Contributions to Level Set Estimation, Nonparametric Regression, Confidence Intervals from Imputed Data and Marginal Logistic Regression Models for Longitudinal Survey Data

by

Qunshu Ren, B.Sc., M.Sc.

A thesis submitted to

the Faculty of Graduate Studies and Research

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Ottawa-Carleton Institute for Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

December, 2007

Your file Votre référence
ISBN: 978-0-494-40533-8
Our file Notre référence
ISBN: 978-0-494-40533-8

Table of Contents

Abstract

The objective of the thesis is to develop and extend some new results on four selected topics which are currently widely used and studied in probability and statistics. These topics include:

1. Nonparametric estimators of level sets and their properties under minimal conditions. We establish the same results as in Bailo et al. (2001) without imposing their technical condition. Furthermore, the IID assumption will be relaxed to one based on α-mixing sequence. We require no additional conditions to establish our results.

2. The convergence rates for nonparametric regression. Motivated by the classical results on this topic for the IID case, I extended some of the results to dependent mixing sequences. For this topic, I studied two problems, one is about the rate of convergence for complexity regularization with β-mixing data; the other one is the convergence rate in nonparametric regression for an α-mixing setup. Our technical arguments are based on the theory of function-indexed empirical processes and a strong coupling lemmas.

3. Confidence intervals with fractional imputed missing data. For missing data set, fractional imputed data are used and confidence intervals are constructed for population parameters such as mean, distribution and quantile via normal approximation and empirical likelihood methods. To do so, some asymptotic normality theorems need to be set up. These results are extension of Qin, Rao and Ren's (2006) work for random hot deck imputation. We also consider two cases for this work: one is based on simple data sets without any covariate; the other is assuming that observation and some covariates are available and there is a linear relationship between them.

4. Marginal logistic regression for longitudinal complex survey design data. The motivation of the study comes from the requirement of processing of Statistics Canada conducted longitudinal survey data—-the National Population Healthy Survey (NPHS). We proposed a Generalized Estimation Equation (GEE) method for the estimation of model parameters. To avoid the complex derivation of Taylor linearization, one-step estimation function (EF) bootstrap method is used for the variance estimation. Several Goodness-of-fit tests are studied for the model assessment problems.

To verify and support the theoretical derivation in the work, a large amount of computational simulation work has been done with C/C++ and SAS code. Some of the simulation results are listed in the thesis.

Acknowledgements

First of all, I would like to express my sincere gratitude to Professor Majid Mojirsheibani and Professor J. N. K. Rao, my thesis supervisors, for their constant and patient guidance and especially for the invaluable research experience they have given me. I also thank them for their support, and concerns about my life and career during these years.

I would like to thank other professors in the department, especially those I had taken necessary courses from in the related fields which are required for the Ph.D. program. The background knowledge and research skills I gained from those courses are very important to this thesis research.

I would like to thank Dr. Georgia Roberts of Statistics Canada, who was my supervisor when I worked as an internship research student in Statistics Canada. She supported me substantially when I conducted the research program, Chapter 6 of the thesis is based on my internship work at Statistics Canada.

I am indebted to Professor Yongsong Qin, who was a visiting scholar of Professor Rao. When he visited our department, Dr. Rao asked me to cooperate with him

on problems of construction of confidence intervals using empirical likelihood and normal approximation based methods under imputation for missing data. I did some simulation work with him and learned many tools and methods from him. That helped me in extending results on random imputation to fractionally imputed data sets. The extension is included in Chapter 5.

I obtained financial support from different resources, including NSERC PGS-D scholarship (2004-2007), TA and other funding from the Department, RA funds from both of my supervisors, and MITACS and Statistics Canada jointly sponsored internship research program(July 2004-June 2005). With those supports, I was able to concentrate on my research.

I would like to thank the School of Mathematics and Statistics, Carleton University, for providing the necessary research facilities for completing this thesis. My thanks also goes to the Faculty members of the School, who provided the excellent academic environment with many interesting courses and seminars. Special thanks are given to Dr. Gang Li for his assistance with Latex.

I wish to thank my parents for their love and concerns. I wish to give my heartfelt thanks to my wife, Cathy Zhang. She has been patiently supporting me for many years and taking care of my sons, Mark and Anthony. Mark also helped me checking some typing errors. My sincere thanks are given to my Mom, who helps us with every aspect of my family.

Ottawa, Ontario Qunshu Ren

December 7, 2007

Chapter 1

Introduction

In this thesis, I will focus on several different statistical topics. Those topics include two major parts, which cover four topics of my research interest in statistics. Part 1 (Chapters 2-4) is conducted under the guidance of Professor M. Mojirsheibani. In this part, I focus on the convergence (in an L_2 sense) problems of nonparametric regression under a mixing dependent setup. It contains the following three chapters.

In Chapter 2, I study properties of level set estimators under minimal assumptions. Bailo *et al.* (2001) established L_1-convergence results for nonparametric estimators of level sets, for *i.i.d.* sequences, under the technical assumption that there exist constants $(r, R), R > r > 0$ such that $m_{f,R+r}(\mathbf{0}) > 0$ and

$$\int_{B^c(\mathbf{0},R)} (M_{f,h}(\mathbf{z})/m_{f,h}(\mathbf{z}))^{3/2} f^{1/2}(\mathbf{z}) d\mathbf{z} < \infty,$$

where, f is the underlying density, $M_{f,h}(\mathbf{z}) := \sup\{f(\mathbf{x}) : \mathbf{x} \in B(\mathbf{z}, h)\}$, $m_{f,h}(\mathbf{z}) := \inf\{f(\mathbf{x}) : \mathbf{x} \in B(\mathbf{z}, h)\}$, and $B(\mathbf{a}, r)$ denotes the closed ball with center at $\mathbf{a} \in \mathcal{R}^d$ and radius r. In Chapter 2, we establish the same result(with the same rates of convergence) without imposing the above technical condition. Furthermore, the *i.i.d.* assumption will be relaxed to one based on α-mixing sequences. We require no additional conditions to establish our results.

In Chapter 3, I start by reviewing some concepts and important results about the problems of nonparametric regression, especially the consistency and rates of convergence properties under the *i.i.d.* conditions. The main result for this chapter is an almost sure performance bound on the nonparametric regression estimator of $E(Y|\mathbf{X})$.

In Chapter 4, upper bounds are found for the statistical risks of least-squares estimates of a regression function (possibly nonlinear), under a β-mixing setup. The assumptions are minimal and are virtually the same as those of Kohler (2000) for the *i.i.d.* case. Furthermore, unlike the bounds obtained by Baraud *et al.* (2001), our results are not confined to compact sets. Similar results are also obtained for complexity regularized estimates via penalty functions.

The second part of my thesis (Chapters 5-6) is conducted under the guidance of Professor J. N. K. Rao. There are two major problems discussed in these two chapters. Chapter 5 focuses on the problems of confidence intervals with imputed data. The missing data (or non-response) problem has been studied extensively in recent years. To achieve a higher efficiency and make the data analysis process simpler, imputation is an effective tool for treating missing data problems. Previously, some results on normal approximation based and empirical likelihood confidence intervals of marginal population parameters (including mean, distribution function and quantile) under random imputation have been obtained by Qin, Rao and Ren (2006a). In Chapter 5, we extend those results using a fractional random imputation setup, to achieve gain in efficiency. Asymptotically valid confidence intervals on population parameters (mean, distribution and quantile) are constructed. The

confidence intervals are constructed via both asymptotic normality and empirical likelihood methods. Extensive simulation results to validate the asymptotic theory are also reported.

In Chapter 6, I report some of my work on marginal logistic regression models for longitudinal complex survey data. Generalized Estimating Equation (GEE) method is used to estimate the model parameters. Bootstrap methods, especially one-step Estimating Function (EF) bootstrap methods, are used to estimate the variance of the estimators. For detecting model deviations, a number of goodness-of-fit test methods are studied. Among those methods, Horton's (1999) quasi-score method and adjusted Hosmer-Lemeshow method are studied. All those methods are currently applied to data from the National Population Healthy Survey (NPHS), a longitudinal complex survey conducted by Statistics Canada. Results of data analyses are reported.

Simulation work was done, mostly in C/C++, SAS and R/Splus. Those results are listed in the tables in last sections of Chapters 2, 5 and 6. The program codes are available upon request.

Chapter 2

On Nonparametric Estimation of Level Sets

2.1 Introduction

In this chapter we study applications of kernel density estimators to the problem of level set estimation. Level set estimation and its applications to other fields have been extensively studied in recent decades. For example, Muller and Sawitzki (1991) and Polonik (1995) used level sets to define tests of multimodality. Devroye and Wise (1980) employed level sets to develop a nonparametric quality control tool for the detection of abnormal behavior in a system. Hartigan (1975) defined the connected components as clusters in a population based on level sets. Level sets can also be used to define an α-outlier region; see, for example, Davies and Gather (1993). In other applications, Polonik (1997) used level sets to construct robust estimators of location and dispersion parameters. Level sets are also used in medical imaging, which allow a physician to check and locate tumors, as mentioned by Muller (1993).

It is the estimation of level sets that is of particular interest in all the mentioned

examples. Nonparametric density estimators can be used to define level sets. If f is a univariate or multivariate density function, the level set $\{\mathbf{z} : f(\mathbf{z}) > c\} =: \{f > c\}$, for a suitably chosen small constant c, can be estimated by $\{\mathbf{z} : f_n(\mathbf{z}) > c\} =: \{f_n > c\}$, where $f_n(\mathbf{z})$ is the nonparametric kernel estimator of $f(\mathbf{z})$ based on a random sample $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ from f. That is:

$$f_n(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{z} - \mathbf{X}_i) := \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{z} - \mathbf{X}_i}{h}\right),$$

where $K : \Re^d \to \Re$ is the kernel function, $K_h(.) = (1/h^d)K(./h)$ and $h \equiv h_n \to 0$ is the bandwidth, $nh_n^d \to \infty$. In the context of quality control, proposed by Devroye and Wise (1980), one new observation $\mathbf{Z} = \mathbf{z}$ is defined as out of control if it belongs to the level set $\{f \leq c\}$. In a different sense, Hartigan (1975) defined the level set $\{f \leq c\}$ as *not classified* when classifying in a two-cluster problem. Using the empirical level set $\{f_n \leq c\}$, define

$$\begin{aligned} P_n(\mathbf{z}) &:= P\{f_n(\mathbf{Z}) \leq c | \mathbf{Z} = \mathbf{z}\} \\ &= \int_{\{(\mathbf{x}_1,...,\mathbf{x}_n):f_n(\mathbf{Z})\leq c\}} f(\mathbf{x}_1, ..., \mathbf{x}_n) d\mathbf{x}_1 ... d\mathbf{x}_n, \end{aligned} \tag{2.1.1}$$

which may also be interpreted as the probability of not classifying a new datum $'\mathbf{Z}'$. Denote the indictor function of a set A by $\mathbf{I}_A$ and let $P_\infty(\mathbf{Z}) = \mathbf{I}_{\{f(\mathbf{Z})\leq c\}}$, where $\mathbf{Z}$ is a new random observation drawn from f.

With the above definition of $P_n(\mathbf{Z})$, Baillo *et al.* (2001) studied the rate of convergence (to zero) of the L_1-distance

$$E|P_n(\mathbf{Z}) - P_\infty(\mathbf{Z})|$$

under rather technical conditions for *i.i.d.* sequences. In the first part of our contributions, we study the rate of convergence (to zero) of $E|P_n(\mathbf{Z}) - P_\infty(\mathbf{Z})|$ under much

weaker conditions than those stated in Baillo's result, for α-mixing sequences with both exponential and polynomial mixing rates. Here our proofs are very different from those of Baillo *et al.* (2001) and, in fact, do not require the technical condition imposed by these authors on the underlying density (*cf.* Remark A that appears after our Theorem 2.2.1).

In Davies and Gather (1993), an α-outlier region is defined as: $\{f \leq c\}$, such that $P(\mathbf{Z} \in \{f \leq c\}) = \alpha$. Instead of using a constant c, one can also define the level sets via a random α-quantile, c_n. More specifically, let $c_n = c_n(\mathbf{X}_1, ..., \mathbf{X}_n)$ be a random value satisfying

$$P_{f_n}\{f_n \leq c_n\} := \int_{\{\mathbf{z}: f_n(\mathbf{z}) \leq c_n\}} f_n(\mathbf{z})d\mathbf{z} = \alpha, \tag{2.1.2}$$

where $\alpha \in (0, 1)$. Let

$$\begin{aligned} P_f\{f_n \leq c_n\} &= \int_{\{\mathbf{z}: f_n(\mathbf{z}) \leq c_n\}} f(\mathbf{z})d\mathbf{z} \\ &= P\{f_n(\mathbf{Z}) \leq c_n | \mathbf{X}_1, ..., \mathbf{X}_n\}. \end{aligned} \tag{2.1.3}$$

Note that in the context of quality control, the quantile c_n is chosen to make the *estimated false alarm* probability in (2.1.2) to be equal to α. Baillo (2001) has proved that the *real* false alarm probability in (2.1.3) converges to (2.1.2), for *i.i.d.* observations. But the results are setup under strict technical assumption imposed on certain functionals of the underlying density f. In Section 2.3 we establish the same results(with the same rates of convergence) without imposing such technical conditions. Furthermore, the *i.i.d.* assumption will be relaxed to one based on α-mixing sequences. We require no additional conditions to establish our results.

In addition to our theoretical results, simulation studies are also curried out in Section 2.4 to verify some of the convergence results of this chapter. Two algorithms

are given for generating α-mixing sequences with correlation coefficient decaying at exponential order, as well as an approximately polynomial order.

2.2 Convergence rates for estimators of level sets: Case of c constant

Let $\{\mathbf{X}_n\}$ be a $\Re^d$-valued stochastic sequence defined on an underlying probability space $(\Omega, \mathcal{A}, P)$. Let

$$\alpha(k) = \sup_{n \in \mathcal{N}} \sup_{A \in \sigma(\mathbf{X}_i, i \leq n), B \in \sigma(\mathbf{X}_i, i \geq n+k)} |P(A \cap B) - P(A)P(B)|$$

be the α-mixing coefficient of the sequence, where $\sigma(\mathbf{X}_i, i \leq n)$ is the σ field generated by $\{..., \mathbf{X}_{n-1}, \mathbf{X}_n ...\}$. We first state the following assumption(which were also used by Bailo *et. al.*(2001)):

$(F1)$ $P\{|c - f(\mathbf{X})| \leq \epsilon\} = O(\epsilon)$, as $\epsilon \to 0$, where $\mathbf{X}$ denotes a random vector with density f, where c is as in $\{f \leq c\}$.

$(K1)$ the kernel K is a compactly supported density function, and $\|K\|_\infty \leq M < \infty$.

$(H1)$ the bandwidth h is $O(n^{-s})$ for some $s \in (0, 1/(d+2))$.

Note that condition $(F1)$, $(K1)$ and $(H1)$ are not very restrictive. Condition $(F1)$ requires that the set $\{f = c\}$ contains no isolated point, and holds for all continuous random variables; Condition $(K1)$ holds for most common kernels such as naive, Exponechnikov and the truncated Gaussian kernels,

Theorem 2.2.1. *Suppose that conditions* $(F1), (K1), (H1)$ *hold and that* f *satisfies a Lipschitz condition. If* $\{\boldsymbol{X}_n\}$ *is an* α*-mixing sequence with mixing coefficients satisfying* $\alpha(n) \leq Cn^{-t}$ *for some* $t > \frac{7s}{2(1-2s)}$, $0 < s < \frac{1}{d+2}$ *and* $C > 0$, *then* $E|P_n(\boldsymbol{Z}) - P_\infty(\boldsymbol{Z})| = O(n^{-s})$ *as* $n \to \infty$.

Remark A. The above theorem was originally proved by Baillo *et al.* (2001), for *i.i.d.* sequences, and under the additional technical condition that there exist constants $R > r > 0$ such that

$$\int_{B^c(0,R)} \left(\frac{M_{f,h}(z)}{m_{f,h}(z)} \right)^{3/2} f^{1/2}(z) dz < \infty,$$

where $M_{f,h}(z) := \sup\{f(x) : x \in B(z,h)\}$, and $m_{f,h}(z) := \inf\{f(x) : x \in B(z,h)\}$; here $B(a,r)$ denotes the closed ball with center at a and radius r. The fact that our Theorem 2.2.1 avoids such analytic conditions on the underlying density f, clearly makes our results more appealing and optimal. Also, note that the mixing rate in Theorem 2.2.1, (which replaces the *i.i.d.* assumption used in Baillo *et al.*), is only of polynomial order.

Proof of Theorem 2.2.1:

Let $E_n(\mathbf{z}) = E(K_h(\mathbf{z} - \mathbf{X}))$, where $\mathbf{X}$ is a random vector with density f. Since f is Lipschitz and $\int \|\mathbf{t}\| K(\mathbf{t}) d\mathbf{t} < \infty$, it is straightforward to see that for some constant $C_0 > 0$, $|f(\mathbf{z}) - E_n(\mathbf{z})| \leq C_0 h$, $\forall \mathbf{z} \in \Re^d$. Next, observe that,

$$\begin{aligned} & E|P_n(\mathbf{Z}) - P_\infty(\mathbf{Z})| \\ \leq\ & E\left[\left| P\{f_n(\mathbf{Z}) \leq c|\mathbf{Z}\} - \mathbf{I}\{f(\mathbf{Z}) \leq c\} \right| \mathbf{I}\{|c - f(\mathbf{Z})| \leq 2C_0 h\} \right] \\ & + E\left[\left| P\{f_n(\mathbf{Z}) \leq c|\mathbf{Z}\} - \mathbf{I}\{f(\mathbf{Z}) \leq c\} \right| \mathbf{I}\{|c - f(\mathbf{Z})| > 2C_0 h\} \right] \\ =:\ & I_1 + I_2, \end{aligned} \tag{2.2.1}$$

By assumption (F1), $I_1 = O(h)$. As for the term I_2, first note that

$$\begin{aligned} I_2 \ =\ & E\left[\left| P\{f_n(\mathbf{Z}) \leq c|\mathbf{Z}\} - \mathbf{I}\{f(\mathbf{Z}) \leq c\} \right| \mathbf{I}\Big(\{|c - f(\mathbf{Z})| > 2C_0 h\} \bigcap \{f(\mathbf{Z}) > c\} \Big) \right] \\ & + E\left[\left| P\{f_n(\mathbf{Z}) \leq c|\mathbf{Z}\} - \mathbf{I}\{f(\mathbf{Z}) \leq c\} \right| \mathbf{I}\Big(\{|c - f(\mathbf{Z})| > 2C_0 h\} \bigcap \{f(\mathbf{Z}) \leq c\} \Big) \right] \end{aligned}$$

$$
\begin{aligned}
&= E\Big[P\{f_n(\mathbf{Z}) \le c|\mathbf{Z}\}\mathbf{I}\{|c - f(\mathbf{Z})| > 2C_0h \bigcap f(\mathbf{Z}) > c\}\Big] \\
&\quad + E\Big[P\{f_n(\mathbf{Z}) > c|\mathbf{Z}\}\mathbf{I}\{|c - f(\mathbf{Z})| > 2C_0h \bigcap f(\mathbf{Z}) \le c\}\Big] \\
&:= I_3 + I_4. \qquad (2.2.2)
\end{aligned}
$$

Since

$$
\begin{aligned}
&\Big\{\mathbf{Z} : |c - f(\mathbf{Z})| > 2C_0h, |f(\mathbf{Z}) - E_n(\mathbf{Z})| \le C_0h, f(\mathbf{Z}) > c, f_n(\mathbf{Z}) \le c\Big\} \\
\subseteq\ &\Big\{\mathbf{Z} : f_n(\mathbf{Z}) - E_n(\mathbf{Z}) < -C_0h\Big\},
\end{aligned}
$$

one finds

$$
I_3 \le E\Big[P\{f_n(\mathbf{Z}) - E_n(\mathbf{Z}) < -C_0h|\mathbf{Z}\}\Big].
$$

Similarly, one can show that

$$
I_4 \le E\Big[P\{f_n(\mathbf{Z}) - E_n(\mathbf{Z}) > C_0h|\mathbf{Z}\}\Big].
$$

Putting the above bounds together, one has

$$
I_2 \le E\Big[P\{|f_n(\mathbf{Z}) - E_n(\mathbf{Z})| > C_0h|\mathbf{Z}\}\Big].
$$

On the other hand, since $\{K_h(\mathbf{z}-\mathbf{X}_i)-E_n(\mathbf{z}), i = 1, ..., n\}$ is a zero-mean real-valued α-mixing sequence and is bounded by $2\|K\|_\infty$, one can apply Theorem 1.3 of Bosq (1998) to conclude that for any integer $q \in [1, n/2]$, one has

$$
\begin{aligned}
&P\Big\{|f_n(\mathbf{Z}) - E_n(\mathbf{Z})| > C_0h\Big|\mathbf{Z} = \mathbf{z}\Big\} \le 4\exp\Big(-\frac{C_0^2h^2}{8\|K\|_\infty^2}q\Big) + \\
&22\Big(1 + \frac{4\|K\|_\infty}{h}\Big)^{1/2} q\alpha\Big(\Big[\frac{n}{2q}\Big]\Big). \qquad (2.2.3)
\end{aligned}
$$

Clearly, $t > \frac{7s}{2(1-2s)}$ implies that $2s < \frac{2t-3s}{2(t+1)}$. Now choose any r such that $2s < r < \frac{2t-3s}{2(t+1)}$ and put $q = \lfloor \frac{n^r}{2} \rfloor$. Then the first term on the *r.h.s.* of (2.2.3) is of order

$O\left(e^{-c_1 n^{r-2s}}\right)$, where $c_1 > 0$ does not depend on n. As for the second term on the *r.h.s.* of (2.2.3), it becomes $O(n^{r+\frac{s}{2}-(1-r)t}) = O(n^{-s})$, because $r < \frac{2t-3s}{2(t+1)}$. This completes the proof of theorem 2.2.1. □

It is possible to improve the conclusion of the above theorem under sufficient smoothness conditions on f. More specifically, let $\mathcal{C}_{2,d}(b)$ be the space of twice continuously differentiable real valued functions f, defined on $\Re^d$, satisfying $\|f\|_\infty \leq b$ and $\|f^{(2)}\|_\infty \leq b$ where $f^{(2)}$ denotes any second-order partial derivative of the function f. To state our next result we need the following condition on h:

$(H1')$ the bandwidth h is $O(n^{-s})$ for some $s \in (0, 1/(d+4))$.

Theorem 2.2.2. *Suppose that conditions of* $(F1), (K1), (H1')$ *hold. Let* $\{\boldsymbol{X}_n\}$ *be an* α*-mixing sequence with mixing coefficients satisfying* $\alpha(n) \leq C_1 n^{-t}$ *for some* $t > \frac{7s}{1-4s}$, *where* $C_1 > 0$ *is a constant and* s *is as in condition*$(H1')$. *If* $f \in \mathcal{C}_{2,d}(b)$, *then* $E|P_n - P_\infty| = O(n^{-2s})$ *as* $n \to \infty$, *where* s *is as in* $(H1)$.

Proof of Theorem 2.2.2:

Since $f \in \mathcal{C}_{2,d}(b)$, one finds

$$Ef_n - f = \frac{h^2}{2} \int \sum_{1 \leq i,j \leq d} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x} - \theta h_n \mathbf{v}) v_i v_j K(\mathbf{v}) d\mathbf{v}, \tag{2.2.4}$$

where $\theta = \theta(f, \mathbf{x}, h, \mathbf{v}) \in (0,1)$. That is, $|f - Ef_n| = O(h^2)$. Thus, $\forall\ \mathbf{z} \in \Re^d$, there is a constant $C_2 > 0$, such that $|f(\mathbf{z}) - Ef_n(\mathbf{z})| \leq C_2 h^2$.

Now, observe that

$$\begin{aligned} & E|P_n(\mathbf{Z}) - P_\infty(\mathbf{Z})| \\ \leq\ & E\left[\left|P(\{f_n(\mathbf{Z}) \leq c\}|\mathbf{Z}) - \mathbf{I}(\{f(\mathbf{Z}) \leq c\})\right| \mathbf{I}\{|c - f(\mathbf{Z})| \leq 2C_2 h^2\}\right] \end{aligned}$$

$$+E\left[\left|P\{f_n(\mathbf{Z}) \le c|\mathbf{Z}\} - \mathbf{I}\{f(\mathbf{Z}) \le c\}\right|\mathbf{I}\{|c - f(\mathbf{Z})| > 2C_2h^2\}\right]$$
$$:= \quad I_5 + I_6, \tag{2.2.5}$$

But $I_5 = O(h^2)$ and (using the arguments leading to the bound on I_2),

$$I_6 \le E\left[P\{|f_n(\mathbf{Z}) - E_n(\mathbf{Z})| > C_2h^2\Big|\mathbf{Z}\}\right]. \tag{2.2.6}$$

Furthermore, one more application of Theorem 1.3 of Bosq (1998) yields

$$P\left\{|f_n(\mathbf{Z}) - E_n(\mathbf{Z})| > C_2h^2\Big|\mathbf{Z} = \mathbf{z}\right\} \le 4\exp\left(-\frac{C_2^2h^4}{8\|K\|_\infty^2}q\right) +$$
$$22\left(1 + \frac{4\|K\|_\infty}{h^2}\right)^{1/2} q\alpha\left(\left[\frac{n}{2q}\right]\right). \tag{2.2.7}$$

Clearly, $t > \frac{7s}{1-4s}$ implies $4s < \frac{t-3s}{t+1}$. Now choose any r such that $4s < r < \frac{t-3s}{t+1}$, and put $q = \lfloor \frac{n^r}{2} \rfloor$. Then the first term on the r.h.s. of (2.2.7) is of order $O\left(e^{-c_3n^{r-4s}}\right)$, where $c_3 > 0$. It is straightforward to see that the second term on the r.h.s. of (2.2.7) is of order $O(n^{r+s-(1-r)t}) = O(n^{-2s})$.

This completes the proof of Theorem 2.2.2. □

2.3 Convergence rates for estimators of probability of false alarm: Case of random c

In this section, we consider the convergence rates under the case of a random α-quantile, c_n, instead of a constant c. Let $c_n = c_n(\mathbf{X}_1, ..., \mathbf{X}_n)$ be a random value satisfying

$$P_{f_n}\{f_n \leq c_n\} = \int_{\{\mathbf{z}: f_n(\mathbf{z}) \leq c_n\}} f_n(\mathbf{z})d\mathbf{z} = \alpha, \tag{2.3.1}$$

where $\alpha \in (0, 1)$. Also, let

$$P_f\{f_n \leq c_n\} = \int_{\{\mathbf{z}: f_n(\mathbf{z}) \leq c_n\}} f(\mathbf{z})d\mathbf{z} = P\{f_n(\mathbf{Z}) \leq c_n | \mathbf{X}_1, ..., \mathbf{X}_n)\}, \tag{2.3.2}$$

that is also called the probability of false alarm. The aim of this section is to obtain almost sure convergence results for $P_f\{f_n \leq c_n\} - \alpha$. First we state the following assumptions.

$(F2)$ $f(\mathbf{x}) = O(\|\mathbf{x}\|^{-\gamma})$ as $\|\mathbf{x}\| \to \infty$ for some $\gamma > d$.

$(K2)$ K is a compactly supported density of bounded variation.

Condition $(F2)$ requires the tails of the distribution vanishing fast enough, which should be satisfied for most distributions(except for heavy-tailed distributions such as Cauchy). First we state the following technical lemma due to Doukhan (1994).

Lemma 2.3.1. *Assume that (X_i) is a mixing sequence of centered random variables with $|X_i| \leq 1$ and such that the strong mixing sequence (resp. the uniform mixing sequence) satisfies $\alpha_n \leq u\nu^n$ (resp. $\phi_n \leq u\nu^n$) for some $u > 0, 0 \leq \nu < 1$. If, moreover, n satisfies $n \sup_i(\boldsymbol{E}|X_i|^2)^{2/(2+\varepsilon)} \geq 1$ (resp. $n \sup_i(\boldsymbol{E}|X_i|^2 \geq 1$) then there*

are some constants $a, b > 0$ *such that for any* $x \geq 0$,

$$\boldsymbol{P}\Big(|\sum_{i=1}^{n} X_i| \geq x\sqrt{n}\sigma \log \sigma^{-1}\Big) \leq a \exp\{-b\sqrt{x}\}$$

$$resp. \boldsymbol{P}(|\sum_{i=1}^{n} X_i| \geq x\sqrt{n}\sigma) \leq a \exp\{-b\sqrt{x}\},$$

where σ *is a constant such that* $\sigma^2 = \sup_i(\boldsymbol{E}|X_i|^{2+\varepsilon})^{\frac{2}{2+\varepsilon}}$.

For the proof of the lemma one may refer to Doukhan (1994) pp. 33.
We have the following result.

Theorem 2.3.1. *Let* $\{\boldsymbol{X}_n\}$ *be an* α*-mixing sequence with mixing coefficient satisfying* $\alpha(n) \leq K_1 e^{-\lambda n}$ *for some* $\lambda > 0$ *and* $K_1 > 0$. *Suppose that conditions* $(F2)$ *and* $(K2)$ *hold and that* $f \in \mathcal{C}_{2,d}(b)$. *Set* $h = \left(\frac{n}{\log n}\right)^{-\frac{1}{d+4}}$. *Then*

$$|P_f\{f_n \leq c_n\} - \alpha| \overset{a.s.}{=} O(h^s), \tag{2.3.3}$$

where $s = 2(1 - \frac{d}{\gamma})$. *If, in addition, the density* f *is compactly supported, then (2.3.3) holds for any* $s \in (0, 2)$.

Proof:

Let $B(\mathbf{x}, h^{-t})$ be a closed ball with center at $\mathbf{x}$ and radius h^{-t}, where t is a positive number to be fixed later. Put $B(h^{-t}) = B(\mathbf{0}, h^{-t})$. Observe that (as in Theorem 2

of Baillo *et al.* (2001)),

$$\begin{aligned}
&\left|P_f\{f_n \le c_n\} - P_{f_n}\{f_n \le c_n\}\right| = \left|\int_{\{\mathbf{z}:f_n(\mathbf{z})\le c_n\}} (f_n(\mathbf{z}) - f(\mathbf{z}))d\mathbf{z}\right| \\
\le\ & \int_{\{\mathbf{z}:f_n(\mathbf{z})\le c_n\}\bigcap B(h^{-t})} \left|f_n(\mathbf{z}) - f(\mathbf{z})\right|d\mathbf{z} + \int_{\{\mathbf{z}:f_n(\mathbf{z})\le c_n\}\bigcap B^c(h^{-t})} \left|f_n(\mathbf{z}) - f(\mathbf{z})\right|d\mathbf{z} \\
\le\ & \int_{B(h^{-t})} |f_n(\mathbf{z}) - Ef_n(\mathbf{z})|d\mathbf{z} + \int_{B(h^{-t})} |f(\mathbf{z}) - Ef_n(\mathbf{z})|d\mathbf{z} \\
&+ \int_{B^c(h^{-t})} f(\mathbf{z})d\mathbf{z} + \int_{B^c(h^{-t})} f_n(\mathbf{z})d\mathbf{z} \\
:=\ & I_n(1) + I_n(2) + I_n(3) + I_n(4) \qquad (2.3.4)
\end{aligned}$$

where $B^c(h^{-t})$ is the complement of $B(h^{-t})$. We start by bounding the first term on the *r.h.s.* of (2.3.4),

$$I_n(1) \le \sup_{\|\mathbf{z}\|<h^{-t}} |f_n(\mathbf{z}) - Ef_n(\mathbf{z})| \int_{B(h^{-t})} d\mathbf{z}. \qquad (2.3.5)$$

But $\sup_{\|\mathbf{z}\|<h^{-t}} |f_n(\mathbf{z}) - Ef_n(\mathbf{z})| \overset{a.s.}{=} o\left(log_k n\left(\frac{\log n}{n}\right)^{\frac{2}{d+4}}\right)$ for every positive integer k, where $log_k n$ is the kth iterated logarithm; see, for example, Bosq (1998, page 51). Therefore

$$\begin{aligned}
I_n(1) &\overset{a.s.}{\le} o\left(log_k n\left(\frac{\log n}{n}\right)^{\frac{2}{d+4}}\right)O(h^{-td}) \\
&\overset{a.s.}{=} O\left(h^{2-td} log_k n\right). \qquad (2.3.6)
\end{aligned}$$

As for $I_n(2)$, the assumption that $f \in C_{2,d}(b)$ immediately yields $|f - Ef_n| = O(h^2)$ and thus

$$I_n(2) = O(h^{2-td}). \qquad (2.3.7)$$

To deal with the terms $I_n(3)$ and $I_n(4)$, first note that by $(F2)$

$$I_n(3) \le C_3 h^{t(\gamma-d)}, \qquad (2.3.8)$$

for some constant C_3. Next, we may assume, without loss of generality, that the kernel K is supported in the ball $B(1)$ (which can be obtained by a simple transformation). Then,

$$\begin{aligned} I_n(4) &\leq \frac{\|K\|_\infty}{nh^d}\sum_{i=1}^{n}\int_{B^c(h^{-t})}\mathbf{I}_{B(\mathbf{X}_i,h)}(\mathbf{z})d\mathbf{z} \\ &\leq \frac{\|K\|_\infty}{nh^d}\sum_{i=1}^{n}(2h)^d\mathbf{I}_{B^c(h^{-t}-h)}(\mathbf{X}_i) \\ &= 2^d\|K\|_\infty\mathcal{P}_n(B^c(h^{-t}-h)), \end{aligned}$$

where $\mathcal{P}_n$ denotes the empirical probability distribution based on the sample $\mathbf{X}_1$... $\mathbf{X}_n$. Note that by $(F2)$, for large n (small h), one has

$$E\Big(\mathbf{I}_{B^c(h^{-t}-h)}(\mathbf{X}_i)\Big) = P\Big\{\mathbf{X}_i \in B^c(h^{-t}-h)\Big\} \leq C_4 h^{t(\gamma-d)},$$

for some positive constant C_4. Now a result of Lemma 2.3.1 yields

$$\begin{aligned} & P\{\mathcal{P}_n(B^c(h^{-t}-h)) \geq 2C_4h^{t(\gamma-d)}\} \\ \leq\ & P\Big\{\sum_{i=1}^{n}\Big[\mathbf{I}_{B^c(h^{-t}-h)}(\mathbf{X}_i) - P\{\mathbf{X}_i \in B^c(h^{-t}-h)\}\Big] \geq C_4nh^{t(\gamma-d)}\Big\} \\ \leq\ & C_6\exp\Big\{-C_5h^{t(\gamma-d)/2}n^{1/4}\Big\}, (\text{for } n \text{ large enough}), \end{aligned}$$

for some constants $C_5, C_6 > 0$. Since $h = \left(\frac{n}{\log n}\right)^{-\frac{1}{d+4}}$, the above bound will be summable in n for all $t < \frac{d+4}{2(\gamma-d)}$. Thus, by the Borel-Cantelli lemma,

$$\mathcal{P}_n(B^c(h^{-t}-h)) \overset{a.s.}{=} O(h^{t(\gamma-d)}) \tag{2.3.9}$$

To choose t in an optimal way, so that the largest of the four bounds in (2.3.6), (2.3.7), (2.3.8) and the (2.3.9) is as small as possible, one has to solve the equation $2 - td = t(\gamma - d)$ for t. This would immediately give $t = \frac{2}{\gamma}$. The theorem is now

proved upon replacing t by $\frac{2}{\gamma}$ in (2.3.6)-(2.3.9). When f is assumed to be compactly supported, the terms $I_n(3)$ and $I_n(4)$ disappear for n large enough. That completes the proof. □

Note that Theorem 2.3.1 assumes that the mixing coefficient decays at an exponential order. One can naturally ask if there is a similar result for an α-mixing sequence with a polynomial mixing rates. The answer is yes. However, we need to make one additional assumption: for each pair of indices i and j, $i \neq j$, let $g(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i, \mathbf{x}_j) - f(\mathbf{x}_i)f(\mathbf{x}_j)$, where $f(\mathbf{x}_i, \mathbf{x}_j)$ is the joint density of $(\mathbf{x}_i, \mathbf{x}_j)$. The new assumption is:

$(H2)$. $|g(\mathbf{z}') - g(\mathbf{z})| \leq l\|\mathbf{z}' - \mathbf{z}\|$; $\mathbf{z}', \mathbf{z} \in \Re^{2d}$ for some constant $l > 0$.

