} \{\eta_i I(D_i = 1) + (1 - \eta_i) v_i\} I(T_i < c_i) + \sum_{j=n_1+1}^n w_j I(T_j > c_j)\}.$$

Again PAVA can be used to estimate G_1 and G_2 . Numerical results can be found in Huang et al. (2016).

(4) Transformation Probit Model with Current Status Data

Other than Cox regression model, the transformation model and the accelerated failure time model, the transformation normal model is also popular in survival analysis. Denote the lifetime as T^* . Let α be an unspecified monotone nondecreasing function. We assume that

$$\alpha(T^*)|x \sim N(-x\beta, 1).$$

Then the cumulative distribution is

$$F(t|x) = P(T^* < t|x) = P(\alpha(T^*) < \alpha(t)|x) = \Phi(\alpha(t) + x\beta),$$

where $\Phi(\cdot)$ is the standard normal distribution. The corresponding density is

$$f(t|x) = \phi(\alpha(t) + x\beta)\alpha'(t), \quad \phi(x) = \frac{d\Phi(x)}{dx}.$$

Suppose we have current status data ($\delta_i = I(T_i^* \leq t_i), t_i$), $i = 1, 2, \dots, n$, where t_i is the examining time for individual i . The log-likelihood is

$$\ell = \sum_{i=1}^n [\delta_i \log \Phi(\alpha(t_i) + x_i\beta) + (1 - \delta_i) \log\{1 - \Phi(\alpha(t_i) + x_i\beta)\}].$$

Without loss of generality, we assume

$$t_1 < t_2 < \dots < t_n; \quad \alpha_1 = \alpha(t_1) \leq \alpha_2 = \alpha(t_2) \leq \dots \leq \alpha_n = \alpha(t_n).$$

It is challenging to maximize this log-likelihood subject to the monotonic constraints. To circumvent this problem, we create an artificial data set that generates the same likelihood as the original current status data. To obtain the maximum likelihood estimate, however it is more straightforward to use the EM algorithm for the artificial data.

Consider a Probit model

$$Y_i|x_i \sim N(-\alpha_i - x_i\beta, 1), \quad \alpha_i = \alpha(t_i).$$

Then

$$\begin{aligned} P(Y_i < 0|x_i) &= P(Y_i + \alpha_i + x_i\beta < \alpha_i + x_i\beta) = \Phi(\alpha_i + x_i\beta), \\ P(Y_i > 0|x_i) &= 1 - \Phi(\alpha_i + x_i\beta). \end{aligned}$$

Instead of observing $Y_i > 0$ or $Y_i < 0$, if Y_i can be observed, then the complete log-likelihood is

$$\ell_C = - \sum_{i=1}^n (y_i + \alpha_i + x_i\beta)^2 = - \sum_{i=1}^n (\tilde{y}_i - \alpha_i)^2, \quad \tilde{y}_i = -y_i - x_i\beta,$$

Equivalently we need to minimizing

$$h = \sum_{i=1}^n (\tilde{y}_i - \alpha_i)^2 = \sum_{i=1}^n \{\tilde{y}_i^2 - 2\tilde{y}_i\alpha_i + \alpha_i^2\}.$$

This is an isotonic regression problem. The PAVA can be applied to estimate α_i for fixed β . Since Y_i is not observable, we can impute its value by using conditional expectation. Note that

$$E[Y|Y > 0] = \mu + \frac{\phi(-\mu)}{1 - \Phi(-\mu)}, \quad E[Y|Y < 0] = \mu - \frac{\phi(-\mu)}{\Phi(-\mu)}.$$

Let $O = I(Y > 0)$ or $I(Y < 0)$ be the observed quantity. The conditional expectation of log-likelihood is

$$\begin{aligned} E[h|O] &= \sum_{i=1}^n \{-2E[\tilde{Y}_i|O]\alpha_i + \alpha_i^2\} + \sum_{i=1}^n E[\tilde{Y}_i^2|O] \\ &= \sum_{i=1}^n \{E[\tilde{Y}_i|O] - \alpha_i\}^2 + \sum_{i=1}^n \{E[\tilde{Y}_i^2|O_i] - E^2[\tilde{Y}_i|O_i]\} \\ &= \sum_{i=1}^n \{E[\tilde{Y}_i|O] - \alpha_i\}^2 + \sum_{i=1}^n V[\tilde{Y}_i|O_i] \\ &= \sum_{i=1}^n \{E[\tilde{Y}_i|O] - \alpha_i\}^2 + \sum_{i=1}^n V[Y_i|O_i] \\ &= \sum_{i=1}^n \{E[\tilde{Y}_i|O] - \alpha_i\}^2 + \text{constant}(\alpha_i^0, x_i\beta^0). \end{aligned}$$

This becomes a weighted isotonic regression problem.

$$\begin{aligned} E[\tilde{Y}_i|O_i] &= -E[Y_i|O_i] - x_i\beta \\ &= -x_i\beta - \begin{cases} -\alpha_i^0 - x_i\beta^0 + \frac{\phi(\alpha_i^0 + x_i\beta^0)}{1 - \Phi(\alpha_i^0 + x_i\beta^0)}, & \text{if } Y_i > 0 \\ -\alpha_i^0 - x_i\beta^0 - \frac{\phi(\alpha_i^0 + x_i\beta^0)}{\Phi(\alpha_i^0 + x_i\beta^0)}, & \text{if } Y_i < 0. \end{cases} \end{aligned}$$

Note that $\Phi(\cdot)$ is a convex function. It is known that composition with an affine mapping perseveres the convex property (Boyd and Vandenberghe 2004, Sect. 3.2.2 on p. 79). Therefore we can claim that the log-likelihood is a convex function. Csiszar and Tusnady's (1984) result indicates convergence of the EM algorithm. More details can be found in a recent work by Liu and Qin (2017). They also have discussed bivariate current status data problems.

26.7 Application in Receiver Operating Characteristic (ROC)

The accuracy of a medical diagnostic test is typically assessed by its sensitivity and specificity, defined respectively as the probability that a truly diseased subject has a positive test result and the probability that a truly non-diseased subject has a negative test result. When the sampling distribution of a diagnostic variable is continuous, it is common to use a receiver operating characteristic (ROC) curve, a plot of the sensitivity versus one minus the specificity across all possible cut-points of the variable. There are many excellent books on this topic, among others, for example, Pepe (2004) and Zhou et al. (2011a,b).

In clinical practice, several medical diagnostic tests are often available, but may not be perfect in the sense that no single test is sufficiently sensitive and specific on its own for the purpose of population screening for a disease. Different diagnostic tests are usually sensitive to different aspects of the disease. One approach to improve the performance is to combine multiple diagnostic tests to obtain an optimal composite diagnostic test with higher sensitivity that detects the presence of disease more accurately. Under a multivariate normal assumption for different test results, Su and Liu (1993) found the best linear combination of multiple tests. By using the Neyman–Pearson fundamental lemma, Eguchi and Copas (2002, 2006) and McIntosh and Pepe (2002) constructed the best ROC for different diagnostic tests using the log-likelihood ratio statistic for diseased and nondiseased populations. Consequently, it is natural to find the best combination of multiple diagnostic tests by directly modeling the log-likelihood ratio function.

Let \mathbf{X} be a high dimensional marker vector. We are interested in finding the optimal combination of \mathbf{X} such that the false positive probability is under control for some specified precision and the 1-false negative probability is maximized. False positive probability and false negative probability are defined, respectively,

$$FP(u) = P(\psi(\mathbf{X}) \geq u | D = 0), \quad FN(u) = P(\psi(\mathbf{X}) < u | D = 1),$$

where $D = 1$ or 0 denote disease or disease free, respectively, and ψ is a one dimensional function of \mathbf{X} . For different u , we may plot $FP(u)$ versus $1 - FN(u)$ to form the ROC curve.

Our goal is to find the optimal $\psi_0(\mathbf{X})$ such that

$$P\{\psi_0(\mathbf{X}) \geq u | D = 0\} \leq \alpha(u)$$

and

$$\max_{\psi} P\{\psi(\mathbf{X}) \geq u | D = 1\} = P\{\psi_0(\mathbf{X}) \geq u | D = 1\},$$

where $\alpha(u)$ is a specified function of u . By using the Neyman–Pearson Lemma, the optimal solution is

$$\psi_0(\mathbf{X}) = \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})},$$

where $f_1(\mathbf{X})$ and $f_0(\mathbf{X})$ are the marker densities in the disease group and disease free group, respectively.

More formally we present the following useful result.

Neyman–Pearson Lemma

Let $\pi_i = P(D = i)$, $i = 0, 1$. Denote $f_i(x)$, $i = 0, 1$ as the densities of \mathbf{X} for $D = 0$ or 1, respectively. Define

$$\lambda(x) = \log(\pi_1/\pi_0) + \log(f_1/f_0),$$

$$A(u) = I(\lambda(x) > u).$$

Let

$$P_0\{\lambda(x) > u\} = \alpha$$

be the probability evaluated under $D = 0$. Write

$$B(u) = I(\tilde{\lambda}(x) > u), \quad P_0\{B(u)\} \leq \alpha$$

for any given function $\tilde{\lambda}(x)$. We would like to show if both $P_0(A)$ and $P_0(B)$ are less than α , then

$$P_1(A) \geq P_1(B)$$

In fact we can use the following arguments.

$$\begin{aligned} \pi_0 \exp(u) \int_{A-B} f_0(x) dx &\leq \pi_0 \int_{A-B} \exp\{\lambda(x)\} f_0(x) dx \\ &= \pi_0 \int_{A-B} \frac{\pi_1}{\pi_0} \frac{f_1(x)}{f_0(x)} f_0(x) dx \\ &= \pi_1 \int_{A-B} f_1(x) dx. \end{aligned}$$

Similarly

$$\begin{aligned} \pi_0 \exp(u) \int_{B-A} f_0(x) dx &\geq \pi_0 \int_{B-A} \exp\{\lambda(x)\} f_0(x) dx \\ &= \pi_0 \int_{B-A} \frac{\pi_1}{\pi_0} \frac{f_1(x)}{f_0(x)} f_0(x) dx \\ &= \pi_1 \int_{B-A} f_1(x) dx. \end{aligned}$$

Taking the difference,

$$\begin{aligned} &\pi_0 \exp(u) \left[\int_{A-B} f_0(x) dx - \int_{B-A} f_0(x) dx \right] \\ &\leq \pi_1 \left[\int_{A-B} f_1(x) dx - \int_{B-A} f_1(x) dx \right], \end{aligned}$$

or

$$0 \leq \pi_0 \exp(u) [P_0(A) - P_0(B)] \leq \pi_1 [P_1(A) - P_1(B)].$$

Since $P_0(A) = \alpha \geq P_0(B)$, the left side is nonnegative. As a consequence $P_1(A) \geq P_1(B)$. This result is given by Eguchi and Copas (2006).

