Parameter estimation and bias correction for diffusion processes

and

a nonparametric approach to census population size estimation

by

Chengyong Tang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Song X. Chen, Major Professor
Krishna B. Athreya
Wayne A. Fuller
Justin Tobias
Cindy L. Yu

Iowa State University

Ames, Iowa

2008

To my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my major professor Dr. Song Xi Chen for his inspiring guidance, constructive suggestion and enthusiastic encouragement during my graduate study. I am very thankful to my committee members, Dr. Krishna Athreya, Dr. Wayne Fuller, Dr. Justin Tobias and Dr. Cindy Yu for their precious help. Graduate assistantship from National Science Foundation grant SES-0518904 is acknowledged. I am also very grateful to the US Census Bureau for providing data analyzing platform and travel assistance. Thanks also go to all my friends for their support and inspiration.

CHAPTER 1. Introduction

This dissertation consists of two individual components. The first one, Chapter 2 in particular, considers the parameter estimation problem of a continuous-time stochastic process, whose focus is on the bias properties of the estimators and the remedy for correcting the bias. The second component, consisting of Chapters 3 and 4, concentrates on the application of nonparametric method in the estimation of population size. Chapter 3 considers the nonparametric estimation of enumeration probability functions in the census, accounting for various features of the human population census. And Chapter 4 focuses on the population size estimation, with the application of the nonparametric estimation introduced in Chapter 3.

1.1 Diffusion Processes and Parameter Estimation

Diffusion process has been used to model stochastic dynamics arising in physics, biology and other natural sciences. As a continuous-time stochastic process, it has a very long history. Starting from (Brown, 1828) or earlier, the continuous-time stochastic process was firstly used to model the motion of particles in liquid. Physician, the great Einstein (1956) for instance, provides the early background of the properties of such process. One latest surge of interest on these processes comes from molecular biology in modeling the dynamics of proteins as part of an effort to understand how energy transfer and conversion happen within biological cells (Fricks, 2004). Perhaps the most eminent use of these continuous time stochastic processes in the last three decades has

been in finance following the works of Merton (1971) and Black and Scholes (1973) which established the foundation of option pricing theory in finance. Since then, there has been phenomenal growth of financial products and instruments whose theoretical background is powered by these processes as documented in Sundaresan (2000).

A d-dimensional time homogeneous parametric diffusion process $\{X_t \in R^d; t \geq 0\}$ is defined by the following stochastic differential equation

$$dX_t = \mu(X_t; \theta)dt + \sigma(X_t; \theta)dB_t, \tag{1.1}$$

where θ is a q-dimensional parameter, $\mu(\cdot; \theta) : \mathcal{R}^d \to \mathcal{R}^d$ and $\sigma(\cdot; \theta) = (\sigma_{ij})_{d\times p} > 0 : \mathcal{R}^d \to \mathcal{R}^{d\times p}$ are respectively drift and diffusion functions representing respectively the conditional mean and variance of the infinitesimal change of X_t at time t, and B_t is a p-dimensional Brownian motion. The existence and uniqueness of the process $\{X_t; t \geq 0\}$ satisfying (1.1) and its probability properties are given in Stroock and Varadhan (1979). Extensions of diffusion processes with Lévy driven processes have been proposed which allow modeling of jumps. Some discussions are given in Barndorff-Nielsen and Shephard (2001 and 2002).

A unique feature of statistical inference for diffusion processes is that despite these processes are continuous-time stochastic models, their observations are made only at discrete time points, say at n equally spaced $\{t\delta\}_{t=0}^n$. Here δ is the sampling interval and can be either fixed or very small corresponding to high-frequency data. Therefore, the estimation of the parameter in the continuous-time process is based discrete observations. See Lo (1988), Bibby and Sørensen (1995), Aït-Sahalia (2002), Aït-Sahalia and Mykland (2003) and Fan (2005) for discussions and overviews for estimation of diffusion processes based on discrete observations.