Now we have the following theorem.

Theorem 2.3.2. *Let $\{\boldsymbol{X}_n\}$ be an α-mixing sequence with mixing coefficient satisfying $\alpha(n) \leq K_2 n^{-\lambda_2}$ for some $\lambda_2 > \frac{2d+1}{d+1}$ and $K_2 > 0$. Suppose that conditions $(K2)$ and $(H2)$ hold, and that $f \in \mathcal{C}_{2,d}(b)$ is a compactly supported density. Set $h = \left(\frac{n}{\log n}\right)^{-\frac{1}{d+4}}$. Then,for any $s > 0$,*

$$E\left|P_f\{f_n \leq c_n\} - \alpha\right| = O(h^{2-s}). \tag{2.3.10}$$

Proof:

Let $t > 0$ and observe that,

$$\begin{aligned} & E\left|P_f\{f_n \leq c_n\} - P_{f_n}\{f_n \leq c_n\}\right| \\ \leq\ & E\left[\int_{B(h^{-t})} |f(\mathbf{z}) - f_n(\mathbf{z})|d\mathbf{z} + \int_{B^c(h^{-t})} (f(\mathbf{z}) + f_n(\mathbf{z}))d\mathbf{z}\right] \\ =\ & E\left[\int_{B(h^{-t})} |f(\mathbf{z}) - f_n(\mathbf{z})|d\mathbf{z}\right], \text{for large } n \end{aligned}$$

$$
\begin{aligned}
& \text{(because both } f \text{ and } K \text{ are compactly supported)} \\
\leq & \left[E\left(\int_{B(h^{-t})} |f(\mathbf{z}) - f_n(\mathbf{z})| d\mathbf{z} \right)^2 \right]^{1/2} \\
\leq & \left[E \int_{B(h^{-t})} |f(\mathbf{z}) - f_n(\mathbf{z})|^2 d\mathbf{z} \right]^{1/2} \\
& \text{(by Jensen and Cauchy-Schwarz inequalities)} \\
\leq & \left[\sup_{\mathbf{z} \in \Re^d} E|f(\mathbf{z}) - f_n(\mathbf{z})|^2 \int_{B(h^{-t})} d\mathbf{z} \right]^{1/2}.
\end{aligned}
$$

By Corollary 2.1 of Bosq (1998, pp. 46), and under the stated conditions, one finds

$$\sup_{\mathbf{z} \in \Re^d} E|f(\mathbf{z}) - f_n(\mathbf{z})|^2 = O(h^4).$$

Since $\int_{B(h^{-t})} d\mathbf{z} = O(h^{-td})$, one concludes that

$$E\left\{ |P_f\{f_n \leq c_n\} - \alpha| \right\} \leq O(h^{2-dt/2}). \tag{2.3.11}$$

Choosing $s = dt/2$ completes the proof. □

In summary, in the first part of this chapter, we have obtained the same rate of convergence (to zero) of $E|P_n(\mathbf{Z}) - P_\infty(\mathbf{Z})|$, under much weaker conditions than those of Baillo *et al.*(2001), for α-mixing sequences with both exponential and polynomial mixing rates. Here our proofs are very different from those of Baillo *et al.* (2001), requiring no technical conditions imposed on the underlying density. For the case of false alarm probability problem (random c case in section 2.3), we established the same results(with the same rates of convergence) without imposing the strict technical condition on the density function as in Baillo *et al.* (2001). Furthermore, the *i.i.d.* assumption was relaxed to one based on α-mixing sequences. We required no additional conditions to establish our results.

2.4 Simulation study

In this section, simulation studies are preformed in order to verify (numerically) the convergence of level sets, discussed in the previous sections. These studies are carried out for 6 different sample sizes, $n = 50, 100, 200, 500, 1000$ and 2000. The simulations are only designed for the case of convergence rates for false alarm probability (Theorem 2.3.1 and 2.3.2).

In the following subsections, we first discuss the methods for generating α-mixing sequences with the mixing coefficient decay rates at exponential and polynomial order.

2.4.1 α-mixing sequence generation

The following algorithm generates α-mixing sequences with exponentially decreasing mixing coefficients. The marginal distribution of the variables is taken to be a mixture of two normals: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where the values of μ_1, μ_2, σ_1^2 and σ_2^2 are given in the following section.

Algorithm 1:

Choose $0 < \alpha < 1$, $\beta = \sqrt{1-\alpha^2}$, $0 < p < 1$, $-\infty < \mu_1 < \infty$, $-\infty < \mu_2 < \infty$, and $\sigma_1 > 0, \sigma_2 > 0$.

Step 1. Generate $\xi_1 \sim N(0, \sigma_1^2)$, $\eta_1 \sim N(0, \sigma_2^2)$ and $\delta_1 \sim B(p)$, where $B(p)$ is the Bernoulli distribution with parameter p.

Step 2. Put $X_1 = \mu_1 + \xi_1$, $Y_1 = \mu_2 + \eta_1$ and $Z_1 = \delta_1 X_1 + (1-\delta_1)Y_1$.

For $i = 2, ..., n$, perform steps 3-4:

Step 3. Generate $\xi_i \sim N(0, \sigma_1^2)$, $\eta_i \sim N(0, \sigma_2^2)$ and $\delta_i \sim B(p)$ independently.

Step 4. Put $X_i = (1-\alpha)\mu_1 + \alpha X_{i-1} + \beta\xi_i$, $Y_i = (1-\alpha)\mu_2 + \alpha Y_{i-1} + \beta\eta_i$ and

$Z_i = \delta_i X_i + (1 - \delta_i) Y_i$.

In the following theorem, we will prove that the sequence generated by the above algorithm is an α-mixing one.

Theorem 2.4.1. . *Algorithm 1 generates an α-mixing sequence $\{Z_i\}$, where each $Z_i, i = 1, ..., n$ has a marginal distribution: $Z_i \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$ and the mixing coefficient between Z_i and Z_j is:*

$$C\alpha^{|i-j|},$$

where $C = \frac{p^2\sigma_1^2 + (1-p)\sigma_2^2}{p^2\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(\mu_1 - \mu_2)^2}$.

Proof:

Clearly, $Z_1 \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$. For $i = 2, .., .n$, recursively, one can prove that $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_i \sim N(\mu_2, \sigma_2^2)$, therefore $Z_i \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$. As for the correlation coefficient, note that $E(Z_i) = E[E(Z_i|\delta_i)] = p\mu_1 + (1-p)\mu_2$ and

$$\begin{aligned} V(Z_i) &= E(Z_i^2) - (EZ_i)^2 = E[E(Z_i^2|\delta_i)] - (p\mu_1 + (1-p)\mu_2)^2 \\ &= p(\mu_1^2) + (1-p)(\mu_2^2 + \sigma_2^2) - (p\mu_1 + (1-p)\mu_2)2 \\ &= p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(\mu_1 - \mu_2)^2. \end{aligned}$$

Without loss generality, let $i < j$, then $E(X_i X_j) = (1-\alpha)\mu_1^2 + \alpha E(X_i X_{j-1})$. Recursively, one has $E(X_i X_j) = \mu_1^2 + \alpha^{j-i}\sigma_1^2$, and, $E(Y_i Y_j) = \mu_2^2 + \alpha^{j-i}\sigma_2^2$. Therefore $E(Z_i Z_j) = p^2 E(X_i X_j) + p(1-p)E(X_i Y_j + X_j Y_i) + (1-p)^2 E(Y_i Y_j) = [p\mu_1 + (1-p)\mu_2]^2 + \alpha^{j-i}[p^2\sigma_1^2 + (1-p)\sigma_2^2]$. Hence, $Cov(Z_i, Z_j) = E(Z_i Z_j) - (EZ_i EZ_j) = \alpha^{j-i}[p^2\sigma_1^2 + (1-p)\sigma_2^2]$, and the correlation between Z_i and Z_j is $\rho(Z_i, Z_j) = C\alpha^{j-i}$,

where $C = \frac{p^2\sigma_1^2+(1-p)\sigma_2^2}{p^2\sigma_1^2+(1-p)\sigma_2^2+p(1-p)(\mu_1-\mu_2)^2}$. This proves $\{Z_i, i = 1, ..., n\}$ is ρ-mixing and hence α-mixing with an exponential mixing coefficient decay rate. (Note that the ρ-mixing is defined in Section 4.2)

That completes the proof. □

Motivated by the above algorithm, we can design the following algorithm to generate a sequence which is an approximate α-mixing sequence with polynomial order of decay rate.

Algorithm 2.

Step 1. Generate $\xi_1 \sim N(0, \sigma_1^2)$, $\eta_1 \sim N(0, \sigma_2^2)$ and $\delta_1 \sim B(p)$, where $B(p)$ is the Bernoulli distribution with parameter, p.

Step 2. Put $X_1 = \mu_1 + \xi_1$, $Y_1 = \mu_2 + \eta_1$, $Z_1 = \delta_1 X_1 + (1 - \delta_1)Y_1$, $S = X_1, T = Y_1$ and $k = 1$,

For $i = 2, ..., n$, perform steps 3-5:

Step 3. Generate $\xi_i \sim N(0, \sigma_1^2)$, $\eta_i \sim N(0, \sigma_2^2)$ and $\delta_i \sim B(p)$ independently.

Step 4. Put $X_i = (1 - \alpha)\mu_1 + \alpha S + \beta\xi_i$, $Y_i = (1 - \alpha)\mu_2 + \alpha T + \beta\eta_i$, and $Z_i = \delta_i X_i + (1 - \delta_i)Y_i$.

Step 5. If $i = 2k$, then $S = X_i, T = Y_i, k = i$, else skip this step and go back to step 3 until all $\{Z_i\}, i = 1, ..., n$ are generated.

Note that the sequence $\{Z_n\}$ generated by this algorithm is not stationary. However, its subsequence, $\{Z_{i_k}, i_k = 2^k, k = 0, 1, ...,\}$, is an α-mixing sequence with an exponential mixing rate according to Theorem 2.4.1. The fact that for $2^k \leq j < 2^{k+1}$, $\rho(Z_1, Z_j) = O(\rho(Z_1, Z_{2^k})) = O(\alpha^k)$ implies that $\rho(Z_1, Z_j) = O(\alpha^{\log_2 j}) = O(j^{\log_2 \alpha})$, and $\rho(Z_i, Z_j) \geq \rho(Z_1, Z_{j-i}) = O(|j-i|^{\log_2 \alpha})$. Algorithm 2 generates a sequence with the same normal mixtures as its marginal distribution. Note that the correlation

coefficient decreases at a polynomial rate. We have proved in Theorem 2.3.2 that the error term (2.3.10) still converges when the mixing coefficient decays polynomially with orders lower than $\frac{2d+1}{d+1}$. Our simulation results in Tables 1, 2 and 3 in the next section also verify our results.

2.4.2 Simulation Results and Analysis

Algorithms 1 and 2 are used to generate samples with both exponential and polynomial mixing rates. An *i.i.d.* sample is also generated by taking $\alpha = 0$ in Algorithm 1. The initial parameters are chosen as: $p = 0.3, \mu_1 = 0, \mu_2 = 1, \sigma_1 = 1, \sigma_2 = 5$ and $\alpha = 0.2$ (in Table 1), $p = 0.3, \mu_1 = 0, \mu_2 = 1, \sigma_1 = 1, \sigma_2 = 5$ and $\alpha = 0.4$ (in Table 2) and $\alpha = 0.9$ (in Table 3). In these simulations, we study the convergence (to zero) of the error $|P_f\{f_n \leq c_n\} - \alpha|$ in (2.3.3) or (2.3.10), as the sample size n increases. Both *i.i.d.* and mixing sequences are considered. Here, c_n is chosen to satisfy $P_{f_n}\{f_n \leq c_n\} \simeq 0.05$ in (2.3.1). The Gaussian kernel is used for the density estimation and the bandwidth is taken to be $h = (\log n/n)^{1/5}$. The results are summarized in Table 1 for 6 different sample sizes. For every case, the simulations are repeated 500 times and the average 'error' and its standard error are listed in the tables. The results show that, except for some sample-to-sample fluctuations, the 'error' tends to zero (as n gets larger) for *i.i.d.* and mixing sequences with exponential rates. For mixing sequences with polynomial rates, the errors are convergent when $\alpha < 0.5$ and the smaller the value of α is, the faster the estimated level set converges to the true one with the sample size increasing.

Table 1. The error of level sets vs. number of sample($\alpha = 0.2$).

α	SamSize	h	s	c_n	t	error	std	type of Sam
0.2	50	0.601	0.05	0.015	0.103	0.053	0.029	Indep
0.2	50	0.601	0.05	0.015	0.106	0.057	0.030	Mix-e
0.2	50	0.601	0.050	0.015	0.109	0.059	0.031	Mix-p
0.2	100	0.54	0.050	0.014	0.079	0.030	0.018	Indep
0.2	100	0.54	0.050	0.013	0.078	0.029	0.020	Mix-e
0.2	100	0.54	0.050	0.013	0.081	0.0031	0.019	Mix-p
0.2	200	0.484	0.05	0.013	0.067	0.017	0.013	Indep
0.2	200	0.484	0.050	0.013	0.064	0.015	0.012	Mix-e
0.2	200	0.484	0.050	0.013	0.069	0.019	0.014	Mix-p
0.2	500	0.416	0.05	0.012	0.056	0.006	0.005	Indep
0.2	500	0.416	0.050	0.013	0.054	0.004	0.005	Mix-e
0.2	500	0.416	0.050	0.012	0.059	0.010	0.007	Mix-p
0.2	1000	0.37	0.05	0.012	0.053	0.003	0.003	Indep
0.2	1000	0.37	0.050	0.012	0.051	0.002	0.003	Mix-e
0.2	1000	0.37	0.050	0.012	0.056	0.006	0.005	Mix-p
0.2	2000	0.328	0.050	0.012	0.051	0.001	0.002	Indep
0.2	2000	0.328	0.050	0.012	0.051	0.001	0.002	Mix-e
0.2	2000	0.328	0.050	0.012	0.055	0.005	0.004	Mix-p

Table 2. The error of level sets vs. number of sample($\alpha = 0.4$).

α	SamSize	h	s	c_n	t	error	std	type of Sam
0.4	50	0.601	0.05	0.015	0.103	0.053	0.029	Indep
0.4	50	0.601	0.05	0.015	0.106	0.057	0.033	Mix-e
0.4	50	0.601	0.050	0.016	0.124	0.074	0.037	Mix-p
0.4	100	0.54	0.050	0.014	0.079	0.030	0.018	Indep
0.4	100	0.54	0.050	0.013	0.079	0.030	0.021	Mix-e
0.4	100	0.54	0.050	0.014	0.095	0.0045	0.025	Mix-p
0.4	200	0.484	0.05	0.013	0.067	0.017	0.013	Indep
0.4	200	0.484	0.050	0.012	0.065	0.016	0.013	Mix-e
0.4	200	0.484	0.050	0.013	0.081	0.031	0.021	Mix-p
0.4	500	0.416	0.05	0.012	0.056	0.007	0.005	Indep
0.4	500	0.416	0.050	0.012	0.055	0.006	0.006	Mix-e
0.4	500	0.416	0.050	0.012	0.074	0.024	0.015	Mix-p
0.4	1000	0.37	0.05	0.012	0.053	0.004	0.003	Indep
0.4	1000	0.37	0.050	0.012	0.052	0.003	0.003	Mix-e
0.4	1000	0.37	0.050	0.012	0.070	0.020	0.014	Mix-p
0.4	2000	0.328	0.050	0.012	0.051	0.002	0.002	Indep
0.4	2000	0.328	0.050	0.012	0.051	0.002	0.001	Mix-e
0.4	2000	0.328	0.050	0.012	0.069	0.019	0.014	Mix-p

Table 3. The error of level sets vs. number of sample($\alpha = 0.9$).

α	SamSize	h	s	c_n	t	error	std	type of Sam
0.9	50	0.601	0.05	0.015	0.103	0.053	0.029	Indep
0.9	50	0.601	0.05	0.021	0.183	0.133	0.061	Mix-e
0.9	50	0.601	0.050	0.028	0.308	0.258	0.089	Mix-p
0.9	100	0.54	0.050	0.014	0.079	0.030	0.018	Indep
0.9	100	0.54	0.050	0.018	0.133	0.083	0.046	Mix-e
0.9	100	0.54	0.050	0.026	0.293	0.243	0.084	Mix-p
0.9	200	0.484	0.05	0.013	0.067	0.017	0.013	Indep
0.9	200	0.484	0.050	0.015	0.098	0.048	0.030	Mix-e
0.9	200	0.484	0.050	0.025	0.292	0.242	0.088	Mix-p
0.9	500	0.416	0.05	0.012	0.056	0.007	0.005	Indep
0.9	500	0.416	0.050	0.013	0.075	0.025	0.019	Mix-e
0.9	500	0.416	0.050	0.024	0.298	0.248	0.091	Mix-p
0.9	1000	0.37	0.05	0.012	0.053	0.004	0.003	Indep
0.9	1000	0.37	0.050	0.013	0.067	0.017	0.011	Mix-e
0.9	1000	0.37	0.050	0.024	0.301	0.251	0.093	Mix-p
0.9	2000	0.328	0.050	0.012	0.051	0.002	0.002	Indep
0.9	2000	0.328	0.050	0.013	0.064	0.014	0.007	Mix-e
0.9	2000	0.328	0.050	0.023	0.305	0.255	0.092	Mix-p

It is not surprising that in the above tables, the convergence rates for α-mixing sequences with exponential order are higher than those with polynomial order, but are lower than those of independent samples. The errors all decrease to zero as sample size increase, except for the case of polynomial order with $\alpha = 0.9$. We have shown that the correlation coefficient of the sequence generated by algorithm 2 is

$O(n^{\log_2 \alpha})$. For example, for the case where $\alpha = 0.2$, the order of the correlation coefficient is lower than $O(n^{-\log_2 5}) = O(n^{-2.3})$.However, when $\alpha = 0.9 > 0.5$, then the order will be larger than $O(n^{-1})$, which results in the errors to diverge as shown in the above simulation.

In the next simulation, different bandwidths, h are chosen to show the relationship between the convergence rates and the bandwidths. Here the sample size is taken to be $n = 1000$.

The results suggest that a bandwidth of order $(\log n/n)^{1/5}$ typically performs better in reducing the error rates.

Table 4. The error of level sets vs. bandwidths($\alpha = 0.4$).

bandwidth	SamSize	h	s	c_n	t	error	std	Sample
$(\log n/n)^{1/3}$	1000	0.190	0.050	0.012	0.058	0.008	0.006	Indep
$(\log n/n)^{1/3}$	1000	0.190	0.050	0.012	0.056	0.006	0.005	Mix-e
$(\log n/n)^{1/3}$	1000	0.190	0.051	0.012	0.075	0.025	0.014	Mix-p
$(\log n/n)^{1/5}$	1000	0.37	0.05	0.012	0.053	0.004	0.003	Indep
$(\log n/n)^{1/5}$	1000	0.37	0.050	0.012	0.052	0.003	0.003	Mix-e
$(\log n/n)^{1/5}$	1000	0.37	0.050	0.012	0.070	0.020	0.014	Mix-p
$2(\log n/n)^{1/5}$	1000	0.739	0.05	0.012	0.049	0.001	0.002	Indep
$2(\log n/n)^{1/5}$	1000	0.739	0.05	0.012	0.049	0.001	0.003	Mix-e
$2(\log n/n)^{1/5}$	1000	0.739	0.05	0.012	0.065	0.016	0.013	Mix-p
$3(\log n/n)^{1/5}$	1000	1.109	0.050	0.011	0.046	0.004	0.003	Indep
$3(\log n/n)^{1/5}$	1000	1.109	0.050	0.011	0.046	0.004	0.004	Mix-e
$3(\log n/n)^{1/5}$	1000	1.109	0.050	0.012	0.062	0.012	0.012	Mix-p

Chapter 3

Rates of Convergence of Nonparametric Regression: *i.i.d.* Cases

3.1 Introduction

In regression analysis, a sample with n random vectors $\{\mathbf{X}_i, Y_i\}, i = 1, ..., n$ is given, where n is the size of the sample, X_i is $\Re^d$-valued and Y is $\Re$-valued. The main objective is to use the sample to find a function $f \in \mathcal{F} : \Re^d \to \Re$ such that $f(X)$ is a "good approximation to Y". When the structure of the regression function is assumed to be known and depends only on a number of parameters, the method is sometimes called parametric regression; when the structure of regression function is unknown, it is called nonparametric regression.

The early theoretical work on nonparametric regression, in the general context of nonparametric function estimation, includes those of Vapnik and Chervonenkis (1974), Bosq and Lecoutre (1987), Devroye, Gyorfi and Lugosi (1996), Lee, Bartlett and Williamson (1996), Bosq (1998), Gyorfi, Kohler, Krzyzak and Walk (2002).

In this chapter, I will focus on convergence properties of nonparametric regression

under the *i.i.d.* case. I will start with reviewing some important results which provide the background for our work in the subsequent sections.

3.2 Basic Concepts and Consistency of Least Squares Estimates

Consider the model

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i, \tag{3.2.1}$$

where $f \in \mathcal{F}$ and $\mathcal{F} : \Re^d \to \Re$ is a given class of functions. Then, the regression function $m(\mathbf{x}) = \mathbf{E}\{Y|\mathbf{X} = \mathbf{x}\}$ is the solution (in the L_2 sense) of the following problem

$$\mathbf{E}|m(\mathbf{X}) - Y|^2 = \min_{f\in\mathcal{F}} \mathbf{E}|f(\mathbf{X}) - Y|^2. \tag{3.2.2}$$

Clearly, to find the exact regression function, $m(\mathbf{X})$, which is the solution to (3.2.2), one has to know the joint distribution of $(\mathbf{X}, Y)$. When the joint distribution is completely unknown, which is the case in real applications, one has to use the observed data $\mathcal{D}_n = \{\mathbf{Z}_i\}_{i=1}^n = \{\mathbf{X}_i, Y_i\}_{i=1}^n$, to construct a nonparametric estimate of $m(.)$. The nonparametric estimate $m_n(.)$ is defined as the solution of the empirical L_2 risk minimization problem:

$$m_n(.) = \arg\min_{f\in\mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2 \right\} \tag{3.2.3}$$

where $\mathcal{F}_n$ is a class of functions on $\Re^d$ which could be the original class $\mathcal{F}$. Here the index n in $\mathcal{F}_n$ implies that the cardinality of $\mathcal{F}$ is allowed to increase with n. The function $m_n(.)$ is called the least-squared estimate. Note that

$$\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2|\mathcal{D}_n\} = \mathbf{E}\{|[m_n(\mathbf{X}) - m(\mathbf{X})] + [m(\mathbf{X}) - Y]|^2|\mathcal{D}_n\}$$
$$= \mathbf{E}\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2|\mathcal{D}_n\} + \mathbf{E}\{|m(\mathbf{X}) - Y|^2\},$$

where we have used the fact that $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$ is independent of $\mathcal{D}_n$ and

$$\begin{aligned}
&\mathbf{E}\{(m_n(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)|\mathcal{D}_n\} \\
&= \mathbf{E}\Big\{\mathbf{E}\{(m_n(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)|\mathbf{X}\}\Big|\mathcal{D}_n\Big\} \\
&= \mathbf{E}\Big\{(m_n(\mathbf{X}) - m(\mathbf{X}))\mathbf{E}\{m(\mathbf{X}) - Y)|\mathbf{X}\}\Big|\mathcal{D}_n\Big\} \\
&= \mathbf{E}\{(m_n(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - m(\mathbf{X}))\}|\mathcal{D}_n\} = 0.
\end{aligned}$$

Hence,

$$\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2|\mathcal{D}_n\} = \mathbf{E}\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2|\mathcal{D}_n\} + \mathbf{E}|m(\mathbf{X}) - Y|^2. \qquad (3.2.4)$$

Thus the L_2 risk of an estimate m_n is close to the optimal value if and only if the L_2 error

$$\mathbf{E}\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2|\mathcal{D}_n\} = \int_{\mathcal{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2\mu(d\mathbf{x}) \qquad (3.2.5)$$

is close to zero, where μ is the probability measure of $\mathbf{X}$. To study the consistency properties of $m_n(.)$, we first state a number of standard definitions:

Definition 3.2.1. *A sequence of regression estimates $\{m_n\}$ is called* ***weakly consistent (in L_2)*** *for a certain distribution of $(\boldsymbol{X}, Y)$ if*

$$\lim_{n\to\infty} \boldsymbol{E}\{|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2\} = 0.$$

Furthermore, $\{m_n\}$ is called ***weakly universally consistent (in L_2)*** *if it is weakly consistent for all distributions of $(\boldsymbol{X}, Y)$ with $\boldsymbol{E}\{Y^2\} < \infty$.*

Definition 3.2.2. *A sequence of regression function estimates $\{m_n\}$ is called* ***strongly consistent (in L_2)*** *for a certain distribution of $(\boldsymbol{X}, Y)$ if*

$$\lim_{n\to\infty} \boldsymbol{E}\{|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2\Big|\mathcal{D}_n\} = 0. \quad \text{a.s.}$$

Again, $\{m_n\}$ is called ***strongly universally consistent (in L_2)*** *if it is strongly consistent for all distributions of $(\boldsymbol{X}, Y)$ with $\boldsymbol{E}\{Y^2\} < \infty$.*

The main step to obtain the consistency of the L_2 error in (3.2.5) is to split it into two terms. We have the following fundamental lemma.

Lemma 3.2.1. *Let m and m_n be as above. Then*

$$E\left[|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \middle| \mathcal{D}_n\right] \leq 2 \sup_{f \in \mathcal{F}_n} |\overline{g_f} - \boldsymbol{E}g_f| + \inf_{f \in \mathcal{F}_n} \boldsymbol{E}\{|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2\}, \quad (3.2.6)$$

where $g_f(\boldsymbol{Z}) = |f(\boldsymbol{X}) - Y|^2$, $\boldsymbol{Z} = \{\boldsymbol{X}, Y\}$, $\boldsymbol{E}g_f =: \boldsymbol{E}g_f(\boldsymbol{Z})$ and $\overline{g_f} =: \frac{1}{n}\sum_{i=1}^n g_f(\boldsymbol{Z}_i)$. Here, $g_f : \Re^d \times [-B, B] \to R^+$.

Proof:

Using the decomposition in (3.2.4), one has,

$$\begin{aligned}
E\left[|m_n(\mathbf{X}) - m(\mathbf{X})|^2 \middle| \mathcal{D}_n\right] &= E\{m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - E\{|m(\mathbf{X}) - Y|^2\} \\
&= \left(E\{m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \inf_{f \in \mathcal{F}_n} E|f(\mathbf{X}) - m(\mathbf{X})|^2\right) \\
&\quad + \left(\inf_{f \in \mathcal{F}_n} E|f(\mathbf{X}) - m(\mathbf{X})|^2 - E\{|m(\mathbf{X}) - Y|^2\}\right) \\
&\leq \left(E\{m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \inf_{f \in \mathcal{F}_n} E|f(\mathbf{X}) - m(\mathbf{X})|^2\right) \\
&\quad + \inf_{f \in \mathcal{F}_n} E|f(\mathbf{X}) - m(\mathbf{X})|^2
\end{aligned}$$

However,

$$\begin{aligned}
&E\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \inf_{f \in \mathcal{F}_n} E|f(\mathbf{X}) - m(\mathbf{X})|^2 \\
&= \sup_{f \in \mathcal{F}_n} \left(E\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \frac{1}{n}\sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2\right.
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n}|m_n(\mathbf{X}_i)-Y_i|^2-\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{X}_i)-Y_i|^2\\
&+\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{X}_i)-Y_i|^2-E\{|m_n(\mathbf{X}_i)-Y|^2\}\Bigg)\\
&\leq \sup_{f\in\mathcal{F}_n}\Bigg(E\{|m_n(\mathbf{X})-Y|^2|\mathcal{D}_n\}-\frac{1}{n}\sum_{i=1}^{n}|m_n(\mathbf{X}_i)-Y_i|^2\\
&+\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{X}_i)-Y_i|^2-E\{|m_n(\mathbf{X}_i)-Y|^2\}\Bigg)\\
&\leq 2\sup_{f\in\mathcal{F}_n}|\overline{g_f}-\mathbf{E}g_f|
\end{aligned}
$$

□

Hence to get the necessary condition for consistency of m_n, one has to assume the second term on the right side of (3.2.6) to converge to zero as $n \to \infty$. That is true when the class of functions, $\mathcal{F}_n$ is dense in $L_2(\mu)$ for any probability measure μ on $\Re^d$. In general, we do not have any control on the second term. As for the first term on the right side of (3.2.6), if the cardinality of $\mathcal{F}_n$ is finite, it can be shown that it will converge to zero almost surely by Hoeffding's inequality:

Lemma 3.2.2. *For any $\epsilon > 0$, one has,*

$$
\boldsymbol{P}\left\{\left|\frac{1}{n}\sum_{j=1}^{n}g_f(\boldsymbol{Z}_j)-\boldsymbol{E}\{g_f(\boldsymbol{Z})\}\right|>\epsilon\right\}\leq 2e^{-\frac{2n\epsilon^2}{B^2}}, \tag{3.2.7}
$$

where $f \in \mathcal{F}_n$, $\mathcal{F}_n$ is the class of functions bounded by B and $g_f(\boldsymbol{Z})$ is defined as above.

In our least squares problem the cardinality of $\mathcal{F}_n$ is always infinite, to ensure the first term of (3.2.6) converges under this condition, one needs to introduce the notion of ϵ-covering number, which is also used to study the rates of convergence of the L_2 error of least squares estimates in (3.2.5).

Definition 3.2.3. *Let $\epsilon > 0$ and let $\mathcal{F}$ be a class of functions $\Re^d \to \Re$. Let ν be a probability measure on $\Re^d$. For a function $f : \Re^d \to \Re$ set*

$$\|f\|_{L_p(\nu)} := \left\{ \int |f(\boldsymbol{z})|^p d\nu \right\}^{\frac{1}{p}}, \text{ where } 1 \leq p < \infty.$$

(a) Every finite collection of functions $f_1, ..., f_N : \Re^d \to \Re$ with the property that for every $f \in \mathcal{F}$ there is a $j = j(f) \in \{1, ..., N\}$ such that $\|f - f_j\|_{L_p(\nu)} < \epsilon$, is called an ϵ-cover of $\mathcal{F}$ with respect to $\|.\|_{L_p(\nu)}$.

(b) Let $\mathcal{N}(\epsilon, \mathcal{F}, \|.\|_{L_p(\nu)})$ be the size (number) of smallest ϵ-cover of $\mathcal{F}$ w.r.t $\|.\|_{L_p(\nu)}$. Take $\mathcal{N}(\epsilon, \mathcal{F}, \|.\|_{L_p(\nu)}) = \infty$ if no finite ϵ-cover exist. Then $\mathcal{N}(\epsilon, \mathcal{F}, \|.\|_{L_p(\nu)})$ is called an ϵ-cover number of $\mathcal{F}$ w.r.t. $\|.\|_{L_p(\nu)}$.

(c) When $p = \infty$, $\mathcal{N}_\infty(\epsilon, \mathcal{F})$ is the notation for the ϵ-cover number of $\mathcal{F}$ w.r.t. the sup-norm.