Therefore the ROC constructed from the likelihood ratio statistic is the optimal combination of different biomarkers. Copas and Corbett (2002) used the logistic regression method to combine ROCs. Instead of modelling F_0 and F_1 separately, Qin and Zhang (2010) used the exponential tilting model

$$f_1(\mathbf{X}) = \frac{\exp(\mathbf{X}^T \beta) f_0(\mathbf{X})}{\int \exp(\mathbf{X}^T \beta) f_0(\mathbf{X}) d\mathbf{X}}$$

to find the ROC based on the optimal combination of $\mathbf{X}^T \beta$, where the baseline density $f_0(\cdot)$ is not specified. Based on the results in Chap. 11, it is not difficult to estimate β and the underlying distributions of F_0 and F_1 . Note the semiparametric model based estimates of F_0 and F_1 are much smoother than the nonparametric counterparts since the model based estimates use both case and control data.

To further improve the robustness of the ROC estimation in multiple covariates or biomarker problems, Chen et al. (2016) proposed directly modeling the density ratio as a nonparametric function of a combination of multiple diagnostic tests. Specifically, they assumed the density ratio between the diseased and non-diseased populations is a monotonic nonparametric functional form of a combination of multiple diagnostic tests, i.e.,

$$f_1(\mathbf{X}) = \frac{\bar{G}(\mathbf{X}^T \beta) f_0(\mathbf{X})}{\int \bar{G}(\mathbf{X}^T \beta) f_0(\mathbf{X}) d\mathbf{X}},$$

where $\bar{G}(\cdot)$ is a monotonic non-increasing function (survival function). Through this method, they were able to find the optimal combination over the monotonic functions of all linear combinations of multiple diagnostic tests. Their semiparametric maximum likelihood estimation procedure can also be implemented using PAVA (Ayer et al. 1955), which is available in the R packages “Iso” and “isotone” (R Development Core Team 2011).

More specifically, for fixed β , we can treat $x_i\beta, i = 1, 2, \dots, n$ as the observed data and then use the same algorithm as in the estimation of the monotone density ratio model (Sect. 24.3). We can then use the “optim” function in R to search for β . This procedure can be iterated until convergence. It may be desirable to try many initial values for β since the objective function is very rough. Theoretical results and extensive numerical results were given in Chen et al. (2016).

26.8 Panel Count Data and Simplex Constraints

In this section, we discuss estimation problems under simplex constraints. This method has potential applications in interval censoring problems in survival analysis and panel count data. More discussions on the simplex algorithm with application in statistical model for positron emission tomography can be found in Vardi et al. (1985).

Liu (2000) considered the constrained Poisson model

$$n_i|\theta_i \sim \text{Poisson}(\theta_i), \quad i = 1, 2, \dots, m,$$

where

$$\theta_i = \sum_{j=1}^q b_{ij}\alpha_j, \quad \alpha_j \geq 0, j = 1, 2, \dots, q$$

and $b_{ij} \geq 0$ are known scalars with $\max_{1 \leq i \leq m} b_{ij} > 0$ for $j = 1, 2, \dots, q$, and $\alpha_j \geq 0$ are unknown parameters. Denote $\mathbf{b} = (b_{ij})$ as the design matrix. Because the elements of \mathbf{b} are non-negative, the constrained $\theta = (\theta_1, \dots, \theta_m)$ lies in the convex cone

$$\mathcal{C} = \{\theta : \theta = \mathbf{b}^T \alpha, \quad \alpha = (\alpha_1, \dots, \alpha_p) \geq 0 = (\alpha_1, \dots, \alpha_q) \geq 0\}.$$

It is a challenging task to directly maximize the likelihood subject to the non-negative constraints $\alpha_i, i = 1, 2, \dots, m$.

As an alternative approach, the EM algorithm may be used by creating some latent variables.

Let

$$Z_{ij} \sim \text{Poisson}(b_{ij}\alpha_j), \quad i = 1, 2, \dots, m, j = 1, 2, \dots, q.$$

The observed data are

$$n_i = \sum_{j=1}^q Z_{ij} = Z_{i+} \sim \text{Poisson}(\sum_{j=1}^q b_{ij}\alpha_j).$$

Clearly if Z_{ij} are available, the complete data log-likelihood is

$$\ell = \sum_{ij} \{Z_{ij} \log(b_{ij}\alpha_j) - b_{ij}\alpha_j\} = \sum_{j=1}^q (Z_{+j} \log \alpha_j - b_{+j}\alpha_j) + C.$$

we can estimate α_j by

$$\hat{\alpha}_j = \frac{\sum_{i=1}^m Z_{ij}}{\sum_{i=1}^m b_{ij}}.$$

Since only n_i 's are available, we impute Z_{ij} for given n_1, \dots, n_m .

It is well known that conditioning on $Z_{i+}, Z_{ij}, (j = 1, 2, \dots, q)$ has a multinomial distribution with probability

$$p_{ij} = \frac{b_{ij}\alpha_j}{\sum_{j=1}^q b_{ij}\alpha_j}.$$

Therefore

$$E[Z_{ij}|Z_{i+} = n_i] = n_i p_{ij}, \quad j = 1, 2, \dots, q.$$

An immediate application of this algorithm is for the panel count data in Poisson process. Suppose each process can be observed in interval (a_i, b_i) , $i = 1, 2, \dots, n$. We only observed the number of events n_i occurred in interval (a_i, b_i) . Suppose the cumulative intensity function for the Poisson process is $\Lambda(t)$. Then the likelihood is

$$L = \prod_{i=1}^n \{\Lambda(b_i) - \Lambda(a_i)\}^{n_i} \exp[-\{\Lambda(b_i) - \Lambda(a_i)\}]/n_i!$$

Let

$$t_1 < t_2 < t_3 < \dots < t_N$$

be the ordered $a_i, b_i, i = 1, 2, \dots, n$. Since $\Lambda(t)$ is a monotonic non-decreasing function, we only need to consider piecewise constant $\Lambda(t)$, say, λ_i in each interval $(t_i, t_{i+1}]$, $i = 1, 2, \dots, N - 1$. Then

$$\Lambda(b_i) - \Lambda(a_i) = \sum_{k=1}^{N-1} \lambda_k I(a_i < t_k \leq b_i).$$

We can create a latent variable

$$Z_{ij} \sim \text{Poisson}(\theta_{ij}), \quad \theta_{ij} = \lambda_j I(a_i < t_j \leq b_i).$$

Following the same approach as before, we can use an EM algorithm to find the maximum likelihood estimate subject to constraints $\lambda_j \geq 0, j = 1, 2, \dots, N - 1$.

Moreover, it is not difficult to incorporate covariate information in the Poisson process. For example we may assume

$$\Lambda(t|x) = \xi \Lambda(t) \exp(x\beta),$$

where $\xi \sim g(\xi)$ is a random frailty. Currently Diao et al. (2017) are studying this method. Yao et al. (2016) used a spline method to attach the same problem. Additional discussion on the panel count data can be found in the excellent monograph by Sun and Zhao (2013).

References

- Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113, 231–263.
- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72, 1–19.
- Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1), 91–117.
- Abadie, A., & Imbens, G. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of the American Statistical Association*, 29(1), 1–11.
- Abrevaya, J. (1999). Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics*, 93, 203–228.
- Abrevaya, J. (2003). Pairwise-difference rank estimation of the transformation model. *Journal of Business and Economics Statistics*, 21, 437–447.
- Aerts, M., Claeskens, G., Hens, N., & Molenberghs, G. (2002). Local multiple imputation. *Biometrika*, 89, 375–388.
- Aitchison, J., & Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika*, 57, 253–262.
- Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 813–828.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Cski (Eds.), *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2–8, 1971* (pp. 267–281). Budapest: Akadmai Kiad.
- Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1–10.
- Albert, P., Ratnasinghe, D., Tangrea, J., & Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154, 687–693.
- Alho, J. M. (1990a). Adjusting for nonresponse bias using logistic regression. *Biometrika*, 77, 617–624.
- Alho, J. M. (1990b). Logistic regression in capture-recapture models. *Biometrics*, 46, 623–635.
- Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.
- Andersen, E. (1970). Asymptotic properties of conditional maximum likelihood estimators. *JRSSB*, 32, 283–301.
- Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B*, 46(1), 1–30.
- Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19–35.

- Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, 66, 17–26.
- Anderson, J. A., & Philips, P. R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics*, 30, 22–31.
- Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Berlin: Springer.
- Angrist, J., & Krueger, A. B. (2001). *Instrumental variables and the search for identification: From supply and demand to natural experiments (No. w8456)*. Cambridge: National Bureau of Economic Research.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–455.
- Arnold, B. C., Castillo, E., & Sarabia, J. M. (2007). Distributions with generalized skewed conditionals and mixtures of such distributions. *Communications in Statistics: Theory and Methods*, 36, 1493–1503.
- Arnold, B. C., & Strauss, D. (1988). Bivariate distributions with exponential conditionals. *Journal of the American Statistical Association*, 83(402), 522–527.
- Arratia, R., Goldstein, L., & Kochman, F. (2015). Size bias for one and all. <https://arxiv.org/abs/1308.2729>.
- Asgharian, M., M'Lan, C. E., & Wolfson, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *JASA*, 97, 201–209.
- Asgharian, M., & Wolfson, D. B. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *Annals of Statistics*, 33, 2109–2131.
- Atkinson, A. C. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society: Series B*, 32, 323–353.
- Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26, 641–647.
- Azzalini, A. (2013). *The skew-normal and related families*. Cambridge: Cambridge University Press.
- Back, K., & Brown, D. P. (1992). GMM, maximum likelihood, and nonparametric efficiency. *Economics Letters*, 39, 23–28.
- Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, 85, 535–547.
- Bailey, K. R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *The Annals of Statistics*, 12, 730–736.
- Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33, 2297–2340.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. New York: Wiley.
- Bartlett, M. S. (1953). Approximate confidence intervals. II. *More than one unknown parameter*. *Biometrika*, 40, 306–317.
- Begg, C. B. (1994). Methodological issues in studies of the treatment, diagnosis, and etiology of prostate cancer. *Seminars in Oncology*, 21, 569–579.
- Begg, C., & Gray, R. (1987). Methodology for case-control studies with prevalent cases. *Biometrika*, 74, 191–195.
- Begg, C., & Mazumder, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50, 1088–1101.
- Begg, C. B., & Zhang, Z. F. (1994). Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiology Biomarkers and Prevention*, 3, 173–175.
- Begun, J. M., Hall, W. J., Huang, W. M., & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11, 432–452.
- Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2, 273–277.

- Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5, 445–463.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36, 192–236.
- Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24, 179–195.
- Betensky, R. A., Christian, C. K., Gustafson, M. L., Daley, J., & Zinner, M. J. (2006). Hospital volume versus outcome: An unusual example of bivariate association. *Biometrics*, 62, 598–604.
- Bhattacharya, P. K., Chernoff, H., & Yang, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Annals of Statistics*, 11, 505–511.
- Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, 10, 647–671.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation in semiparametric models*. Baltimore: Johns Hopkins University Press.
- Bickel, P. J., Nair, V. N., & Wang, P. C. (1992). Nonparametric inference under biased sampling from a finite population. *The Annals of Statistics*, 20, 853–878.
- Bickel, P. J., & Ritov, J. (1991). Large sample theory of estimation in biased sampling regression models. I. *The Annals of Statistics*, 19, 797–816.
- Bickel, P. J., & Ritov, Y. (1993). Efficient estimation using both direct and indirect observations. *Theory of Probability and its Applications*, 38, 194–213.
- Bickel, P. J., Ritov, Y., & Wellner, J. A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Annals of Statistics*, 19, 1316–1346.
- Boden, L. I., & Ozonoff, A. L. (2008). Capture-recapture estimates of nonfatal workplace injuries and illnesses. *Annals of Epidemiology*, 18, 500–506.
- Bohning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C., & Arnold, M. (2013). A generalization of Chaos estimator for Covariate information. *Biometrics*, 69, 1033–1042.
- Bohning, D., Rocchetti, I., Aflo, M., & Holling, H. (2016). Flexible ratio regression approach for zero-truncated capture-recapture counts. *Biometrics*, 72, 697–706.
- Boos, D. D., & Brownie, C. (1986). Testing for a treatment effect in the presence of nonresponders. *Biometrics*, 42, 191–197.
- Borchers, D. L., Buckland, S. T., & Zucchini, W. (2002). *Estimating animal abundance closed populations*. Berlin: Springer.
- Borgan, O., & Langholz, B. (1993). Nonparametric estimation of relative mortality from nested case-control studies. *Biometrics*, 49, 593–602.
- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.
- Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society: Series B*, 34, 216–217.
- Breslow, N. E. (1976). Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics*, 32, 409–416.
- Breslow, N. E. (1996). Statistics in epidemiology: The case-control study, Fisher lecture. *JASA*, 91, 14–28.
- Breslow, N. E. (2003). Are statistical contributions to medicine undervalued? *Biometrics*, 59, 1–8.
- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research. Volume I - The analysis of case-control studies* (Vol. 32). Lyon: IARC Scientific Publications.
- Breslow, N. E., & Day, N. E. (1986). *Statistical methods in cancer research. Volume II - The design and analysis of cohort studies* (Vol. 82). Lyon: IARC Scientific Publications.
- Breslow, N. E., Robins, J. M., & Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6, 447–455.
- Britt, H. I., & Luecke, R. H. (1973). The estimation of parameters in nonlinear, implicit models. *Technometrics*, 15, 233–247.
- Brookmeyer, R., & Gail, M. H. (1994). *AIDS epidemiology: A quantitative approach*. Oxford: Oxford University Press.
- Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, 66, 429–436.

- Cao, W. H., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*, 723–734.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement error in non-linear models: A modern perspective* (2nd ed.). Champan and Hall: Boca Raton.
- Casella, G., & Berger, R. L. (2008). *Statistical inference*. Duxbury advanced series.
- Cavanagh, C., & Sherman, R. P. (1998). Rank estimators for monotonic index models. *Journal of Econometrics*, *84*, 351–381.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*, 305–334.
- Chan, K. C. G. (2013). Survival analysis without survival data: connecting length-biased and case-control data. *Biometrika*, *100*, 764–770.
- Chan, N. H., Chen, S. X., Peng, L., & Yu, C. L. (2009). Empirical likelihood methods based on characteristic functions with applications to Lévy processes. *Journal of the American Statistical Association*, *104*(488), 1621–1630.
- Chan, G., & Qin, J. (2015). Rank-based testing of equal survivorship based on cross-sectional survival data with or without prospective follow-up. *Biostatistics*, *16*, 772–784.
- Chan, G., & Qin, J. (2016). Nonparametric maximum likelihood estimation of the multi-sample Wicksells corpuscle problem. *Biometrika*, *103*, 273–286.
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*, 783–791.
- Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, *45*, 427–438.
- Chatterjee, N., & Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, *92*, 399–418.
- Chatterjee, N., Chen, Y.-H., Maas, P., & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *JASA*, *111*, 107–117.
- Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2008). Generalized linear models incorporating population level information: An empirical likelihood based approach. *Journal of the Royal Statistical Society: Series B*, *70*, 311–328.
- Chaudhuri, S., Drton, M., & Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, *94*(1), 199–216.
- Chen, B., Li, P. F., Qin, J., & Yu, T. (2016). Using a monotonic density ratio model to find the optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, *111*, 861–874.
- Chen, B. J., Li, P. F. & Qin, J. (2017). Generalization of Heckman selection model to nonignorable nonresponse using call-back information. *Statistica Sinica*.
- Chen, B. J., & Qin, J. (2013). A new estimation with minimum trace of asymptotic covariance matrix for incomplete longitudinal data with a surrogate process. *Statistics in Medicine*, *32*, 4763–4780.
- Chen, B. J., & Qin, J. (2014). Use empirical likelihood to calibrate auxiliary information in partly linear monotone regression models. *Statistics in Medicine*, *10*, 1713–1722.
- Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, *63*, 413–421.
- Chen, H. Y. (2015). A note on convergence of an iterative algorithm for semiparametric odds ratio model. *Biometrika*, *102*, 747–751.
- Chen, H. Y., & Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *JASA*, *94*, 896–908.
- Chen, H. Y., Rader, D. E., & Li, M. (2015a). Likelihood inferences on semiparametric odds ratio model. *Journal of the American Statistical Association: Theory and Methods*, *110*, 1125–1135.
- Chen, J. (2016). Consistency of the MLE under mixture models. To appear in *Statistical Science*.
- Chen, J., Variyath, M. A., & Abraham, B. (2008). Adjusted empirical likelihood and its Properties. *Journal of Computational and Graphical Statistics*, *17*, 426–443.

- Chen, J., & Liu, Y. (2013). Quantile and quantile function estimations under density ratio model. *The Annals of Statistics*, 41, 1669–1692.
- Chen, J., & Qin, J. (1993). Empirical likelihood method in finite population and the effective usage of auxiliary information. *Biometrika*, 80, 107–116.
- Chen, J., Sitter, R. R., & Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230–237.
- Chen, K. (2001). Generalized case cohort sampling. *JRSSB*, 63, 791–809.
- Chen, K., & Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika*, 86, 755–764.
- Chen, K., Sun, L. Q., & Tong, X. W. (2012). Analysis of cohort survival data with transformation model. *Statistica Sinica*, 22, 489–508.
- Chen, K., Yao, Y., & Zhou, C. X. (2014). Regression analysis with response-biased sampling. *Statistica Sinica*. To appear.
- Chen, M. H., Dey, D. K., & Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94, 1172–1186.
- Chen, S., Witten, D., & Shojaie, A. (2015b). Selection and estimation for mixed graphical models. *Biometrika*, 102, 47–64.
- Chen, S. X., & Lloyd, C. J. (2000). A nonparametric approach to the analysis of two-stage mark-recapture experiments. *Biometrika*, 87, 633–649.
- Chen, S. X., Qin, J., & Tang, C. Y. (2013). Mann-Whitney test with adjustments to pre-treatment variables for missing values and observational study. *JRSSB*, 75, 81–102.
- Chen, X. H., Dempster, A. P., & Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457–469.
- Chen, Y. G., & Liu, J. S. (2007). Sequential Monte Carlo methods for permutation tests on truncated data. *Statistica Sinica*, 17, 857–872.
- Chen, Y. Q. (2010). Semiparametric regression in size-biased sampling. *Biometrics*, 66, 149–158.
- Chen, Z. H., Bai, Z. D., & Sinha, B. (2004). Ranked set sampling: Theory and applications Berlin: Springer.
- Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89, 81–87.
- Cheng, J., Qin, J., & Zhang, B. (2009). Semiparametric estimation and inference for distributional and general treatment causal effects. *JRSSB*, 71, 881–904.
- Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. (2013). *Stochastic geometry and its applications* (3rd ed.). Chichester: Wiley.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
- Copas, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. *Biometrika*, 62, 701–704.
- Copas, J. B., & Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89, 315–331.
- Copas, J., & Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4), 459–513.
- Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85, 967–972.
- Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica*, 49, 1289–1316.
- Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, 51, 765–782.
- Cosslett, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica*, 55, 559–585.
- Cosslett, S. R. (2013). Efficient semiparametric estimation for endogenously stratified regression via smoothed likelihood. *Journal of Econometrics*, 177, 116–129.
- Cox, D. R. (1961). Test of separate families of hypothesis. In: *Proceedings of the 4th Berkeley Symposium* (Vol. 1, pp. 105–123).