An important application of diffusion processes is in modeling short-term interest rates, which are fundamental quantities in finance as they define excess asset returns and risk premiums of other assets and their derivative prices. A commonly used one

dimensional family of diffusion processes for the interest rates dynamics is

$$dX_t = \kappa(\alpha - X_t)dt + \sigma X_t^{\rho} dB_t, \tag{1.2}$$

where α, κ, σ and ρ are positive parameters. The linear drift prescribes a mean-reversion of X_t toward the long term mean α at a speed κ. The diffusion function σX_t^{ρ} can accommodate a range of pattens in volatility when X_t varies, which reflects the so-called "level-effect" on the volatility as commonly observed in the return data. Important members of this family are the Vasicek model (Vasicek, 1977) with $\rho = 0$ and the CIR model (Cox, Ingersoll, and Ross, 1985) with $\rho = 1/2$. Both Vasicek and CIR models are commonly used in finance due to (i) both have simple and attractive financial interpretations; and (ii) both admit close-form solutions. The latter facilitates explicit calculations of various option prices.

Despite the critical roles played by these interest-rate processes it is well known empirically that estimation of the drift parameters κ and α can incur large bias and/or variability, see for instance Ball and Torous (1996) and Yu and Phillips (2001). This is the case for virtually all the commonly used estimation approaches including the maximum likelihood estimation. The problem exasperates when the process is lack of dynamics which happens when κ is small. Interest rates typically exhibits less amount of changes than stocks, and is typically lack of dynamics. Indeed, as reported in Phillips and Yu (2005) and our simulation study, the maximum likelihood estimator for κ can have more than 200% relative bias even the processes are observed monthly for more than 10 years. This is rather serious as poor qualitative estimates can produce severely biased option prices and serious financial consequences.

This background motivates the analysis of the parameter estimators to develop their bias properties. Two commonly used diffusion processes are considered in Chapter 2. We develop the bias property of the estimators and a parametric bootstrap procedure is proposed for bias correction.

1.2 Population Size Estimation

The second component of the dissertation is regarding the population size estimation, which is an important issue in various studies.

The total numbers of certain target populations are of great interest in various areas. In wildlife studies, for instance the animal residency and abundance of some rare species, the size of living population is essential for the purpose of protection. For some commercial activities, fisheries for instance, the size of the targets is quite an important information of great relevance on profit.

A direct approach to obtain the population size is through census. However, in most cases in animal science and other ecological studies, taking census on the target is neither realistic nor feasible. The facts are the budget of the studies is limited and the target population frame is variable and not well identifiable. To overcome the difficulty and fulfill the purpose of obtaining the census count, various efforts, mainly through estimating the census count, have been conducted. Therefore, the techniques of estimating the population sizes have a long history in Ecology and animal science. A comprehensive introduction in this area is given by (Seber, 2002).

Essentially, the estimation of the population size is through the so-called capture-recapture design. In particular, individuals of the target populations are captured and then released, usually are tagged, at one occasion. This step constructs one sample of the population. Then at the other occasion, usually at a later time and a different location, another capturing is conducted and results in the second sample. Among the captures of same target population in the new sample, individuals are identified as a re-capture or not, commonly by tag. The portion of the recaptures in these two samples contains information which be used to estimate the population size. The comprehensive review of the capture-recapture method is available in (Seber, 2002).

The census of human population is familiar to us. For social, economical and polit-