(d) Let $\boldsymbol{z}_1^n = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ be n fixed points in $\Re^d$. Let ν_n be the corresponding empirical measure, i.e.,

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(\boldsymbol{z}_i) \quad (A \subseteq \Re^d).$$

Then

$$\|f\|_{L_p(\nu_n)} = \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(\boldsymbol{z}_i)|^p \right\}^{\frac{1}{p}}$$

and any ϵ-cover of $\mathcal{F}$ w.r.t. $\|.\|_{L_p(\nu_n)}$ will be called an $L_p(\nu_n)$ ϵ-cover of $\mathcal{F}$ on $\boldsymbol{z}_1^n$ and the ϵ-covering number of $\mathcal{F}$ w.r.t. $\|.\|_{L_p(\nu_n)}$ will be denoted by $\mathcal{N}_p(\epsilon, \mathcal{F}, \boldsymbol{z}_1^n)$.

In the case of sup norm, we have the following examples. (See Sara Van de Geer, (2000) Chapter 2)

Example 1. Let $\mathcal{C}$ be a class of increasing functions $g : \Re \to [0,1]$ and let $\mathcal{X} \subset \Re$ be a finite set with cardinality n, then $\log \mathcal{N}_\infty(\epsilon, \mathcal{C}) \leq \frac{1}{\epsilon} \log\left(n + \frac{1}{\epsilon}\right)$, for all $\epsilon > 0$.

Example 2. Let $\mathcal{C} = \{g : [0,1] \to [0,1], |g| \leq 1\}$. Then for some constant A, $\log \mathcal{N}_\infty(\epsilon, \mathcal{C}) \leq A\frac{1}{\epsilon}$, for all $\epsilon > 0$.

Pollard (1984) proved the following fundamental theorem, which is useful in obtaining consistency results for nonparametric regression estimates.

Theorem 3.2.4. *Let $\mathcal{G}$ be a set of functions $g : \Re^d \to [0, B]$. For any n, and any $\epsilon > 0$,*

$$\boldsymbol{P}\left\{ \sup_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} g(\boldsymbol{Z}_j) - \boldsymbol{E}\{g(\boldsymbol{Z})\} \right| > \epsilon \right\} \leq 8\boldsymbol{E}\mathcal{N}_1(\epsilon/8, \mathcal{G}, \boldsymbol{Z}_1^n) e^{-\frac{n\epsilon^2}{128B^2}}.$$

For the proof of Theorem 3.2.4, one may refer to Pollard (1984). There are various extensions of this inequality; See, for example, Devroye (1982), Alexander (1984), Massart (1990), Van der Vaart and Wellner (1996), etc. Gyorfi *et al.* (2002) presented many examples of estimates that are weakly and strongly universally consistent. The necessary conditions for the consistency of least squares estimates can be found in Van de Geer and Wellner (1996).

Before we finish this section, we need to define the concept of VC (Vapnik-Chervonenkis) dimension, which is used to bound the L_2 errors of nonparametric regression in the following section.

Definition 3.2.5. *Let $\mathcal{A}$ be a class of subsets of $\Re^d$ and $n \in \mathcal{N}$.*

(a) For $z_1, ..., z_n \in \Re^d$, *define*

$$s(\mathcal{A}, \{z_1, ..., z_n\}) = \left|\{A \bigcap \{z_1, ..., z_n\} : A \in \mathcal{A}\}\right|,$$

that is, $s(\mathcal{A}, \{z_1, ..., z_n\})$ *is the number of different subsets of* $\{z_1, ..., z_n\}$ *of the form* $A \bigcap \{z_1, ..., z_n\}$, $A \in \mathcal{A}$.

(b) Let G *be a subset of* $\Re^d$ *of size* n. *One says that* $\mathcal{A}$ *shatters* G *if* $s(\mathcal{A}, G) = 2^n$, *i.e., if each subset of* G *can be represented in the form* $A \bigcap G$ *for some* $A \in \mathcal{A}$.

(c) The n*th shatter coefficient of* $\mathcal{A}$ *is*

$$S(\mathcal{A}, n) = \max_{\{z_1, ..., z_n\} \subseteq \Re^d} s(\mathcal{A}, \{z_1, ..., z_n\}).$$

That is, the shatter coefficient is the maximal number of different subsets of n *points that can be picked out by sets from* $\mathcal{A}$.

(d) Let $\mathcal{A} \neq \emptyset$. *The VC dimension* $V_{\mathcal{A}}$ *of* $\mathcal{A}$ *is defined by*

$$V_{\mathcal{A}} = \sup\{n \in \mathcal{N} : \ S(\mathcal{A}, n) = 2^n\},$$

i.e., the VC dimension $V_{\mathcal{A}}$ *is the largest integer* n *such that there exists a set of* n *points in* $\Re^d$ *which can be shattered by* $\mathcal{A}$.

Example 3. The class of all intervals in $\Re$ of the form $(-\infty, b]$ $(b \in \Re)$ fails to pick out the largest of any two distinct points, hence its VC dimension is 1. Note that $s(\mathcal{A}, 2) = 3 < 2^2$. The class of all intervals in $\Re$ of the form $(a, b]$ $(a, b \in \Re)$ shatters every two-point set but cannot pick out the largest and smallest point of any set of three distinct points. Thus its VC dimension is 2.

One can also study various almost sure properties of the L_2 error given data $\mathcal{D}_n$ by making some assumptions on the rate of growth of the entropy of the class $\mathcal{F}$. More specifically, suppose that there exists constants $s > 0, t \geq 0$ and $M > 0$ such that for every $x > 0$,

$$\log \mathcal{N}_\infty(x, \mathcal{F}) \leq H_t(x) := \begin{cases} Mx^{-t}, & \text{if } t > 0, \quad (A) \\ \log(Mx^{-s}), & \text{if } t = 0. \quad (B) \end{cases} \tag{3.2.8}$$

We have the following examples which satisfy the above condition.

Example 4.(Differentiable functions) For $i = 1, ..., s$, let $k_i \geq 0$ be non-negative integers and put $k = k_1 + ... + k_s$. Also, for any $g : \Re^s \to \Re$, define $D^{(k)}g(\mathbf{u}) = \partial^k g(\mathbf{u})/\partial u_1^{k_1}, ..., \partial u_s^{k_s}$. Consider the class of functions with bounded partial derivatives of order r:

$$\Psi = \left\{ g : [0,1]^d \to \Re^1 \middle| \sum_{k \leq r} \sup_u |D^{(k)}g(\mathbf{u})| \leq A < \infty \right\}.$$

Then, for every $\epsilon > 0$, $\log \mathcal{N}_\infty(\epsilon, \Psi) \leq M\epsilon^{-\alpha}$, where $\alpha = d/r$ and $M \equiv M(d, r)$; this is due to Kolmogorov and Tikhomirov (1959).

Example 5. Consider the class Ψ of all convex functions $\psi : \mathcal{C} \to [0, 1]$, where $\mathcal{C} \subset \Re^d$ is compact and convex. If ψ satisfies $|\psi(\mathbf{z}_1) - \psi(\mathbf{z}_2)| \leq L|\mathbf{z}_1 - \mathbf{z}_2|$, for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{C}$, then $\log \mathcal{N}_\infty(\epsilon, \Psi) \leq M\epsilon^{-d/2}$, for every $\epsilon > 0$, where $M \equiv M(d, L)$; see Van der Vaart and Wellner (1996).

We also need the following definition for the VC dimension of subgraphs of functions

Definition 3.2.6. *Let $\mathcal{F}$ be a class of functions on $\Re^d$ taking their values in $[0, B]$. The class of sets*

$$\mathcal{F}^+ := \left\{ \{(\boldsymbol{z}, t) \in \Re^d \times \Re; t \leq f(\boldsymbol{z})\}; f \in \mathcal{F} \right\},$$

is called the class of subgraphs of $\mathcal{F}$. Denote the VC dimension of the set of all subgraphs of functions of $\mathcal{F}$ as V_{F^+}.

3.3 Convergence Rate of Nonparametric Regression

In the previous section, we discussed the problem of consistency of nonparametric regression. We introduced the basic concepts of the ϵ entropy number and the VC dimension, as well as the unform law of large numbers for nonparametric regression. In this section we deal with the convergence rates in these problems. Here the concept of ϵ-covering number is a key technical tool. I will present the following results and some simple extensions for *i.i.d.* sequences as the basic tools for the next chapter's results where I will extend them to mixing sequences. Theorem 3.3.1 is a key result for the derivation.

Theorem 3.3.1. *Let* $\boldsymbol{Z} = (\boldsymbol{X}, Y)$ *be a* $\Re^d \times \Re$*-valued random vector, where* $|Y| \leq B_n$, $B_n \geq 1$. *Let* $\mathcal{F}$ *be a set of functions* $f : \Re^d \longrightarrow \Re$ *satisfying* $|f(\boldsymbol{x})| \leq B_n$. *Let* $\{\boldsymbol{Z}_i\}_{i=1}^n$ *be* i.i.d. $\sim Z$ *and define* $g_f(\boldsymbol{Z}_i) = |f(\boldsymbol{X}_i) - Y_i|^2 - |m(\boldsymbol{X}_i) - Y_i|^2, i = 1, ..., n$. *Also put* $\boldsymbol{E}g_f =: \boldsymbol{E}g_f(\boldsymbol{Z})$ *and* $\overline{g_f} =: \frac{1}{n}\sum_{i=1}^n g_f(\boldsymbol{Z}_i)$. *Then for each* $n \geq 1$,

$$P\{\exists f \in \mathcal{F} : \boldsymbol{E}g_f - \overline{g_f} \geq \varepsilon \cdot (\alpha + \beta + \boldsymbol{E}g_f)\}$$

$$\leq 14\boldsymbol{E}\mathcal{N}_1(\frac{\beta\varepsilon}{20B_n}, \mathcal{F}, \boldsymbol{X}_1^n) \exp(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B_n^4}) \tag{3.3.1}$$

where $\alpha, \beta > 0$ *and* $0 < \varepsilon \leq 1/2$. *Here* $\mathcal{N}_1(\epsilon, \mathcal{F}, \boldsymbol{X}_1^n)$ *is the* L_1 ε*-covering number of* $\mathcal{F}$ *on* $\boldsymbol{X}_1^n$ *and* $\boldsymbol{X}_1^n$ *was defined in the previous section.*

The proof of Theorem 3.3.1 is rather tedious and involves 7 major steps, including symmetrization and randomization by a "ghost" sample, as well as conditioning and

bounding the covering number. The details of the derivation can be found in Kohler (2000), Gyorfi *et al.* (2002) and Lee *et al.* (1996).

Let $\tilde{m}_n(.)$ be the solution of the following empirical risk minimization problem

$$\tilde{m}_n(.) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2, \tag{3.3.2}$$

where $\mathcal{F}_n$ is a class of function $f : \Re^d \to \Re$, which could grow with n. Let $m_n(.)$ be the truncated version of $\tilde{m}_n(.)$:

$$m_n(\mathbf{x}) = T_{B_n}\tilde{m}_n(\mathbf{x}) := \begin{cases} \tilde{m}_n(\mathbf{x}), & \text{if} |\tilde{m}_n(\mathbf{x})| \leq B_n, \\ \pm B_n, & \text{otherwise.} \end{cases} \tag{3.3.3}$$

Then we have the following theorem which gives a non-asymptotic bound on the L_2 risk of m_n.

Theorem 3.3.2. *Let the estimate m_n be as defined above. Then under the conditions of Theorem 3.3.1 and $1 \leq B_n < \infty$, one has, for every $n \geq 1$,*

$$\boldsymbol{E}\{|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \Big| \mathcal{D}_n\}$$

$$\leq \frac{c_1}{n} + \frac{(c_2 + c_3 \log(n)) V_{\mathcal{F}_n^+}}{n} + 2 \inf_{f \in \mathcal{F}_n} \boldsymbol{E}\{|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2\}, \tag{3.3.4}$$

where

$$c_1 = 24 \cdot 214 B_n^4 (1 + \log 42), c_2 = 48 \cdot 214 B_n^4 \log(480 e B_n^2),$$

$$c_3 = 48 \cdot 214 B_n^4,$$

and $V_{\mathcal{F}_n^+}$ is the Vapnik-Chervonenkis dimension of the class of all subgraphs of functions $f \in \mathcal{F}_n$ which was defined at the end of last section.

For the proof of Theorem 3.3.2, one may refer to Theorem 11.5 of Gyorfi *et al.* (2002). The following Theorem 3.3.3 gives the bound from *the other side.* For the proof of Theorem 3.3.3, one may refer to Theorem 11.6 of Gyorfi *et al.* (2002):

Theorem 3.3.3. *Let g_f and $\overline{g_f}$ be as in Theorem 3.3.1. Assume $\delta > 0, 0 < \eta < 1$, and $n \geq 1$. Then under the conditions of Theorem 3.3.1, one has:*

$$P\left\{ \sup_{f\in\mathcal{F}} \frac{\overline{g_f} - \boldsymbol{E}g_f}{\delta + \overline{g_f} + \boldsymbol{E}g_f} > \eta \right\} \leq 4\boldsymbol{E}\mathcal{N}_1(\frac{\delta\eta}{5}, \mathcal{F}_n, \boldsymbol{x}_1^n) \exp(-\frac{3\eta^2\delta n}{40B_n}) \tag{3.3.5}$$

The following corollary is an immediate consequence of the above theorems

Corollary 3.3.4. *Let $B_n \geq 1$ and let $\mathcal{F}_n$ be a set of functions $f : \Re^d \to [0, B_n]$, g_f and $\overline{g_f}$ be as in theorem 3.3.1. Let $\alpha, \beta, \gamma > 0$ and $0 < \varepsilon \leq 1$. Then one has:*

$$P\left\{ \sup_{f\in\mathcal{F}_n} \frac{\boldsymbol{E}g_f - \overline{g_f}}{\alpha + \beta + \boldsymbol{E}g_f} > \varepsilon \right\} \leq 14\boldsymbol{E}\mathcal{N}_1(\frac{\beta\varepsilon}{20B_n}, \mathcal{F}_n, \boldsymbol{x}_1^n) \exp(-\frac{\varepsilon^2\alpha n}{428B_n^4}), \tag{3.3.6}$$

and,

$$P\left\{ \sup_{f\in\mathcal{F}_n} \frac{\overline{g_f} - \boldsymbol{E}g_f}{\gamma + \boldsymbol{E}g_f} > \varepsilon \right\} \leq 4\boldsymbol{E}\mathcal{N}_1(\frac{2\gamma\varepsilon}{15}, \mathcal{F}_n, \boldsymbol{x}_1^n) \exp(-\frac{\varepsilon^2\gamma n}{60B_n}) \tag{3.3.7}$$

Proof: The proof of (3.3.6) is trivial. For (3.3.7), first observe that, for any $0 < \eta \leq 1/3$ and $\delta > 0$, one has

$$\begin{aligned}
& P\{\exists f \in \mathcal{F}_n : \overline{g_f} - \mathbf{E}g_f \geq \eta \cdot (\delta + \overline{g_f} + \mathbf{E}g_f)\} \\
& = P\{\exists f \in \mathcal{F}_n : (1-\eta)(\overline{g_f} - \mathbf{E}g_f) \geq \eta \cdot (\delta + 2\mathbf{E}g_f)\} \\
= \; & P\left\{\exists f \in \mathcal{F}_n : \frac{\overline{g_f} - \mathbf{E}g_f}{\delta/2 + \mathbf{E}g_f} \geq \frac{2\eta}{1-\eta}\right\} \\
\leq \; & 4\mathbf{E}\mathcal{N}_1(\frac{\delta\eta}{5}, \mathcal{F}_n, \mathbf{x}_1^n) \exp(-\frac{3\eta^2\delta n}{40B_n})
\end{aligned}$$

Put $\gamma = \delta/2, \varepsilon = \frac{2\eta}{1-\eta}$, (thus $0 < \varepsilon \leq 1$). Then one has:

$$P\left\{\exists f \in \mathcal{F}_n : \frac{\overline{g_f} - \mathbf{E}g_f}{\gamma + \mathbf{E}g_f} \geq \varepsilon\right\} \leq 4\mathbf{E}\mathcal{N}_1(\frac{2\gamma\varepsilon}{15}, \mathcal{F}_n, \mathbf{x}_1^n) \exp(-\frac{\varepsilon^2\gamma n}{60B_n})$$

□

We first state a technical corollary

Corollary 3.3.5. *Let $\gamma = \alpha+\beta$, where $\alpha > 0, \beta > 0$ are arbitrary. If $B_n \leq B < \infty$, then:*

$$\lim_{n\to\infty} \Delta_n \sup_{f\in\mathcal{F}_n} \left|\frac{\overline{g_f} - \boldsymbol{E}g_f}{\gamma + \boldsymbol{E}g_f}\right| = 0, \quad a.s. \tag{3.3.8}$$

$$\Delta_n = \begin{cases} (\frac{n}{log_k n})^{\frac{1}{2+t}}, & \text{for all } k = 1, 2, ..., \text{ if (3.2.8A) holds,} \\ \\ n^{\frac{1-u}{2}}, & \text{for all arbitrary } u > 0, \text{ if (3.2.8B) holds.} \end{cases}$$

Here $log_k n$ is the k-th iterated logarithm.

Proof:

Under condition (3.2.8A), by Corollary 3.3.4, for every $0 < \varepsilon \leq 1/3$ one has

$$\begin{aligned} P\left\{\Delta_n \sup_{f\in\mathcal{F}} \frac{\mathbf{E}g_f - \overline{g_f}}{\gamma + \mathbf{E}g_f} > \varepsilon\right\} &= P\left\{\sup_{f\in\mathcal{F}_n} \frac{\mathbf{E}g_f - \overline{g_f}}{\gamma + \mathbf{E}g_f} > \frac{\varepsilon}{\Delta_n}\right\} \\ &\leq 14\mathbf{E}\mathcal{N}_1(\frac{\beta\varepsilon}{20\Delta_n B_n}, \mathcal{F}_n, \mathbf{x}_1^n) \exp(-\frac{\varepsilon^2\alpha n}{428\Delta_n^2 B_n^4}) \\ &\leq 14\exp\left\{M\left(\frac{\beta\varepsilon}{20\Delta_n B_n}\right)^{-t} - \frac{\varepsilon^2\alpha n}{428\Delta_n^2 B_n^4}\right\} \\ &= 14\exp\left\{-(Log_k n)^{\frac{2}{2+t}} n^{\frac{t}{2+t}}\left[\frac{\varepsilon^2\alpha}{428B_n^4} - \frac{M(20B_n)^t}{(\beta\varepsilon)^t Log_k n}\right]\right\}. \end{aligned}$$

Similarly, one has

$$P\left\{\Delta_n \sup_{f\in\mathcal{F}_n} \frac{\overline{g_f} - \mathbf{E}g_f}{\gamma + \mathbf{E}g_f} > \varepsilon\right\} \leq$$
$$4\exp\left\{-(log_k n)^{\frac{2}{2+t}} n^{\frac{t}{2+t}}\left[\frac{3\varepsilon^2\gamma}{40B_n} - M(\frac{15B_n}{2\gamma\varepsilon})^t/log_k n\right]\right\}.$$

Since $B_n \leq B < \infty$, and $\frac{1}{log_k^n} \to 0$ as $n \to \infty$, an application of the B-C lemma yields (3.3.8), where $\Delta_n = (\frac{n}{log_k n})^{\frac{1}{2+t}}$. Similarly one can show that (3.3.8) also holds with $\Delta_n = n^{\frac{1-u}{2}}$, for arbitrary small $u > 0$, under (3.2.8B). □

The above corollary can be used to study the consistency properties of m_n. The following almost sure bound is the immediate result of Lemma 3.2.1 and Corollary 3.3.5.

Theorem 3.3.6. *Let m and m_n be as before. Let $\mathcal{F}_n$ be a class of functions, $f : \Re^d \to \Re$, satisfying condition (3.2.8). Then*

$$E\left[|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \middle| \mathcal{D}_n\right] \leq$$
$$o(\Delta_n^{-1}) + \inf_{f \in \mathcal{F}_n} E|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \quad a.s., \tag{3.3.9}$$

where Δ_n is defined in Corollary 3.3.5.

3.4 Complexity-Regularized Least Squares

In the previous section we reviewed the main results on the convergence rates of nonparametric regression and presented some extensions (Corollary 3.3.4, Corollary 3.3.5 and Theorem 3.3.6). In this section, we will discuss the problem of complexity-regularized nonparametric regression. It is clear that increasing the sample size n also increases the cardinality of the class $\mathcal{F}_n$, which in turn increases the complexity of the problem and results in estimates that poorly fit new data. To control the complexity of the regression estimation, a penalized term is sometimes introduced into (3.2.3) which monotonically increases with the complexity of $\mathcal{F}_n$. That is the so called "Complexity Regularization" of the least squares estimates.

Let $\mathcal{P}_n$ denote a finite set of parameters. With $p \in \mathcal{P}_n$, let $\mathcal{F}_{n,p}$ represent a class of functions $f : \Re^d \to \Re$. Also let $pen_n(p) \in \mathcal{R}_+$ be a complexity penalty for $\mathcal{F}_{n,p}$. Note that the penalty is monotonically increasing with the complexity of $\mathcal{F}$. The penalty depends on the class of functions from which one chooses the estimate. Let $\tilde{m}_{n,p}(.) = \tilde{m}_{n,p}(., D_n) \in \mathcal{F}_{n,p}$ be the solution of the following minimization problem

$$\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_{n,p}(X_i) - Y_i|^2 = \min_{f \in \mathcal{F}_{n,p}} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{3.4.1}$$

Let $m_{n,p}(.) = T_{B_n}\tilde{m}_{n,p}(.)$ be the truncated least squares estimate of m in $\mathcal{F}_{n,p}$ where

$$m_{n,p}(\mathbf{x}) = T_{B_n}\tilde{m}_{n,p}(\mathbf{x}) \quad := \quad \begin{cases} \tilde{m}_{n,p}(\mathbf{x}), & \text{if } |\tilde{m}_{n,p}(\mathbf{x})| \le B_n, \\ \pm B_n, & \text{otherwise.} \end{cases} \tag{3.4.2}$$

Next choose an estimate m_{n,p^*} minimizing the sum of the empirical L_2 risk of m_{n,p^*} and $pen_n(p^*)$, that is, choose $p^* = p^*(D_n) \in \mathcal{P}_n$ such that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} |m_{n,p^*}(\mathbf{X}_i) - Y_i|^2 + pen_n(p^*) \\ &= \min_{p \in \mathcal{P}_n} \left\{ \frac{1}{n}\sum_{i=1}^{n} |m_{n,p}(\mathbf{X}_i) - Y_i|^2 + pen_n(p) \right\} \end{aligned} \tag{3.4.3}$$

and set

$$m_n(., D_n) = m_{n,p^*(D_n)}(., D_n). \tag{3.4.4}$$

One example of such complexity penalty is given by Gyorfi *et al.* (2002): Let

$$\mathcal{P}_n = \{(M, K) \in \mathcal{N}_0 \times \mathcal{N} : 0 \le M \le \log n, 1 \le K \le n\}.$$

For $(M, K) \in \mathcal{P}_n$ let $\mathcal{F}_{n,(M,K)}$ be the set of all piecewise polynomials of degree M (or less) w.r.t. an equidistant partition of $[0, 1]$ into K intervals. One penalty we can use as the upper bound on the right-hand side of (3.4.5) is

$$pen_n((M, K)) = log(n)^2 \cdot \frac{K(M+1)}{n}.$$

We have the following results

Theorem 3.4.1. *Let m_n be the minimum complexity regression estimate with a penalty term $pen_n(p)$ satisfying*

$$pen_n(p) \geq 2 \cdot 2568 \frac{B_n^4}{n} \left(\log(120 e B_n^4 n) V_{\mathcal{F}_{n,p}^+} + \frac{c_p}{2} \right), \tag{3.4.5}$$

for some $c_p \in \mathcal{R}_+$ satisfying

$$\sum_{p \in \mathcal{P}_n} e^{-c_p} \leq 1. \tag{3.4.6}$$

Then under the conditions of Theorems 3.3.1 and 3.3.2, one has

$$\boldsymbol{E}\left\{ |m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \middle| \mathcal{D}_n \right\}$$

$$\leq 2 \inf_{p \in \mathcal{P}_n} \left\{ pen_n(p) + \inf_{f \in \mathcal{F}_{n,p}} \boldsymbol{E}\{|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2\} \right\} + 5 \cdot 2568 \frac{B_n^4}{n}. \tag{3.4.7}$$

For the proof of Theorem 3.4.1, one may refer to Theorem 12.1 of Gyorfi et. al (2002).

Chapter 4

On Nonparametric Regression with β-Mixing Sequence

4.1 Introduction

The study of the convergence rate of the L_2 error has been extended to dependent cases by a number of authors. Modha and Masry (1996) studied the minimum complexity regression estimates for m-dependent and strongly mixing sequences. Their proposed estimators are adapted to a list of parametric models based on neural networks. The L_2 convergence rates they obtained are of order $O((\log n/n)^{1/2})$. Baraud, *et al.* (2001) studied the problem of estimating an unknown regression function for β-mixing sequences. They establish non-asymptotic error bounds under some restrictive assumptions. They show that their bounds are nearly optimal as compared with the rates in the *i.i.d.* case. However, their results hold on compact sets only. Furthermore, the density of the $\mathbf{X}$'s is assumed to be bounded away from zero on such compact sets. Their conditions are still restrictive, the process of derivation is too complex and the results are not easy to apply.

In this chapter we extend the results of Lee *et al.* (1996) to a β-mixing framework.

With a relatively straightforward derivation and under mild conditions, a bound of order $n^{b-1}\log n$ is obtained ($b > 0$ can be arbitrarily small) if the β-mixing rate has a polynomial order. The bound becomes of order $\log^2 n/n$ if an exponential mixing rate is assumed. Unlike the results of Baraud *et al.* (2001), our results are not confined to compact sets. Similar results are also obtained for the complexity-regularized estimators in section 3.

4.2 Background for β-Mixing Processes

In this section we start with some definitions of dependent sequences and introduce an important technique called the coupling technique. The following definitions can be found, for example, in Bosq (1998, Chapter 2).

Definition 4.2.1. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $\mathcal{B}, \mathcal{C}$ be two sub σ fields of $\mathcal{A}$. Define*

$$(a)\quad \alpha = \alpha(\mathcal{B},\mathcal{C}) = \sup_{B\in\mathcal{B},\ C\in\mathcal{C}} |P(BC) - P(B)P(C)|.$$

$$(b)\quad \beta = \beta(\mathcal{B},\mathcal{C}) = \boldsymbol{E}\sup_{C\in\mathcal{C}} |P(C) - P(C|\mathcal{B})|.$$

$$(c)\quad \varphi = \varphi(\mathcal{B},\mathcal{C}) = \sup_{B\in\mathcal{B},C\in\mathcal{C},P(B)>0} |P(C) - P(C|B)|.$$

$$(d)\quad \rho = \rho(\mathcal{B},\mathcal{C}) = \sup_{X\in L^2(\mathcal{B}),\ Y\in L^2(\mathcal{C})} |corr(X,Y)|.$$

Let $\mathcal{F}_a^b = \sigma(\mathbf{Z}_i, a \le i \le b)$.

Definition 4.2.2. *A sequence* $\{\boldsymbol{Z}_n, n \geq 1\}$ *is said to be* α*-mixing if*

$$\alpha(n) = \sup_{k \in \mathcal{N}} \alpha(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \longrightarrow 0 \quad as \ n \longrightarrow \infty$$

A sequence $\{\boldsymbol{Z}_n, n \geq 1\}$ *is said to be* β*-mixing if*

$$\beta(n) = \sup_{k \in \mathcal{N}} \beta(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \longrightarrow 0 \quad as \ n \longrightarrow \infty$$

A sequence $\{\boldsymbol{Z}_n, n \geq 1\}$ *is said to be* φ*-mixing if*

$$\varphi(n) = \sup_{k \in \mathcal{N}} \varphi(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \longrightarrow 0 \quad as \ n \longrightarrow \infty$$

A sequence $\{\boldsymbol{Z}_n, n \geq 1\}$ *is said to be* ρ*-mixing if*

$$\rho(n) = \sup_{k \in \mathcal{N}} \rho(\mathcal{F}_1^k, \mathcal{F}_{k+n}^{\infty}) \longrightarrow 0 \quad as \ n \longrightarrow \infty$$

For the above 4 coefficients we have the following inequalities:

$$2\alpha \leq \beta \leq \varphi \tag{4.2.1}$$

$$4\alpha \leq \rho \leq 2\varphi^{1/2}. \tag{4.2.2}$$

Hence we have the following scheme:

$$\begin{array}{ccc} \varphi\text{-mixing} & \Rightarrow & \beta\text{-mixing} \\ \Downarrow & & \Downarrow \\ \rho\text{-mixing} & \Rightarrow & \alpha\text{-mixing} \end{array}$$

The following lemma, due to Berbee (1979), is an important and useful coupling technique for handling β-mixing sequences via *i.i.d.* random vectors.

Lemma 4.2.1. *(Berbee's lemma) Let (X, Y) be a $\Re^d \times \Re^{d'}$-valued random vector. Then there exists a $\Re^{d'}$-valued random vector Y^* such that*

(1) $P_{Y^*} = P_Y$ and Y^* is independent of X,(where P is a probability measure)

(2) $P(Y^* \neq Y) = \beta(\sigma(X), \sigma(Y))$, where $\sigma(X)$ is the σ-field generated by random variable X

The following examples can be seen in Bosq (1998).

Example 4.1. Consider the linear process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}, t \in \mathbb{Z}$$

where the ε_t's are independent zero-mean real random variables with a common density and finite second moment. Then the series above converges in quadratic mean, and (X_t) is ρ-mixing and therefore α-mixing with coefficients which decrease to zero at an exponential rate.

The following example of a β-mixing sequence is given by Baraud *et al.* (2001).

Example 4.2. Consider a sequence ε_i for $i \in \mathbb{Z}$ and we take the X_i's to be generated by a standard time series model:

$$\mathbf{X}_i = \sum_{k=0}^{+\infty} a_k \varepsilon_{i-1-2k}.$$

Also, we make the following assumptions: The ε_i's are *i.i.d.* Gaussian random variables. The a_j's are such that $a_0 = 1$, $\sum_{j=0}^{+\infty} a_j z^{2j} \neq 0$ for all$z$ with $|z| \leq 1$ and for all $j \geq 1, |a_j| \leq Cj^{-d}$ for some constants $C > 0$ and $d > 17$, $Y_i = f(\mathbf{X}_i) + \varepsilon_i, i = 1, ..., n$ is a regression model. Then the sequence $(\mathbf{X}_i, Y_i)$'s is a β-mixing sequence.

4.3 Empirical Risk for β-Mixing Framework

In this section we extend the main results of Chapter 3 on the rates of convergence of nonparametric regression under the *i.i.d.* assumption to mixing sequences. Let $\{\mathbf{Z}_i,\ 1 \leq i \leq n\}$ be a β-mixing sequence, where $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \Re^{d+1}$, $f \in \mathcal{F}_n$ and $\mathcal{F}_n : \Re^d \to \Re$ is a class of functions satisfying $\left|f(\mathbf{x})\right| \leq B_n$, $|Y_i| < B_n$, $B_n \geq 1$. Define $g_{i,f} := g_f(\mathbf{Z}_i) = |f(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2$, $i = 1, ..., n$. Also put $\overline{g_f} =: \frac{1}{n}\sum_{i=1}^{n} g_f(\mathbf{Z}_i)$. We start by stating the following technical lemma, which may be viewed as an extension of the results of Kohler (2000), Gyorfi *et al.* (2002) and Lee *et al.* (1996). First, note that for any n, there exists m and $l \in \mathcal{N}$, such that $ml \leq n < (l+1)m$. Therefore, one may write $n = ml + r$, where $0 \leq r < m$. The choice of m and l will be given later. The main results about the convergence rate of nonparametric regression under the β-mixing setup are established based on the following lemma.

Lemma 4.3.1. *Let $\mathcal{D}_n = \{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n\}$ be a stationary β-mixing sequence, with the mixing coefficient $\beta(n)$. Also let $\mathcal{F}_n$ be the class of functions defined above. Then, for every $\alpha > 0, \beta > 0$, and $0 < \varepsilon \leq 1/2$, one finds, for large n*

$$\begin{aligned} & P\left\{ \sup_{f \in \mathcal{F}_n} \frac{\boldsymbol{E}g_f - \overline{g_f}}{\alpha + \beta + \boldsymbol{E}g_f} > \varepsilon \right\} \\ \leq\ & 14m\boldsymbol{E}\mathcal{N}_1\left(\frac{\beta\varepsilon}{25B_n}, \mathcal{F}_n, \mathcal{X}_1\right) \exp\left(-\frac{16\varepsilon^2(1-\varepsilon)\alpha l}{25 \cdot 214(1+\varepsilon)B_n^4}\right) + n\beta(m). \quad (4.3.1) \end{aligned}$$

Here $\mathcal{N}_1(\varepsilon, \mathcal{F}_n, \mathcal{X}_1)$ is the L_1 ε-covering number of $\mathcal{F}$ on $\mathcal{X}_1 = \left\{\boldsymbol{X}_j, j \in \{1, m+1, ..., ml_1+1\}\right\}$, and $\boldsymbol{E}g_f = \boldsymbol{E}g_f(\boldsymbol{Z}_1) = \boldsymbol{E}\left[|f(\boldsymbol{X}_1) - Y_1|^2 - |m(\boldsymbol{X}_1) - Y_1|^2\right]$.

Proof: Consider the following sequence of partial sums

$$V_1 = g_{1,f} + g_{m+1,f} + ... + g_{ml_1+1,f}$$

$$V_2 = g_{2,f} + g_{m+2,f} + ... + g_{ml_2+2,f}$$

$$\vdots \qquad\qquad \vdots$$

$$V_r = g_{r,f} + g_{m+r,f} + ... + g_{ml_r+r,f}$$

$$\vdots \qquad\qquad \vdots$$

$$V_m = g_{m,f} + g_{m+m,f} + ... + g_{ml_m+m,f}$$

where $l_1 = l_2 = \cdots = l_r = l$ and $l_{r+1} = \cdots = l_m = l - 1$. If $r = 0$, then $l_1 = l_2 = ... = l_m = l - 1$. Clearly, $\sum_{i=1}^{n} g_{i,f} = \sum_{i=1}^{m} V_i$. Using Berbee's Lemma (1979) recursively, for each $i = 1, ..., m$, there exists an iid sequence $\{\mathbf{Z}_i^*, \mathbf{Z}_{m+i}^*, ..., \mathbf{Z}_{ml_i+i}^*\}$, such that

1. $P\mathbf{Z}_{jm+i}^* = P\mathbf{Z}_{jm+i}$, for $j = 0, 1, ..., l_i$, and $i = 1, 2, ..., m$.