- Cox, D. R. (1962). Further results on tests of separate families of hypothesis. *Journal of the Royal Statistical Society Series B*, 24, 406–424.
- Cox, D. R. (1969). Some sampling problems in technology. In N. L. Johnson & H. Smith Jr. (Eds.), *New developments in survey sampling* (pp. 506–527). New York: Wiley.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34, 187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.
- Cox, D. R., & Isham, V. (1980). *Point processes*. London: Chapman and Hall.
- Cox, D. R., & Miller, H. D. (1977). *The theory of stochastic processes*. London: Chapman & Hall.
- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data* (Vol. 21). CRC Press.
- Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B*, 49, 1–39.
- Csiszar, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12, 768–793.
- Csiszar, I., & Shield, P. C. (2004). *Information theory and statistics: A tutorial*. Foundations and Trends in Communications and Information Theory Published, sold and distributed by: now Publishers Inc. PO Box 1024, Hanover, MA 02339.
- Csiszar, I., & Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, 1, 205–237.
- Daskalakis, C., Diakonikolas, I., & Servedio, R. A. (2011). Learning poisson binomial distributions. [arXiv:1107.2702](https://arxiv.org/abs/1107.2702).
- Daudin, J. J. (1980). Partial association measures and an application to qualitative regression. *Biometrika*, 67, 581–590.
- Davidov, O., & Iliopoulos, G. (2009). On the existence and uniqueness of the NPMLE in biased sampling models. *Journal of Statistical Planning and Inference*, 139, 176–183.
- Davidov, O., & Iliopoulos, G. (2013). Convergence of Luo and Tsais iterative algorithm for estimation in proportional likelihood. *Biometrika*.
- Davies, P., & Phillips, A. J. (1988). Nonparametric tests of population differences and estimation of the probability of misidentification with unidentified paired data. *Biometrika*, 75, 753–760.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41, 1–31.
- De Carvalho, M., & Davison, A. C. (2014). Spectral density ratio models for multivariate extremes. *Journal of The American Statistical Association*, 109, 764–776.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39, 1–38.
- Devile, J. C., & Sarndai, C. E. (1992). *Calibration estimation in survey sampling*. JASA, 87, 376–382.
- Devlin, B., Roeder, K., & Wasserman, L. (2000). Genomic control for association studies: a semi-parametric test to detect excess-haplotype sharing. *Biostatistics*, 1, 369–387.
- Dewanji, A., & Kalbfleisch, J. D. (1987). Estimation of sojourn time distributions for cyclic semi-Markov processes in equilibrium. *Biometrika*, 74, 281–288.
- Diao, G. Q., Ning, J., & Qin, J. (2012, June). Maximum likelihood estimation for semiparametric density ratio model. *The International Journal of Biostatistics*. 8(1), Article 16, ISSN (Online) 1557–4679. doi:[10.1515/1557-4679.1372](https://doi.org/10.1515/1557-4679.1372).
- Diao, G. Q., Qin, J., & Yuan, A. (2017). Maximum semiparametric likelihood estimation of panel count data using an EM algorithm. Manuscript.
- DiCiccio, T. J., Hall, P., & Romano, J. P. (1991). Bartlett adjustment for empirical likelihood. *Annals of Statistics*, 19, 1053–1061.
- DiCiccio, T. J., & Romano, J. (1989). On adjustments for the signed root of the empirical likelihood ratio statistic. *Biometrika*, 7, 447–456.
- Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *JASA*, 81, 328–336.

- Doksum, K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *The Annals of Statistics*, 15, 325–345.
- Duan, X., Qin, J., & Wang, Q. H. (2010). Optimal estimation in surrogate outcome regression problems. *Canadian Journal of Statistics*, 38, 633–646.
- Dykstra, R., Kocher, S., & Robertson, T. (1995). Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association*, 90(431), 1034–1040.
- Efron, B. (1967). The two sample problem with censored data. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 831–853). Berkeley, CA: University of California Press.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70, 892–898.
- Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics*, 9, 139–172.
- Efron, B., & Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, 94, 824–834.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton: CRC Press.
- Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6), 2431–2461.
- Eguchi, S., & Copas, J. (2002). A class of logistic type discriminant functions. *Biometrika*, 89, 1–22.
- Eguchi, S., & Copas, J. (2006). Interpreting kullback-leibler divergence with the neyman-pearson lemma. *Journal of Multivariate Analysis*, 97, 2034–2040.
- Elston, R. C., & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, 21, 523–542.
- Epstein, M. P., & Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *The American Journal of Human Genetics*, 73, 1316–1329.
- Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, 66, 27–32.
- Farewell, V. T. (1982). A note on regression analysis of ordinal data with variability of classification. *Biometrika*, 69, 533–538.
- Feller, W. (1965). *An introduction to probability theory and its applications*. New York: Wiley.
- Fewster, R. M., & Jupp, P. E. (2009). Inference on population size in binomial detectability models. *Biometrika*, 96, 805–820.
- Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Human Genetics*, 6(1), 13–25.
- Flehinger, B. J., Reiser, B., & Yashchin, E. (1998). Survival with competing risks and masked causes of failures. *Biometrika*, 151–164.
- Fleming, T. R., & Harrington, D. P. (1994). *Counting processes and survival analysis*. New York: Wiley.
- Fokianos, K., Kedem, B., Qin, J., & Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56–65.
- Friedman, J., Hastie, T., Hofling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1, 302–332.
- Gail, M. H., & Benichou, J. (2000). *Encyclopedia of epidemiologic methods*. New York: Wiley.
- Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI*, 81, 1879–1886.
- Gan, L., & Jiang, J. (1999). A test for global maximum. *Journal of the American Statistical Association*, 94, 847–854.
- Gao, L., et al. (2011). Length bias correction for RNA-seq gene set analyses. *Bioinformatics*, 27, 662–669.
- Gilbert, P., Lele, S., & Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86, 27–43.

- Gill, R. D., Vardi, Y., & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, 16, 1069–1112.
- Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208–1212.
- Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63, 277–284.
- Godambe, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika*, 67, 155–162.
- Godambe, V. P. (1984). On ancillarity and Fisher information in the presence of a nuisance parameter. *Biometrika*, 71, 626–629.
- Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72, 419–428.
- Godambe, V. P., & Thompson, M. E. (1974). Estimating equations in the presence of a nuisance parameter. *The Annals of Statistics*, 2, 568–571.
- Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Journal Statistical Review, Revue Internationale de Statistique*, 127–138.
- Godambe, V. P., & Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal of Statistical Planning and Inference*, 22, 137–172.
- Godambe, V. P., & Vijayan, K. (1996). Optimal estimation for response-dependent retrospective sampling. *Journal of the American Statistical Association*, 91(436), 1724–1734.
- Godley, P. A., & Schell, M. J. (1999). Adjusted odds ratios under nondifferential misclassification: Application to prostate cancer. *Journal of Clinical Epidemiology*, 52, 129–136.
- Golan, A., Judge, G., & Miller, D. (1996). *Maximum entropy econometrics: Robust estimation with limited data*. New York: Wiley.
- Goldstein, L., & Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *The Annals of Statistics*, 20, 1903–1928.
- Good, P. I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics*, 35, 483–489.
- Gould, A., & Lawless, J. F. (1988). Consistency and efficiency of regression coefficient estimates in location-scale models. *Biometrika*, 75, 535–540.
- Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models*. I and II. Translated by Quang Vuong. Cambridge: Cambridge University Press.
- Gourieroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, 50, 63–80.
- Gourieroux, C., Monfort, A., & Trogon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52, 681–700.
- Grenander, U. (1956a). On the theory of mortality measurement. I. *Skand. Aktuarieridskr.*, 39, 70–96.
- Grenander, U. (1956b). On the theory of mortality measurement. II. *Skand. Aktuarieridskr.*, 39, 125–153.
- Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints estimators, algorithms and asymptotics*. New York: Cambridge University Press.
- Groeneboom, P., Maathuis, M. H., & Wellner, J. A. (2008a). Current status data with competing risks: Consistency and rates of convergence of the MLE. *The Annals of Statistics*, 36, 1031–1063.
- Groeneboom, P., Maathuis, M. H., & Wellner, J. A. (2008b). Current status data with competing risks: Limiting distribution of the MLE. *Annals of Statistics*, 36, 1064–1089.
- Guan, Z., & Qin, J. (2016). Empirical likelihood method for non-ignorable missing data problems. To appear in *Lifetime Data Analysis*.
- Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, 12, 971–988.