ical purposes, governments conduct census periodically, say every ten years in the US. The census of human populations has some different features than that of ecological interest. First of all, the target is quite well defined and framed from the governmental records. And hence the survey of the population is easier. Further, the data collected from sampling the human population usually contain more information from various variables recorded, which might be utilized to evaluate and improve the quality of census count. On the other hand, high quality of the census count is required. For instance, the US census count estimation is concerning undercount at 1% level or less. As the census of the human population can not be perfect, the evaluation of the census count is necessary. This means some later studies are required to evaluate the census procedure. In US, the study is called accuracy and coverage evaluation(ACE). This is through a similar procedure to the capture-recapture study. The data collected from the census are treated as components of one sample. A second sample is collected later to evaluate the quality the census. The studies based on the two samples share the same flavor as that of the capture-recapture in Ecology and animal abundance estimation. Usually, the two samples in the human populations are collected based on frames of two different social systems and the procedure is called dual-system estimation. As an example for the two samples based studies of the census count conducted by the US Census Bureau, see (US Census Bureau, 2004). In the US Census studies, the first sample is named E-sample (enumeration sample), and the second sample is called P-sample (post-enumeration sample). In the rest parts of the paper, we will use the E- and P-sample representing the samples from the first and second surveys with general sense.

As a large scale survey effort, the human population census usually encounters missing values in the data collection. For the study of interest in this paper, one important type of missing values arises from un-resolved recapture status. Ideally, data in the P-sample are either match or not match to those in the E-sample. The un-resolved enumeration status takes place when an individual in the P-sample can be neither match

nor not match to one in the E-sample, usually due to in-sufficient information. The proportion of such un-resolved cases in the 2000 Census data is about 1%. Considering the requirement of high quality of study of this type and the overall estimated level of undercounts in the 2000 Census (US Census Bureau, 2004), the un-resolved cases may not be ignorable. Therefore, fully utilize information from missing data is necessary in order to improve the dual system estimation.

Another feature from large scale data collections is the erroneous enumerations. Erroneous Enumerations are invalid records in the Census and typically lead to over-estimation of the population size. For the US Census, there are two main sources of EEs as described in Hogan (1993) and Haberman, Jiang, and Spencer (1998). One is caused by persons enumerated that should not have been, which includes duplicated or fictitious records, and people born after or died before the census. Another source of EEs is due to enumerations at wrong locations, for instance those enumerations that should be included in the census but not at the location they were counted. The identification of erroneous enumerations is through steps of studies based on the E-sample. In some cases, whether an individual in the E-sample is correct enumerated or not can not be determined. This results in missing data as well. In the 2000 US Census data, un-resolved correct enumeration is about 3%. Therefore, information from this proportion of data is also of great interest. More on erroneous enumeration and its effect on the estimation are considered in Chapters 3 and 4.

The probability that one individual is being enumerated is crucial in performing population size estimation. The Horvitz-Thompson type estimator of the population size is given by

$$\hat{N} = \sum_{i \in \mathcal{S}} \frac{1}{p_i},$$

where $\mathcal{S}$ is the collection of one sample and p_i is the probability that the i^{th} individual being enumerated. However, p_i is unknown. The two sample capture-recapture study

essentially provides information to estimate p_i. Chapter 3 concentrates on the estimation of enumeration probability function, taking account into the features of human census. Efficient ways of utilizing the information from available records are explored. And the studies of the resulting population size estimator is the focus of Chapter 4.

CHAPTER 2. Parameter Estimation and Bias Correction for Diffusion Processes

Let $X_0, X_\delta, \cdots X_{n\delta}$ be discrete observations from process (1.1) at equally spaced time points $\{t\delta\}_{t=0}^n$ over a time interval $[0, T]$ where $T = n\delta$. To simplify notation, we write these observations as $\{X_t\}_{t=0}^n$ by hiding δ whenever doing so does not lead to confusion. The objectives of this study are (i) to understand the above empirical phenomena by developing expansions to the bias and variance of estimators for the Vasicek and CIR processes; and (ii) to propose a bias correction approach that is applicable to general diffusion processes. Two regimes of asymptotic are considered in our analysis. One has δ (the sampling interval) fixed while the sample size $n \to \infty$. The other has δ converges to zero as $n \to \infty$. The latter corresponds to high frequency data, and allows simplification of results as compared to results for the fixed-δ case.