2. $P(\mathbf{Z}_{jm+i}^* \neq \mathbf{Z}_{jm+i}) = \beta(m)$.

Let $W_i = \sum_{j=0}^{l_i} g_f(\mathbf{Z}_{jm+i}^*)$, and put $\mathcal{I}_i = \{i, m+i, ..., ml_i + i\}$. Then

$$\begin{aligned}
& P\left\{ \sup_{f \in \mathcal{F}_n} \frac{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{E}g_f - g_{i,f})}{\alpha + \beta + \mathbf{E}g_f} \geq \varepsilon \right\} \\
= \; & P\left\{ \sup_{f \in \mathcal{F}_n} \frac{\sum_{i=1}^{m}(\mathbf{E}V_i - V_i)}{\alpha + \beta + \mathbf{E}g_f} \geq n\varepsilon \right\} \\
\leq \; & \sum_{i=1}^{m} P\left\{ \sup_{f \in \mathcal{F}_n} \frac{\mathbf{E}V_i - V_i}{\alpha + \beta + \mathbf{E}g_f} \geq \frac{n\varepsilon}{m} \right\} \\
& \text{(note that } \mathbf{E}V_i = l_i \mathbf{E}g_f\text{)} \\
\leq \; & \sum_{i=1}^{m} P\left\{ \sup_{f \in \mathcal{F}_n} \frac{\mathbf{E}V_i - W_i}{\alpha + \beta + \mathbf{E}g_f} \geq \frac{n\varepsilon}{m} \right\} + \sum_{i=1}^{m} P\left\{ \bigcup_{j \in \mathcal{I}_i} (\mathbf{Z}_j \neq \mathbf{Z}_j^*) \right\}
\end{aligned}$$

$$
\begin{aligned}
& \text{(by Berbee's lemma)} \\
\leq & \sum_{i=1}^{m} P\left\{\sup_{f\in\mathcal{F}_n} \frac{\mathbf{E}V_i - W_i}{\alpha+\beta+\mathbf{E}g_f} \geq \frac{n\varepsilon}{m}\right\} + \sum_{i=1}^{m}\sum_{j\in\mathcal{I}_i} P(\mathbf{Z}_j \neq \mathbf{Z}_j^*) \\
:= & \ I_1 + I_2.
\end{aligned}
$$

Put $\varepsilon_i = \frac{n\varepsilon}{m(l_i+1)}$ and let $\delta_n = \frac{n\varepsilon}{m(l+1)}$. Clearly, $\delta_n < \varepsilon_i$ (because $l_i = l-1 < l$ for $i = 1, ..., m$). Also, $\delta_n < \varepsilon$ because $\frac{n}{m(l+1)} < 1$. Now let $g_{i,f}^* = g_f(\mathbf{Z}_i^*)$ and observe that

$$
\begin{aligned}
I_1 &= \sum_{i=1}^{m} P\left\{\sup_{f\in\mathcal{F}_n} \frac{\frac{1}{l_i+1}\sum_{j=0}^{l_i}(\mathbf{E}g_f - g_{jm+i,f}^*)}{\alpha+\beta+\mathbf{E}g_f} \geq \varepsilon_i\right\} \\
&\leq \sum_{i=1}^{m} P\left\{\sup_{f\in\mathcal{F}_n} \frac{\frac{1}{l_i+1}\sum_{j=0}^{l_i}(\mathbf{E}g_f - g_{jm+i,f}^*)}{\alpha+\beta+\mathbf{E}g_f} \geq \delta_n\right\} \\
&\leq 14\sum_{i=1}^{m}\mathbf{E}\mathcal{N}_1\left(\frac{\beta\delta_n}{20B_n}, \mathcal{F}_n, \mathcal{X}_i\right)\exp\left(-\frac{\delta_n^2(1-\delta_n)\alpha(l_i+1)}{214(1+\delta_n)B_n^4}\right) \qquad (4.3.2)
\end{aligned}
$$

(by Theorem 3.3.1 and the fact that $\delta_n \in (0, \frac{1}{2})$),

where $\mathcal{X}_i = \{\mathbf{X}_j^*, j \in \mathcal{I}_i\}$. Since for large n, (or equivalently m and l), one can have $\frac{4}{5}\varepsilon \leq \delta_n \leq \varepsilon$, one immediately concludes that

$$
I_1 \leq 14m\mathbf{E}\mathcal{N}_1\left(\frac{\beta\varepsilon}{25B_n}, \mathcal{F}_n, \mathcal{X}_1\right)\exp\left(-\frac{16\varepsilon^2(1-\varepsilon)\alpha l}{25\cdot 214(1+\varepsilon)B_n^4}\right).
$$

On the other hand, with

$$
I_2 = nP(\mathbf{Z}_1 \neq \mathbf{Z}_1^*) = n\beta(m), \qquad (4.3.3)
$$

one finds,

$$
P\left\{\exists f \in \mathcal{F}_n : \mathbf{E}g_f(\mathbf{Z}) - \frac{1}{n}\sum_{i=1}^{n} g_f(\mathbf{Z}_i) \geq \varepsilon\cdot(\alpha+\beta+\mathbf{E}g_f(\mathbf{Z}))\right\}
$$

$$
\leq 14m\mathbf{E}\mathcal{N}_1\left(\frac{\beta\varepsilon}{25B_n}, \mathcal{F}_n, \mathcal{X}_1\right)\exp\left(-\frac{16\varepsilon^2(1-\varepsilon)\alpha l}{25\cdot 214(1+\varepsilon)B_n^4}\right) + n\beta(m) \qquad (4.3.4)
$$

This completes the proof of the lemma. □

Note that we have defined the nonparametric empirical regression and its truncated version in (3.3.2) and (3.3.3) as $\widetilde{m}_n$ and m_n. In what follows, we consider both polynomial and exponential mixing conditions.

$$\textbf{Condition (P)}: \beta(k) = C_k k^{-\frac{1+a}{b}}, \text{ for any } 0 < b \leq 1 \text{ and } a \geq 1 - b \qquad (4.3.5)$$

$$\textbf{Condition (E)}: \beta(k) = C_k e^{-(1+a)k}, \text{ for any } a \geq 1, \qquad (4.3.6)$$

where $C_k \to C$ as $k \to \infty$. Then we have the following results which is the counterpart of Theorem 3.3.2 of *i.i.d.* case.

Theorem 4.3.1. *Let $\mathcal{D}_n$ be a stationary β-mixing sequence. Suppose that $|Y| \leq B_n$. Then for large n,*

$$\begin{aligned} \boldsymbol{E}\left\{ |m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \,\middle|\, \mathcal{D}_n \right\} &\leq 2 \inf_{f \in \mathcal{F}_n} \boldsymbol{E}|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \\ &+ \begin{cases} \frac{c_{1n}}{n^{1-b}} + \frac{(c_{2n}+c_{3n}\log n)K_n}{n^{1-b}} + \frac{12B_n^2}{n^a}, & \text{if (4.3.5) holds,} \\ \frac{c_{4n}\log n}{n} + \frac{(c_{5n}+c_{6n})K_n \log n}{n} + \frac{12B_n^2}{n^a}, & \text{if (4.3.6) holds,} \end{cases} \end{aligned} \qquad (4.3.7)$$

where $K_n = V_{\mathcal{F}_{n_n^+}}$ is the Vapnik-Chervonenkis dimension of the class of all subgraphs of functions $f \in \mathcal{F}_n$,

$$c_{1n} = 8025 B_n^4 (1 + \log 42 + b \log n),$$

$$c_{2n} = 2 \cdot 8025 B_n^4 \log(600 e B_n^2),$$

$$c_{3n} = 2 \cdot 8025 B_n^4 (1 - b),$$

$$c_{4n} = 8025 B_n^4 \{1 + \log 42 + \log\log n\}$$

$$c_{5n} = 2 \cdot 8025 B_n^4 \log(600 e B_n^2),$$

and

$$c_{6n} = 2 \cdot 8025 B_n^4 \{\log n - \log\log n\}$$

Proof:

Using the fact that $\mathbf{E}\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2 | \mathcal{D}_n\} = \mathbf{E}[|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n] - \mathbf{E}|m(\mathbf{X}) - Y|^2$, one may consider the error decomposition

$$\begin{aligned}
\mathbf{E}\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2 | \mathcal{D}_n\} = & \left\{ \mathbf{E}\left\{ |m_n(\mathbf{X}) - Y|^2 \middle| \mathcal{D}_n \right\} - \mathbf{E}|m(\mathbf{X}) - Y|^2 \right. \\
& \left. - \frac{2}{n} \sum_{i=1}^{n} \left(|m_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right) \right\} \\
& + \frac{2}{n} \sum_{i=1}^{n} \left(|m_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right) \\
:= \; & T_{1n} + T_{2n}
\end{aligned}$$

Given the definition of m_n in (3.3.2) and (3.3.3), and the condition $|Y| \leq B_n$ *a.s.*, one can use Kohler's arguments (Kohler 2000) to write

$$\begin{aligned}
\mathbf{E}T_{2n} = \quad & \mathbf{E}\left\{ \frac{2}{n} \sum_{i=1}^{n} |m_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right\} \\
& \leq \mathbf{E}\left\{ \frac{2}{n} \sum_{i=1}^{n} |\widetilde{m}_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right\} \\
& \leq \mathbf{E}\left\{ \frac{2}{n} \inf_{f \in \mathcal{F}_n} \sum_{i=1}^{n} \{|f(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2\} \right\} \\
& \leq \inf_{f \in \mathcal{F}_n} \mathbf{E}\left\{ \frac{2}{n} \sum_{i=1}^{n} \{|f(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2\} \right\} \\
& = \left\{ \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(\mathbf{X}) - Y|^2 - \mathbf{E}|m(\mathbf{X}) - Y|^2 \right\} \\
& = 2 \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(\mathbf{X}) - m(\mathbf{X})|^2 \qquad (4.3.8)
\end{aligned}$$

As for the term T_{1n}, let $t = 1/l$ and observe that $t \geq \frac{1}{l_i+1}$, one has

$$\begin{aligned}
P\{T_{1n} \geq t\} \leq\ & P\left\{\mathbf{E}\left\{|m_n(\mathbf{X}) - Y|^2 \middle| \mathcal{D}_n\right\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}\right. \\
& -\frac{1}{n}\sum_{i=1}^{n}\{|m(\mathbf{X}_i) - Y_i|^2 - |m_n(\mathbf{X}_i) - Y_i|^2\} \geq \\
& \left.\frac{1}{2}(t + \mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 \middle| \mathcal{D}_n\} - \mathbf{E}|m(\mathbf{X}) - Y|^2)\right\} \\
\leq\ & P\left\{\exists f \in \mathcal{F}_n : \mathbf{E}g_f(\mathbf{Z}) - \frac{1}{n}\sum_{i=1}^{n} g_f(\mathbf{Z}_i) \geq \frac{1}{2}(\frac{t}{2} + \frac{t}{2} + \mathbf{E}g_f(\mathbf{Z}))\right\} \\
\leq\ & 14m\mathbf{E}\mathcal{N}_1\left(\frac{\frac{t}{2}\cdot 0.5}{25B_n}, \mathcal{F}_n, \mathcal{X}_1\right)\exp\left(-\frac{16\cdot 0.5^2(1-0.5)\frac{t}{2}l}{25\cdot 214(1+0.5)B_n^4}\right) + n\beta(m) \\
& \text{(by 4.3.1)} \\
=\ & 14m\mathbf{E}\mathcal{N}_1\left(\frac{t}{100B_n}, \mathcal{F}_n, \mathcal{X}_1\right)\exp\left(-\frac{tl}{8025B_n^4}\right) + n\beta(m) \qquad (4.3.9)
\end{aligned}$$

By Lemma 9.2 and Theorem 9.4 of Gyorfi *et al.* (2002), the expected covering number, $\mathcal{N}_1\left(\frac{t}{100B_n}, \mathcal{F}_n, \mathcal{X}_1\right)$ in the above expression can be bounded by

$$3\left\{\frac{2e(2B_n)}{\frac{1}{100B_nl}}\log\left(\frac{3e(2B_n)}{\frac{1}{100B_nl}}\right)\right\}^{V_{\mathcal{F}_n^+}} \leq 3\left(600eB_n^2l\right)^{2V_{\mathcal{F}_n^+}}$$

Since $T_{1n} \leq \mathbf{E}\left\{|m_n(X) - Y|^2 | \mathcal{D}_n\right\} + \frac{2}{n}\sum_{i=1}^{n}|m(\mathbf{X}_i) - Y_i|^2$, $\forall f \in \mathcal{F}_n$, where $|f| \leq B_n$ and $|Y| \leq B_n$, one concludes that $|m_n(X) - Y| \leq 2B_n$. Therefore $T_{1n} \leq 12B_n^2$. Now observe that

$$\begin{aligned}
& \mathbf{E}|T_{1n}| = \mathbf{E}\{T_{1n}^+ + T_{1n}^-\} \\
& \leq \mathbf{E}T_{1n}^+ = \int_0^\infty P\{T_{1n} > t\}dt \\
& = \int_0^{12B_n^2} P\{T_{1n} > t\}dt \\
& \leq \left[\int_0^\varepsilon + \int_\varepsilon^{12B_n^2}\right] P\{T_{1n} > t\}dt \\
& \leq \varepsilon + \int_\varepsilon^\infty 14m\cdot 3(600eB_n^2l)^{2V_{\mathcal{F}_n^+}}\exp(-\frac{l}{8025B_n^4})dt + 12B_n^2n\beta(m) \qquad (4.3.10)
\end{aligned}$$

In passing, we also note that there is a minor error in the proof of Theorem 11.4 in Gyorfi *et al.* 2002 as they treated T_{1n} as a nonnegative random variable, which is not true in general. In fact T_{1n} will be strictly nonnegative only when $m \in \mathcal{F}_n$ (in which case, $T_{2n} = 0$). Now minimizing $\mathbf{E}T_{1n}$ *w.r.t* ε, it is straightforward to show that the minimizer is obtained when

$$\varepsilon = \frac{8025B_n^4}{l} \log(42m(600eB_n^2 l)^{2V_{\mathcal{F}_n^+}}).$$

Putting the above together, one finds

$$\begin{aligned} \mathbf{E}\{T_{1n}\} &\leq \frac{8025B_n^4}{l}\left\{ \log(42m) + 2V_{\mathcal{F}_n^+} \log(600eB_n^2 l) \right\} \\ &+ \frac{8025B_n^4}{l} + 12B_n^2 n\beta(m). \end{aligned}$$

When Condition P (4.3.5) holds (i.e., a polynomial mixing rate), then upon choosing $m = \lfloor n^b \rfloor$ and $l = \lfloor n^{1-b} \rfloor$, where $\lfloor . \rfloor$ is the floor function, one finds, for large n,

$$\begin{aligned} &\mathbf{E}\left\{ |m_n(\mathbf{X}) - m(\mathbf{X})|^2 \middle| \mathcal{D}_n \right\} \\ &\qquad \leq \frac{c_{1n}}{n^{1-b}} + \frac{(c_{2n} + c_{3n}\log n)V_{\mathcal{F}_n^+}}{n^{1-b}} + \frac{12B_n^2}{n^a} \\ &\qquad +2 \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(\mathbf{X}) - m(\mathbf{X})|^2, \end{aligned}$$

On the other hand, if condition E (4.3.6) holds (i.e., an exponential mixing rate), then by choosing $m = \lfloor \log n \rfloor, l = \lfloor \frac{n}{\log n} \rfloor$, one has for large n,

$$\begin{aligned} &\mathbf{E}\left\{ |m_n(\mathbf{X}) - m(\mathbf{X})|^2 \middle| \mathcal{D}_n \right\} \\ &\leq \frac{c_{4n} cdot \log n}{n} + \frac{(c_{5n} + c_{6n)} \log n V_{\mathcal{F}_n^+}}{n} + \frac{12B_n^2}{n^a} \\ &+2 \inf_{f \in \mathcal{F}_n} \mathbf{E}|f(\mathbf{X}) - m(\mathbf{X})|^2, \end{aligned}$$

This completes the proof of the theorem. □

Remark. Observe that if $B_n \leq B < \infty$, then, under the exponential mixing rate (4.3.6) the above bound, (which is $O(n^{-1}V_{\mathcal{F}_n^+}\log^2 n)$), will be nearly optimal, as compared to the *i.i.d.* case which is of order $O(n^{-1}V_{\mathcal{F}_n^+}\log n)$. On the other hand, for the polynomial mixing rate (4.3.5), the convergence is of order $O(n^{-1+b}V_{\mathcal{F}_n^+}\log n)$.

As in Lemma 4.3.1, we can also apply Berbee's Lemma to Theorem 3.3.3. That yields the following result, which gives a one-sided bound.

Lemma 4.3.2. *Let $\delta > 0, 0 < \eta < 1$, and $n \geq 1$. Then Under the condition of Lemma 4.3.1, one has:*

$$P\left\{\sup_{f\in\mathcal{F}_n}\frac{\overline{g_f}-\boldsymbol{E}g_f}{\delta+\overline{g_f}+\boldsymbol{E}g_f}>\eta\right\}\leq 4m\boldsymbol{E}\mathcal{N}_1(\frac{\delta\eta}{5},\mathcal{F}_n,\boldsymbol{x}_1^n)\exp(-\frac{3\eta^2\delta n}{40B_n})+n\beta(m) \quad (4.3.11)$$

The proof of Lemma 4.3.2 is trivial and thus omitted. Combining Lemma 4.3.1 and Lemma 4.3.2, one can conclude,

Lemma 4.3.3. *Under the conditions of Lemma 4.3.2, one has:*

$$P\left\{\sup_{f\in\mathcal{F}_n}\frac{\boldsymbol{E}g_f-\overline{g_f}}{\alpha+\beta+\boldsymbol{E}g_f}>\varepsilon\right\}\leq 14m\boldsymbol{E}\mathcal{N}_1(\frac{\beta\varepsilon}{20B_n},\mathcal{F}_n,\boldsymbol{x}_1^n)\exp(-\frac{\varepsilon^2\alpha n}{428B_n^4})+n\beta(m), \quad (4.3.12)$$

and,

$$P\left\{\sup_{f\in\mathcal{F}_n}\frac{\overline{g_f}-\boldsymbol{E}g_f}{\gamma+\boldsymbol{E}g_f}>\varepsilon\right\}\leq 4m\boldsymbol{E}\mathcal{N}_1(\frac{2\gamma\varepsilon}{15},\mathcal{F}_n,\boldsymbol{x}_1^n)\exp(-\frac{\varepsilon^2\gamma n}{60B_n})+n\beta(m) \quad (4.3.13)$$

for all $\alpha, \beta, \gamma > 0$ and $0 < \varepsilon \leq 1/2$.

The Borel-Cantelli Lemma in conjunction with Lemma 4.3.3 immediately yields the following result:

Theorem 4.3.2. *Suppose that the entropy satisfies (3.2.8). Let $\mathcal{D}_n$ be a stationary β-mixing sequence. Suppose that $|Y| \leq B_n$. Let $u > 0$ be arbitrary, and put $\gamma = \alpha + \beta$, $\alpha > 0, \beta > 0$. Then,*

$$\lim_{n\to\infty} \Delta_n \sup_{f\in\mathcal{F}_n} \left| \frac{\overline{g_f} - \boldsymbol{E}g_f}{\gamma + \boldsymbol{E}g_f} \right| \stackrel{a.s.}{=} 0, \tag{4.3.14}$$

where

$$\Delta_n = \begin{cases} \left(\frac{n^{1-b}}{log_k n}\right)^{\frac{1}{2+t}}, & \text{if (3.2.8A) and (4.3.6)hold,} \\ n^{\frac{(1-u)(1-b)}{2}}, & \text{if (3.2.8A) and (4.3.5) hold,} \\ \left(\frac{n}{\log n log_k n}\right)^{\frac{1}{2+t}}, & \text{if (3.2.8B) and (4.3.6) hold,} \\ \left(\frac{n}{\log n}\right)^{\frac{1-u}{2}}, & \text{if (3.2.8B) and (4.3.5) hold,} \end{cases} \tag{4.3.15}$$

where t is defined in (3.2.8).

The proof of Theorem 4.3.2 is similar to that of Corollary 3.3.5 and is therefore omitted. Lemma 3.2.1, continues to hold for stationary β-mixing sequences. Thus we have the following result:

Theorem 4.3.3. *Let $m(.)$ and $m_n(.)$ be defined as in Section 3.2. Let $\mathcal{F}_n$ be a class of functions. Then*

$$E\left[|m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \middle| \mathcal{D}_n \right] \leq$$
$$o(\Delta_n^{-1}) + \inf_{f\in\mathcal{F}_n} E|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \quad a.s., \tag{4.3.16}$$

where Δ_n is defined in Theorem 4.3.2.

4.4 Complexity Regression Estimates with β-Mixing Sequences

In the previous section, we extended the results of Kohler (2000), for the convergence rate of the L_2 error of least squares estimators, under *i.i.d.* condition, to a β-mixing setup. In this section we continue to extend our results to least squares estimation with complexity regularization. First let $\mathcal{P}_n$ denote a finite set of parameters. Define m_n as (3.4.1)-(3.4.4) in Section 3.4. The following result provides performance bounds on the L_2-error of m_n:

Theorem 4.4.1. *Let m_n be the minimum complexity regression estimate with a penalty term $pen_n(p)$ satisfying*

$$pen_n(p) \geq 2 \cdot 8025 \frac{B_n^4}{l} V_{\mathcal{F}_{n,p}^+} \log(150 e B_n^2 l) + \frac{c_p}{2}. \tag{4.4.1}$$

Here, $p \in \mathcal{P}_n$, is a class of penalty functions, and the constants $c_p \in \mathcal{R}_+$ satisfy $\sum_{p \in \mathcal{P}_n} e^{-c_p} \leq 1$, l is chosen as in the previous section, depending on the decay rate of the β-mixing sequence. Then

$$\begin{aligned} & \boldsymbol{E}\left\{ |m_n(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \middle| \mathcal{D}_n \right\} \\ \leq\ & 2 \inf_{p \in \mathcal{P}_n} \left\{ pen_n(p) + \inf_{f \in \mathcal{F}_{n,p}} \boldsymbol{E}|f(\boldsymbol{X}) - m(\boldsymbol{X})|^2 \right\} \\ & + \begin{cases} (5 + b \cdot \log n) \cdot 8025 \frac{B_n^4}{n^{1-b}}, & \text{if (4.3.5) holds,} \\ (5 + \log n - \log\log n) \cdot 8025 \frac{B_n^4 \log n}{n}, & \text{if (4.3.6) holds,} \end{cases} \end{aligned} \tag{4.4.2}$$

Proof:

Start with the error decomposition

$$\begin{aligned} E\{|m_n(\mathbf{X}) - m(\mathbf{X})|^2 \Big| \mathcal{D}_n\} &= \left\{ \mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 \Big| \mathcal{D}_n\} - \mathbf{E}|m(\mathbf{X}) - Y|^2 \right. \\ &\quad \left. - \frac{2}{n} \sum_{i=1}^{n} \left(|m_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right) - 2pen_n(p^*) \right\} \\ &\quad + \left\{ \frac{2}{n} \sum_{i=1}^{n} \left(|m_n(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2 \right) + 2pen_n(p^*) \right\} \\ &=: T_{1,n} + T_{2,n} \end{aligned}$$

It is shown in Gyorfi *et al.* (2002) that

$$\mathbf{E}T_{2,n} = 2 \inf_{p \in \mathcal{P}_n} \left\{ \inf_{f \in \mathcal{F}_{n,p}} E\{|f(\mathbf{X}) - m(\mathbf{X})|^2\} + pen_n(p) \right\}.$$

To deal with the term $T_{1,n}$, fix any $t > 0$ and put $\alpha = t + pen_n(p), \beta = pen_n(p)$, and $\varepsilon = 1/2$. Also let $g_{i,f}, V_i$, and W_i be as in Section 3. Then an application of the bound in (4.3.1) yields

$$\begin{aligned} & P\{T_{1,n} > t\} \\ = \; & P\left\{ \exists p \in \mathcal{P}_n, \exists f \in T_{B_n}\mathcal{F}_{n,p}, \frac{1}{n}\sum_{i=1}^{n}(\mathbf{E}g_f - g_{i,f}) \geq \frac{1}{2}\Big((t + pen_n(p)) + pen_n(p) + \mathbf{E}g_f \Big) \right\} \\ & (T_{B_n} \text{ is defined in Section 3.3}) \\ \leq \; & P\left\{ \sup_{p \in \mathcal{P}_n} \sup_{f \in T_{B_n}\mathcal{F}_{n,p}} \sum_{i=1}^{m} \frac{\mathbf{E}V - V_i}{\alpha + \beta + \mathbf{E}g_f} \geq n\varepsilon \right\} \\ \leq \; & \sum_{i=1}^{m} P\left\{ \sup_{p \in \mathcal{P}_n} \sup_{f \in T_{B_n}\mathcal{F}_{n,p}} \frac{\mathbf{E}V - V_i}{\alpha + \beta + \mathbf{E}g_f} \geq \frac{n\varepsilon}{m} \right\} \\ \leq \; & \sum_{i=1}^{m} \left\{ P\left[\sup_{p \in \mathcal{P}_n} \sup_{f \in T_{B_n}\mathcal{F}_{n,p}} \frac{\mathbf{E}V - W_i}{\alpha + \beta + \mathbf{E}g_f} \geq \frac{n\varepsilon}{m} \right] + P\Big(\bigcup_{j \in I_i} (\mathbf{Z}_j \neq \mathbf{Z}_j^*) \Big) \right\} \\ \leq \; & \sum_{i=1}^{m} \sum_{p \in \mathcal{P}_n} P\left\{ \sup_{f \in T_{B_n}\mathcal{F}_{n,p}} \frac{\mathbf{E}V - W_i}{\alpha + \beta + \mathbf{E}g_f} \geq \frac{n\varepsilon}{m} \right\} + n\beta(m) \\ \leq \; & 14m \sum_{p \in \mathcal{P}_n} \mathbf{E}\mathcal{N}_1\left(\frac{\beta\varepsilon}{25B_n}, T_{B_n}\mathcal{F}_{n,p}, \mathcal{X}_1 \right) \exp\left(- \frac{16\varepsilon^2(1-\varepsilon)\alpha l}{25 \cdot 214(1+\varepsilon)B_n^4} \right) + n\beta(m) \end{aligned}$$

(by (4.3.2))

$$\leq 14m \sum_{p\in\mathcal{P}_n} \mathbf{E}\mathcal{N}_1\left(\frac{pen_n(p)}{50B_n}, T_{B_n}\mathcal{F}_{n,p}, \mathcal{X}_1\right) \exp\left(-\frac{(t+pen_n(p))l}{8025B_n^4}\right) + n\beta(m)$$

(Since $pen_n(p) \geq \frac{2}{l}$)

$$\leq 42m \sum_{p\in\mathcal{P}_n} \left(\frac{2e(2B_n)}{\frac{1}{25B_nl}} \log \frac{3e(2B_n)}{\frac{1}{25B_nl}}\right)^{V_{\mathcal{F}^+_{n,p}}} \exp\left(-\frac{pen_n(p)l}{8025B_n^4}\right) \exp\left(-\frac{tl}{8025B_n^4}\right) + n\beta(m)$$

(By Lemma 9.2 and Theorem 9.4 of Gyorfi et al.(2002))

$$\leq \left\{42m \sum_{p\in\mathcal{P}_n} \left(150eB_n^2l\right)^{2V_{\mathcal{F}^+_{n,p}}} \exp\left(-\frac{pen_n(p)l}{8025B_n^4}\right)\right\} \exp\left(-\frac{tl}{8025B_n^4}\right) + n\beta(m)$$

$$:= \{I(n)\} \exp\left(-\frac{tl}{8025B_n^4}\right) + n\beta(m).$$

Note that

$$\begin{aligned} I(n) &\leq 42m \sum_{p\in\mathcal{P}_n} \left(150eB_n^2l\right)^{2V_{\mathcal{F}^+_{n,p}}} \exp\left\{-2V_{\mathcal{F}^+_{n,p}} \log(150eB_n^2l) - c_p\right\} \\ &\leq 42m \sum_{p\in\mathcal{P}_n} \exp(-c_p) \leq 42m. \end{aligned}$$

Since $T_{1,n} < 12B_n^2$ as shown in Theorem 4.3.1, one finds,

$$\begin{aligned} E(T_{1,n}) &= \int_0^\infty P(T_{1,n} > t)dt = \int_0^{12B_n^2} P(T_{1,n} > t)dt \\ &= \int_0^u P(T_{1,n} > t)dt + \int_u^{12B_n^2} P(T_{1,n} > t)dt \\ &\leq u + \int_u^\infty 42m \exp(-\frac{tl}{8025B_n^4}) + 12B_n^2 n\beta(m) \\ &= u + 42m \cdot 8025B_n^4 \frac{1}{l} \exp(-\frac{ul}{8025B_n^4}) + 12B_n^2 n\beta(m). \end{aligned}$$

Let $u = 8025 \cdot \log(42m) \cdot \frac{B_n^4}{l}$. Then

$$ET_{1,n} \leq 8025(1 + \log(42) + \log(m))\frac{B_n^4}{l} + 12B_n^2 n\beta(m)$$

$$\leq 8025(5 + \log(m))\frac{B_n^4}{l} + 12B_n^2 n\beta(m)$$

Hence, when condition (4.3.5) holds, one may choose $m = \lfloor n^b \rfloor$, (which yields $\beta(m) = n^{-(1+a)}$), and, as a result,

$$ET_{1,n} \leq 8025(5 + a\log(n))\frac{B_n^4}{n^{1-b}} + \frac{12B_n^2}{n^a}.$$

On the other hand, if (4.3.6) holds, then the choice $l = \lfloor \frac{n}{\log n} \rfloor$ gives

$$ET_{1,n} \leq (5 + \log n - \log\log n) \cdot 8025\frac{B_n^4 \log n}{n} + \frac{12B_n^2}{n^a}.$$

This completes the proof of the theorem. □

Chapter 5

Confidence Intervals for Population Parameters with Fractional Imputed Data

5.1 Introduction

Item nonresponse occurs frequently in sample surveys for various reasons such as unwillingness of sampled units to supply the desired information, loss of information caused by uncontrollable factors, failure on the part of investigator to gather correct information and so on. Item nonresponse is usually handled by some form of imputation to fill in missing item values. Commonly used deterministic imputation methods, such as mean imputation, lead to inconsistent estimators of the distribution function $\theta = F(a)$, unlike random imputation methods. The latter, however, induces imputation variance due to random selection. Fractional random imputation (Kim and Fuller, 2004), which involves the creation of more than one imputed value in the data file for each missing value, offers a compromise solution. It reduces the imputation variance and also leads to consistent estimators of the parameters. In the data file, J imputed values are created for each missing value Y_i and the

fraction J^{-1} is attached to each imputed value. Typically, $J = 5$ to 10 should be adequate in controlling the imputation variance, as shown in our simulation results.

Suppose that in a random sample $\{Y_i, i = 1, \cdots, n\}$ some Y_i's are missing completely at random (MCAR). Qin, Rao and Ren (2006a) studied mean, random and adjusted random imputation methods, and established the asymptotic normality of the imputed estimators of the mean μ, distribution $\theta = F(a)$ and q-th quantile θ_q, as $n \to \infty$. Based on this result, normal approximation based confidence intervals (CI) on μ, θ and θ_q were constructed. Furthermore, the log-empirical likelihood (EL) ratios for μ, θ and θ_q were also obtained and shown to be asymptotically scaled χ^2_1. This result was used to obtain asymptotically correct EL based confidence intervals on μ, θ and θ_q. In this chapter, we extend the above results to the fractional random imputation. Similar asymptotical results are derived and two types of CI under normality and empirical likelihood are constructed. Results of a simulation study on the finite sample performance of normal approximation based and EL based confidence intervals are reported.

Consider the following *i.i.d.* sample of incomplete data generated from the random vector (Y, δ):

$$(Y_i, \delta_i),\ i = 1, 2, \cdots, n$$

where $\delta_i = 0$ if Y_i is missing; and $\delta_i = 1$ otherwise. Throughout this chapter, we assume that no auxiliary complete data $\{X_i; i = 1, ..., n\}$ is available and Y is missing completely at random (MCAR), i.e. $P(\delta = 1|Y) = P(\delta = 1) = p$. We also assume that $0 < p \leq 1$ and $0 < Var(Y) = \sigma^2 < \infty$. We use $F(\cdot)$ to denote the distribution function of Y. No parametric structure for the distribution of Y is assumed. Our aim is to construct asymptotically valid confidence intervals for the

mean $\mu = EY$, distribution function $\theta = F(a) = P(Y \leq a)$ for given $a \in R$, and q-th quantile $\theta_q = F^{-1}(q),\ 0 < q < 1$.