- Hahn, J. (1998). On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–332.
- Hainmueller, J. (2011). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20, 25–46.
- Hall, A. R. (2005). *Generalized method of moments. Advanced texts in econometrics*. Oxford: Oxford University Press.
- Hall, P. (1900). Pseudo-likelihood theory for empirical likelihood. *Annals of Statistics*, 18, 121–140.
- Hall, P., & La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58, 109–127.
- Hall, P., & Titterington, D. M. (1984). Efficient nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society B*, 46, 465–473.
- Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35, 303–316.
- Han, P., & Lawless, J. F. (2016). Discussion of "Constrained maximum likelihood estimation for model calibration using summary-level information from external big data source" by Chatterjee, Chen, Maas and Carroll. *Journal of the American Statistical Association*, 111, 118–121.
- Han, P., & Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100, 417–430.
- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95, 481–488.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.
- Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, 14, 262–280.
- Hartely, H. O., & Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547–557.
- Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In L. LeCam & R. A. Olshen (Eds.), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (pp. 807–810). Monterey, CA: Wadsworth.
- Hasminskii, R. Z., & Ibragimov, I. A. (1983). In K. Ito & J. V. Prohorov (Eds.), *On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter*. Lecture notes in mathematics (Vol. 1021, pp. 195–229). New York: Springer.
- Hausman, J. A., & Wise, D. A. (1981). Stratification on endogenous variables and estimation: The Gary income maintenance experiment. *Structural analysis of discrete data with econometric applications* (pp. 365–391). Cambridge: MIT Press.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Hedges, L. V., & Olkins, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press. INC.
- Henmi, M., & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating equations. *Biometrika*, 91, 929–941.
- Hinkley, D. V. (1973). Two-sample tests with unordered pairs. *Journal of the Royal Statistical Society. Series B*, 35, 337–346.
- Hjort, N. L., & Pollard, D. (1997). Asymptotics for minimisers of convex processes. <http://www.stat.yale.edu/~pollard/>.
- Hoeffding, W. (1951). A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, 22, 558–566.
- Hogan, H. (1993). The 1990 post-enumeration survey: Operations and results. *JASA*, 88, 1047–1060.
- Holt, J. D., & Prentice, R. L. (1974). Survival analyses in twin studies and matched pair experiments. *Biometrika*, 61, 17–30.
- Hosmer, D. W. Jr. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29, 761–770.
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

- Hsieh, D. A., Manski, C. F., & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, 80, 651–662.
- Hu, Z., Follmann, D. A., & Qin, J. (2011). Dimension reduced kernel estimation for distribution function with incomplete data. *Journal of Statistical Planning and Inference*, 141, 3084–3093.
- Hu, Z., Follmann, D. A., & Qin, J. (2012). Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *JASA*, 107, 247–257.
- Huang, A. (2014a). Joint estimation of the mean and error distribution in generalized linear models. *JASA*, 109, 186–196.
- Huang, C. Y., & Qin, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *JASA*, 107, 946–957.
- Huang, C. Y., & Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100, 877–888.
- Huang, C. Y., Ning, J., & Qin, J. (2015). Semiparametric likelihood inference for left-truncated and right censored data. *Biostatistics*, 16, 785–798.
- Huang, C. Y., Qin, J., & Tsai, H. T. (2016). Efficient estimation of the Cox model with auxiliary subgroup survival information. *JASA*, 514, 787–799.
- Huang, C. Y., Qin, J., & Wang, M. C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics*, 66, 39–49.
- Huang, C. Y., Qin, J., & Zou, F. (2007). Empirical likelihood-based inference in a genetic mixture model. *Canadian Journal of Statistics*, 35, 563–574.
- Huang, J. (1997). Asymptotic properties of the NPMLE of a distribution function based on ranked set samples. *The Annals of Statistics*, 25, 1036–1049.
- Huang, J. (2002). A note on estimating a partly linear model under monotonicity constraint. *Journal of Statistical Planning and Inference*, 107, 343–351.
- Huang, J., & Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics* (pp. 123–169). Berlin: Springer.
- Huang, Y. (2014b). Corrected score with sizable covariate measurement error: Pathology and remedy. *Statistica Sinica*, 24, 357–374.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–233).
- Hudgens, M. G., Satten, G. A., & Longini, I. M. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, 57, 74–80.
- Imbens, G., & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61, 655–680.
- Imbens, G., & Lancaster, T. (1996). Efficient estimation and stratified sampling. *Journal of Econometrics*, 74, 289–318.
- Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25, 305–327.
- Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64, 555–574.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.
- Imbens, G. W., & Spady, R. H. (2001). The performance of empirical likelihood and its generalizations. *Manuscript*.
- Imbens, G. W., Spady, R. H., & Johnson, P. (1998). Information theoretic approaches. *Manuscript*.
- Ibragimov, I. A., Has'minskii, It., & Z., (1981). *Statistical estimation. Asymptotic theory*. New York: Springer.
- Ireland, C. T., & Kullback, S. (1968a). Contingency tables with given marginals. *Biometrika*, 55, 179–188.

- Ireland, C. T., & Kullback, S. (1968b). Minimum discrimination information estimation. *Biometrics*, 24, 707–713.
- Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109–117.
- Jager, L., & Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35, 2018–2053.
- Jagers, P., Oden, A., & Trulsson, L. (1985). Post-stratification and ratio estimation: Usages of auxiliary information in survey sampling and opinion polls. I. *International Statistical Review*, 53, 221–238.
- Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, 10, 479–484.
- Jewell, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11–21.
- Jewell, N., & Kalbfleisch, J. (2004). Maximum likelihood estimation of ordered multinomial parameters. *Biostatistics*, 2, 291–306.
- Jewell, N., Van der Lann, M., & Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, 90, 183–197.
- Ji, S., Ning, J., Qin, J., & Follmann, D. (2017). Conditional independence test by generalized Kendalls tau with generalized odds ratio. To appear in *Statistical Methods in Medical Research*.
- Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, 5, 195–199.
- Johnson, R. A., Verrill, S., & Moore, D. H. (1987). Two-sample rank tests for detecting changes that occur in a small proportion of the treated population. *Biometrics*, 43, 641–655.
- Jupp, P. E., Kim, P. T., Koo, J.-Y., & Wiegert, P. (2003). The intrinsic distribution and selection bias of long-period cometary orbits. *Journal of the American Statistical Association*, 98, 515–521.
- Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association*, 73, 167–170.
- Kalbfleisch, J. D., & Lawless, J. F. (1987). Likelihood analysis of multistate models for disease incidence and mortality. *Statistics in Medicine*, 1, 149–160.
- Kalbfleisch, J. D., & Lawless, J. F. (1988). Estimation of reliability in field performance studies. *Technometrics*, 30, 365–388.
- Kalbfleisch, J. D., & Lawless, J. F. (1989). Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, 84, 360–372.
- Kalbfleisch, J. D., & Lawless, J. F. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1, 19–32.
- Kalbfleisch, J. D., & Prentice, R. L. (1973). Marginal likelihood based on Coxs regression and life model. *Biometrika*, 60, 267–278.
- Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.
- Kalbfleisch, J. D., & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *Journal of the Royal Statistical Society. Series B*, 32, 175–208.
- Kagan, A. M., Rao, C. R., & Linnik, Y. V. (1973). *Characterization problems in mathematical statistics. Wiley series in probability & mathematical statistics*. New York: Wiley.
- Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22, 523–539.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *American Statistical Association*, 53, 457–481.
- Kaufman, G. M., Balcer, Y., & Krutik, D. (1975). A probabilistic model of oil and gas discovery. A chapter. In J.W. Harbaugh, J.C. Davis & J. Wendebourg (Eds.), *Computing risk for oil prospects: Principles and programs*.

- Kay, R., & Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74(3), 495–501.
- Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27, 887–906.
- Kim, J., & Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18, 191–219.
- Kim, J. K., & Shao, J. (2013). *Statistical methods for handling incomplete data*. Boca Raton: Chapman & Hall/CRC.
- Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2004). *Loss models: From data to decisions*. New York: Wiley.
- Koenker, R. (2005). *Quantile Regression. Econometric society monographs*. Cambridge: Cambridge University Press.
- Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.
- Konishi, S., & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83, 875–890.
- Korn, E. L. (1984). Kendall's tau with a blocking variable. *Biometrics*, 40(1), 209–214.
- Kosorok, M. R. (2008a). *Introduction to empirical processes and semiparametric inference*. New York: Springer.
- Kosorok, M. R. (2008b). Bootstrapping the Grenander estimator. *IMS Collections beyond parametrics in interdisciplinary research: Festschrift in Honor of Professor Pranab K. Sen* (Vol. 1, pp. 282–292).
- Kou, S. G., & Ying, Z. (1996). Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica*, 6, 809–829.
- Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.
- Kullback, S. (1968). Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39, 1236–1243.
- Kvam, P. H., & Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association*, 89, 526–537.
- Lagakos, S. W., Barraj, L. M., & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75, 515–523.
- Lai, T. L., & Ying, Z. L. (1994). A missing information principle and M-estimators in regression analysis with censored and truncated data. *Annals of Statistics*, 22, 1222–1255.
- Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805–811.
- Lancaster, T. (1992). *The Econometric analysis of transition data*. Cambridge: Cambridge University Press.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95, 391–413.
- Lancaster, T., & Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, 71, 145–160.
- Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121, 185–199.
- Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84, 2363–2367.
- Lange, K. (2010). *Numerical analysis for statisticians* (2nd ed.). Berlin: Springer.
- Lange, K., & Elston, R. C. (1975). Extensions to pedigree analysis. *Human Heredity*, 25, 95–105.
- Langholz, B., & Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science*, 11, 35–53.
- Langholz, B., & Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology*, 131(1), 169–176.
- Langholz, B., & Thomas, D. C. (1991). Efficiency of cohort sampling designs: some surprising results. *Biometrics*, 1563–1571.