The bias and variance expansions reveal that regardless δ is fixed or diminishing to zero, the bias of the κ estimators and the variances of the two drift parameters estimators are effectively at the order of $T^{-1} = (n\delta)^{-1}$. Our analysis also reveals that the bias and variance of the estimators for the diffusion parameter σ^2 basically enjoys much smaller orders at n^{-1}. These explain why estimation of κ incurs more bias than the other parameters and why the drift parameter (κ and α) estimates are more variable than that of the diffusion parameter σ^2.

We then propose a parametric bootstrap procedure for bias correction in parameter estimation of general diffusion processes. Both theoretical and empirical analysis

show that the proposed bias correction effectively reduces the bias without inflating the variance. We demonstrate in numerical simulations that the proposed bootstrap procedure can be combined with a range of parameter estimators including the approximate likelihood estimation of Aït-Sahalia (2002).

This Chapter is structured as follows. Section 2.1 outlines parameter estimators used in our analysis. The expansions on the bias and variance of the estimators for Vasicek and CIR processes are presented in Section 2.2. Section 2.3 outlines the bootstrap bias correction with justifications. Simulation results are reported in Section 2.4. Section 2.5 analyzes a data set of Fed fund rates and we use it to demonstrate (i) the effect of parameter estimation on option prices and (ii) how to carry out bias correction for option prices. All technical details are deferred to the last section.

2.1 Parameter Estimation for Diffusion Processes

2.1.1 A General Overview

As a diffusion process is Markovian, if its transitional density is known, the maximum likelihood estimation (MLE) is the natural choice for parameter estimation. However, for most of diffusion processes, their transitional distributions are not explicitly known which prevents the use of the MLE. In these cases, several methods are available, which include the martingale estimating equation approach by Bibby and Sørensen (1995); the pseudo-Gaussian likelihood approach of Nowman (1997); the Generalized Method of Moments (GMM) estimator of Hansen and Scheinkman (1995); and the approximate likelihood approach of Aït-Sahalia (2002). Aït-Sahalia and Mykland(2003 and 2004) consider likelihood and the GMM based estimation when δ is random and quantify its impacts on parameter estimation.

Nonparametric estimators for the drift and diffusion functions have been also proposed, which include the kernel estimator by Aït-Sahalia (1996) and Stanton (1997),

and the semiparametric estimators of Jiang and Knight (1997). Fan and Zhang (2003) examine the estimators of Stanton (1997) and analyze the effects of high order stochastic expansions on estimation. Bandi and Phillips (2003) consider two stage kernel estimation of the drift and diffusion functions, replacing the strictly stationary assumption with weaker recurrent Markov processes. See Cai and Hong (2003) and Fan (2005) for reviews.

We carry out our analysis under two regimes of asymptotics. It is assumed, in the first regime that $n \to \infty$ while δ is a fixed constant; and in the second regime that

$$n \to \infty, \quad \delta \to 0, \quad T = n\delta \to \infty \quad \text{and for some } k > 2 \quad T\delta^{1/k} \to \infty. \tag{2.1}$$

The second regime the sampling interval diminishes to zero while the total observational time goes to infinity as $n \to \infty$. This is the so called double asymptotics. The last part of (2.1) is used to bound various remainder terms in moment expansions. We note that $T \to \infty$ mimics the standard asymptotic of $n \to \infty$, and as shown in our analysis is the main driving force in determining the bias and variance properties in the drift parameter estimation.

The motivations for assuming $\delta \to 0$ besides $n \to \infty$ are two folds. One is that high frequency financial data are increasingly available. The other is to accommodate discretization based estimators which normally requires $\delta \to 0$ to make the discretization error to diminish to zero faster enough so that the estimators are consistent. The estimator based on the conventional Euler discretization as well as the approximate likelihood of Aït-Sahalia (2002) with a finite number of terms are examples of such type of estimators.