Let $r = \sum_{i=1}^{n} \delta_i$ and $m = n - r$. Denote the set of respondents as s_r, the set of nonrespondents as s_m, and the mean of respondents as

$$\bar{Y}_r = \frac{1}{r} \sum_{i \in s_r} Y_i.$$

In Qin *et al.* (2006a), three imputation methods are considered: mean imputation (M), random imputation (R) and adjusted random imputation (A). In this chapter, we will focus on random fractional imputation (We denote it as I in whole chapter). Fractional random imputation (I), uses J random imputed values to replace the missed value Y_i (Kalton and Kish (1984), Fay (1996), and Kim and Fuller (2004)). The major objective of using fractional imputation is to reduce the imputation variance. Similar to random imputation (R), fractional imputation selects a simple random sample of size $m \times J$, $\{Y_{ij}^{(R)}; i \in s_m, j = 1, ..., J\}$ with replacement from s_r and then uses the associated Y-values as donors for missing $Y_i, i \in s_m$. For estimating the mean μ, $Y_i^{(I)} = \frac{1}{J} \sum_{j=1}^{J} Y_{ij}^{(R)}$ is used as the imputed value for missing Y_i. However, for θ and θ_q, one has to keep all the J donors for each missing value to reserve the structure of distribution and quantile. The imputed (or completed) data file consists of $\{(Y_i, \delta_i = 1), (Y_{ij}^{(R)}, \delta_i = 0, j = 1, ..., J\}$.

Standard estimators based on the completed data are used to estimate μ, θ and θ_q. The following properties of imputed estimators are well known. Mean imputation eliminates the variance due to imputation, but the distribution of item values is not preserved in the sense that the imputed estimator of θ is inconsistent. Random imputation preserves the distribution of item values, but leads to imputation

variance which can be a significant component of the total variance if the item response rate is not high. Fractional imputation reduces the imputation variance, and at the same time preserves the distribution of item values. By taking $J = \infty$ and $J = 1$, mean imputation and random imputation can be thought of as special cases of fractional imputation. A well chosen value J achieves balance between random and mean imputation. Moreover, the estimator of the distribution function θ is consistent, as shown in this chapter.

The original idea of empirical likelihood dates back to Hartley and Rao (1968) in sample survey context, and to Thomas and Grunkemeier (1975) in survival analysis context. Owen (1988, 1990) made a systematic study of the empirical likelihood (EL) method in the complete data settings. It has several advantages over the normal-approximation based methods and the bootstrap in constructing confidence intervals. EL intervals are range preserving and transformation respecting. Also, the shape and orientation of EL intervals are determined entirely by the data. However, the EL requires modifications in the case of data with imputed values. Wang and Rao (2002a, 2002b) studied the asymptotic properties of EL intervals for the mean μ under deterministic regression imputation. In this Chapter, we obtain asymptotically correct EL intervals for μ, θ and θ_q under R and I, and compare their performance with normal approximation based intervals in a simulation study.

In Section 5.2, we study simple random sample(*i.i.d.* case) and establish the asymptotic normality of the imputed estimators under I and construct normal approximation based confidence intervals for the population parameters μ, θ and θ_q. In Section 5.3, empirical likelihood ratio statistics are constructed, limiting distributions of these statistics are given, and empirical likelihood based confidence intervals

for the population parameters are constructed. In Section 5.4, some simulation results are presented to validate the theoretical results in this chapter. In Section 5.5, results are extended to stratified random sample. In Section 5.6, I summarize the idea of extension to the fractional random linear regression imputation for missing responses and the reason why I did not include it into my thesis as a separate chapter.

5.2 Normal Approximation

5.2.1 Mean μ

Estimator of μ after imputation under I is given by

$$\bar{Y}_I = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i Y_i + (1-\delta_i)Y_i^{(I)}\} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_{I,i}$$

The result on the asymptotic normality of above $\bar{Y}_I$ is studied in Theorem 5.2.1.

Theorem 5.2.1. *Assume that* $0 < p \leq 1$ *and* $0 < Var(Y) = \sigma^2 < \infty$. *Assume that there exists an* $\alpha > 0$ *such that* $E|Y|^{2+\alpha} < \infty$. *Then, as* $n \to \infty$,

$$\sqrt{n}(\bar{Y}_I - \mu) \xrightarrow{d} N(0, (\frac{1-p}{J} + p^{-1})\sigma^2). \tag{5.2.1}$$

For the proof of Theorem 5.2.1, we need the following technical lemma, (Chen and Rao (2007)):

Lemma 5.2.1. *Let* U_n, V_n *be two sequences of random variables and* $\mathcal{B}_n$ *be a* σ*-algebra. Assume that:*

1. *There exists* $\sigma_{1n} > 0$ *such that*

$$\sigma_{1n}^{-1} V_n \xrightarrow{d} N(0,1)$$

as $n \to \infty$, *and* V_n *is* $\mathcal{B}_n$ *measurable.*

2. $E[U_n|\mathcal{B}_n] = 0$ *and* $Var(U_n|\mathcal{B}_n) = \sigma_{2n}^2$ *such that*

$$\sup_t |P(\sigma_{2n}^{-1}U_n \leq t|\mathcal{B}_n) - \Phi(t)| = o_p(1),$$

where $\Phi(\cdot)$ *is the distribution function of the standard normal random variable.*

3. $\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 = \gamma^2 + o_p(1)$.

Then, as $n \to \infty$,

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \xrightarrow{d} N(0,1).$$

Proof of Theorem 5.2.1:

Let $\mathcal{B}_n = \sigma((\delta_i, Y_i), i = 1, ..., n)$, then $E(Y_{I,i}|\mathcal{B}_n) = Y_i\delta_i + \overline{Y}_I(1 - \delta_i)$. Clearly, $E(\overline{Y}_I|\mathcal{B}_n) = \overline{Y}_r$. Let $\sqrt{n}(\overline{Y}_I - \mu) = \sqrt{n}(\overline{Y}_r - \mu) + \sqrt{n}(\overline{Y}_I - \overline{Y}_r) := V_n + U_n$. Then, $Var(V_n) = \frac{1}{p}\sigma^2 := \sigma_{1n}^2$ and $\sigma_{1n}^{-1}V_n \xrightarrow{d} N(0,1)$ by the Central Limit Theorem (CLT) for the *i.i.d.* case. Next, because $E(Y_{i,j}^{(R)}|\mathcal{B}_n) = \overline{Y}_r$ and $Var(Y_{i,j}^{(R)}|\mathcal{B}_n) = E(Y_{i,j}^{2(R)}|\mathcal{B}_n) - \overline{Y}_r^2 = \frac{1}{r}\sum_{i \in s_r}(Y_i - \overline{Y}_r)^2 = S_r^2$, one has: if $i \in S_r$ then $Var(Y_{I,i}|\mathcal{B}_n) = 0$; if $i \in S_m$, then $Var(Y_{I,i}|\mathcal{B}_n) = \frac{1}{J^2}Var(\sum_{j=1}^J Y_{i,j}^{(R)}|\mathcal{B}_n) = \frac{1}{J^2}\sum_{j=1}^J Var(Y_{i,j}^{(R)}|\mathcal{B}_n) = \frac{1}{J}S_r^2$. Hence, letting $p = E(\frac{r}{n})$ be the response rate, one has

$$Var\{\sqrt{n}(\overline{Y}_I - \mu)|\mathcal{B}_n\} = \frac{1}{n}Var\{\sum_{i=1}^n Y_{I,i}|\mathcal{B}_n\}$$
$$= \frac{1}{n}\sum_{i \in S_m} Var(Y_{I,i}|\mathcal{B}_n) = \frac{1}{J}(1 - \frac{r}{n})S_r^2 = \frac{1-p}{J}\sigma^2 + o_p(1) = \sigma_{2n}^2.$$

Thus, one has, $E(U_n|\mathcal{B}_n) = 0$ and $Var(U_n|\mathcal{B}_n) = \frac{1-p}{J}\sigma^2 + o_p(1)$.

Further,

$$\sum_{i=1}^n E(Y_{I,i}^2|\mathcal{B}_n) = \sum_{i \in s_r} Y_i^2 + \sum_{i \in s_m} E((Y_i^{(I)})^2|\mathcal{B}_n)$$
$$= \sum_{i \in s_r} Y_i^2 + \sum_{i \in s_m} \frac{1}{J^2}E\left[(\sum_{j=1}^J Y_{i,j}^{(R)})^2|\mathcal{B}_n)\right]$$

$$= \sum_{i \in s_r} Y_i^2 + \frac{n-r}{rJ} \sum_{i \in s_m} Y_i^2 = n\left(p + \frac{1-p}{J}\right)\left(EY^2 + o_p(1)\right). \qquad (5.2.2)$$

On the other hand, by using induction and the following C_r-inequality:

$$E|X+Y|^r \leq C_r(E|X|^r + E|Y|^r),$$

(where C_r is a constant), one can show that there exists a constant $C_{J,\alpha}$, ($C_{J,\alpha}$ is free of n), such that

$$E\left\{\left|\sum_{j=1}^{J} Y_{i,j}^{(R)}\right|^{2+\alpha} |\mathcal{B}_n\right\} \leq C_{J,\alpha}\left\{\sum_{j=1}^{J} E|Y_{i,j}^{(R)}|^{2+\alpha}|\mathcal{B}_n\right\}.$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^{n} E(|Y_{I,i}|^{2+\alpha}|\mathcal{B}_n) = \sum_{i \in s_r} |Y_i|^{2+\alpha} + \sum_{i \in s_m} E\left\{|Y_i^{(I)}|^{2+\alpha}|\mathcal{B}_n\right\} \\
&= \sum_{i \in s_r} |Y_i|^{2+\alpha} + \sum_{i \in s_m} E\{|\frac{1}{J}\sum_{j=1}^{J} Y_{i,j}^{(R)}|^{2+\alpha}|\mathcal{B}_n\} \\
&\leq \sum_{i \in s_r} |Y_i|^{2+\alpha} + \frac{C_{J,\alpha}}{J^{1+\alpha}} \sum_{i \in s_m} E\{|Y_{i,j}^{(R)}|^{2+\alpha}|\mathcal{B}_n\} \\
&= \sum_{i \in s_r} |Y_i|^{2+\alpha} + \frac{C_{J,\alpha}}{J^{1+\alpha}} \sum_{i \in s_m} \{\frac{1}{r}\sum_{k \in s_r} |Y_k|^{2+\alpha}\} = \sum_{i \in s_r} |Y_i|^{2+\alpha}\left(1 + \frac{n-r}{r}\frac{C_{J,\alpha}}{J^{1+\alpha}}\right) \\
&= n\left(p + \frac{C_{J,\alpha}}{J^{1+\alpha}}(1-p)\right)(E|Y|^{2+\alpha} + o_p(1)). \qquad (5.2.3)
\end{aligned}$$

Thus, by (5.2.2) and (5.2.3),

$$\frac{\sum_{i=1}^{n} E\left(|Y_{I,i}|^{2+\alpha}|\mathcal{B}_n\right)}{\left\{\sum_{i=1}^{n} E(|Y_{I,i}|^2|\mathcal{B}_n)\right\}^{\frac{2+\alpha}{2}}} \leq \frac{n\left[p + \frac{C_{J,\alpha}}{J^{1+\alpha}}(1-p)\right][E|Y|^{2+\alpha} + o_p(1)]}{n^{\frac{2+\alpha}{2}}\left[(p + \frac{1-p}{J})EY^2\right]^{\frac{2+\alpha}{2}}} \to 0,$$

as $n \to \infty$, for any $\alpha > 0$. Thus the Liapunov condition for CLT of independent random variables holds:

$$\sup_t \left|P(\sigma_{2n}^{-1}U_n \leq t|\mathcal{B}_n) - \Phi(t)\right| = o_p(1).$$

Hence, by Lemma 5.2.1, one has,

$$\sqrt{n}(\bar{Y}_I - \mu) \xrightarrow{d} N(0, \sigma^2_{I,u}),$$

where $\sigma^2_{I,u} = (\frac{1-p}{J} + p^{-1})\sigma^2$. □

Using Theorem 5.2.1, normal approximation based confidence intervals for μ may be constructed. Let $X \sim N(0,1)$, and Z_α be such that $P(|X| \leq Z_\alpha) = 1 - \alpha$. Throughout this chapter, we take the observed response rate $\hat{p} = \frac{r}{n} = \sum_{i=1}^{n} \frac{\delta_i}{n}$ as the estimator of p, which is a consistent estimator of p. From Theorem 5.2.1, $\bar{Y}_I$ is a consistent estimator of μ.

To obtain consistent variance estimators under different imputations, we examine the sample variances. Under random fractional imputation, the sample variance is

$$s_I^2 = \frac{1}{n-1}\{\sum_{i \in s_r}(Y_i - \bar{Y}_I)^2 + \sum_{i \in s_m}(Y_i^{(I)} - \bar{Y}_I)^2\}.$$

From the proof of Theorem 5.3.2 one can show that $s_I^2 = (\frac{1-p}{J} + p)\sigma^2 + o_p(1)$. Hence $(1-\alpha)$-level CI for the mean μ under the fractional imputation is given by

$$\left\{\bar{Y}_I - \left(\frac{1-\hat{p}+J\hat{p}^{-1}}{n[(1-\hat{p})+\hat{p}J]}\right)^{1/2} s_I Z_\alpha, \bar{Y}_I + \left(\frac{1-\hat{p}+J\hat{p}^{-1}}{n[(1-\hat{p})+\hat{p}J]}\right)^{1/2} s_I Z_\alpha\right\}. \quad (5.2.4)$$

Random imputation (R) intervals are obtained from (5.2.5) by letting $J = 1$. (Qin *et al.*, 2006a).

5.2.2 Distribution function: $\theta = F(a)$

The estimator of $\theta = F(a)$ after fractional random imputation is given by

$$\hat{F}_I(a) = \frac{1}{n}\left[\sum_{i \in s_r} I(Y_i \leq a) + \frac{1}{J}\sum_{i \in s_m}\sum_{j=1}^{J} I(Y_{ij}^{(R)} \leq a)\right]. \quad (5.2.5)$$

We have the following asymptotic normality result.

Theorem 5.2.2. *Assume that $F(a) > 0$ and a is known. Then,*

$$\sqrt{n}(\hat{F}_I(a) - \theta) \xrightarrow{d} N(0, (\frac{1-p}{J} + p^{-1})\theta(1-\theta)), \tag{5.2.6}$$

as $n \to \infty$.

Proof of Theorem 5.2.2:

Let $\hat{F}_r(a) = \frac{1}{r}\sum_{i \in s_r} I(Y_i \leq a)$. Then,

$$\sqrt{n}(\hat{F}_r(a) - \theta) = \frac{n}{r}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \delta_i (I(Y_i \leq a) - \theta).$$

By CLT for the *i.i.d.* case, one has,

$$(p^{-1}\theta(1-\theta))^{-1/2}\sqrt{n}(\hat{F}_r(a) - \theta) \xrightarrow{d} N(0,1).$$

Let $V_n = \sqrt{n}(\hat{F}_r(a) - \theta)$, $U_n = \sqrt{n}((\hat{F}_I(a) - \hat{F}_r(a))$, then $E(V_n|\mathcal{B}_n) = 0$, $Var(V_n|\mathcal{B}_n) := \sigma_{1n}^2 = \frac{1}{p}\theta(1-\theta) + o_p(1)$, and $\sqrt{n}(\hat{F}_I(a) - \theta) = V_n + U_n$. Note that,

$$\begin{aligned}\hat{F}_I(a) &= \frac{1}{n}\left[\sum_{i \in s_r} I(Y_i \leq a) + \frac{1}{J}\sum_{i \in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \leq a)\right] \\ &= \frac{r}{n}\hat{F}_r(a) + \frac{1}{Jn}\sum_{i \in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \leq a),\end{aligned}$$

$$E\left[I(Y_{ij}^R \leq a)|\mathcal{B}_n\right] = \frac{1}{r}\sum_{i \in s_r} I(Y_i \leq a) = \hat{F}_r(a),$$

and

$$\begin{aligned}Var\left[I(Y_{ij}^R \leq a)|\mathcal{B}_n\right] &= E\left[I^2(Y_{ij}^R \leq a)|\mathcal{B}_n\right] \\ &- \left[E(I(Y_{ij}^{(R)} \leq a)|\mathcal{B}_n)\right]^2 = \hat{F}_r(a)(1 - \hat{F}_r(a)).\end{aligned}$$

Hence, $E(U_n|\mathcal{B}_n) = F_r(a)$ and

$$
\begin{aligned}
Var(U_n|\mathcal{B}_n) &= \frac{1}{n} Var\left\{ \left(\sum_{i \in s_r} I(Y_i \le a) + \frac{1}{J} \sum_{i \in s_m} \sum_{j=1}^{J} I(Y_{ij}^{(R)} \le a) \right) | \mathcal{B}_n \right\} \\
&= \frac{1}{nJ^2} Var\left\{ \sum_{i \in s_m} \sum_{j=1}^{J} I(Y_{ij}^{(R)} \le a) | \mathcal{B}_n \right\} \\
&= \frac{1}{nJ^2} \sum_{i \in s_m} \sum_{j=1}^{J} Var\{ I(Y_{ij}^{(R)} \le a) | \mathcal{B}_n \} \\
&= \frac{1-p}{J} \hat{F}_r(a)(1 - \hat{F}_r(a)) := \sigma_{2n}^2.
\end{aligned}
$$

Because $\hat{F}_r(a) = \theta + o_p(1)$, one has $Var(U_n|\mathcal{B}_n) = \frac{1-p}{J}\theta(1-\theta) + o_p(1)$. By CLT for the independent random variables $I(Y_{ij}^{(R)} \le a)$,

$$
\sup_t \left| P(\sigma_{2n}^{-1} U_n \le t | \mathcal{B}_n) - \Phi(t) \right| = o_p(1).
$$

Hence by Lemma 5.2.1, $\sqrt{n}(\hat{F}_I(a) - \theta) \xrightarrow{d} N(0, \sigma_n^2)$, where $\sigma_n^2 = (\frac{1-p}{J} + p^{-1})\theta(1 - \theta) + o_p(1)$. □

Note that similar to Lemma 2.1 in Qin *et al.* (2006a), $F_I(a)$ is a consistent estimator of θ. Let $\hat{\sigma}_{I,d}^2 = (\frac{1-\hat{p}}{J} + \hat{p}^{-1})\hat{F}_I(a)(1 - \hat{F}_I(a))$, then $(1-\alpha)$-level CI for θ under fractional random imputation is given by

$$
\left\{ \hat{F}_I(a) - n^{-1/2}\hat{\sigma}_{I,d} Z_\alpha, \quad \hat{F}_I(a) + n^{-1/2}\hat{\sigma}_{I,d} Z_\alpha \right\}.
$$

For $J = 1$ the above CI reduces to the CI for random imputation(Qin *et al.* (2006a)).

5.2.3 q-th Quantile, θ_q

We now consider fractional random imputation for estimating the quantile θ_q. Let

$$
\widehat{\theta}_q^F = \inf_u \{ \hat{F}_I(u) \le q \} := \hat{F}_I^{-1}(q),
$$

be the estimated quantile from the complete data file, where $\hat{F}_I(u)$ is defined in (5.2.5). Then we have the following asymptotic normality result:

Theorem 5.2.3. *Assume that there exists an $\alpha > 0$ such that $E|Y|^{2+\alpha} < \infty$, and that the density function $f(\cdot)$ of Y exists and is continuous in a neighborhood of θ_q with $f(\theta_q) > 0$. Then as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_q^{(F)} - \theta_q) \xrightarrow{d} N(0, \sigma_{Iq}^2), \tag{5.2.7}$$

and,

$$\hat{\theta}_q^{(F)} = \theta_q - \frac{\hat{F}_I(\theta_q) - F(\theta_q)}{f(\theta_q)} + o_p(n^{-1/2}), \tag{5.2.8}$$

where $\sigma_{Iq}^2 = (\frac{1-p}{J} + p^{-1})q(1-q)/f^2(\theta_q)$.

Proof of Theorem 5.2.3:

From Theorem 5.2.2, one has,

$$\sqrt{n}(\hat{F}_I(a) - F(a)) \xrightarrow{d} N(0, (\frac{1-p}{J} + p^{-1})F(a)(1 - F(a))).$$

Since $f(\theta_q) > 0$ and $F(a)$ is continuous, for any fixed $u \in \Re$, we have

$$\sqrt{n}\Big[\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\Big] \xrightarrow{d} N(0, (\frac{1-p}{J} + p^{-1})q(1-q)), \tag{5.2.9}$$

as $n \to \infty$. Hence

$$\begin{aligned}
&P\left\{\frac{\sqrt{n}}{\sigma_{I,q}}(\widehat{\theta}_q^F - \theta_q) \le u\right\} = P\{\widehat{\theta}_q^F \le \theta_q + n^{-1/2}\sigma_{F,q}u\} = P\{q \le \hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u)\} \\
&= P\left\{\sqrt{n}\Big[\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\Big]\right. \\
&\quad \left. \ge \sqrt{n}\Big[F(\theta_q) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\Big]\right\} \\
&= P\left\{\sqrt{n}\Big[\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\Big] \ge -\sigma_{I,q}uf(\theta_q) + o(1)\right\} \\
&= P\left\{\frac{\sqrt{n}}{-\sigma_{I,q}f(\theta_q)}\Big[\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\Big] \le u + o(1)\right\} \\
&= \phi(u) + o(1).
\end{aligned}$$

Thus, one can conclude that

$$\sqrt{n}(\hat{\theta}_q^{(F)} - \theta_q) \xrightarrow{d} N(0, \sigma_{Iq}^2),$$

as $n \to \infty$, where, $\sigma_{Iq}^2 = (\frac{1-p}{J} + p^{-1})q(1-q)/f^2(\theta_q)$. Further, the result

$$\frac{\sqrt{n}(\hat{\theta}_q^{(F)} - \theta_q)}{\sigma_{Iq}} = -\frac{\sqrt{n}\left[\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)\right]}{\sigma_{I,q}f(\theta_q)} + o_p(1)$$

gives

$$\begin{aligned}\hat{\theta}_q^{(F)} &= \theta_q - \frac{\hat{F}_I(\theta_q + n^{-1/2}\sigma_{I,q}u) - F(\theta_q + n^{-1/2}\sigma_{I,q}u)}{f(\theta_q)} + o_p(n^{-1/2}) \\ &= \theta_q - \frac{\hat{F}_I(\theta_q) - F(\theta_q)}{f(\theta_q)} + o_p(n^{-1/2}).\end{aligned}$$

□

Hence, $(1-\alpha)$-level CI under fractional random imputation for q-th quantile θ_q is given by

$$[\hat{\theta}_q^{(F)} - n^{-1/2}\hat{\sigma}_{Iq}Z_\alpha, \hat{\theta}_q^{(F)} + n^{-1/2}\hat{\sigma}_{Iq}Z_\alpha],$$

where $\hat{\sigma}_{Iq}^2$ is a consistent estimator of σ_{Iq}^2. Here we take

$$\hat{\sigma}_{Iq}^2 = (\frac{1-\hat{p}}{J} + \hat{p}^{-1})q(1-q)/\hat{f}_I^2(\hat{\theta}_q^{(F)}).$$

where $\hat{f}_I(\hat{\theta}_q^{(I)}) = \frac{F_I(\hat{\theta}_q^{(I)}+n^{-1/2}) - F_I(\hat{\theta}_q^{(I)}-n^{-1/2})}{2n^{-1/2}}$.

For $J = 1$, the above CI reduces to the CI under random imputation (Qin *et al.* (2006a).

From the following Lemma 5.2.2, we see that $\hat{\sigma}_{Iq}^2$ is a consistent estimator of σ_{Iq}^2.

Lemma 5.2.2. *Under the conditions of Theorem 5.2.3,*

$$F_I(\hat{\theta}_q^{(I)}) = f(\theta_q) + o_p(1),$$

The proof of Lemma 5.2.2 follows along the lines of Qin *et. al* (2006a).

5.2.4 Woodruff-type Quantile CI under fractional random imputation

A Woodruff (1952) type CI for θ_q under fractional random imputation is given by

$$[\hat{F}_I^{-1}(q - n^{-1/2}Z_\alpha\{q(1-q)(\frac{1-\hat{p}}{J} + \hat{p}^{-1})\}^{1/2}),$$
$$\hat{F}_I^{-1}(q + n^{-1/2}Z_\alpha\{q(1-q)(\frac{1-\hat{p}}{J} + \hat{p}^{-1})\}^{1/2})].$$

Note that $s_n^2 := q(1-q)(\frac{1-\hat{p}}{J} + \hat{p}^{-1})$ above is a consistent estimator of the variance of $F_I(\theta_q)$. It can be shown that, as $n \to \infty$,

$$P\left[\hat{F}_I^{-1}(q - n^{-1/2}Z_\alpha s_n) \le \theta_q \le \hat{F}_I^{-1}(q + n^{-1/2}Z_\alpha s_n)\right] \to 1 - \alpha. \tag{5.2.10}$$

One can show that, for any $\epsilon_{jn} = O_p(n^{-1/2}), j = 1, 2,$

$$F_I(\theta_q + \epsilon_{1n}) - F_I(\theta_q - \epsilon_{2n}) = f(\theta_q)(\epsilon_{1n} + \epsilon_{2n}) + o_p(n^{-1/2}).$$

Then by Theorem 5.2.2 and following the proof of Theorem 4 in Francisco and Fuller (1991), we have

$$\hat{F}_I^{-1}(q) \pm n^{-1/2}Z_\alpha s_n\{f(\theta_q)\}^{-1} = \hat{F}_I^{-1}(q \pm n^{-1/2}Z_\alpha s_n) + o_p(n^{-1/2}).$$

Therefore, to prove (5.2.10), we only need to show that

$$P\left[\hat{F}_I^{-1}(q) - n^{-1/2}Z_\alpha s_n\{f(\theta_q)\}^{-1} \le \theta_q \le \hat{F}_I^{-1}(q) + n^{-1/2}Z_\alpha s_n\{f(\theta_q)\}^{-1}\right]$$
$$\to 1 - \alpha,$$

which is implied by Theorem 5.2.3.

5.3 Empirical Likelihood CI for Fractional Imputation

5.3.1 Empirical Likelihood CI for Mean

Let

$$Z_{I,i}(\mu) = Y_{I,i} - \mu = \begin{cases} Y_i - \mu, & \text{if } i \in s_r \\ \frac{1}{J}\sum_{j=1}^{J} Y_{ij}^R - \mu, & \text{if } i \in s_m, \end{cases}$$

$\{p_i, i = 1, ..., n\}$ is the empirical distribution. Then the empirical log-likelihood ratio for μ under fractional random imputations (I) is defined as

$$\ell_I(\mu) = -2\sum_{i=1}^{n} \log\{np_{iI}(\mu)\} \tag{5.3.1}$$

where $p_{iI}(\mu), i = 1, ..., n$, maximize $\sum_{i=1}^{n} \log(np_i)$ subject to constraints $\sum_{i=1}^{n} p_i = 1$, $\sum_{i=1}^{n} p_i Z_{I,i}(\mu) = 0$. Using the Lagrange multiplier method, we get

$$p_{iI}(\mu) = \frac{1}{n}\frac{1}{1 + \lambda_I Z_{I,i}(\mu)}$$

where λ_I is the solution of the equation

$$\sum_{i=1}^{n} \frac{Z_{I,i}(\mu)}{1 + \lambda_I Z_{I,i}(\mu)} = 0.$$

Hence,

$$\ell_I(\mu) = 2\sum_{i=1}^{n} \log\{1 + \lambda_I Z_{I,i}(\mu)\},$$

Result on the asymptotic distribution of the above empirical log-likelihood ratio for μ is summarized in Theorem 5.3.1.

Theorem 5.3.1. *Under the conditions that* $0 < p \leq 1$ *and that there exists an* $\alpha > 0$ *such that* $E|Y|^{2+\alpha} < \infty$,

$$\ell_I(\mu) \xrightarrow{d} \frac{1-p+Jp^{-1}}{1-p+Jp}\chi_1^2 \tag{5.3.2}$$

as $n \xrightarrow{d} \infty$.

Proof of Theorem 5.3.1:

Following Owen (1990), under the condition $EY^2 < \infty$, we have

$$\max_{1\leq i\leq n} |Z_{I,i}(\mu)| = o_p(n^{1/2}). \tag{5.3.3}$$

On the other hand,

$$\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i(Y_i-\mu)^2 + (1-\delta_i)\left[\frac{1}{J}\sum_{j=1}^{J}(Y_{ij}^R-\mu)\right]^2\right\} \tag{5.3.4}$$

Since $\frac{1}{r}\sum_{i\in S_r}(Y_i-\mu)^2 = \sigma^2 + o_p(1)$, we have $\frac{1}{n}\sum_{i\in S_r}(Y_i-\mu)^2 = p\sigma^2 + o_p(1)$. Let $W_i = \left(\sum_{j=1}^{J}(Y_{ij}^R-\mu)\right)^2$, for $i \in s_m$, then $\{W_i\}$ are *i.i.d.* given the data set $\mathcal{B}_n$ and because $E\{(Y_{ij}^R-\mu)|\mathcal{B}_n\} = \overline{Y}_r - \mu$, and $E\left(\sum_{j=1}^{J}(Y_{ij}^R-\mu)|\mathcal{B}_n\right) = J(\overline{Y}_r-\mu)$, we get

$$\begin{aligned}
E(W_i|\mathcal{B}_n) &= E\left\{\left(\sum_{j=1}^{J}(Y_{ij}^R-\mu)\right)^2|\mathcal{B}_n\right\} \\
&= E[\sum_{j=1}^{J}(Y_{ij}^R-\overline{Y}_r+\overline{Y}_r-\mu)^2|\mathcal{B}_n] = Var(\sum_{j=1}^{J}Y_{ij}^R|\mathcal{B}_n) + J^2E\{(\overline{Y}_r-\mu)^2|\mathcal{B}_n\} \\
&= \sum_{j=1}^{J}Var(Y_{ij}^R|\mathcal{B}_n) + J^2(\overline{Y}_r-\mu)^2 = J\sigma^2 + o_p(1).
\end{aligned}$$

Noting that $(\overline{Y}_r-\mu)^2 = o_p(1)$, we have

$$\begin{aligned}
\frac{1}{nJ^2}\sum_{i\in s_m}\left[\sum_{j=1}^{J}(Y_{ij}^R-\mu)\right]^2 &= \frac{n-r}{n}\frac{1}{(n-r)J^2}\sum_{i\in s_m}W_i \\
&= \frac{1-p}{J}(\sigma^2+o_p(1)).
\end{aligned}$$

Thus

$$\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\mu) = p\sigma^2 + \frac{1-p}{J}\sigma^2 + o_p(1).$$

On the other hand, by Theorem 5.2.1,

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\mu)\right)^2 = (\sqrt{n}(\overline{Y}_I - \mu))^2 = (\frac{1-p}{J} + p^{-1})\sigma^2 + o_p(1).$$

Now using the same argument used in Theorem 1 of Owen (1990), we conclude that,

$$\ell_I(\mu) = \{\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\mu)\}^{-1}\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\mu)\}^2 + o_p(1) \xrightarrow{d} \frac{1-p+Jp^{-1}}{1-p+Jp}\chi_1^2,$$

where χ_1^2 is a χ^2 variable with one degree of freedom. □

Let $\chi_{1,\alpha}^2$ be the α-quantile of χ_1^2, i.e., $P(\chi_1^2 \leq \chi_{1,\alpha}^2) = 1-\alpha$. Then $(1-\alpha)$-level CI for mean is obtained as

$$\left\{\mu : \frac{1-\widehat{p}+J\widehat{p}}{1-\widehat{p}+J\widehat{p}^{-1}}\ell_I(\mu) \leq \chi_{1,\alpha}^2\right\}. \tag{5.3.5}$$

For $J = 1$ the above CI reduces to the CI for random imputation(Qin *et al.* (2006a)).

5.3.2 Empirical Likelihood CI for Distribution function

Let

$$Z_{I,i}(\theta) = \begin{cases} I(Y_{I,i} \leq y) - \theta, & \text{if } i \in s_r \\ \frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \leq a) - \theta, & \text{if } i \in s_m \end{cases}$$

Then the empirical log-likelihood ratios for θ under I is defined as

$$\ell_I(\theta) = -2\sum_{i=1}^{n} \log\{np_{iI}(\theta)\},$$

where $p_{iI}(\theta), i = 1, ..., n$ maximize $\sum_{i=1}^{n} \log(np_i)$ subject to $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i Z_{I,i}(\theta) = 0$.

It can be shown that

$$\ell_I(\theta) = 2\sum_{i=1}^{n} \log\{1 + \lambda_{I,d} Z_{I,i}(\theta)\},$$

where $\lambda_{I,d}$ is the solution of the equation

$$\sum_{i=1}^{n} \frac{Z_{I,i}(\theta)}{1 + \lambda_{I,d} Z_{I,i}(\theta)} = 0,$$

Result on the asymptotic distribution of the above empirical log-likelihood ratio for θ is summarized in Theorem 5.3.2 below.