- Lauder, I. J. (1977). Tracing quantitative measurements on human chromosomes in family studies. *Annals of Human Genetics*, 41, 77–86.
- Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82, 808–815.
- Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (Vol. 2). Wiley series in probability and statistics. New York: Wiley.
- Lawless, J. F., Kalbfleisch, J. D., & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society B*, 61, 413–438.
- Lee, A. J., Scott, A. J., & Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 93, 385–397.
- Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York: Wiley.
- Lehmann, E. L., & D'abreia, H. J., (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.
- Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses. Springer texts in statistics*. New York: Springer.
- Leigh, G. M. (1988). A comparison of estimates of natural mortality from fish tagging experiments. *Biometrika*, 75, 347–354.
- Lemdani, M., & Pons, O. (1995). Tests for genetic linkage and homogeneity. *Biometrics*, 1033–1041.
- Leung, D., & Qin, J. (2006). Semi-parametric inference in a bivariate (multivariate) mixture model. *Statistica Sinica*, 16, 153–164.
- Li, G., & Qin, J. (1998). Semiparametric likelihood based inferences for biased and truncated data when total sample size is known. *Journal of the Royal Statistical Society: Series B*, 60, 243–254.
- Li, G., & Qin, J. (2006). Analysis of two-sample truncated data using generalized logistic model. *Journal of Multivariate Analysis*, 97, 675–697.
- Li, H., & Gail, M. H. (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Human Heredity*, 73, 159–173.
- Li, H., & Yin, G. (2009). Generalized method of moments estimation for linear regression with clustered failure time data. *Biometrika*, 96, 293–306.
- Li, P. F., & Qin, J. (2011). A new nuisance-parameter elimination method with application to the unordered homologous chromosome pairs problem. *JASA*, 496, 1476–1484.
- Liang, K. Y. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika*, 70, 607–612.
- Liang, K. Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika*, 74, 259–264.
- Liang, K. Y., & Rathouz, P. J. (1999). Hypothesis testing under mixture models: Application to genetic linkage analysis. *Biometrics*, 55, 65–74.
- Liang, K. Y., & Qin, J. (2000). Regression analysis under non-standard situations: A pairwise pseudo-likelihood approach. *Journal of the Royal Statistical Society. Series B*, 62, 773–786.
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.
- Lim, J., Wang, X. L., & Choi, W. (2009). Maximum likelihood estimation of ordered multinomial probabilities by geometric programming. *Computational Statistics and Data Analysis*, 53, 889–893.
- Lin, Y., & Chen, K. (2013). Efficient estimation of the censored linear regression model. *Biometrika*, 100(2), 525–530.
- Lin, D., & Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33, 256–265.
- Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 296(1427), 639–662.
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 221–239.

- Lindsay, B. G. (1995, January). *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics: NSF-CBMS regional conference series in probability and statistics. Hayward.
- Lipsitz, S. R., Zhao, L. P., & Molenberghs, G. (1998). A semiparametric method of multiple imputation. *Journal of the Royal Statistical Society B*, 60, 127–144.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *JASA*, 88, 125–134.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Little, R. J. A., & Wu, M. M. (1991). Models for contingency table with known margins when target and sampled populations differ. *JASA*, 6, 87–95.
- Liu, C. H. (2000). Estimation of discrete distributions with a class of simplex constraints. *Journal of the American Statistical Association*, 95, 109–120.
- Liu, D., Zheng, Y., Prentice, R. L., & Hsu, L. (2014). Estimating risk with time-to-event data: An application to the Womens Health Initiative. *Journal of American Statistical Association*, 109, 514–524.
- Liu, H., & Qin, J. (2017). Semiparametric probit models with univariate or bivariate current status data. To appear in *Biometrics*.
- Liu, Y., Li, P., & Qin, J. (2016). Maximum empirical likelihood estimation for abundance in a closed population from capture-recapture data. To appear in *Biometrika*.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 44, 226–233.
- Luo, X., & Tsai, W. Y. (2009). Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. *Biometrika*, 96(4), 873–886.
- Luo, X., & Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika*, 99, 211–222.
- Luo, X., & Tsai, W. Y. (2015). Moment-type estimators for the proportional likelihood ratio model with longitudinal data. *Biometrika*, 102, 121–134.
- Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices of the Royal Astronomical Society*, 155, 95–118.
- Ma, Y., Genton, M. G., & Tsiatis, A. A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *Journal of the American Statistical Association*, 100, 980–989.
- Ma, Y., Kim, M., & Genton, M. G. (2013). Semiparametric efficient and robust estimation of an unknown symmetric population under arbitrary sample selection bias. *Journal of the American Statistical Association*, 108, 1090–1104.
- Ma, Y., & Wang, Y. (2012). Efficient semiparametric estimation for mixture data. *Electronic Journal of Statistics*, 6, 710–737.
- Machado, M. P. (2004). A consistent estimator for the binomial distribution in the presence of incidental parameters: An application to patent data. *Journal of Econometrics*, 119, 73–98.
- Mandel, M. (2007). Nonparametric estimation of a distribution function under biased sampling and censoring - a unified approach. In R. Liu, W. Strawderman & C.-H Zhang (Eds.), *Complex datasets and inverse problems, the IMS lecture notes monograph series* (Vol. 54, pp. 224–238). Institute of Mathematical Statistics.
- Mandel, M., & Ritov, Y. (2010). The accelerated failure time model under biased sampling. *Biometrics*, 66, 1306–1308.
- Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3, 205–228.
- Manski, C., & McFadden, D. (1981). Structure analysis of discrete data with economic applications. *Econometric models of probabilistic choice* (pp. 198–272). Cambridge: MIT Press.
- Mao, C. X., & Lindsay, B. G. (2007). Estimating the number of classes. *The Annals of Statistics*, 35, 917–930.
- Marchenko, Y. V., & Genton, M. G. (2012). A Heckman selection-t model. *Journal of the American Statistical Association*, 107, 304–317.

- Mazumdar, M., & Jefferson, T. R. (1983). Maximum likelihood estimates for multinomial probabilities via geometric programming. *Biometrika*, 70, 257–261.
- McClean, S., & Devine, C. (1995). A nonparametric maximum likelihood estimator for incomplete renewal data. *Biometrika*, 82, 791–803.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, 109–142.
- McCullagh, P. (1984). On the elimination of nuisance parameters in the proportional odds model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 250–256.
- McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman and Hall.
- McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, 53, S13–S29.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 995–1026.
- McIntosh, M. W., & Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, 58, 657–664.
- McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York: Wiley.
- McLeish, D. L., & Small, C. G. (1992). A projected likelihood function for semiparametric models. *Biometrika*, 79(1), 93–102.
- Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, 14, 2143–2160.
- Miao, W., Ding, P., & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111, 1673–1683.
- Molanes Lopez, E. M., Keilegom, I. V., & Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics*, 36, 413–432.
- Murcray, C. E., Lewinger, J. P., & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169(2), 219–226.
- Murphy, S. A., & Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95, 449–465.
- Murray, G. D., & Titterington, D. M. (1978). Estimation problems with data from a mixture. *Applied Statistics*, 325–334.
- Nagelkerke, N. J., Borgdorff, M. W., & Kim, S. J. (2001). Logistic discrimination of mixtures of M. tuberculosis and non-specific tuberculin reactions. *Statistics in Medicine*, 20, 1113–1124.
- Nair, V. N., & Wang, P. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics*, 31, 423–436.
- Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77, 127–137.
- Nan, B., & Wellner, J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica*, 23, 1155.
- Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *JRSS(A)*, 135, 370–384.
- Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.
- Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12, 217–229.
- Newey, W. K., & Smith, R. J. (2001). *Asymptotic bias and equivalence of GMM and GEL estimators*. Department of Economics: University of Bristol.
- Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1–32.
- Niemiro, W. (1992). Asymptotics for M-estimators defined by convex minimization. *The Annals of Statistics*, 20, 1514–1533.
- Ning, J., Qin, J., & Shen, Y. (2014). Score estimating equations from embedded likelihood functions under accelerated failure time. *JASA*, 109, 1625–1635.
- Ning, Y., Zhao, T., & Liu, H. (2017). A likelihood ratio framework for high dimensional semiparametric regression. To appear in *Annals of Statistics*.

- Oakes, D. (1981). Survival times: aspects of partial likelihood. *International Statistical Review*, 49, 235–252.
- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 697–715). Berkeley, CA: University of California Press.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237–249.
- Owen, A. B. (1990). Empirical likelihood confidence regions. *Annals of Statistics*, 18, 90–120.
- Owen, A. B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19, 1725–1747.
- Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association*, 90, 516–521.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton: Chapman and Hall.
- Pakes, A. G., Sapatinas, T., & Fosam, E. B. (1996). Characterizations, length-biasing, and infinite divisibility. *Statistical Papers*, 37, 53–69.
- Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179–189.
- Pearson, K., & Lee, A. (1901). On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society of London*, A, 195, 79–150.
- Peng, L., & Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 482, 637–649.
- Pepe, S. M. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series*. New York: Oxford University Press.
- Peyhardi, J., Trottier, C., & Guedon, Y. (2015). A new specification of generalized linear models for categorical responses. *Biometrika*, 102, 889–906.
- Pollard, D. (1984). *Convergence of stochastic processes*. Berlin: Springer.
- Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhya Series A*, 31, 23–36.
- Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 167–179.
- Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73, 1–11.
- Prentice, R. L., & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, 65, 153–158.
- Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.
- Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14, 53–68.
- Qin, J. (1992). Empirical likelihood and semiparametric models. University Waterloo Ph.D. thesis.
- Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21, 1182–1196.
- Qin, J. (1998a). Semiparametric likelihood based method for goodness of fit test and estimation in upgraded mixture models. *The Scandinavian Journal of Statistics*, 25, 681–691.
- Qin, J. (1998b). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85, 619–630.
- Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics*, 27, 1368–1384.
- Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, 87, 484–490.
- Qin, J., & Follmann, D. (2014). Semiparametric maximum likelihood inference by using number of failed contact attempts to adjust for non-ignorable non-response. *Biometrika*, 101, 985–991.
- Qin, J., Garcia, T. P., Ma, Y. Y., Tang, M. X., Marderz, K., & Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *Annals of Applied Statistics*, 8, 1182–1208.
- Qin, J., & Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300–325.