2.1.2 Estimation for Vasicek Process

The Vasicek process satisfies the univariate stochastic differential equation

$$dX_t = \kappa(\alpha - X_t)dt + \sigma dB(t). \tag{2.2}$$

It is the Ornstein-Uhlenbeck process and was proposed by Vasicek (1977) for interest rate dynamics. The conditional distribution of X_t given X_{t-1} is

$$X_t|X_{t-1} \sim N\left\{X_{t-1}e^{-\kappa\delta} + \alpha(1-e^{-\kappa\delta}), \tfrac{1}{2}\sigma^2\kappa^{-1}(1-e^{-2\kappa\delta})\right\}$$

and the stationary distribution of is $N(\alpha, \frac{1}{2}\sigma^2\kappa^{-1})$. The conditional mean and variance of X_t given X_{t-1} are

$$\begin{aligned} E\left(X_t|X_{t-1}\right) &= X_{t-1}e^{-\kappa\delta} + \alpha(1-e^{-\kappa\delta}) =: \mu(X_{t-1}) \quad \text{and} && (2.3)\\ Var(X_t|X_{t-1}) &= \tfrac{1}{2}\sigma^2\kappa^{-1}(1-e^{-2\kappa\delta}). && (2.4) \end{aligned}$$

Let $\phi(x)$ be the density function of the standard normal distribution $N(0,1)$. Then, the likelihood function of $\theta = (\kappa, \alpha, \sigma^2)$ is

$$L(\theta) = \phi\left(\sigma^{-1}\sqrt{2\kappa}(X_0-\alpha)\right)\prod_{t=1}^{n}\phi\left(\sigma^{-1}\sqrt{2\kappa(1-e^{-2\kappa\delta})^{-1}}\{X_t-\mu(X_{t-1})\}\right).$$

The maximum likelihood estimators (MLE) are

$$\hat{\kappa} = -\delta^{-1}\log(\hat{\beta}_1), \quad \hat{\alpha} = \hat{\beta}_2 \quad \text{and} \quad \hat{\sigma}^2 = 2\hat{\kappa}\hat{\beta}_3(1-\hat{\beta}_1^2)^{-1} \tag{2.5}$$

where

$$\begin{aligned} \hat{\beta}_1 &= \frac{n^{-1}\sum_{i=1}^n X_iX_{i-1} - n^{-2}\sum_{i=1}^n X_i\sum_{i=1}^n X_{i-1}}{n^{-1}\sum_{i=1}^n X_{i-1}^2 - n^{-2}(\sum_{i=1}^n X_{i-1})^2},\\ \hat{\beta}_2 &= \frac{n^{-1}\sum_{i=1}^n (X_i - \hat{\beta}_1 X_{i-1})}{1-\hat{\beta}_1} \quad \text{and}\\ \hat{\beta}_3 &= n^{-1}\sum_{i=1}^n \{X_i - \hat{\beta}_1X_{i-1} - \hat{\beta}_2(1-\hat{\beta}_1)\}^2. \end{aligned} \tag{2.6}$$

The conditional mean and variance (2.3) and (2.4) suggest that the discrete observations $\{X_t\}_{t=0}^n$ follow an AR(1) process with $\beta_1 = e^{-\kappa\delta}$ as the auto-regressive coefficient. As $\beta_1 \to 1$ when $\delta \to 0$, we are having a near unit root situation. Our analysis shows that

$$E(\hat{\beta}_1) = \beta_1 - \frac{4}{n} + \frac{3\kappa\delta}{n} + \frac{7}{n^2\kappa\delta} + o(n^{-2}\delta^{-1} + n^{-1}\delta). \tag{2.7}$$

Here the bias of $\hat{\beta}_1$ is controlled by two forces of asymptotic: δ and n, due to the continuous-time nature of the process. The expansion (2.7) echoes an expansion

$$E(\hat{\beta}_1) = \beta_1 - \frac{1+3\beta_1}{n} + O(n^{-2}) \tag{2.8}$$

given by Marriott and Pope (1954) and Kendall (1954) for discrete-time AR(1).

Although (2.3) and (2.4) suggest a link with AR(1) time series, a key difference between our current study and the conventional AR(1) time series is that δ may diminish to 0 in the sampling of a continuous-time processes. Hence the existing theory on β_1 from time series is not directly applicable to continuous time diffusion processes when we consider the diminishing δ asymptotic for the Vasicek process.