Theorem 5.3.2. *Assume that $F(a) > 0$, and that there exists an $\alpha > 0$ such that $E|Y|^{2+\alpha} < \infty$. Then as $n \to \infty$,*

$$\ell_I(\theta) \xrightarrow{d} \frac{1 - p + Jp^{-1}}{1 - p + Jp} \chi_1^2 \tag{5.3.6}$$

Proof of Theorem 5.3.2:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\theta) &= \frac{1}{n}\Big\{ \sum_{i\in s_r} (I(Y_i \le a) - \theta)^2 + \sum_{i\in s_m} [\frac{1}{J}[\sum_{j=1}^{J} I(Y_{ij}^R \le a) - \theta]^2 \Big\} \\
&= \frac{1}{n}\Big\{ \sum_{i\in s_r} I(Y_i \le a) - 2\theta \sum_{i\in s_r} I(Y_i \le a) + r\theta^2 + \sum_{i\in s_m} \Big[\frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le a)\Big]^2 \\
&\quad - 2\theta \sum_{i\in s_m} \frac{1}{J}[\sum_{j=1}^{J} I(Y_{ij}^R \le a) + (n-r)\theta^2] \Big\} \\
&= \frac{1}{n}\Big\{ \Big[\sum_{i\in s_r} I(Y_i \le a) + \frac{1}{J}\sum_{i\in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \le a)\Big] \\
&\quad - 2\theta\Big[\sum_{i\in s_r} I(Y_i \le a) + \frac{1}{J}\sum_{i\in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \le a)\Big] \\
&\quad + n\theta^2 - \sum_{i\in s_m}\Big[\frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le a)\Big]\Big[1 - \frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le a)\Big]\Big\} \\
&= F_I(a) - 2\theta F_I(a) + \theta^2 - \frac{1}{n}\sum_{i\in s_m} \hat{F}_J^{(i)}(1 - \hat{F}_J^{(i)}),
\end{aligned}$$

where $F_I(a)$ is defined as in theorem 5.2.2: $\hat{F}_J^{(i)} = \frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \leq a)$. As $F_I(a) = \theta + o_p(1)$, $F_I(a) - 2\theta F_I(a) + \theta^2 = \theta(1-\theta) + o_p(1)$. Further, let $V^i = \hat{F}_J^{(i)}(1-\hat{F}_J^{(i)})$, then $\{V^i\}$ are conditionally *i.i.d.*. Let $E^* = E(.|\mathcal{B}_n)$, then

$$E^*(V^i) = E^*(F_J^{(i)}) - E^*(F_J^{(i)})^2 = \theta - Var^*(F_J^{(i)}) - \theta^2 + o_p(1)$$
$$= \theta(1-\theta) - \frac{1}{J^2}\sum_{j=1}^{J} Var^*(I(Y_{ij}^R \leq a)) + o_p(1) = (1-\frac{1}{J})\theta(1-\theta) + o_p(1).$$

Further,

$$\frac{1}{n}\sum_{i\in s_m} \hat{F}_J^{(i)}(1-\hat{F}_J^{(i)}) = \frac{n-r}{n}\left[\frac{1}{n-r}\sum_{i\in s_m} V^i\right] = (1-p)\frac{J-1}{J}\theta(1-\theta) + o_p(1).$$

Hence,

$$\frac{1}{n}\sum_{i\in s_m} Z_{I,i}^2(\theta) = (\frac{1-p}{J} + p)\theta(1-\theta) + o_p(1).$$

On the other hand, by Theorem 5.2.2,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\theta) = \sqrt{n}(F_I(a) - \theta) \xrightarrow{d} N(0, \sigma_{d,I}^2), \quad \text{as } n \to \infty,$$

where $\sigma_{d,I}^2 = (\frac{1-p}{J} + p^{-1})\theta(1-\theta)$. Hence, by Owen (1990),

$$\ell_I(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\theta)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\theta)\right\}^2 + o_p(1) \xrightarrow{d} \frac{1-p+Jp^{-1}}{1-p+Jp}\chi_1^2.$$

□

Using Theorem 5.3.2, $(1-\alpha)$-level empirical likelihood based confidence intervals on θ under I are obtained as

$$\{\theta : \frac{1-\hat{p}+J\hat{p}}{1-\hat{p}+J\hat{p}^{-1}}\ell_I(\theta) \leq \chi_{1,\alpha}^2\},$$

For $J = 1$, the above interval reduces to the interval under random imputation (Qin *et al.* 2006a).

5.3.3 Empirical Likelihood CI for Quantile

Let

$$Z_{I,i}(\theta_q) = \begin{cases} I(Y_{I,i} \leq \theta_q) - q, & \text{if } i \in s_r \\ \frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \leq \theta_q) - q, & \text{if } i \in s_m \end{cases}$$

Then the empirical log-likelihood ratio for θ_q under I is defined as

$$\ell_I(\theta_q) = -2\sum_{i=1}^{n} \log\{np_i(\theta_q)\},$$

where $p_{iI}(\theta_q), i = 1, ..., n$ maximize $\sum_{i=1}^{n} \log(np_i)$ subject to constrains $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i Z_{I,i}(\theta_q) = 0$.

Using Lagrange multipler method,

$$\ell_I(\theta_q) = 2\sum_{i=1}^{n} \log\{1 + \lambda_{I,q} Z_{I,i}(\theta_q)\},$$

where $\lambda_{I,q}$ is the solution of the equation

$$\sum_{i=1}^{n} \frac{Z_{I,i}(\theta_q)}{1 + \lambda_{I,q} Z_{I,i}(\theta_q)} = 0,$$

Result on the asymptotic distribution of the above empirical log-likelihood ratio for θ_q is summarized in Theorem 5.3.3.

Theorem 5.3.3. *Under conditions of Theorem 5.2.3, as* $n \to \infty$,

$$\ell_I(\theta_q) \xrightarrow{d} \frac{1 - p + Jp^{-1}}{1 - p + Jp}\chi_1^2. \tag{5.3.7}$$

Proof of Theorem 5.3.3:

$$\frac{1}{n}\sum_{i=1}^{n} Z_{I,i}^2(\theta_q) = \frac{1}{n}\Big\{ \sum_{i\in s_r}(I(Y_i \le \theta_q) - q)^2 + \sum_{i\in s_m}[\frac{1}{J}[\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q) - q]^2 \Big\}$$

$$= \frac{1}{n}\Big\{ \sum_{i\in s_r} I(Y_i \le \theta_q) - 2q\sum_{i\in s_r} I(Y_i \le \theta_q) + rq^2 + \sum_{i\in s_m}\Big[\frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q)\Big]^2$$

$$- 2q\sum_{i\in s_m}\frac{1}{J}[\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q) + (n-r)q^2\Big\}$$

$$= \frac{1}{n}\Big\{\Big[\sum_{i\in s_r} I(Y_i \le \theta_q) + \frac{1}{J}\sum_{i\in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q)\Big] -$$

$$2q\Big[\sum_{i\in s_r} I(Y_i \le a) + \frac{1}{J}\sum_{i\in s_m}\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q)\Big]$$

$$+ nq^2 - \sum_{i\in s_m}\Big[\frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q)\Big]\Big[1 - \frac{1}{J}\sum_{j=1}^{J} I(Y_{ij}^R \le \theta_q)\Big]\Big\}$$

$$= F_I(\theta_q) - 2\theta F_I(\theta_q) + q^2 - \frac{1}{n}\sum_{i\in s_m}\hat{F}_J^{(i)}(1-\hat{F}_J^{(i)}),$$

where F_I and $\hat{F}_J^{(i)}$ are defined as in Theorem 5.3.2. As $F_I(\theta_q) = q + o_p(1)$, we have $F_I(\theta_q) - 2qF_I(\theta_q) + q^2 = q(1-q) + o_p(1)$. Further, let $V^i = \hat{F}_J^{(i)}(1-\hat{F}_J^{(i)})$, then $\{V^i\}$ are conditionally *i.i.d.* and

$$E^*(V^i) = E^*(F_J^{(i)}) - E^*(\hat{F}_J^{(i)})^2 = \frac{J-1}{J}q(1-q) + o_p(1).$$

Further,

$$\frac{1}{n}\sum_{i\in s_m}\hat{F}_J^{(i)}(1-\hat{F}_J^{(i)}) = (1-p)\frac{J-1}{J}q(1-q) + o_p(1).$$

Hence,

$$\frac{1}{n}\sum_{i\in s_m} Z_{I,i}^2(\theta_q) = (\frac{1-p}{J} + p)q(1-q) + o_p(1).$$

On the other hand, by Theorem 5.2.2,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\theta_q) = \sqrt{n}(F_I(\theta_q) - q) \xrightarrow{d} N(0, \sigma_{q,I}^2), \quad \text{as } n \to \infty,$$

where $\sigma^2_{q,I} = (\frac{1-p}{J} + p^{-1})q(1-q)$. Hence, by Owen (1990),

$$\ell_I(\theta_q) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z^2_{I,i}(\theta_q)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{I,i}(\theta_q)\right\}^2 + o_p(1) \xrightarrow{d} \frac{1-p+Jp^{-1}}{1-p+Jp}\chi^2_1.$$

□

Using Theorem 5.3.3, $(1-\alpha)$-level empirical likelihood based CI on θ_q is obtained as:

$$\{\theta_q : \frac{1-\hat{p}+J\hat{p}}{1-\hat{p}+J\hat{p}^{-1}}\ell_I(\theta_q) \leq \chi^2_{1,\alpha}\}.$$

For $J = 1$, the above CI reduces to the interval under random imputation (Qin *et al.*(2006a)).

5.4 Simulation Study

We conducted a small simulation study on the finite sample performance of normal approximation and empirical likelihood based confidence intervals on the mean $\mu = E(Y)$, distribution function $\theta = F(a)$ for fixed a and quantile $\theta_q = F^{-1}(q)$.

The computations required for conducting simulations are quite extensive, especially for constructing CIs using empirical likelihood method, in which case a two-dimensional optimization method is applied. That is one of the reasons I selected the C/C++ as the simulation environment, which is much faster than SAS or Splus/R. Another reason is that for SAS, R or other statistical software packages, it is not easy to develop new algorithms if there is no existing algorithm available. Besides, a number of utility C codes are available from my previous research projects, such as matrix inversion, random number generator and the two-dimensional optimization algorithm, etc. For constructing the empirical likelihood based CIs, a new double bi-section method is designed, which is stable, robust and converges

fast. SAS and R are also used for some simple tasks. All codes and algorithms are available upon request.

In this simulation study, we generated the values Y_i from the exponential distribution exp(1) and three cases of response probabilities $0.7, 0.8$ and 0.9. For each of the three cases, we generated $10,000$ random samples of incomplete data $\{Y_i, \delta_i,\ i = 1, \cdots, n\}$ for $n = 60, 120$ and 180. For nominal confidence level $1 - \alpha = 0.95$, using the simulated samples, we evaluated the coverage probability (CP), lower tail error rate (L), upper tail error rate(U) and the average length of the interval (AL) of the normal approximation based (NA) and empirical likelihood based (EL) intervals. In the case of quantiles, we denote the Woodruff type confidence interval as W. All results on CIs for mean, distribution and quantile constructed under both normal theories and empirical likelihood methods are listed in Tables 1-5, analysis and discussion follow thereafter.

Table 1 reports the simulation results for the mean $\mu = E(Y)$. The number J for fractional imputation (I) is chosen as 1 (the same as random imputation) or 5. We choose the fractional number $J = 5$ as our simulation study shows that the efficiency under this case has increased significant comparing to the random imputation case and there is no much of improvement if the fractional number is increased further. It is seen from Table 1 that EL provides more balanced error rates (L and U) than NA under I. In the case of NA, L is significantly lower and U is significantly higher than the nominal 2.5%. For example, for $n = 60$, $p = 0.7$, L$= 1.2\%$ and U$= 6.1\%$ for NA compared to L$= 2.6\%$ and U$= 3.7\%$ for EL. The imbalance in error rates decreases as n increases. The performance of EL in terms of CP is also better than NA. But for average length (AL), NA has a slightly shorter average length than EL,

as expected. For example, for $n = 60, p = 0.7$ and EL, we have $AL = 0.636$ for $J = 5$ compared to $AL = 0.661$ for $J = 1$. Similarly, for NA, we have $AL = 0.60$ for $J = 5$ compared to $AL = 0.642$ for $J = 1$.

TABLE 1

Confidence interval coverage probability(CP), average length(AL) and lower(L) and upper(U) tail error rates for the mean $\mu = EY$ under different response probabilities p and sample sizes n and fractional number J=1 and 5

p	n	CI	CP	L	U	AL	CP	L	U	AL
			J=1				J=5			
0.7	60	NA	0.927	0.012	0.061	0.642	0.922	0.015	0.063	0.600
		EL	0.937	0.026	0.037	0.661	0.938	0.029	0.034	0.636
	120	NA	0.935	0.012	0.053	0.461	0.938	0.012	0.049	0.429
		EL	0.944	0.025	0.030	0.471	0.946	0.027	0.027	0.448
	180	NA	0.941	0.016	0.043	0.380	0.938	0.017	0.045	0.353
		EL	0.946	0.026	0.028	0.386	0.946	0.028	0.027	0.364
0.8	60	NA	0.925	0.014	0.061	0.590	0.928	0.013	0.058	0.559
		EL	0.934	0.028	0.038	0.606	0.938	0.026	0.037	0.583
	120	NA	0.936	0.012	0.051	0.423	0.937	0.013	0.050	0.400
		EL	0.941	0.026	0.033	0.432	0.945	0.025	0.030	0.412
	180	NA	0.944	0.014	0.043	0.348	0.941	0.015	0.044	0.329
		EL	0.947	0.025	0.028	0.353	0.944	0.026	0.029	0.336
0.9	60	NA	0.928	0.011	0.061	0.542	0.928	0.012	0.061	0.525
		EL	0.937	0.024	0.039	0.555	0.937	0.026	0.038	0.541
	120	NA	0.938	0.013	0.049	0.388	0.941	0.015	0.0448	0.375
		EL	0.948	0.023	0.029	0.394	0.946	0.023	0.031	0.383
	180	NA	0.940	0.017	0.043	0.318	0.942	0.016	0.042	0.308
		EL	0.946	0.026	0.028	0.322	0.945	0.026	0.029	0.313

One can say that NA has a slightly shorter CI than EL but at the expense of undercoverage and imbalance in the tail error rates. Hence, EL exhibits better performance than NA. In terms of the fractional number, J, it seems that no significant change occurs in terms of CP, L or U, but AL is shorter for $J = 5$ than for $J = 1$. That situation is true for all cases of NA and EL. One can conclude that fractional imputation decreases AL as J increases but there is no significant change in terms of CP.

TABLE 2

Confidence interval coverage probability(CP), average length(AL) and lower(L) and upper(U) tail error rates for the distribution function $\theta = F(a)$ with $a = 1$ under different response probabilities p, sample sizes n, and fractional number, J=1 and 5

p	n	CI	CP	L	U	AL	CP	L	U	AL
			J=1				J=5			
0.7	60	NA	0.932	0.052	0.016	0.272	0.932	0.052	0.016	0.254
		EL	0.951	0.023	0.026	0.266	0.954	0.018	0.028	0.250
	120	NA	0.936	0.046	0.018	0.194	0.936	0.046	0.018	0.180
		EL	0.948	0.026	0.026	0.191	0.950	0.024	0.027	0.179
	180	NA	0.943	0.037	0.020	0.159	0.944	0.038	0.018	0.148
		EL	0.947	0.025	0.028	0.157	0.950	0.024	0.026	0.147
0.8	60	NA	0.928	0.053	0.018	0.250	0.936	0.048	0.016	0.236
		EL	0.947	0.024	0.028	0.245	0.951	0.021	0.027	0.232
	120	NA	0.940	0.042	0.017	0.178	0.939	0.043	0.018	0.168
		EL	0.946	0.029	0.025	0.176	0.947	0.026	0.027	0.166
	180	NA	0.947	0.034	0.018	0.146	0.944	0.037	0.018	0.138
		EL	0.949	0.025	0.026	0.144	0.949	0.024	0.027	0.137

Table2(cont'd)

0.9	60	NA	0.941	0.0423	0.016	0.229	0.936	0.048	0.016	0.221
		EL	0.947	0.028	0.025	0.224	0.949	0.026	0.026	0.218
	120	NA	0.940	0.042	0.018	0.163	0.944	0.040	0.017	0.157
		EL	0.948	0.028	0.024	0.161	0.949	0.026	0.025	0.156
	180	NA	0.945	0.038	0.017	0.133	0.944	0.038	0.019	0.129
		EL	0.948	0.028	0.025	0.132	0.947	0.027	0.026	0.128

Tables 2 and 3 report the simulation results for the distribution function $\theta = F(a)$ when $a = 1$ (Table 2) and $a = 2$ (Table 3). It is clear from Table 2 and Table 3 that EL outperforms NA in terms of CP, with values closer to nominal 95% even for $n = 60$, and balanced error rates L and U. For example, with $n = 60$, $p = 0.7$, $a = 1$ and $J = 5$, CP = 93.2%, L = 5.2% and U = 1.6% for NA compared to CP = 95.4%, L = 1.8% and U = 2.8% for EL. The performance of EL in terms of CP is also better than NA. But for average length (AL), NA has a slightly shorter average length than EL, as expected. For example, for $n = 60, p = 0.7$ and EL, we have $AL = 0.25$ for $J = 5$ compared to $AL = 0.266$ for $J = 1$. Similarly, for NA, we have $AL = 0.254$ for $J = 5$ compared to $AL = 0.272$ for $J = 1$. One can say that NA has a slightly shorter CI than EL but at the expense of undercoverage and imbalance in the tail error rates. Again, for the fractional number, J, it seems that no significant change occurs in terms of CP, L or U but AL is shorter for $J = 5$ case than for $J = 1$. That situation is true for all cases of NA and EL. One can conclude that fractional imputation decreases AL but there is no significant change in terms of CP.

TABLE 3

Confidence interval coverage probability(CP), average length(AL) and lower(L) and upper(U) tail error rates for the distribution function $\theta = F(a)$ with $a = 2$

p	n	CI	CP	L	U	AL	CP	L	U	AL
			J=1				J=5			
0.7	60	NA	0.939	0.024	0.038	0.320	0.936	0.023	0.041	0.298
		EL	0.951	0.024	0.025	0.311	0.949	0.024	0.027	0.291
	120	NA	0.941	0.025	0.035	0.228	0.942	0.025	0.033	0.212
		EL	0.947	0.026	0.027	0.225	0.949	0.027	0.025	0.209
	180	NA	0.953	0.023	0.023	0.184	0.949	0.022	0.029	0.174
		EL	0.951	0.02	0.029	0.169	0.952	0.024	0.024	0.172
0.8	60	NA	0.939	0.024	0.037	0.294	0.939	0.022	0.039	0.278
		EL	0.949	0.024	0.027	0.286	0.950	0.023	0.027	0.272
	120	NA	0.945	0.023	0.032	0.209	0.945	0.023	0.032	0.198
		EL	0.949	0.025	0.026	0.206	0.948	0.025	0.026	0.195
	180	NA	0.941	0.038	0.021	0.173	0.948	0.022	0.030	0.162
		EL	0.949	0.025	0.026	0.169	0.951	0.024	0.025	0.160

Table3(Cont'd)

$p = 0.9$	60	NA	0.940	0.023	0.038	0.269	0.941	0.022	0.037	0.260
		EL	0.950	0.025	0.026	0.263	0.949	0.024	0.027	0.255
	120	NA	0.945	0.024	0.030	0.191	0.947	0.023	0.030	0.185
		EL	0.948	0.026	0.026	0.189	0.949	0.025	0.026	0.183
	180	NA	0.950	0.021	0.030	0.157	0.950	0.022	0.028	0.151
		EL	0.951	0.023	0.026	0.155	0.952	0.024	0.025	0.150

TABLE 4

Confidence interval coverage probability(CP), average length(AL) and lower(L) and upper(U) tail error rates for the quantile θ_q with $q = 0.25$

p	n	CI	CP	L	U	AL	CP	L	U	AL
			J=1				J=5			
0.7	60	NA	0.890	0.034	0.076	0.382	0.901	0.043	0.056	0.361
		W	0.951	0.017	0.032	0.390	0.950	0.020	0.030	0.362
		EL	0.947	0.028	0.025	0.389	0.950	0.023	0.026	0.366
	120	NA	0.909	0.028	0.063	0.272	0.920	0.032	0.048	0.254
		W	0.953	0.017	0.030	0.275	0.956	0.018	0.026	0.255
		EL	0.950	0.024	0.026	0.274	0.955	0.021	0.024	0.256
	180	NA	0.917	0.026	0.057	0.220	0.921	0.030	0.049	0.205
		W	0.955	0.019	0.026	0.223	0.955	0.020	0.025	0.207
		EL	0.952	0.024	0.024	0.222	0.954	0.022	0.024	0.207
0.8	60	NA	0.894	0.0315	0.075	0.350	0.907	0.040	0.053	0.333
		W	0.951	0.019	0.031	0.355	0.949	0.020	0.031	0.337
		EL	0.948	0.028	0.024	0.358	0.947	0.025	0.027	0.337
	120	NA	0.913	0.026	0.061	0.248	0.921	0.033	0.045	0.236
		W	0.955	0.019	0.026	0.251	0.954	0.018	0.028	0.237
		EL	0.953	0.025	0.023	0.251	0.952	0.021	0.027	0.237
	180	NA	0.921	0.025	0.054	0.202	0.925	0.029	0.045	0.191
		W	0.952	0.021	0.027	0.204	0.951	0.023	0.026	0.192
		EL	0.950	0.026	0.024	0.203	0.950	0.025	0.024	0.192

TABLE 4(Cont'd)

0.9	60	NA	0.906	0.025	0.069	0.318	0.913	0.034	0.052	0.311
		W	0.956	0.017	0.027	0.328	0.953	0.020	0.027	0.314
		EL	0.949	0.028	0.023	0.322	0.952	0.025	0.023	0.314
	120	NA	0.917	0.026	0.057	0.220	0.921	0.031	0.048	0.220
		W	0.956	0.017	0.027	0.231	0.954	0.019	0.027	0.222
		EL	0.951	0.025	0.025	0.227	0.953	0.022	0.025	0.221
	180	NA	0.923	0.026	0.051	0.185	0.929	0.027	0.043	0.179
		W	0.951	0.023	0.026	0.186	0.950	0.024	0.026	0.180
		EL	0.948	0.027	0.025	0.185	0.949	0.027	0.024	0.179

Tables 4 and 5 report the simulation results for the median θ_q with $q = 0.25$ and 0.5 respectively. Here NA leads to severe undercoverage whereas the Woodruff (W) method and EL leads to CP closer to nominal 95%. For example, with $n = 60$, $p = 0.7$, the median ($q = 0.5$) and $J = 5$ we have CP = 88.2% for NA compared to CP = 95.2% for EL, and CP = 95.1% for W. Also, EL and W provide similar results in terms of CP, L, U and AL, although AL is slightly smaller for EL. Our results suggest that NA is not recommended for quantiles, and either EL or W should be used in practice. However, EL provides a unified method for all the parameters μ, θ and θ_q, whereas W is tailor-made for θ_q.

TABLE 5

Confidence interval coverage probability(CP), average length(AL) and lower(L) and upper(U) tail error rates for the quantile θ_q with $q = 0.5$

p	n	CI	CP	L	U	AL	CP	L	U	AL
			J=1				J=5			
0.7	60	NA	0.873	0.040	0.087	0.651	0.882	0.051	0.067	0.624
		W	0.954	0.024	0.022	0.698	0.951	0.026	0.023	0.644
		EL	0.950	0.025	0.024	0.683	0.952	0.025	0.022	0.646
	120	NA	0.886	0.039	0.075	0.467	0.895	0.040	0.065	0.435
		W	0.947	0.028	0.025	0.480	0.949	0.025	0.026	0.445
		EL	0.945	0.029	0.027	0.0.476	0.948	0.025	0.026	0.444
	180	NA	0.899	0.038	0.063	0.380	0.905	0.039	0.056	0.353
		W	0.952	0.024	0.024	0.390	0.950	0.025	0.024	0.361
		EL	0.951	0.025	0.024	0.388	0.951	0.024	0.025	0.360
0.8	60	NA	0.873	0.041	0.086	0.598	0.890	0.042	0.068	0.575
		W	0.955	0.023	0.023	0.632	0.952	0.024	0.025	0.593
		EL	0.953	0.023	0.024	0.626	0.952	0.024	0.025	0.592
	120	NA	0.896	0.037	0.068	0.427	0.903	0.036	0.061	0.407
		W	0.949	0.026	0.025	0.436	0.950	0.025	0.024	0.412
		EL	0.947	0.027	0.025	0.4341	0.950	0.026	0.025	0.411
	180	NA	0.906	0.034	0.059	0.351	0.905	0.039	0.056	0.331
		W	0.948	0.026	0.026	0.356	0.952	0.024	0.025	0.335
		EL	0.946	0.027	0.027	0.353	0.951	0.024	0.025	0.334

TABLE 5(Cont'd)

0.9	60	NA	0.885	0.032	0.083	0.544	0.901	0.038	0.061	0.538
		W	0.955	0.023	0.022	0.579	0.952	0.024	0.025	0.554
		EL	0.954	0.023	0.023	0.576	0.950	0.023	0.026	0.551
	120	NA	0.903	0.031	0.066	0.390	0.906	0.035	0.059	0.380
		W	0.949	0.026	0.025	0.436	0.949	0.026	0.025	0.386
		EL	0.946	0.027	0.026	0.395	0.948	0.026	0.026	0.384
	180	NA	0.910	0.032	0.058	0.319	0.915	0.033	0.052	0.311
		W	0.948	0.026	0.026	0.356	0.950	0.025	0.025	0.313
		EL	0.951	0.025	0.025	0.324	0.948	0.026	0.026	0.311

5.5 CI under Stratified Sampling and Fractional Random Imputation within Strata

5.5.1 Normal approximation intervals

Suppose that the population is divided into H strata with known relative sizes W_h, $h = 1, \ldots, H$; $\sum_{h=1}^{H} W_h = 1$. Independent simple random samples of sizes n_h, $h = 1, \ldots, H$ are drawn from the strata, and the strata sampling fractions, n_h/N_h, are assumed to be negligible. We express μ, θ and θ_q as $\mu = \Sigma W_h \mu_h$, $F(a) = \Sigma W_h F_h(a)$ and $\theta_q = F^{-1}(q)$ respectively. We regard the sample of incomplete data in stratum h, $\{(Y_{hi}, \delta_{hi}),\ i = 1, \ldots, n_h\}$ as an i.i.d. sample generated from the random vector (Y_h, δ_h). We assume MCAR mechanism within each stratum, i.e., $P(\delta_h = 1|Y_h) = P(\delta_h = 1) = p_h$, $0 < p_h \leq 1$. Assuming that fractional random imputation is performed separately in each stratum, h and for i'th missing data, J values are randomly imputed with replacement for the missing Y_{hi}, denoted as $\{Y_{hij}^{(R)}, j = 1, ..., J\}$ from the donor set $\{Y_{hi}, i \in s_{hr}\}$. We have $\bar{Y}_I = \Sigma W_h \bar{Y}_{Ih}$ as the imputed estimator of μ, where $\bar{Y}_{Ih} = \frac{1}{n_h} \sum_{i \in s_h} (\delta_{hi} Y_{hi} + (1 - \delta_{hi}) \frac{1}{J} \sum_{j=1}^{J} Y_{hij}^{(R)} \equiv \frac{1}{n_h} \sum_{i \in s_h} Y_{I,hi}$. We obtain an extension of Theorem 2.1 by letting $n_h \to \infty$ for each h with fixed H and assuming that $0 < Var(Y_h) = \sigma_h^2 < \infty$.

Normal approximation based $(1 - \alpha)$–level intervals on μ are given by $\bar{Y}_I \pm z_{\alpha/2} \left[\Sigma W_h^2 n_h^{-1} \left(\frac{1 - \hat{p}_h}{J} + \hat{p}_h^{-1} \right) s_{mh}^2 \right]^{1/2}$, where $s_{mh}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (Y_{hi}^{(I)} - \bar{Y}_{Ih})^2$, is the estimator of σ_h^2 and $Y_{hi}^{(I)} = \delta_{hi} Y_{hi} + \frac{1 - \delta_{hi}}{J} \sum_{j=1}^{J} Y_{hij}^{(R)}$.

Estimator of $F(a)$ under fractional random imputation is given by $F_I(a) = \Sigma W_h \hat{F}_{Ih}(a)$, where $\hat{F}_{Ih} = \frac{1}{n_h} \sum_{i \in s_{hr}} I(Y_i \leq a) + \frac{1}{J n_h} \sum_{i \in s_{hm}} \sum_{j=1}^{J} (Y_{hij}^{(R)} \leq a)$. Normal approximation based $(1 - \alpha)$-level intervals are given by

$F_I(y) \pm z_{\alpha/2}\hat{\sigma}_{dI}$, where $\hat{\sigma}^2_{dI}(y) = \sum_h W_h^2 n_h^{-1}(\frac{1-\hat{p}_h}{J} + \hat{p}_h^{-1})\hat{\sigma}^2_{Ih}(y)$ and $\hat{\sigma}^2_{Ih}(y)$ is estimated as $\hat{F}_{Ih}(y)\{1 - \hat{F}_{Ih}(y)\}$.

We focus only on the Woodruff intervals for quantile under fractional random imputation (I) because normal approximation based intervals for quantiles did not perform well under simple random sampling in our simulation study (section 5.4). The $(1-\alpha)$–level Woodruff interval on θ_q under I is given by $\hat{F}_I^{-1}(q \pm z_{\alpha/2}\hat{\sigma}_{qI})$, where $\hat{\sigma}_{qI}$ is estimated by $\hat{\sigma}^2_{qI} = \sum_h W_h^2 n_h^{-1}(\frac{1-\hat{p}_h}{J} + \hat{p}_h^{-1})q(1-q)$.

5.5.2 EL intervals

We now sketch the EL intervals under stratified random sampling and fractional random imputation within strata. For EL based CI on μ under I, we maximize $\Sigma_h \Sigma_i \log(n_h p_{hi})$ subject to $\Sigma_i p_{hi} = 1$, $h = 1, \ldots, H$ and $\Sigma_h W_h \Sigma_i p_{hi}(Y_{I,hi} - \mu) = 0$, leading to empirical log-likelihood ratio

$$\ell_I(\mu) = 2\Sigma_h \Sigma_i \log\{1 + m_h t(\mu)(Y_{I,hi} - \psi_h(\mu))\}, \tag{5.5.1}$$

where $\mathbf{n} = (n_1, \ldots, n_H)'$, $m_h = nW_h n_h^{-1}$, and $\psi_h(\mu), t(\mu)$ satisfy

$$\sum_i \frac{Y_{I,hi} - \psi_h(\mu)}{1 + m_h t(\mu)(Y_{I,hi} - \psi_h(\mu))} = 0, h = 1, \cdots, H, \sum_h W_h \psi_h(\mu) = \mu. \tag{5.5.2}$$

Theorem 5.5.1. *Under the assumption that* $n_h^{-1}\sum_{h=1}^H n_h \to \lambda_h (0 < \lambda_h < \infty)$, $\ell_I(\mu)$ *has limiting distribution estimated by* $\left(\sum_h W_h^2 \lambda_h (\frac{1-\hat{p}_h}{J} + \hat{p}_h^{-1}) s^2_{mh}\right)$ $\left(\sum_h W_h^2 \lambda_h ((\frac{1-\hat{p}_h}{J} + \hat{p}_h) s^2_{mh} + \hat{\Delta}^2_{mh})\right)^{-1} \chi^2_1$, $\hat{p}_h$ *is the estimated response rate in stratum* h, $\psi_h(\mu)$ *is the solution of the corresponding equation (5.5.2),* $\hat{\Delta}_{mh}$ *is an estimator of* $\bar{Y}_{Ih} - \psi_h(\mu)$ *and* s^2_{mh} *is the sample variance of fractional imputed data in* h*-th stratum.*

Similar convenient notations are employed below in this section without further mention. Thus EL based $(1-\alpha)$–level intervals on μ are given by

$$\left\{\mu : \left(\sum_h W_h^2 \lambda_h (\frac{1-\hat{p}_h}{J} + \hat{p}_h^{-1}) s_{mh}^2\right)^{-1} \times \left(\sum_h W_h^2 \lambda_h ((\frac{1-\hat{p}_h}{J} + \hat{p}_h) s_{mh}^2 + \hat{\Delta}_{mh}^2)\right) \ell_I(\mu) \leq \chi_{1,\alpha}^2\right\}.$$

Zhong and Rao (2000) and Wu (2004) have given algorithms for evaluating empirical log-likelihood ratio for the complete data case. Here the same algorithms are applied to the data file to calculate $\ell_I(\mu)$.

Proof of Theorem 5.5.1:

Under the assumption that $\hat{\lambda}_h = n/n_h \to \lambda_h (0 < \lambda_h < \infty)$, one has

$$\begin{aligned} &\sqrt{n} \sum_h \sum_i W_h n_h^{-1} \{Y_{I,hi} - \psi_{I,h}(\mu)\} \\ &= \sum_h \sum_i W_h (n/n_h)^{1/2} n_h^{-1/2} (Y_{I,hi} - \mu_h) \\ &= \sum_h W_h (n/n_h)^{1/2} n_h^{-1/2} \sum_i (Y_{I,hi} - \mu_h) \\ &\xrightarrow{d} N(0, \sigma_1^2), \end{aligned} \tag{5.5.3}$$

where $\sigma_1^2 = \sum_h W_h^2 \lambda_h (\frac{1-p_h}{J} + p_h^{-1}) \sigma_h^2$. Further

$$\begin{aligned} &\sum_h \sum_i W_h m_h n_h^{-1} \{Y_{I,hi} - \psi_{I,h}(\mu)\}^2 \\ &= \sum_h \sum_i W_h m_h n_h^{-1} \{(Y_{I,hi} - \mu_h)^2 + (\mu_h - \psi_{I,h}(\mu))^2\} \\ &= \sum_h W_h m_h n_h^{-1} \sum_i \{(Y_{I,hi} - \mu_h)^2 + (\mu_h - \psi_{I,h}(\mu))^2\} + o_p(1) \\ &= \sum_h W_h^2 \lambda_h \{(\frac{1-p_h}{J} + p_h) \sigma_h^2 + (\mu_h - \psi_{I,h}(\mu))^2\} + o_p(1), \end{aligned} \tag{5.5.4}$$

Similar to the proof of Theorem 1 in Owen (1990), (5.5.3) to (5.5.4) imply that

$$t(\mu) = O_p(n^{-1/2}),$$

and that

$$\begin{aligned} & t(\mu)\sum_h\sum_i W_h m_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\}^2 \\ &= \sum_h\sum_i W_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\} + O_p(n^{-1}). \end{aligned} \tag{5.5.5}$$

From (5.5.3),(5.5.4), (5.5.5) and Taylor expansion, we have

$$\begin{aligned} \ell_I(\mu) &= 2\Sigma_h\Sigma_i \log\{1 + m_h t(\mu)(Y_{I,hi} - \psi_{I,h}(\mu))\} \\ &= 2nt(\mu)\Sigma_h\Sigma_i W_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\} \\ &\quad -nt^2(\mu)\sum_h\sum_i W_h m_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\}^2 + o_p(1) \\ &= nt(\mu)\Sigma_h\Sigma_i W_h n_h^{-1}\{Y_{M,hi} - \psi_{I,h}(\mu)\} + o_p(1) \\ &= n\left[\Sigma_h\Sigma_i W_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\}\right]^2 \\ &\quad \times\left[\sum_h\sum_i W_h m_h n_h^{-1}\{Y_{I,hi} - \psi_{I,h}(\mu)\}^2\right]^{-1} + o_p(1). \end{aligned}$$

It follows that $\ell_I(\mu)$ has limiting distribution estimated by

$$\left(\sum_h W_h^2\lambda_h(\frac{1-p_h}{J} + p_h^{-1})s_{mh}^2\right)\left(\sum_h W_h^2\lambda_h((\frac{1-p_h}{J} + p_h)s_{mh}^2 + \hat{\Delta}_{mh}^2)\right)^{-1}\chi_1^2.$$

Thus the theorem is proved. □

We now study the EL based CI on $\theta = F(a)$ under I. Following the above lines for μ, the EL based $(1-\alpha)$-level intervals on θ are given by

$$\begin{aligned} &\left\{\theta : \left(\sum_h W_h^2\hat{\lambda}_h(\frac{1-\hat{p}_h}{J} + \hat{p}_h^{-1})\hat{\sigma}_{dh}^2\right)\right)^{-1} \\ &\quad \times\sum_h W_h^2\hat{\lambda}_h\left((\frac{1-\hat{p}_h}{J} + \hat{p}_h)\hat{\sigma}_{dh}^2 + \hat{\Delta}_{dh}^2\right)\ell_I(\theta) \le \chi_{1,\alpha}^2\right\}, \end{aligned} \tag{5.5.6}$$

where $\hat{\sigma}^2_{dh}$ is an estimator of $F_h(a)\{1 - F_h(a)\}$, which is estimated as $\hat{F}_{Ih}(a)(1 - \hat{F}_{Ih}(a))$ under fractional random imputation, and $\hat{\Delta}_{dh}$ is an estimator of $F_h(a) - \psi_h(\theta)$. We are not going to give the proof for this result and the following result for quantile as they are virtually similar to the mean case.