- Qin, J., & Leung, D. (2005). A semiparametric two-component “compound” mixture model and its application to estimating malaria attributable fractions. *Biometrics*, 61, 456–464.
- Qin, J., & Liang, K. Y. (1999). Generalized odds ratio model and pairwise conditional likelihood. Unpublished manuscript.
- Qin, J., & Liang, K. Y. (2011). Hypothesis testing in a mixture case-control model. *Biometrics*, 67, 182–193.
- Qin, J., Ning, J., Liu, H., & Shen, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *JASA*, 96, 1434–1449.
- Qin, J., Shao, J., & Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *JASA*, 103, 797–810.
- Qin, J., & Wang, M. C. (2001). Semiparametric analysis of truncated data. *Lifetime Data Analysis*, 7, 225–242.
- Qin, J., Yu, T., Li, P. F., Liu, H., & Chen, B. J. (2017a). Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference. Submitted manuscript.
- Qin, J., & Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609–618.
- Qin, J., & Zhang, B. (2005). Marginal likelihood, conditional likelihood and empirical likelihood: Connections and applications. *Biometrika*, 92, 251–270.
- Qin, J., & Zhang, B. (2007). Empirical likelihood-based inference in missing response problems and its application in observational studies. *Journal of Royal Statistical Society, Series B*, 69, 101–122.
- Qin, J., & Zhang, B. (2008). Empirical likelihood-based difference-in-differences estimators. *Journal of Royal Statistical Society, Series B*, 70, 329–349.
- Qin, J., & Zhang, B. (2010). Best combination of multiple diagnostic tests for screening purposes. *Statistics in Medicine*, 29, 2905–2919.
- Qin, J., & Zhang, B. (2011). Optimal estimating functions in incomplete data and length biased sampling data problems. *Canadian Journal of Statistics*, 39, 510–518.
- Qin, J., Zhang, B., & Leung, D. (2009). Empirical likelihood in missing data problems. *JASA*, 104, 1492–1503.
- Qin, J., Zhang, B., & Leung, D. (2017b). Efficient augmented inverse probability weighted estimation in missing data problems. *Journal of Business & Economic Statistics*, 35, 86–97.
- Qin, J., Zhang, H., Landi, M. T., Caporaso, N. E., & Yu, K. (2016). A hybrid parametric and empirical likelihood model for evaluating interactions in case-control studies. *Statistics and Its Interface*, 9, 147–158.
- Qin, J., Zhang, H., Li, P. F., Albanes, D., & Yu, K. (2015). Using covariate specific disease prevalence information to increase the power of case-control study. *Biometrika*, 102, 169–180.
- Qu, A., Lindsay, B. G., & Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823–836.
- Quesenberry, C. P., & Jewell, N. P. (1986). Regression analysis based on stratified samples. *Biometrika*, 73, 605–614.
- Rao, B. P. (1969). Estimation of a unimodal density. *Sankhya Series A*, 31, 23–36.
- Rao, C. R. (1973). *Linear statistical inference and its application*. New York: Wiley.
- Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In G. P. Patil (Ed.), *Classical and contagious discrete distributions* (pp. 320–332). Calcutta: Pergamon Press and Statistical Publishing Society.
- Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In A. C. Atkinson & S. F. Fienberg (Eds.), *A celebration of statistics*. Berlin: Springer.
- Rao, J. N. K., Kovar, J. G., & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 826–831.
- Ren, J. J., & Zhou, M. (2011). Full likelihood inferences in the Cox model: An empirical likelihood approach. *Annals of the Institute of Statistical Mathematics*, 63, 1005–1018.

- Rabinowitz, D. (1997). A note on efficient estimation from case-control data. *Biometrika*, 84, 486–488.
- Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18, 303–328.
- Robert, C. Y. (2013). Automatic declustering of rare events. *Biometrika*, 100, 587–606.
- Robertson, T., Wright, F. T., & Dykstra, R. (1988). *Order restricted statistical inference*. Chichester, New York, Brisbane, Toronto, Singapore: Wiley.
- Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48, 479–495.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology*. Philadelphia: Lippencott-Raven.
- Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, 54, 221–226.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B., & van der Laan, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4, 1–41.
- Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *Annals of Statistics*, 23, 1160–1174.
- Samaniego, F. J., & Jones, L. E. (1981). Maximum likelihood estimation for a class of multinomial distributions arising in reliability. *Journal of the Royal Statistical Society B*, 43, 46–52.
- Samuelson, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84, 379–394.
- Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 42, 142–152.
- Sanathanan, L. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association*, 72, 669–672.
- Santner, T. J., & Duffy, D. E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73, 755–758.
- Sarndal, C. E., Swensson, B., & Wretman, J. (1991). *Model assisted survey sampling*. New York: Springer.
- Satten, G. A., & Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55, 207–210.
- Satten, G. A., & Epstein, M. P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology*, 27, 192–201.
- Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussions). *Journal of the American Statistical Association*, 101, 1619–1637.
- Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92, 31–46.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35, 634–672.
- Scott, A. J., & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57–71.
- Self, S. G., & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16, 64–81.
- Sham, P. (1998). *Statistics in human genetics*. London: Arnold.
- Shao, J. (2003). *Mathematical statistics*. Berlin: Springer.

- Sheehy, A. (1988). Kullback-Leibler constrained estimation of probability measures. Stanford technical report No. 137.
- Shen, P. S. (2007). A general semiparametric model for left-truncated and right-censored data. *Journal of Nonparametric Statistics*, 19, 113–129.
- Shen, P. S. (2009). Semiparametric analysis of survival data with left truncation and right censoring. *Computational Statistics and Data Analysis*, 53, 4417–4432.
- Sigman, K. (2009). *Lecture notes*. <http://www.columbia.edu/~ks20/stochastic-I/stochastic-I.html>.
- Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43, 310–313.
- Simon, R. (1980). Length-biased sampling in etiologic studies. *American Journal of Epidemiology*, 111, 444–452.
- Small, C. G., & McLeish, D. L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, 76, 693–703.
- Small, C. G., & McLeish, D. L. (1994). *Hilbert space methods in probability and statistical inference*. New York: Wiley.
- Small, C. G., & Murdoch, D. J. (1993). Nonparametric Neyman-Scott problems: Telescoping product methods. *Biometrika*, 80, 763–769.
- Song, R., Zhou, H., & Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 19, 221–228.
- Sherman, R. P. (1993). The Limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61, 123–137.
- Sprott, D. A. (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, 60, 457–465.
- Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement-error models. *Communications in Statistics-Theory and Methods*, 18, 4335–4358.
- Stern, H., & Cover, T. M. (1989). Maximum entropy and the lottery. *Journal of the American Statistical Association*, 84, 980–985.
- Strawderman, R. L., & Wells, M. T. (1998). Approximately exact inference for the common odds ratio in several 2×2 tables. *Journal of the American Statistical Association*, 93, 1294–1307.
- Struewing, J. P., Harge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., et al. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine*, 336, 1401–1408.
- Su, J. Q., & Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88, 1350–1355.
- Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Berlin: Springer.
- Sun, J., & Zhao, X. (2013). *The statistical analysis of panel count data*. Berlin: Springer.
- Sun, Y., Qin, J., & Huang, C.-Y. (2016). Missing information principle: A unified approach for general left-truncated and/or right-censored survival data problems. *Manuscript*.
- Tan, K. M., Ning, Y., Witten, D., & Liu, H. (2016). Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103, 761–777.
- Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101, 1619–1637.
- Tang, G., Little, R. J., & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90, 747–764.
- Tanner, M. A. (1996). *Tools for statistical inference*. New York: Springer.
- Tao, R., Zeng, D., Franceschini, N., North, K. E., Boerwinkle, E., & Lin, D. Y. (2015). Analysis of sequence data under multivariate trait-dependent sampling. *Journal of the American Statistical Association*, 110, 560–572.
- Terwilliger, J., Shannon, W., Lathrop, G., Nolan, J., Goldin, L., Chase, G., et al. (1997). True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics*, 61, 430–438.