2.1.3 Estimation for CIR Process

A CIR (Cox et al., 1985) diffusion process satisfies

$$dX_t = \kappa(\alpha - X_t)dt + \sigma\sqrt{X_t}dB(t), \tag{2.9}$$

with $2\kappa\alpha/\sigma^2 > 1$. Let $c = 4\kappa\sigma^{-2}(1-e^{-\kappa\delta})^{-1}$, the transitional distribution of cX_t given X_{t-1} is non-central $\chi^2_\nu(\lambda)$ with the degree of freedom $\nu = 4\kappa\alpha\sigma^{-2}$ and the non-central component $\lambda = cX_{t-1}e^{-\kappa\delta}$.

The conditional mean is the same with (2.3) of the Vasicek process. However, due to the heteroscedasticity in the diffusion function, the conditional variance becomes

$$Var(X_t|X_{t-1}) \;=\; \tfrac{1}{2}\alpha\sigma^2\kappa^{-1}(1-e^{-\kappa\delta})^2 + X_{t-1}\sigma^2\kappa^{-1}(e^{-\kappa\delta} - e^{-2\kappa\delta}). \tag{2.10}$$

Since the non-central χ^2-density function is an infinite series involving central χ^2 densities, explicit expression of the MLEs for $\theta = (\kappa, \alpha, \sigma^2)$ is not available. To gain insight on the parameter estimation, we consider pseudo-likelihood estimators proposed by Nowman (1997), which admit close form expressions. Nowman employed a method of Bergstrom (1984) that approximates the CIR process by

$$dX_t = \kappa(\alpha - X_t)dt + \sigma\sqrt{X_{m\delta}}dB(t) \quad \text{for} \quad t \in [m\delta, (m+1)\delta) \tag{2.11}$$

which discretizes the diffusion function within each $[m\delta, (m+1)\delta)$ by its value at the left end point of the interval while keeping the drift unchanged, instead of discretizing the Brownian motion as in the conventional Euler approximation.

Without confusion in the notation, let $\{X_t\}_{t=0}^n$ be observations from process (2.11). Then, they satisfy the following discrete time series model

$$X_t = e^{-\kappa\delta} X_{t-1} + \alpha(1 - e^{-\kappa\delta}) + \eta_t, \tag{2.12}$$

where $E(\eta_t) = 0$, $E(\eta_t \eta_s) = 0$ if $t \neq s$ and $E(\eta_t^2) = \frac{1}{2}\sigma^2\kappa^{-1}(1 - e^{-2\kappa\delta})X_{t-1} =: \xi(X_{t-1}, \theta)$. However, unlike the Vasicek case, where the discrete-time Gaussian model carries the same amount of information as the original continuous-time model, the discrete model (2.12) does not contain the same amount of information as the original continuous-time process (2.9). Hence, results from discrete time series are not applicable even for the case of fixed δ asymptotic.

By pretending η_t to be Gaussian distributed, a pseudo log-likelihood

$$\ell(\theta) = -\sum_{t=1}^{n} \left[\tfrac{1}{2} \log\{\xi(X_{t-1}, \theta)\} + \tfrac{1}{2}\xi^{-1}(X_{t-1}, \theta) \left\{ X_t - e^{-\kappa\delta} X_{t-1} - \alpha(1 - e^{-\kappa\delta}) \right\}^2 \right] \tag{2.13}$$

is obtained which leads to pseudo-MLEs

$$\hat{\kappa} = -\delta^{-1} \log(\hat{\beta}_1), \quad \hat{\alpha} = \hat{\beta}_2 \quad \text{and} \quad \hat{\sigma}^2 = \frac{2\hat{\kappa}\hat{\beta}_3}{1 - \hat{\beta}_1^2} \tag{2.14}$$