Finally, we investigate the EL based CI on $\theta_q = F^{-1}(q)$. Under fractional random imputation within strata, the EL based $(1 - \alpha)$-level intervals on θ_q are given by

$$\left\{\theta_q : \left(\sum_h W_h^2 \hat{\lambda}_h (\frac{1 - \hat{p}_h}{J} + \hat{p}_h^{-1}) q(1-q)\right)^{-1} \right.$$
$$\left. \times \sum_h W_h^2 \hat{\lambda}_h \left((\frac{1 - \hat{p}_h}{J} + \hat{p}_h) q(1-q) + \hat{\Delta}^2_{qh} \right) \ell_I(\theta_q) \leq \chi^2_{1,\alpha} \right\}$$

where $\hat{\Delta}_{qh}$ is the estimators of $F_h(\theta_q) - \psi_h(\theta_q)$.

5.6 Extension to Fractional Random Linear Regression Imputation

In a joint paper (Qin, Rao and Ren (2006b)), normal approximation and empirical likelihood intervals on μ, θ and θ_q are studied under missing at random (MAR) assumption and random linear regression imputation for missing responses $Y_i, i \in s_m$. The linear regression model is of the form $Y_i = X_i'\beta + \varepsilon_i$ with zero mean independent model errors ε_i and the covariates X_i are not missing. Under random linear regression imputation, random residuals $e_i^t (i \in s_m)$ are first generated from the estimated residuals $e_i = Y_i - X_i'\hat{\beta}_r, i \in s_r$ obtained by fitting the linear regression model to $\{(Y_i, X_i), i \in s_r\}$. The imputed values are then given by $Y_i^t = X_i'\hat{\beta}_r + e_i^t, i \in s_m$. I have extended this work to fractional random linear regression imputation under MAR, i.e., probability of response on Y depends on X but not on Y. However, the reduction in confidence interval length relative to random linear regression imputation turned out to be small in a simulation study I conducted, unlike in the case of no covariates studied in the previous sections of this chapter. Reason for this small reduction is that the imputation variance due to random drawing of residuals is small relative to the total variance, unlike in the case of no covariates where imputed Y- values are randomly drawn from $\{Y_i, i \in s_r\}$. In view of this result, details of the extension to fractional random linear regression imputation are not included in my thesis.

Chapter 6

Marginal Logistic Regression Models for Longitudinal Complex Survey Data

In this Chapter we focus on marginal logistic modelling of binary response data from a longitudinal survey with a complex sampling design. Design-weighted Generalized Estimating Equations (GEE) are used to estimate the model parameters. Odds ratios are used as measures of association between pairs of binary responses in the working covariance matrix. A one-step estimating function (EF) bootstrap method is used for variance estimation. Several goodness-of-fit tests are studied for the model assessment problem. The methods are illustrated through their application to data from Statistics Canada's National Population Health Survey.

6.1 Introduction.

In recent years, longitudinal surveys, where sample subjects are observed over two or more time points, are being undertaken by government agencies in order to provide longitudinal data for analytic studies and to aid in the development of public policy. At Statistics Canada, for example, the National Population Health Survey (NPHS), the National Longitudinal Survey of Children and Youth (NLSCY) and the Survey of Labor and Income Dynamics (SLID) were all launched in the mid 1990's for this purpose and now several cycles of data on the same samples of individuals are available. Data from such surveys can be used for a variety of purposes including (a) gross flows estimation, (b) event history modelling, (c) conditional modelling of the response at a given time point as a function of past responses and present and past co-variables, and (d) modelling of marginal means of responses as functions of co-variables. Binder (1998) gives a good account of the possible uses of longitudinal survey data.

In this chapter, we focus on some aspects of marginal mean modelling with survey data, in particular binary responses and marginal logistic regression models. The first problem that we discuss here is the estimation of the model parameters. The case of a simple random sample, where individuals are considered to be independent and to have equal chances of being selected, has been studied extensively in the literature, especially in applications in biomedical and health science. For longitudinal studies, in order to avoid underestimating the standard errors of parameters and p-values of tests in making inferences, the within subject correlation among repeated measures must be accounted for during statistical analysis. For a marginal model, the mean response of a subject at a given time point depends only

on the associated covariates and not on any subject-specific random effects or previous response. The primary interest of using a marginal model is to take account of within-subject dependence among the repeated response. Assumptions about the full joint distribution of repeated observations are not needed and only the model for the marginal mean response needs to be correctly specified.

The case of marginal modelling with a simple random sample, where individuals are considered to be independent and to have equal chances of being selected, has been studied extensively in the literature, especially for applications in biomedical and health sciences; see Diggle *et al.* (2002) and Fitzmaurice *et al.* (2004). Liang and Zeger (1986) used generalized estimating equations (GEE) to estimate the model parameters, assuming a "working" correlation structure for the repeated measurements. They also obtained standard errors of parameter estimators, based on "sandwich" variance estimators that are "robust" in the sense of validity even if the working correlation is not equal to the true correlation structure. For the binary response case, Lipsitz, *et al.* (1991) used odds ratios to model the working covariance structure in GEE, instead of working correlations used by Liang and Zeger (1986) and others. An odds ratio is a natural measure for capturing association between binary repeated outcomes from the same individual. Moreover, for binary responses the odds ratio approach is less seriously constrained than the correlation approach (Liang *et al.* 1992).

When data are obtained from a longitudinal survey with a complex sampling design that involves clustering of subjects, there may be cross-sectional correlations

among subjects in addition to within-individual dependencies. Because of this additional complexity, new methods are needed to replace the methods used for longitudinal data where subjects are assumed to be independent. Methods that account for within individual dependence but ignore clustering of subjects and other design features could also lead to erroneous inferences from underestimation of standard errors of parameter estimators and p-values of tests. In Section 6.2, we adapt the GEE method for binary responses and marginal logistic regression models to the case of longitudinal survey data and estimate the model parameters as the solution to survey-weighted GEE. We use the odds ratio approach to model the working covariance structure. For the case of complex survey data, where the non-independence among individuals must be accounted for, variance estimation techniques must be modified. Rao (1998) explained how to obtain appropriate robust sandwich-type variance estimator for the case of complex survey data, but, in practice, this can be difficult to carry out with Taylor linearization if one wishes to account for non-response adjustment and post-stratification. As an alternative, a design-based bootstrap method becomes a good choice because it can account for clustering and other survey design features as well as non-response adjustment and post-stratification in a straightforward manner. At Statistics Canada, design information for variance estimation is released only in the form of bootstrap survey weights for many of its longitudinal surveys based on stratified multi-stage cluster sampling designs, where the first-stage clusters (or primary sampling units) are selected with probability proportional to size measures attached to the clusters. Bootstrap design weights are first obtained by the method of Rao and Wu (1988) and Rao, Wu and Yue (1992), and then adjusted for unit non-response and post-stratification in the same manner

as the full sample design weights to arrive at the bootstrap survey weights. Rust and Rao (1998) provide an overview of the bootstrap and other re-sampling methods in the context of stratified multi-stage cluster sampling. In Section 6.3, we extend the estimating function (EF) bootstrap method of Hu and Kalbfleisch (2000) to the case of complex longitudinal survey data and obtain EF-bootstrap standard errors that account for within subject dependence of repeated measurements as well as clustering of subjects and other survey design features. Advantages of the EF-bootstrap over the direct bootstrap are discussed in Section 6.3. Rao and Tausi (2004) similarly developed an EF-jackknife method for inference from cross-sectional survey data.

The GEE approach assumes that the marginal means are correctly specified. It is therefore important to check the adequacy of the mean specification, using goodness-of-fit tests and other diagnostics before making inferences on model parameters. Several global goodness-of-fit(GOF) tests have been proposed in the literature for the binary response case and a logistic model, assuming simple random sampling of subjects. In particular, for the case of no repeated measurements on a sample individual, the well-known Hosmer and Lemeshow (1980) test uses a Pearson chi-squared statistic after partitioning the subjects into groups based on the values of estimated response probabilities under the assumed model for the mean. Horton *et al.* (1999) studied the case of longitudinal data under simple random sampling and proposed a score test after partitioning all the response probabilities under the null hypothesis into groups, assuming working independence for the repeated measurements. As noted by Horton *et al.* (1999), global (or omnibus) tests may

not have high power to detect specific alternatives and one should not regard a non-significant goodness-of-fit test as clear evidence that the assumed marginal mean specification gives a good fit. In contrast to most goodness-of-fit tests, which are based on the subjective partitioning, Pan (2002) proposed two residual and normal based GOF tests: the Pearson chi-square and unweighted sum of residual squares for ordinary logistic regression. Unfortunately, Pan's methods work well only when a continuous covariate is present.

For the case of no repeated measurements, Graubard *et al.* (1997) used the survey weights and the associated estimated response probabilities to partition the subjects into groups, and then developed goodness-of-fit Wald tests based on the weighted sum of responses and the weighted sum of estimated response probabilities in the groups, taking account of the survey design.

In Section 6.4.2, we extend the Horton *et al.* (1999) score test to the case of longitudinal survey data and show that it is equivalent to an extension of the Graubard *et al.* (1997) Wald test to the longitudinal case. In Section 6.4.3 we extend the Hosmer and Lemeshow's chi-square test to longitudinal survey data using Rao-Scott adjustments (Rao and Scott, 1981, Roberts *et al.* 1987) that account for the survey design. Pan's (2002) asymptotic normal approach is studied in Section 6.4.4. But it is not applicable to our National Population Health Survey data in which all covariates are binary. We use design based bootstrap methods in developing the above GOF tests.

To illustrate the proposed methods for inference on model parameters and global goodness-of-fit of the marginal mean model, we used longitudinal data from Statistics Canada's National Population Health Survey (NPHS) as an example in Section

6.5. The results form NPHS data and analysis are presented.

6.2 Survey-weighted Estimating Equations (SEE) and Odds Ratio Approach

Suppose a sample, s, of size n, is selected by a complex survey design from a population U of size N, and that the same sampled units are observed for T occasions. Let the data have the form $\{(y_{it}, x_{it}), i \in s, t = 1, ..., T\}$, where y_{it} is the response of the i^{th} individual on occasion t and x_{it} is a $p \times 1$ vector of fixed associated covariates. In the case of a binary response variable (*i.e.* $y_{it} = 0$ or 1), the marginal logistic regression model is a natural choice for describing the relationship between y_{it} and x_{it}. In the marginal logistic regression model, the marginal density of response y_{it} given covariate x_{it} is the Bernoulli density,

$$f(y_{it}|x_{it}) = p_{it}^{y_{it}}(1 - p_{it})^{1-y_{it}}, \tag{6.2.1}$$

where $E(y_{it}|x_{it}) = p_{it}$, $logit(p_{it}) = \beta' X_{it}$, and $X_{it} = \{1, x'_{it}\}'$. Let $X_i = \{X_{i1}, X_{i2}, ..., X_{iT}\}$, $Y_i = \{y_{i1}, y_{i2}, ..., y_{iT}\}'$, $p_i = \{p_{i1}, ..., p_{iT}\}'$, and V be the "working" variance of Y_i. Assuming independence between sample individuals (or simple random sampling with negligible sampling fraction), an estimator of the vector of model parameters β is obtained as the solution of the Generalized Estimating Equations (GEE):

$$\widehat{u}(\beta) = \sum_{i \in s} D_i' V_i^{-1}(Y_i - p_i(\beta)) = 0, \tag{6.2.2}$$

where $D_i = \frac{\partial p_i(\beta)}{\partial \beta}$, (Liang and Zeger, 1986). Note that V_i is the identity matrix under a working independence assumption for the observations from the i^{th} individual, or

is a positive definite matrix under a working correlation assumption. It should be kept in mind that, while V_i may differ from the true covariance matrix of Y_i , we assume that the mean of Y_i is correctly specified, i.e. $EY_i = p_i(\beta)$.

In the case of a complex survey design, let the survey weights be $\{w_i, i \in s\}$. Rao (1998) proposed the following survey-weighted estimating equations (SEEI) for estimating β:

$$\widehat{u}_{1w}(\beta) = \sum_{i \in s} w_i D_i' V_i^{-1}(Y_i - p_i(\beta)) = 0. \tag{6.2.3}$$

Denote the solution of (6.2.3) as $\widehat{\beta}_w$. Note that $\widehat{\beta}_w$ is a survey-weighted estimator of the census parameter, β_N, which is the solution of the census estimating equations, $u_N(\beta) = \sum_{i \in U} D_i' V_i^{-1}(Y_i - p_i(\beta)) = 0$. The census parameter β_N would be a consistent estimator of β if the population, U, of individuals is a simple random sample or more generally a self-weighting sample from a super-population obeying the marginal model. The survey-weighted estimator, $\widehat{\beta}_w$, is consistent for β_N (and hence for β) if $\widehat{u}_{1w}(\beta)$ is design-unbiased or consistent for $u_N(\beta)$. We assume that the sample fractions are negligible so that $\sqrt{n}(\hat{\beta}_w - \beta) \approx \sqrt{n}(\hat{\beta}_w - \beta_N)$, and thus it is not necessary to distinguish β_N from β.

In the case of a marginal model with binary responses, Lipsitz *et al.* (1991) used the odds ratio as a measure of association between pairs of binary responses. The major reason for this is that the odds ratio is not constrained by the means of the two binary variables, which is a problem with the correlation. As well, we can use a working model for the odds ratios to define V_i. If we let $Y_{ist} = Y_{is}Y_{it}$ for all $s = 1, .., T-1$, $t = s+1, , T$, and $p_{ist} = E(Y_{ist}) = P(Y_{is} = 1, Y_{it} = 1)$, then for given $s \neq t$, the odds ratio γ_{ist} is defined as:

$$\gamma_{ist} = \frac{P(Y_{is} = 1, Y_{it} = 1)P(Y_{is} = 0, Y_{it} = 0)}{P(Y_{is} = 1, Y_{it} = 0)P(Y_{is} = 0, Y_{it} = 1)} = \frac{p_{ist}(1 - p_{is} - p_{it} + p_{ist})}{(p_{is} - p_{ist})(p_{it} - p_{ist})}. \tag{6.2.4}$$

Suppose that the odds ratio γ_{ist} is modeled as a function of covariates (e.g., the log odds ratio is the linear function of some covariates), and that α is the vector of parameters in that model, i.e. $\gamma_{ist} = \gamma_{ist}(\alpha)$. Then the elements of the working covariance matrix V_i can be written:

$$\begin{aligned} Var(Y_{it}) &= V_{itt} = p_{it}(\beta)(1 - p_{it}(\beta)) \\ Cov(Y_{it}, Y_{is}) &= V_{ist} = p_{ist}(\beta, \alpha) - p_{is}(\beta)p_{it}(\beta). \end{aligned} \tag{6.2.5}$$

where, from the quadratic equation (6.2.4), p_{ist} can be expressed as $p_{ist} = g(p_{is}, p_{it}, \gamma_{ist})$, which is a function of both β and α. Since β and α are both unknown, we need to use a second set of survey-weighted estimating equations (SEEII). Let $U_i = (Y_{i12}, ..., Y_{(T-1)T})'$ and $\theta_i(\beta, \alpha) = (p_{i12}(\beta, \alpha), p_{i13}(\beta, \alpha), ..., p_{i(T-1)T}(\beta, \alpha))'$. Then SEEII are given by:

$$\widehat{u}_{2w}(\beta, \alpha) = \sum_{i \in s} w_i C_i' F_i^{-1}(U_i - \theta_i(\beta, \alpha)) = 0. \tag{6.2.6}$$

where $C_i = \frac{\partial \theta_i'}{\partial \alpha}$ and $F_i = diag\{p_{ist}(1 - p_{ist})\}$. The Newton-Raphson iterative method may be used to solve (6.2.3) and (6.2.6) simultaneously using initial values $\widehat{\alpha}_0$, $\widehat{\beta}_0$, and then letting

$$\widehat{\beta}_{m+1} = \widehat{\beta}_m - (\sum_{i \in s} w_i D_{i(m)}' V_{i(m)}^{-1} D_{i(m)})^{-1} \left\{ \sum_{i \in s} w_i D_{i(m)}' V_{i(m)}^{-1} (Y_i - p_i(\beta_{(m)})) \right\} \tag{6.2.7}$$

and

$$\widehat{\alpha}_{m+1} = \widehat{\alpha}_m - (\sum_{i \in s} w_i C_{i(m)}' F_{i(m)}^{-1} C_{i(m)})^{-1} \left\{ \sum_{i \in s} w_i C_{i(m)}' F_{i(m)}^{-1} (U_i - \theta_i(\alpha_{(m)}, \beta_{(m+1)})) \right\}, \tag{6.2.8}$$

where the subscripts m and $m+1$ indicate that quantities are evaluated at $\beta = \widehat{\beta}_{(m)}$ and $\alpha = \widehat{\alpha}_{(m)}$ in (6.2.7) and at $\beta = \widehat{\beta}_{(m+1)}$ and $\alpha = \widehat{\alpha}_{(m+1)}$ in (6.2.8). At convergence

of the iterations, we obtain $\widehat{\beta}$ and $\widehat{\alpha}$, where $\widehat{\beta}$ is a consistent estimator of β even under misspecification of the means of the U_i. We assume that $\widehat{\alpha}$ converges in probability to some α^* which agrees with α only when the working model $\gamma_{ist} = \gamma_{ist}(\alpha)$ is correctly specified.

6.3 Variance Estimation: One Step EF-Bootstrap

In order to make inferences from an estimated marginal model, variance estimates are required for the estimated model parameters. Assuming independence between sample individuals, Liang and Zeger (1986) used linearization to derive consistent sandwich-type variance estimators which are widely used. Sandwich-type variance estimators have also been developed, through linearization, for many analytical problems applied to survey data. See, for example, Binder (1983) for an application of this approach to generalized linear models and Rao, Scott and Skinner (1998) for this approach in developing Wald and quasi-score tests. However, as the forms of parameter estimates become more complex and as nonresponse and calibration adjustments to survey weights become more involved, such as in the case of some longitudinal analyses, it becomes more difficult to carry out a full linearization. Because of this difficulty, attention has turned to studying replication methods for design-based variance estimation. As examples, Rao, Yung and Hidiroglou (2002) and Rao and Tausi (2004) have proposed Jackknife and Bootstrap re-sampling approaches for variance estimation. For many of its analytical surveys, Statistics Canada is now releasing design information for variance estimation only in the form of survey

bootstrap weights.

The direct bootstrap method for variance estimation (see, for example, Rust and Rao, 1996) involves obtaining point estimates of the parameters of interest with the full-sample survey weights and then, in an identical fashion, with each set of survey bootstrap weights. This method, consisting of many repetitive operations, can be computationally intensive and time consuming. Furthermore, Binder, Kovacevic and Roberts (2004) found that, when using this approach for logistic regression, it was possible to have many sets of bootstrap weights for which the parameter estimation algorithm would not converge due to ill-conditioned matrices that were not invertible. To overcome these problems, Binder, Kovacevic and Roberts (2004) and Rao and Tausi (2004) proposed estimating function (EF) bootstrap approaches, motivated by the work of Hu and Kalbfleish (2000) for the non-survey case. Here, we extend the one-step EF bootstrap approach of Rao and Tausi (2004) to the marginal logistic regression model.

Let $\{\{w_i^{(b)}, i = 1, ..., n\}, b = 1, ..., B\}$ be B sets of bootstrap weights for the sample s. Let

$$\widehat{u}_{1w}^{(b)}(\widehat{\beta}) = \sum_{i \in s} w_i^{(b)} \hat{D}_i' \hat{V}_i^{-1} (Y_i - p_i(\widehat{\beta})) \tag{6.3.1}$$

and

$$\widehat{u}_{2w}^{(b)}(\widehat{\beta}, \widehat{\alpha}) = \sum_{i \in s} w_i^{(b)} \hat{C}_i' \hat{F}_i^{-1} (U_i - \theta_i(\widehat{\beta}, \widehat{\alpha})) \tag{6.3.2}$$

be the bootstrap EF corresponding to (6.2.3) and (6.2.6) and evaluated at $\beta = \hat{\beta}$ and $\alpha = \hat{\alpha}$, where the matrices $\widehat{D}_i$, $\widehat{V}_i$, $\widehat{C}_i$, and$\widehat{F}_i$ in (6.3.5) and (6.3.6) are obtained by evaluating D_i, V_i, C_i, and F_i at $\widehat{\beta}$ and $\widehat{\alpha}$. Now compute one-step Newton-Raphson solutions to the following EF equations using $\widehat{\beta}$ and $\widehat{\alpha}$ as starting values:

$$\widehat{u}_{1w}(\beta) = \widehat{u}_{1w}^{(b)}(\widehat{\beta}) \tag{6.3.3}$$

and

$$\widehat{u}_{1w}(\beta, \alpha) = \widehat{u}_{1w}^{(b)}(\widehat{\beta}, \widehat{\alpha}) \tag{6.3.4}$$

This is equivalent to Taylor linearization of the left hand sides of (6.3.3) and (6.3.4). The one-step bootstrap estimators for the b-th bootstrap sample are given by:

$$\widetilde{\beta}^{(b)} = \widehat{\beta} - (\sum_{i \in s} w_i \widehat{D}_i' \widehat{V}_i^{-1} \widehat{D}_i)^{-1} \widehat{u}_{1w}^{(b)}(\widehat{\beta}). \tag{6.3.5}$$

and

$$\widetilde{\alpha}^{(b)} = \widehat{\alpha} - (\sum_{i \in s} w_i \widehat{C}_i' \widehat{F}_i^{-1} \widehat{C}_i)^{-1} \widehat{u}_{2w}^{(b)}(\widehat{\beta}, \widehat{\alpha}). \tag{6.3.6}$$

Note that for all $b = 1, ..., B$, the inverse matrices in (6.3.5) and (6.3.6) remain the same, so that no further inversion is needed for each bootstrap sample. Since we only iterate once, there are no convergence problems. The EF-bootstrap variance estimator of $\widehat{\beta}$ is given by

$$v_{BOOT}^{EF}(\widehat{\beta}) = \frac{1}{B} \sum_{b=1}^{B} (\widetilde{\beta}^{(b)} - (\widehat{\beta})(\widetilde{\beta}^{(b)} - (\widehat{\beta})'. \tag{6.3.7}$$

6.4 Goodness-of-fit tests

Inferences on regression parameter β, outlined in Sections 6.2 and 6.3, assume that the mean of y_{it}, is correctly specified as $E(y_{it}|\mathbf{X}_{it})$, with $logit[p_{it}(\beta)] = \mathbf{X}_{it}'\beta$. It is therefore important to check the adequacy of the mean specification using goodness-of-fit tests, before making inferences on β. For ordinary logistic regression, there are many goodness-of-fit tests presented. Can those methods be applied to the

longitudinal complex survey design case? We pursue this problem in the following sections.

6.4.1 Construction of groups

In the case of data $\{y_i, x_i; i \in s\}$ without repeated measurements, Graubard *et al.* (1997) obtained weighted decile groups $\{G_1, G_2, ..., G_{10}\}$: n_1 subjects with the smallest estimated probabilities $\hat{\pi}_i = p_i(\hat{\beta})$ under the null hypothesis $H_0 : p_i = p_i(\beta)$, are in the first group G_1, the next n_2 in the second group G_2, and so forth until the group G_{10} with n_{10} observations is formed. The n_l observations in G_l are chosen such that $\frac{\sum_{i \in G_l} w_i}{\sum_{i \in s} w_i} \simeq 1/10; l = 1, 2, ..., 10$. In the special case of equal weights $w_i = w$, this grouping method reduces to the Hosmer and Lemeshow (1980) method of grouping the subjects.

We now extend the above weighted decile grouping method to the case of longitudinal survey data, following Horton *et al.*(1999). We make use of all the estimated probabilities $\widehat{\pi}_{it} = p_{it}(\widehat{\beta})$, under the null hypothesis $H_0 : p_{it} = p_{it}(\beta)$ with $logit(p_{it}) = \beta_0 + \mathbf{x}'_{it}\beta$ and working independence, to form the weighted decile groups $G_1, G_2, ..., G_{10}$. The weights w_i associated with $\widehat{\pi}_{it}$ are used in forming the groups. Note that a subject's group membership can change for different time points t because $\widehat{\pi}_{it}$ for a subject i can change with time t. In the special case of equal weights $w_{it} = w$, this grouping method reduces to the Horton *et al.* (1999) method of grouping for the longitudinal case.

6.4.2 Quasi-score test

Following Horton *et al.* (1999), we formulate an alternative model, based on the grouping $G_1, G_2, ..., G_{10}$, to test the fit of the null model H0: $logit(p_{it}) = \beta_0 + \mathbf{x}'_{it}\beta$. Let $I_{itl} = 1$ if $\widehat{\pi}_{it}$ is in group G_l and $I_{itl} = 0$, otherwise, $l = 1, ..., 9$. The alternative model is then given by

$$logit(p_{it}) = \beta_0 + \mathbf{x}'_{it}\beta + \gamma_1 I_{it1} +, ..., +\gamma_9 I_{it9}. \tag{6.4.1}$$

We treat the indicator variables as fixed covariates even though they are based on the random $\widehat{\pi}_{it}$. In the case of independence among sample individuals, Spruill (1975) provided asymptotic justification for treating the partition $G_1, G_2, ..., G_{10}$ as though based on the true p_{it} . Under the set-up (6.4.1), H_0 is equivalent to testing $H_0^* : \gamma_1 = ... = \gamma_9$.

Following Rao et al. (1998), we now develop a quasi-score test of H_0^* , taking account of the weights and the design. We assume working independence and obtain survey-weighted estimating equations under (6.4.1) as

$$\widehat{u}_w = \begin{bmatrix} \widehat{u}_{1w}(\beta, \gamma) \\ \widehat{u}_{2w}(\beta, \gamma) \end{bmatrix} = \begin{bmatrix} \sum_{i \in s} \sum_{t=1}^{T} w_i \mathbf{x}_{it}(y_{it} - p_{it}(\beta, \gamma)) \\ \sum_{i \in s} \sum_{t=1}^{T} w_i \mathbf{I}_{it}(y_{it} - p_{it}(\beta, \gamma)) \end{bmatrix}. \tag{6.4.2}$$

where $I_{it} = (I_{it1}, ..., I_{it9})'$ and $\gamma = (\gamma_1, ..., \gamma_9)'$. Under H_0^*, we solve $\widehat{u}_{1w}(\beta, 0) = 0$ to get the estimator $\widetilde{\beta}_w$ of β. Note that $\widetilde{\beta}_w$ is identical to $\widehat{\beta}_w$ obtained from (6.2.3) under working independence. We substitute $\widetilde{\beta}_w$ into the second component $\widehat{u}_{2w}(\beta, \gamma)$ of (6.4.2) to get $\widehat{u}_{2w}(\widetilde{\beta}_w, 0)$ under H_0^*. We have $\widehat{u}_{2w}(\widetilde{\beta}_w, 0) = o_w - e_w$, where $o_w = \sum_{i \in s} \sum_{t=1}^{T} w_i \mathbf{I}_{it} y_{it}$ and $e_w = \sum_{i \in s} \sum_{t=1}^{T} w_i \mathbf{I}_{it} \widehat{\pi}_{it}$, where $\widehat{\pi}_{it} = \widehat{p}_{it}(\widetilde{\beta}_w, 0) = \widehat{p}_{it}(\widehat{\beta}_w)$ and $p_{it}(\beta)$ is the null model. Note that o_w and e_w are the weighted observed and expected counts respectively, under H_0^*.

A quasi-score statistic for testing H_0^* is now given by

$$X^2_{QS} = \widehat{u}_{2w}(\widetilde{\beta}_w, 0)' \Big[var\{\widehat{u}_{2w}(\widetilde{\beta}_w, 0)\} \Big]^{-1} \widehat{u}_{2w}(\widetilde{\beta}_w, 0). \tag{6.4.3}$$

where $var\{\widehat{u}_{2w}(\widetilde{\beta}_w, 0)\}$ is a consistent estimator of variance of $\widehat{u}_{2w}(\widetilde{\beta}_w, 0)$ under H_0^*. For the NPHS data (next section) we have used a Bootstrap variance estimator, using the Rao and Wu (1988) Bootstrap weights to calculate the Bootstrap estimates $\widehat{u}^{(b)}_{2w} = \widehat{u}_{2w}(\widehat{\beta}^{(b)}_w, 0), b = 1, ..., B$, where B is the number of Bootstrap replicates and $\widehat{\beta}^{(b)}_w$ is obtained from

$$\widehat{u}^{(b)}_{2w}(\beta, 0) = \sum_{i \in s} \sum_{t=1}^{T} w_i^{(b)} \mathbf{I}_{it}(y_{it} - p_{it}(\beta)) = 0, \tag{6.4.4}$$

where $w^{(b)}, b = 1, ..., B$ are the Bootstrap weights. To simplify the computations we propose a one-step Newton-Raphson iteration with $\widehat{\beta}_w$ as starting value to obtain $\widehat{\beta}^{(b)}_w$. The Bootstrap variance estimator of $\widehat{u}_{2w}(\widetilde{\beta}_w, 0)$ is then given by

$$v_{BOOT}(\widehat{u}_{2w}) = \frac{1}{B} \sum_{b=1}^{B} (\widehat{u}^{(b)}_{2w} - \widehat{u}_{2w})(\widehat{u}^{(b)}_{2w} - \widehat{u}_{2w})'. \tag{6.4.5}$$

Since the NPHS data file provides the weights $\{w_i, i \in s\}$ and the Bootstrap weights $\{w_i^{(b)}, i \in s, b = 1, ..., B\}$ along with the response variables, it is straightforward to implement $v_{BOOT}(\widehat{u}_{2w})$ from the NPHS data file. The EF method of the previous section was designed for variance estimation of regression parameter estimators $\widehat{\beta}_w$, and further theoretical work is needed to study its applicability to quasi-score tests. For the NPHS data (next section), we did not encounter ill-conditioned matrices in calculating $\widehat{\beta}^{(b)}_w$; therefore, we used the Rao-Wu Bootstrap method to obtain $\widehat{u}^{(b)}_{2w}$ needed in $v_{BOOT}(\widehat{u}_{2w})$ given by (6.4.5). An alternative to (6.4.5) is obtained by substituting $\widehat{u}^{(.)}_{2w} = \frac{1}{B} \sum_{b=1}^{B} \widehat{u}^{(b)}_{2w}$ for $\widehat{u}_{2w}$ in (6.4.5), but the result should be close to (6.4.5).

We can write X^2_{QS} as a Wald statistic based on $o_w - e_w$:

$$X^2_{QS} = (o_w - e_w)' \Big[var(o_w - e_w) \Big]^{-1} (o_w - e_w), \tag{6.4.6}$$

where $var(o_w - e_w)$ is a consistent variance estimator of $o_w - e_w$. Graubard *et al.* (2004) proposed a Wald statistic similar to (6.4.6) for the case of no repeated measurements. Under H_0^*, the statistic X^2_{QS} is asymptotically distributed as a χ^2 variable with 9 degrees of freedom. Hence, the p-value may be obtained as $Pr\Big(\chi^2_9 > X^2_{QS}(obs)\Big)$, where $X^2_{QS}(obs)$ is the observed value of X^2_{QS}. The p-value provides evidence against H_0^*.