- Thas, O., & De Neve, J. (2012). *Probabilistic index models*. *JRSSB*, *74*, 623–671.
- Thomas, D. C. (1977). Addendum to Methods of cohort analysis: Appraisal by application to asbestos mining by F.D.K. Liddell, J.C. McDonald and D.C. Thomas. *Journal of the Royal Statistical Society: Series A*, *140*, 469–491.
- Thomas, D. R., & Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *American Statistical Association*, *70*, 865–871.
- Thompson, M. E. (1997). *Theory of sample surveys. Monographs on statistics and applied probability*. London: Chapman and Hall.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, *58*, 267–288.
- Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.
- Tsai, W. Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika*, *77*, 169–177.
- Tsai, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika*, *96*, 601–615.
- Tsai, W. Y., Jewell, N. P., & Wang, M. C. (1987). The product-limit estimate of a survival curve under right censoring and left truncation. *Biometrika*, *74*, 883–886.
- Tsao, M., & Wu, F. (2013). Empirical likelihood on the full parameter space. *The Annals of Statistics*, *41*, 2176–2196.
- Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*, 475–494.
- Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, *18*, 354–372.
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Berlin: Springer.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B*, *38*, 290–295.
- Umbach, D. M., & Weinberg, C. R. (1997). Designing and analysing case control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, *16*, 1731–1743.
- Van Campenhout, J., & Cover, T. (1981). Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, *27*, 483–489.
- Van der Vaart, A. W., & Wellner, J. A. (1992). Existence and consistency of maximum likelihood in upgraded mixture models. *Journal of Multivariate Analysis*, *43*, 133–146.
- van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes with applications to statistics*. New York: Springer.
- Van Wieringen, W. N., Van De Wiel, M. A., & Van Der Vaart, A. W. (2008). A test for partial differential expression. *Journal of the American Statistical Association*, *103*, 1039–1049.
- Vardi, Y. (1982a). Nonparametric estimation in presence of length bias. *Annals of Statistics*, *10*, 616–620.
- Vardi, Y. (1982b). Nonparametric estimation in renewal processes. *The Annals of Statistics*, *10*, 772–785.
- Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics*, *13*, 178–203.
- Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, *76*, 751–761.
- Vardi, Y., Shepp, L. A., & Kaufman, L. (1985). A statistical model for positron emission tomography. *Journal of the American Statistical Association*, *80*, 8–20.
- Vardi, Y., & Zhang, C. H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Annals of Statistics*, *20*, 1022–1039.
- Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

- Vounatsou, P., Smith, T., & Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Applied Statistics*, 47, 575–587.
- Wacholder, S., Hartge, P., Struwing, J. P., Pee, D., McAdams, M., Brody, L., et al. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology*, 148, 623–630.
- Wacholder, S., Silverman, D. T., McLaughlin, J. K., & Mandel, J. S. (1992). Selection of controls in case-control studies: II-III. Types of controls. *American Journal of Epidemiology*, 135(1029–1041), 1042–1050.
- Wald, A. (1948). Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process. *The Annals of Mathematical Statistics*, 19, 40–46.
- Wang, M. C. (1987). Product-limit estimates: A generalized maximum likelihood study. *Communication in Statistics*, 16, 3117–3132.
- Wang, M. C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, 84, 742–748.
- Wang, M. C. (1996). Hazards regression analysis for length-biased data. *Biometrika*, 83, 343–354.
- Wang, M. C. (1992). The analysis of retrospectively ascertained data in the presence of reporting delays. *Journal American Statistical Association*, 87, 390–400.
- Wang, M. C., Jewell, N. P., & Tsai, W. Y. (1986). Asymptotic properties of the product-limit estimate under random truncation. *Annals of Statistics*, 14, 1597–1605.
- Waterman, R. P., & Lindsay, B. G. (1996). Projected score methods for approximating conditional scores. *Biometrika*, 83, 1–13.
- Watson, G. S. (1971). Estimating functionals of particle size distributions. *Biometrika*, 58, 483–490.
- Wei, L. J., Ying, Z., & Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77, 845–851.
- Weinberg, C. R., & Umbach, D. M. (1999). Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*, 55, 718–726.
- Wellner, J. A. (2015). Maximum likelihood in modern times: The ugly, the bad, and the good. IMS Le Cam Lecture. <https://www.stat.washington.edu/jaw/RESEARCH/TALKS/LeCam-v2.pdf>
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika*, 82, 57–67.
- Wicksell, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika*, 17, 84–99.
- Wicksell, S. D. (1926). The corpuscle problem: Second memoir: Case of ellipsoidal corpuscles. *Biometrika*, 18, 151–172.
- Wild, C. J. (1983). Failure time models with matched data. *Biometrika*, 70, 633–641.
- Wild, C. J. (1991). Fitting prospective regression models to case-control data. *Biometrika*, 78, 705–717.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9, 60–62.
- Wills, A., & Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, 71, 1042–1049.
- Wolfson, C., Wolfson, D. B., Asgharian, M., M'LAN, C. E., Ostbye, T., Rockwood, K., et al. (2001). Clinical progression of dementia study group. A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344, 1111–1116.
- Wood, A. M., White, I. R., & Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A*, 169, 525–542.
- Woodroffe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics*, 13, 163–177.
- Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185–193.
- Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11, 95–103.
- Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937–951.

- Wu, R., Ma, C., & Casella, G. (2007). *Statistical genetics of quantitative traits: Linkage, maps and QTL*. Springer science and business. New York: Springer.
- Xu, L., Lin, N., Zhang, B., & Shi, N. Z. (2012). A finite mixture model for working correlation matrices in generalized estimating equations. *Statistica Sinica*, 22, 755–776.
- Yao, B., Wang, L. M., & He, X. (2016). Semiparametric regression analysis of panel count data allowing for within-subject correlation. *Computational Statistics and Data Analysis*, 97, 47–59.
- Yao, Y. (2015). Maximum likelihood method for linear transformation models with cohort sampling data. *Statistica Sinica*, 25, 1231–1248.
- Yanagawa, T., & Fujii, Y. (1995). Projection-method Mantel-Haenszel estimator for K 2 × J tables. *Journal of the American Statistical Association*, 90, 649–656.
- Yang, D., & Small, D. (2013). An R package and a study of methods for computing empirical likelihood. *Journal of Statistical Computation and Simulation*, 83, 1363–1372.
- Yi, G. Y. (2016). *Statistical analysis with measurement error or misclassification strategy, method and application*. Berlin: Springer.
- Ying, Z. L. (1990). Linear rank statistics for truncated data. *Biometrika*, 77, 909–914.
- Yu, T., Chen, B. J., Li, P. F., & Qin, J. (2017). Pseudo-likelihood-based inference for the semiparametric transformation model. In revision to *JASA*.
- Zelen, M. (1971). The analysis of several 2 × 2 contingency tables. *Biometrika*, 58, 129–137.
- Zelen, M. (2004). Forward and backward recurrence times and length biased sampling: Age specific models. *Lifetime Data Analysis*, 10, 325–334.
- Zeng, D., & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society. Series B*, 69, 507–564.
- Zhang, B. (2002). An EM algorithm for a semiparametric finite mixture model. *Journal of Statistical Computation and Simulation*, 72, 791–802.
- Zhang, Z., & Rockette, H. (2005). On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134, 206–223.
- Zhao, J., & Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear model models with nonignorable missing data. *JASA*, 110, 1577–1590.
- Zhao, L. P., Prentice, R. L., & Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *JRSSB*, 54, 805–811.
- Zhao, Q. Y., & Percival, D. (2015). Entropy balancing is doubly robust. <https://arxiv.org/abs/1501.03571>
- Zhou, H., Song, R., Wu, Y. S., & Qin, J. (2011a). Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, 67, 194–202.
- Zhou, H., Weaver, M., Qin, J., Longnecker, M., & Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome dependent sampling scheme with a continuous outcome. *Biometrics*, 58, 413–421.
- Zhou, M. (2015). *Empirical likelihood method in survival analysis*. Boca Raton: Chapman and Hall.
- Zhou, Q., Zhou, H., & Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika*, 104, 17–29.
- Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2011b). *Statistical methods in diagnostic medicine*. New York: Wiley.
- Zhu, H., & Wang, M.-C. (2012). Analyzing bivariate survival data with interval sampling and application to cancer epidemiology. *Biometrika*, 99, 345–361.
- Zou, F., Fine, J. P., & Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika*, 89, 61–75.

Index

A

- Alternating minimization, 80
Ancillarity and Fisher information with nuisance parameters, 98
Application in ROC, 595
Applications of Godambe's theory in missing covariate problems, 90
Applications of PAVA in exponential families, 560

B

- Basic concepts, 11
Basic properties of length biased sampling problems, 5
Basic results on Poisson process, 18

C

- Case and control study with prevalent cases, 56
Chebyshev correlation inequality, 5
Combine summarized information: A more flexible method in meta analysis, 167
Composite likelihoods and corrected likelihoods, 72
Composite partial likelihood approach, 533
Cosslett's maximum likelihood estimation and related problems, 577
A criterion for the global maximum likelihood estimate, 56

D

- Definition of empirical likelihood and basic properties, 140

Different semiparametric MLE methods for case-Control data, 219

E

- EM algorithm algorithm, 78
Estimating monotonic decreasing density and hazard functions, 570

F

- Family-based case-control studies, 226
Forward and backward recurrence times, 14

G

- Generalized multiplicative censoring and semiparametric truncation model, 542
General theory of empirical likelihood in estimating equations, 143
Genetic liability model or Probit model, 232
Godambe's optimality criterion, 86
Godambe's theory in length biased sampling AFT models , 93

H

- Heckman's selection biased sampling model, 32
Hybrid likelihoods and utilization auxiliary information, 159

I

- I.I.D. representation of the hypergeometric distribution, 297

Information calculation for missing data problems, 123
 Information calculation in over-identified semiparametric models, 121
 Information contained in the conditional expectation model, 115
 Information identity test, 55
 Issues in maximum likelihood estimation, 53

K

Kullback-Leibler information and entropy concepts, 50

L

Length biased sampling examples, 2
 Lorenz curve, 7

M

Maximum binomial likelihood estimation in a genetic mixture model, 581
 Maximum likelihood estimation based on current status data, 587
 Maximum likelihood estimation for length biased sampling problems, 191
 Maximum likelihood estimation for multiple biased sampling problems, 196
 Miscellaneous problems, 234
 Missing information principle, 549
 Modelling based selection biased sampling problems, 40

N

Natural selection biased sampling problems, 23
 Neyman-Scott problem, 60

A non-root n consistent estimator example, 126

P

Pool adjacent violation algorithm (PAVA), 559
 Popular inference methods in the presence of nuisance parameters, 65
 Projection method for the mean estimation and linear regression model, 112
 Projection method in a two sample density ratio model, 120

Q

Quasi-Likelihood methods in linear regression models, 69

S

Semiparametric inference for Logistic regression analysis based on case-control data, 207
 Stochastic ordering, 5

T

Test homogeneous, 56
 Two useful maximization algorithms, 78

V

Variable selection and Akaike Criterion, 76

W

Weight function depends on the underlying distribution problems, 203
 Wicksell corpuscle problem, 35, 547