where

$$\begin{aligned}
\hat{\beta}_1 &= \frac{n^{-2} \sum_{t=1}^n X_t \sum_{t=1}^n X_{t-1}^{-1} - n^{-1} \sum_{t=1}^n X_t X_{t-1}^{-1}}{n^{-2} \sum_{t=1}^n X_{t-1} \sum_{t=1}^n X_{t-1}^{-1} - 1}, \\
\hat{\beta}_2 &= \frac{n^{-1} \sum_{t=1}^n X_t X_{t-1}^{-1} - \hat{\beta}_1}{(1 - \hat{\beta}_1) n^{-1} \sum_{t=1}^n X_{t-1}^{-1}} \quad \text{and} \\
\hat{\beta}_3 &= n^{-1} \sum_{t=1}^{n} \left\{ X_t - X_{t-1}\hat{\beta}_1 - \hat{\beta}_2(1 - \hat{\beta}_1) \right\}^2 X_{t-1}^{-1}.
\end{aligned} \tag{2.15}$$

We emphasize here that the discretized model (2.11) is used only to produce the estimators. It is the original CIR model (2.9) that is used when we analyze their properties.

2.2 Main Results

In this section, we report results from both fixed δ and diminishing δ asymptotic analysis on the bias and variance of the estimators considered in the previous section.

2.2.1 Fixed δ Analysis

The fixed δ results for the maximum likelihood estimators of the Vasicek process are given in the following two theorems.

Let

$$
\begin{aligned}
&B_1(\theta,\delta) = (2 + e^{\kappa\delta} + e^{2\kappa\delta}),\\
&B_2(\theta,\delta) = -\sigma^2\delta^{-1}\left[\kappa^{-1}\left\{2 - \kappa\delta - \tfrac{1}{2}e^{2\kappa\delta}(1 - e^{-\kappa\delta})\right\} - 4\delta(1 - e^{-2\kappa\delta})^{-1}e^{-2\kappa\delta}\right],\\
&V_1(\theta,\delta) = \delta^{-1}(e^{2\kappa\delta} - 1), \quad V_2(\theta,\delta) = \sigma^2(2\kappa)^{-1}\delta(e^{\kappa\delta} - 1)^{-1}(e^{\kappa\delta} + 1) \quad \text{and}\\
&V_3(\theta,\delta) = \sigma^4(\kappa\delta)^{-2}\left\{2(\kappa\delta)^2 + (e^{\kappa\delta} - e^{-\kappa\delta})\left(1 - \frac{2\kappa\delta e^{-2\kappa\delta}}{1 - e^{-2\kappa\delta}}\right)\right\}.
\end{aligned}
$$

Theorem 2.1 *For a stationary Vasicek Process,*

$$
\begin{aligned}
&E(\hat{\kappa}) = \kappa + (n\delta)^{-1}B_1(\theta,\delta) + O(n^{-2}), \quad Var(\hat{\kappa}) = (n\delta)^{-1}V_1(\theta,\delta) + O(n^{-2}),\\
&E(\hat{\alpha}) = \alpha + O(n^{-2}), \quad Var(\hat{\alpha}) = (n\delta)^{-1}V_2(\theta,\delta) + O(n^{-2}),\\
&E(\hat{\sigma}^2) = \sigma^2 + n^{-1}B_2(\theta,\delta) + O(n^{-2}) \quad \textit{and} \quad Var(\hat{\sigma}^2) = n^{-1}V_3(\theta,\delta) + O(n^{-2}).
\end{aligned}
$$

as $n \to \infty$ while δ is fixed.