Following Rao, Scott and Skinner(1998), we can propose a naive score test as

$$Q_N = (\frac{n}{\hat{N}})(\widehat{u}_{2w}(\widetilde{\beta}_w, 0)' \tilde{I}^{-1}_{22.1} \widehat{u}_{2w}(\widetilde{\beta}_w, 0)), \tag{6.4.7}$$

where $\hat{N} = \sum_{i \in s} w_i$,

$$\tilde{I}^{-1}_{22.1} = \tilde{I}_{22} - \tilde{I}_{21} \tilde{I}^{-1}_{11} \tilde{I}_{12}$$

$$\tilde{I} = \begin{bmatrix} \tilde{I}_{11} & \tilde{I}_{12} \\ \tilde{I}_{21} & \tilde{I}_{22} \end{bmatrix} = \sum_{i \in s} w_i \sum_{t=1}^{T} p_{it}(1 - p_{it}) \begin{pmatrix} x_{it} \\ I_{it} \end{pmatrix} (x'_{it} \; I'_{it}). \tag{6.4.8}$$

Here I_{it} is assumed to be a 10×10 vector instead of 9×9. The naive score in (6.4.7) is equivalent to naive Hosmer and Lemeshow statistic in next section. The Rao-Scott (RS) first-order correction uses the factor

$$\widehat{\delta}_0 = \frac{1}{9} \sum \tilde{\delta}_i = tr \Big[\frac{n}{\hat{N}} \tilde{I}^{-1}_{22.1} \tilde{V}_2 \Big], \tag{6.4.9}$$

which represents a generalized design effect, where $\widehat{V}_2 = V_{Boot}(\sum_{i \in s} w_i \sum_{t=1}^{T} z_{it})$, $z_{it=} \widehat{u}_{2it} - \tilde{A} \widehat{u}_{2it}$, $\tilde{A} = \tilde{I}_{21} \tilde{I}^{-1}_{11}$ and $\widehat{u}_{2it} = \mathbf{x}_{it}(y_{it} - p_{it}(\hat{\beta}, 0))$, $\widehat{u}_{1it} = I_{it}(y_{it} - p_{it}(\hat{\beta}, 0))$. The first-order adjusted quasi-score statistic is then given by

$$Q_{RS}(1) = Q_N / \widehat{\delta}., \tag{6.4.10}$$

which is treated as χ^2_9 under H_0.

The statistic $Q_{RS}(1)$ takes account of the survey design, but not as accurately as the second-order adjusted statistic which requires the knowledge of the off-diagonal elements $\widehat{V}_{2,lm}$ of $\widehat{V}_2, l \neq m$. We need the factor

$$\widehat{a}^2 = \widehat{\delta}_{\cdot}^{-2}\left[\frac{1}{9}\sum_{l=1}^{9}(\widehat{\delta}_l - \widehat{\delta}_{\cdot})^2\right] \tag{6.4.11}$$

where

$$\sum_{l=1}^{9}\widehat{\delta}_l^2 = tr\left[\frac{n^2}{\hat{N}^2}\tilde{I}_{22.1}^{-1}\tilde{V}_2^2\tilde{I}_{22.1}^{-1}\right],. \tag{6.4.12}$$

The second-order adjusted quasi-score statistic is given by

$$Q_{RS}(2) = Q_{RS}(1)/(1+\widehat{a}^2), \tag{6.4.13}$$

which is treated as χ^2 with degrees of freedom $9/(1+\widehat{a}^2)$ under H_0. Note that $Q_{RS}(2) \approx Q_{RS}(1)$ if $\widehat{a} \approx 0$.

6.4.3 Adjusted Hosmer-Lemeshow test

We now consider a survey-weighted version of the Hosmer-Lemeshow (HL) chisquared statistic for testing the null hypothesis H_0 in the context of longitudinal data, and then adjust the statistic to account for the survey design, following the method of Rao and Scott (1981). For $l = 1, ..., 10$, let

$$\widehat{p}_{wl} = (\sum_{i\in s}\sum_{t=1}^{T} w_i \mathbf{I}_{it} y_{it})(\sum_{i\in s}\sum_{t=1}^{T} w_i \mathbf{I}_{it})^{-1},$$

$$\widehat{\pi}_{wl} = (\sum_{i\in s}\sum_{t=1}^{T} w_i \mathbf{I}_{it} \pi_{it})(\sum_{i\in s}\sum_{t=1}^{T} w_i \mathbf{I}_{it})^{-1},$$

and

$$\widehat{W}_l = (\sum_{i\in s}\sum_{t=1}^{T} w_i \mathbf{I}_{it})(\sum_{i\in s}\sum_{t=1}^{T} w_i)^{-1}.$$

Then the survey-weighted, longitudinal data version of the HL statistic is given by

$$X^2_{HL} = nT \sum_{l=1}^{10} \widehat{W}_l \frac{(\widehat{p}_{wl} - \widehat{\pi}_{wl})^2}{\widehat{\pi}_{wl}(1 - \widehat{\pi}_{wl})}. \tag{6.4.14}$$

If the weights w_i are equal and $T = 1$, then (6.4.14) reduces to the HL statistic for the case of one time point and simple random sampling. In the latter case, Hosmer and Lemeshow (1980) approximated the null distribution by a χ^2 with $10 - 2 = 8$ degrees of freedom, χ^2_8. This result is not applicable in the survey context and we therefore adjust (6.4.14) using the Rao-Scott (1981) method.

Rao and Scott (1981) proposed a first-order correction and a more accurate second-order correction to a chisquared statistic for categorical data. Roberts, Rao and Kumar (1987) applied the method to logistic regression of estimated cell proportions, and their results are now applied to our grouped proportions $(\widehat{p}_{wl}, \widehat{\pi}_{wl}), l = 1, ..., 10$. Let $\widehat{p}_w = (\widehat{p}_{w,1}, ..., \widehat{p}_{w,10})'$, $(\widehat{\pi}_{w,1}, ..., \widehat{\pi}_{w,10})'$ and $\widehat{V}_r$ be the estimated covariance matrix of $\widehat{p}_w - \widehat{\pi}_w = \widehat{r}_w$. We used the Rao-Wu Bootstrap method to get an estimator using the one-step Newton-Raphson iteration. Let $\widehat{p}_w^{(b)}$ be the b-th Bootstrap version of $\widehat{p}_w$ and $\widehat{\pi}_w^{(b)}$ be the corresponding Bootstrap version of $\widehat{\pi}_w$ using obtained from (6.4.14). Then,

$$v_{BOOT}(\widehat{r}_w) = \frac{1}{B} \sum_{b=1}^{B} (\widehat{r}_w^{(b)} - \widehat{r}_w)(\widehat{r}_w^{(b)} - \widehat{r}_w)', \tag{6.4.15}$$

is a Bootstrap estimator of $cov(\widehat{r}_w)$. The Rao-Scott (RS) first-order correction uses the factor

$$\widehat{\delta}_{\cdot} = \frac{1}{9}(nT)\left[\sum_{l=1}^{10} \widehat{V}_{r,ll} \widehat{W} \{\widehat{\pi}_l(1 - \widehat{\pi}_l)\}^{-1}\right], \tag{6.4.16}$$

which represents a generalized design effect, where $\widehat{V}_{r,ll}$ is the l'th diagonal element

of $\widehat{V}_r$. The first-order adjusted HL statistic is then given by

$$X^2_{RS}(1) = X^2_{HL}/\widehat{\delta}_{\cdot}, \tag{6.4.17}$$

which is treated as χ^2_8 under H_0. Since the choice of degrees of freedom is not clear-cut, we may follow what is done with the Horton *et al.* score statistic and treat (6.4.17) as χ^2_9 instead of χ^2_8. We need new theory on the asymptotic null distribution of (6.4.17).

The statistic $X^2_{RS}(1)$ takes account of the survey design, but not as accurately as the second-order adjusted statistic which requires the knowledge of the off-diagonal elements $\widehat{V}_{r,lm}$ of $\widehat{V}_r, l \neq m$. We need the factor

$$\widehat{a}^2 = \widehat{\delta}_{\cdot}^{-2}\left[\frac{1}{9}\sum_{l=1}^{9}(\widehat{\delta}_l - \widehat{\delta}_{\cdot})^2\right], \tag{6.4.18}$$

where

$$\sum_{l=1}^{9}\widehat{\delta}_l^2 = \sum_{l=1}^{10}\sum_{m=1}^{10}\widehat{V}^2_{r,lm}\frac{(nT\widehat{W}_l)(nT\widehat{W}_m)}{\widehat{\pi}_l\widehat{\pi}_m(1-\widehat{\pi}_l)(1-\widehat{\pi}_m)}. \tag{6.4.19}$$

The second-order adjusted HL statistic is given by

$$X^2_{RS}(2) = \chi^2_{RS}(1)/(1+\widehat{a}^2), \tag{6.4.20}$$

which is treated as χ^2. The Rao-Scott corrected statistics $X^2_{RS}(1)$ and $X^2_{RS}(2)$ both take account of the survey disign unlike X^2_{HL}, but retain the difficulty associated with X^2_{HL} in the sense that the asymptotic distribution is not exactly specified. Hence, we treat $X^2_{RS}(1)$ as χ^2 with degrees of freedom 8 or 9 when $G = 10$, and $X^2_{RS}(2)$ as χ^2 with degrees of freedom $8/(1+\hat{a}^2)$ or $9/(1+\hat{a}^2)$ under H_0. We need new theory on the asymptotic distribution of the Rao-Scott adjustments to X^2_{HL}.

6.4.4 Pan's Normality Based Goodness-of-fit Tests and Other Methods

Some other goodness-of-fit tests are also be studied to see if they are appropriate to longitudinal complex survey design situation. For ordinary logistic regression, Pan (2002) proposed a normal theory based goodness-of-fit test. In contrast to other famous methods, which depend on the subjective partitioning, Pan applies the asymptotic normal property of residual based sum of individual statistic to preform the goodness-of-fit test. In terms of both size and power, simulation results show Pan's method has good performance. The methods can be easily to extend to the survey design case.

The Pearson chi-square for survey-weighted, longitudinal data version of Pan statistic is given by

$$G = \sum_{i \in s_r} \sum_{t=1}^{T} \widehat{w}_i \frac{(y_{it} - \widehat{\pi}_{it})^2}{\widehat{\pi}_{it}(1 - \widehat{\pi}_{it})}, \tag{6.4.21}$$

where w_i and π_{it} are the same as in (6.4.14). The G statistic avoids a subjective partitioning. If the weights w_i are equal, then it reduces to Pan's statistic for the case of simple random sampling. Using Taylor linearization, one can easily to verify that

$$\hat{E}(G) = T\hat{N}, \hat{N} = \sum_{i \in s_r}^{n} w_i \tag{6.4.22}$$

It is hard to get the second moment via Taylor linearization in the survey design case. However, as before, we can always use the EF-Bootstrap method to estimate the variance of G. Let $V_{BOOT}(G)$ is the estimated variance. Similar to Pan (2002), one can argue that G has an approximately normal distribution (as T is bounded and n tends to infinity). The p-value is thus obtained by referring G to a normal

distribution with mean $T\hat{N}$ and variance $\hat{V}_{BOOT}(G)$.

Similarly, a sum of residual squares, U statistic, can also constructed as:

$$U = \sum_{i \in s_r} \sum_{t=1}^{T} \widehat{w}_i (y_{it} - \widehat{\pi}_{it})^2, \tag{6.4.23}$$

with the approximate mean estimated as,

$$\hat{E}(U) = \hat{\pi}' W (1 - \hat{\pi}), \tag{6.4.24}$$

where $W = diag(w_1, ..., w_n), \pi = (\pi_1, ..., \pi_n)'$, and the EF-Bootstrap variance as $\hat{U}_{boot}$. Again the p-value is obtained by referring U to be a normal distribution with above mean and variance.

As mentioned by Pan (2002), these two types of test statistic are applicable when a continuous covariate is present, or in general, when the number of possible combinations of the covariate values is much larger than the sample size. The results presented in Pan (2002) showed that the performance of these tests is satisfactory in terms of both their sizes and power values. However, further simulation results showed that the power of the test is not good if a continuous covatiate is not present.

In addition to the above goodness-of-fit tests, there are some other interesting test methods in the literature which might be applicable to longitudinal survey data. For example, Graubard *et al.* (1997) proposed a simulated Wald test for complex survey data. This approach has three steps: (1) computing W value as in (6.4.14); (2) under this logistic model and using the estimated parameters, repeatedly generate 999 simulated data sets; (3) computing the new W values by using those data sets. The p-value is given by the percentile of the original W value. This approach can be called a parametric Bootstrap method. Fay (1985) proposed a Jackknifed chi-square test statistic for complex survey data. The test statistic has a complex

limiting distribution under the null hypothesis but one can compute the p-value via a simulation method. This method has a close relationship to the Rao-Scott adjustment.

6.5 Application to NPHS data

We applied the marginal logistic regression model and the EF Bootstrap to data from Statistics Canada's National Population Health Survey (NPHS). The NPHS began in 1994/95, and collects information every two years from the same sample of individuals. A stratified multistage design was used to select households within clusters, and then one household member 12 years or older was chosen to be the longitudinal respondent. The longitudinal sample consists of 17,276 individuals. Currently, 5 cycles of data are available.

Motivated by the research of Shields and Shooshtari (2001) who used NPHS data and logistic regression in order to study the relationship between a self-perceived health measure and various socio-economic, lifestyle, physical and psycho-social health variables, we formulated a marginal logistic regression model for T=2 occasions. We took the same sample of 5380 females who were 25+ years of age at the time of sample selection, were respondents in all of the first three cycles of the survey and did not have proxy responses to the health component of the questionnaire. For occasion t our binary response variable is 1 if self-perceived heath of the i'th individual at time t is excellent or very good and is 0 if self-perceived health at time t is good, fair or poor. The associated vector of covariates consists of 41 dichotomous variables similar to those used by Shields and Shooshtari. Some of the covariates describe the status of the individual at the previous survey cycle, while

other covariates describe changes in status between the previous and current survey cycles. For our example, occasion t=1 is 1996/97 (so that data from both 1994/95 and 1996/97 are used to generate) and occasion t=2 is 1998/99 (so that data from both 1996/97 and 1998/99 are used to generate). A survey weight variable appropriate for respondents to the first three cycles of NPHS was chosen, along with a set of $B = 500$ Bootstrap weights.

6.5.1 Parameter estimates and standard errors

We used the following approaches to model our data:

1. Separate: Logistic models were fit separately to the data for each occasion - thus different β's for each occasion

2. $SEE - Ind$: SEE with a working independence assumption.

3. $SEEII - OR_{constant}$: SEE with a constant odds ratio model for the SEE working covariance structure.

4. $SEEII - OR_{f(age)}$: SEE with a working odds ratio modeled as a function of an individual's age group by

$$\log(\gamma_i) = \alpha_0 + \alpha_1 a_i + \alpha_2 a_i^2, \tag{6.5.1}$$

where $a_i = 1$ for age 25-34, $a_i = 2$ for age 35-44, . $a_i = 5$ for age 65-74, and $a_i = 6$ for age greater or equal to 75.

In approaches 3 and 4, a second set of estimating equations, $\widehat{u}_{2w}(\beta, \alpha) = 0$, was used to estimate the unknown parameters α associated with the working odds ratios. Another option is to use empirical odds ratios to estimate α directly from the data, so that $\widehat{u}_{2w}(\beta, \alpha)$ is not needed. The following two approaches use this option:

5. $SEE - OR_{constant} - E$: empirical constant odds ratio; and

6. $SEE - OR_{f(age)} - E$: empirical constant odds ratio within each age group.

For approaches 2 to 6 we fitted the marginal logistic regression model assuming (i) separate β's for each occasion (thus 82 coefficients to be estimated) and also assuming (ii) a common β. Values of β for (i) are denoted as Time1 and Time 2 in Table 1. The one-step EF approach was used to estimate variances for approaches 2 to 6 while the direct survey Bootstrap was used for approach 1.

Table 1 illustrates the coefficient estimates and associated standard errors under the six different approaches for two of the binary explanatory variables, namely "functionally restricted (yes/no)" and "heavy smoker (yes/no)" used in the logistic regression models.

Table 1. Coefficient estimates and their standard errors (in brackets).

Model	Function restrictive			Heavy Smoker		
	Time1	Time2	common β	Time1	Time2	common β
Separate	-0.94	-1.37	-	-0.41	-0.28	-
	(0.18)	(0.17)	-	(0.15)	(0.18)	-
$SEE - Ind$	-0.94	-1.37	-1.13	-0.41	-0.28	-0.32
	(0.18)	(0.17)	(0.14)	(0.15)	(0.18)	(0.12)
$SEEII-$	-0.91	-1.23	-0.99	-0.43	-0.25	-0.34
$OR_{cons.}$	(0.16)	(0.15)	(0.13)	(0.14)	(0.16)	(0.11)
$SEEII-$	-0.90	-1.23	-0.99	-0.43	-0.25	-0.34
$OR_{f(age)}$	(0.16)	(0.15)	(0.13)	(0.14)	(0.16)	(0.11)
$SEE-$	-0.90	-1.19	-0.94	-0.43	-0.24	-0.35
$OR_{cons.} - E$	(0.15)	(0.15)	(0.12)	(0.14)	(0.16)	(0.11)
$SEE-$	-0.88	-1.20	-0.94	-0.42	-0.24	-0.35
$OR_{f(age)} - E$	(0.15)	(0.15)	(0.12)	(0.14)	(0.16)	(0.11)

Table 1 shows that, for our example, standard errors are quite similar under the four different methods of modelling odds ratios. Standard errors under the working independence model (SEE-Ind) are slightly larger than the corresponding values under the odds ratio models when the logistic regression model uses separate β's for each occasion. For example, the standard error for the variable "functionally restricted" under working independence (SEE-Ind) is 0.18 compared to 0.15 under empirical constant odds ratios within age groups ($SEE-OR_{f(age)}-E$). Finally, for a given approach, the standard errors under the common β model are smaller than the corresponding standard errors under the separate β's model. For example, for

$SEE-OR_{f(age)}-E$ and the variable "heavy smoker", the standard error under the common β model is 0.11 compared to 0.16 for Time 2 under the separate β's model. This reduction in standard error is achieved because the common β model borrows more strength from the two time points than the separate β's model; the latter model uses both time points only through the covariance matrix for the two time points. Although we have not shown it here, we could test whether the common β model explains the data as well as the separate β's model.

6.5.2 Goodness-of-fit tests

We tested the goodness-of-fit of our logistic regression model with the 41 dichotomous covariates used by Shields and Shooshtari (2001). Using the quasi-score test (6.4.3) based on the Bootstrap variance estimator (6.4.5), we obtained $X^2_{RS} = 6.57$ and p-value $P(\chi^2_9 > X^2_{RS}) = 0.682$, which suggests that there is no evidence against the assumed logistic regression model. Note that the p-value is calculated under the framework of the alternative model (6.4.1) based on weighted decile groups.

We now turn to the HL chi-squared statistic and its Rao-Scott adjustments for desing effect. For the survey weighted HL statistic (6.4.14), we obtained $X^2_{HL} = 10.11$, but its null distribution is not asymptotically χ^2_8 or χ^2_9. Hence, we use the Rao-Scott adjustments. For the first-order version (6.4.17), we obtained $\widehat{\delta}_{\cdot} = 1.49$, $X^2_{RS}(1) = X^2_{HL}/\widehat{\delta}_{\cdot} = 6.78$ and corresponding p-values $P(\chi^2_8 > 6.78) = 0.56$ and $P(\chi^2_9 > 6.78) = 0.66$, suggesting that there is no evidence against the assumed logistic regression model. Note that X^2_{HL} is substantially larger than $X^2_{RS}(1)$ because it ignores the cross-sectional dependencies. Turning to the more accurate second-order version (6.4.20), we obtained $\widehat{a}^2 = 0.16$, $X^2_{RS}(2) = 6.78/1.16 = 5.86$ and

degrees of freedom $8/(1+\widehat{a}^2) = 6.9$ or $9/(1+\widehat{a}^2) = 6.9$, with corresponding p-values of 0.54 and 0.64 respectively. The above p-values again suggest no evidence against the assumed logistic regression model. The results are listed in the following table.

Table 2. Results of Goodness-of-fit tests

Method	Q-Value	p-Value	Distribution	Note
Horton Score (Naive)	11.81	0.224	$\chi^2(9)$	
Horton Wald (Bootstrap)	6.57	0.682	$\chi^2(9)$	
Horton 1st RS	6.53	0.686	$\chi^2(9)$	$\delta_0 = 1.81$
Horton 2nd RS	4.05	0.62	$\chi^2(5.59)$	$a^2 = 0.61$
HL Naive	10.11	0.196 0.257	$\chi^2(8)$ $\chi^2(9)$	
HL 1st RS	6.78	0.56 0.66	$\chi^2(8)$ $\chi^2(9)$	$\delta_0 = 1.49$
HL 2nd RS	5.86	0.545 0.638	$\chi^2(6.9)$ $\chi^2(7.76)$	$a^2 = 0.16$

Chapter 7

Conclusion and Future Research

7.1 Conclusions

In chapters 2 to 6, I have studied four important topics: Level set estimation, Nonparametric regression, Confidence intervals under imputation for missing data and longitudinal marginal Logistic regression models for survey data. Results I have obtained may be summarized as follows:

For the level set problem, we established the same L_1-convergence bounds for nonparametric level set estimators without imposing the technical condition of Baillo *et al.* (2001). We extended the results from the *i.i.d.* assumption to an α-mixing sequence. We also gave conditions under which various rates can be improved. The results of a simulation study verified those theoretical conclusions. In the simulation study, two algorithms were designed for generating α-mixing sequences with exponential decay rate and approximate α-mixing sequences with polynomial decay rate.

For the nonparametric regression problem, we established the upper bounds for nonparametric regression estimators under a β-mixing setup, without imposing

any assumptions beyond the *i.i.d.* case. Similar results were also obtained for the complexity-regularized estimators. Furthermore, different almost-sure bounds were obtained for the nonparametric regression estimators, under different growth rates of the entropy of the class of regression function and decay rates of the β-mixing sequences.

For the confidence intervals (CI) in the missing data problem, normality theory based CIs and empirical likelihood based CIs were constructed for fractionally imputed data. Simulation results were also given to validate the theory. For the simulation work, a double bi-section algorithm was designed for two dimensional optimization problem which was applied in the Empirical Likelihood (EL) confidence intervals construction.

Under marginal logistic regression models for longitudinal complex survey data, design-weighted Generalized Estimating Equations (GEE) were used for estimating the model parameters. A one-step Estimating Function (EF) bootstrap method was used for variance estimation. Several Goodness-of-Fit (GOF) tests were studied for model assessment. The methods were applied to data from National Population Health Survey (NPHS) of Statistics Canada.

7.2 Future Research

The previous chapters cover four different research topics. There are a number of interesting problems related to those topics, as well as to some other fields. Those problems remain as my future research interest topics.

The first issue is the asymptotic distribution of the L_1 distance between kernel estimators of the level set and the true level set (for a fixed value of the constant

c), i.e., the asymptotic distribution of the conditional expectation

$$E[|P_n(Z) - P_\infty(Z)||X_1, ..., X_n]. \tag{7.2.1}$$

Such results can then be used for goodness-of-fit test purposes. Furthermore, the possibility of using the bootstrap method to study the distributional properties of the statistic (7.2.1) above is also of interest. Also, I would like to study other nonparametric level set estimators such as, nearest neighbor estimators, histogram estimators as well as general partitioning estimators. I am also interested in semi-parametric estimators and parametric estimator (regression model) in level set estimation for my future research.

For the convergence of nonparametric regression problem, the possibility to relax some of the conditions is of interest. For example, β-mixing sequences may be extended to other mixing sequences. We may also study convergence rates of regression for missing and imputation data.

Turing to confidence intervals for imputed data problems, work needs to be done under semi-parametric regression imputation and nonparametric kernel imputation Wang and Rao (2002). Application of imputation methods to health data is also of interest as item non-response problem is a common case in health related databases. Work on applying empirical likelihood (EL) to inference on measures of income inequality, such as low income proportions, Gini coefficient and Lorenz curve also needs to be done.

Missing covariates problem is a difficult but common and important problem. Reilly and Pepe (1995) used a mean score method for imputing the missing covariates. It would be useful to apply the method of Reilly and Pepe to obtain EL conference intervals.

For marginal logistic regression models, more work needs to be done on the asymptotic properties of bootstrap variance estimators and power of GOF tests. It is also of interest to study logistic linear mixed models (or other generalized linear mixed models) with random effects.

Bibliography

[ALE84] Alexander, K. (1984). Probability inequalities for empirical process and a law of the iterated logarithm. *Annals of Probability,* Vol 4 1041-1067.

[BAI01] Baillo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001), Convergence rates in nonparametric estimation of level sets, *Statistics and Probability Letters. 53,* 27-35.

[BA01] Baraud, Y. Comte, F. and Viennet, G. (2001) Adaptive Estimation in Auto-regression or β-Mixing Regression via Model Selection. *Annals of Statistics 29* 10, 839-875

[BE79] Berbee, H. C. P. (1979) Random Walks with Stationary Increments and renewal Theory. *Mathematic Center Tract 112. Mathematic Centrum, Amsterdam.*

[BIN83] Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review 51* 279-292.

[BIN98] Binder, D.A. (1998). Longitudinal surveys. Why are these surveys different from all other surveys? *Survey Methodology, 24* 101-108.

[BIN04] Binder, D.A., Kovacevic, M. and Roberts, G. (2004). Design-based methods for survey data: alternative uses of estimating functions. *American Statistical Association 2004 Proceedings of the Survey Research Methods Section.*

[BO87] Bosq, D, and Lecoutre, J. P. (1987) Theory of estimation function. *Economica* Paris.

[BO98] Bosq, D. (1998) Nonparametric Statistics for Stochastic Processes. 2nd edition. Springer, New York.

[CAV97] Cavalier, L. (1997), Nonparametric Estimation of Regression Level sets, *Statistics 29* 131-160.

[CHE93] Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika.* 80 107-116.

[CHN93] Chen, S. X. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Proceedings of the Institute of Statistical Mathematics* 45 621-637.

[CHN94] Chen, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis* 49 24-40.

[CHE99] Chen, Y. and Shao, J., (1999). Inference with survey data imputed by hot deck when imputed values are nonidentifiable. *Statistical Sinica* 9 361-384.

[CHE00] Chen, J., Rao, J. N. K. and Sitter, R. R., (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica* 10 1153-1169.

[CHE07] Chen, J. and Rao, J. N. K., (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica* 17 1047-1064.

[DAV93] Davis, L. and Gather, U. (1993), The identification of multiple outliers, *Journal of American Statistical Association. 88*, 782-801.

[DE80] Devroys, L. and Wise, G. (1980), Detection of abnormal behavior via nonparametric estimation of the support, *SIAM Journal of Applied Mathematics 38* 480-488

[DE96] Devroye, L., Gyorfi, L., and Lugosi, G. (1996) Probabilistic Theorey of Pattern Recognition. Springer-Verlag, New York. No22. 1371-1385.

[DI91] Diciccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Annals of Statistical* 19 1053-1061.

[DIG02] Diggle, P. J., Heaherty, P., Liang, K. Y. and Zerger, S. L. (2002) Analysis of Longitudinal Data, 2nd edition New York: Oxford University Press.

[DOU94] Doukhan, P. (1994), Mixing Properties and Examples, *Springer-Verlag, New York.*

[FAY85] Fay, R. E. (1985), A Jackknifed Chi-squared test for Complex Sample. *Journal American Statistical Association.* 80 148-157

[FIT04] Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) Applied Longitudinal Analysis. Hoboken, New York, Wiley.

[FRA91] Francisco, C. A. and Fuller, W. A., (1991). Quantile estimation with a complex survey design. *Annals of Statistics* 19 454-469.

[GH71] Ghosh, J. K. (1971). A new proof for the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics* 42 1957-1961.

[GRA97] Graubard, B.I., Korn, E.L. and Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-174.

[GY02] Gyorfi, L., Kohler, M., Krzyzak,A. and Walk,H. (2002) A distribution-free theory of nonparametric regression. Springer, New York.

[HA75] Hartigan, J.A. (1975), Clustering Algorithms. Wiley, New York.

[HAT68] Hartley, H. O. and Rao, J. N. K., (1968). A new estimation theory for sample surveys. *Biometrika* 55 547-557.

[HOR99] Horton, N.J., Bebchuk, J.D., Jones, C.L., Lipsitz, S.R., Catalano, P.J., Zahner, G.E.P. and Fitzmaurice, G.M. (1999). Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine*, 18 213-222.

[HOS80] Hosmer, D.W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression. *Communications in Statistics*, A10 1043-1069.

[HU00] Hu, F. and Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canadian Journal of Statistics*, 28, 449-499.

[KA84] Kalton, G and Kish, L(1984) Some efficient random impuataion methods. *Communications in Statistics* 13(16) 1919-1939.

[KIM04] Kim, J. K. and Fuller, W. (2004) Fractional hot deck imputation, *Biometrika* 91 559-578.

[KO00] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference* 89 10 1-23.

[KOL59] Kolmogorov, A. N. and Tikhomirov, V. M. (1959) Selected Works of A.N. Kolmogorov: Volume I: Mathematics and Mechanics. Publisher: Kluwer Academic Publ.

[LE96] Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1996).Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory. Vol.42 No. 6* 9 2118-2132.

[LI02] Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing data. 2nd edition. John Wiley & Sons, New York.

[LIA86] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73* 13-22.

[LAI92] Liang, K.-Y., Qaqish,B. and Zeger, S. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, B, 54* 3-40.

[LIP91] Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using odds ratios as a measure of association, *Biometrika, 78* 153-160.

[MAS90] Massart, P. (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability,* 18 1269-1283.

[MO96] Modha, D. S. and Masry, E (1996). Minimum Complexity regression Estimation with Weakly Dependent Observations. *Transactions on Information Theory. Vol. 42 No. 6* 9 2133-2145.

[MOO75] Moore, D. S. and Spruill, M. C. (1975) Unified large-sample theory of general chi-squared statistics for tests of fit. *Annals of Statistica* 3 599-616.

[MU91] Muller, D. W. and Sawitzki, G. (1991), Excess mass estimates and tests of multimodality, *Journal of American Statistical Association. 86* 783-746.

[MU93] Muller, D. W. (1993), The excess mass approach in statistics, *Beiträge Zur Statistik 3. Univiversity Heidelberg..*

[OW88] Owen, A. B., (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75 237-249.

[OW90] Owen, A. B., (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics* 18 90-120.

[OW91] Owen, A. B. (1991). Empirical likelihood for linear models. *Annual of Statistics* 19 1725-1747.

[OW01] Owen, A. B. (2001). Empirical likelihood. Chapman and Hall, New York.

[PA02] Pan, W. (2002). Goodncss-of fit tests for GEE with correlated binary data. *Scandinavian Journal of Statistics, Vol 29* 101-110.

[PO95] Polonik, W. (1995), Measuring mass concentration and estimating density contour clusters- an excess mass approach, *Annals of Statistics 23* 855-881.

[PO97] Polonik, W. (1997), Minimum volume sets and generalized quantile processes, *Stochastic Process Application 69* 1-24.

[QR06A] Qin, Y., Rao, J. N. K., and Ren, Q. (2006). Confidence Intervals for Marginal Parameters Under Imputation for Item nonresponse. *Technical Report No. 431 of the Laborstory for Research in Statistics and Probability, Carleton Unversity*

[QIN06B] Qin, Y. S., Rao, J. N. K. and Ren, Q. (2006) Confidence Intervals for Parameters of the Reponse Variable in a linear model with missing data. *Technical Report of the Laborstory for Research in Statistics and Probability, Carleton Unversity*

[RAN82] Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics* 10 462-474.

[RAO81] Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness for fit and independence in two-way tables. *Journal of American Statistical Association, 76* 221-230.

[RAO88] Rao, J. N. K., and Wu, C. F. J. (1988) Resampling inference with complex survey data, *Journal of the American Statistical Association,* 83 231-241.

[RAO92] Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992) Some recent work on resampling methods. *Survey Methodology,* 18 209-217.

[RAO98A] Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998) Quasi-Score Tests with Survey Data. *Statistica Sinica 8* 1059-1070.

[RAO98B] Rao, J. N. K. (1998) Marginal models for repeated observation: Inference with survey data. *American Statistical Association 1998 Proceedings of Survey Research Motheds Section,* 76-82.

[RAO96] Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association.* 91 499-520.

[RAO02] Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002) Estimating equations for the analysis of survey data using poststratification information, *Sankhya 64, Series A* pp364-378.

[RAO04] Rao, J.N.K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics 33 9* 2087-2095.

[RE95] Marie Reilly and Margaret Sullivan Pepe (1995) A mean score method for missing and auxiliary covariate data in regression models *Biometrika* Vol 82 No. 2 299-314.

[ROB87] Roberts, G., Rao, J.N.K. and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika, 74* 1-12.

[RU96] Rust, K. and Rao, J. N. K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research,* 5 283-310.

[SE80] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.

[SH96] Shao, J. and SITTER, R. R. (1996). Bootstrap for imputed survey data. *Journal of American Statistical Association.* 91 1278-1288.

[SHI01] Shields, M and Shooshtari, S. (2001) Determinants of Self-perceived Health. *Health Reports, 13, No. 1* 35-52.

[SP75] Spruill, M. (1975) Comparison of Chi-Square Goodness-of-Fit Tests Based on Approximate Bahadur Slope *Annals of Statistics, 4* 409-412

[THO75] Thomas, D. R. and Grunkemeier, G. L., (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of American Statistical Association.* 70 865-871.

[VAN90] Van de Geer, S. (1990). Estimating a regression function. *Annals of Sataistics* 25 1014-1035.

[VAN96] Van de Geer and Wegkamp(1996). Consistency for the least squares estimator in nonparametric regression. *Annals of Statistics* 24 2513-2523.

[VAA96] Van der Vaart, A. W. and Wellner, J. A. (1996). Weak Convergency and Empirical Processes, with Applications to Statistics. Springer-Verlag, New York.

[VAP71] Vapnik, V. N., Chervonenkis, A. Y., (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probability Application* 16 264-280.

[VAP74] Vapnik, V. N., Chervonenkis, A. Y., 1974. The Theory of Pattern Recognition. Nauka, Moscow.

[WA02A] Wang, Q. and Rao, J. N. K., (2002a). Empirical likelihood-based inference under imputation for missing response data. *Annals Statistics* 30 896-924.

[WA02B] Wang, Q. and Rao, J. N. K., (2002b.) Empirical likelihood-based inference in linear models with missing data. *Scandinavian Journal Statistics* 29 563-576.

[WO52] Woodruff, R. S., (1952). Confidence intervals for medians and other position measures. *Journal of American Statistical Association.* 47 635-646.

[WU04] Wu, C. (2004) Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica* 14 1057-1067.

[ZHO00] Zhong, B. and Rao, J. N. K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika* 87 929-938.