Theorem 2.1 indicates that the estimators for all three parameters have both their bias and variances at the order of n^{-1} when δ is fixed, which are the standard parametric rates. However, a closer examination indicates that the variance of $\hat{\kappa}$ and $\hat{\alpha}$, and the bias of $\hat{\kappa}$ are effectively at the order of T^{-1}. Hence, they are effectively controlled by the total amount of observation time of the process. At the same time, the bias and variance

of $\hat{\sigma}^2$ is indeed n^{-1}, and hence converges to zero much faster. It is not very surprising to see the bias of the long term mean estimator $\hat{\alpha}$ is at n^{-2}, as in the Appendix it is shown that the $\hat{\alpha}$ is asymptotically equivalent to $\bar{X}$ which is unbiased to the long term mean α. We note that $V_1(\theta, \delta)$ and $V_2(\theta, \delta)$ are clearly decreasing functions of δ, while $V_3(\theta)$ is relatively stable against the change in δ. This means that for sample size n fixed, the variabilities of the drift parameter estimates increase as the sampling intervals gets finer. On the other hand, when κ gets smaller for fixed n and δ, the ratio $B_1(\theta, \delta)/\kappa$ is getting larger. This explains why the relative bias increases when the mean reverting is weak (smaller κ).

The following theorem establishes the asymptotic normality of the parameter estimators.

Theorem 2.2 *For a stationary Vasicek process, let* $\hat{\theta} = (\hat{\kappa}, \hat{\alpha}, \hat{\sigma}^2)^T$ *and* $\theta = (\kappa, \alpha, \sigma^2)^T$. *Then, as* $n \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Omega_1),$$

where $\Omega_1 = diag\{\delta^{-1}V_1(\theta, \delta), \delta^{-1}V_2(\theta, \delta), V_3(\theta, \delta)\}$.

Theorem 2.2 illustrates that each component of $\hat{\theta}$ converges at the same rate of $n^{1/2}$ and different estimators are asymptotically uncorrelated. The latter is unique for Vasicek processes, as we will see a similar result does not hold for the CIR processes.

As illustrated by (2.3) and (2.4), given the Vasicek process, the observation $\{X_t\}_{t=1}^n$ is exactly an AR(1) time series. For fixed δ, the asymptotic normalities of the estimators for general AR(p) time series using least squared method and MLE are available in Fuller (1996) and hence can be applied here. We note that Theorem 2.2 is a case with exact expression where the conditional variance $var(X_t|X_{t-1})$ depends on the auto-regressive coefficient $e^{-\kappa\delta}$ and leads to explicit expressions of the MLE.

For sampling interval δ fixed situation,

We need some notations before presenting our analysis for the CIR processes. Let $F(a,b,c,z)$ be a hypergeometric function defined by

$$F(a,b,c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \sum_{k=0}^{\infty} \frac{z^k}{k!} \frac{\Gamma(a+k)\Gamma(b+k)}{\Gamma(c+k)},$$

and $\theta_\alpha = 2\kappa\alpha/\sigma^2$ and $\theta_\beta = 2\kappa/\sigma^2$ which are parameters in the stationary marginal Gamma distribution of a CIR process. And let

$$\begin{aligned}
S_1(\theta,\delta) &= n^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{F(1,1,\theta_\alpha; e^{-(j-i)\kappa\delta}) - 1\}, \\
S_2(\theta,\delta) &= -n^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{F(1,1,\theta_\alpha; e^{-(j-i+1)\kappa\delta}) - 1\}, \\
S_3(\theta,\delta) &= n^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{F(0,1,\theta_\alpha - 1; e^{-(j-i+1)\kappa\delta}) - 1\} \quad \text{and} \\
S_4(\theta,\delta) &= n^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{F(0,1,\theta_\alpha; e^{-(j-i+1)\kappa\delta}) - 1\} \qquad (2.16)
\end{aligned}$$

be series associated with various sums of product moments. It can be shown that $S_i(\theta) = O(1)$ for $i = 1,2,3,4$. We note that $\theta_\alpha \geq 2$ is needed to ensure terms with $X_i^{-1}X_j^{-1}$ having bounded expectations. Furthermore, let

$$\begin{aligned}
B_3(\theta,\delta) &= (1+e^{-\kappa\delta})(1-\theta_\alpha) + 2\theta_\alpha^2 S_1(\theta,\delta)(1-e^{\kappa\delta}) + (1-e^{-\kappa\delta}) \left\{ \theta_\alpha^2 \sum_{i=1}^{4} S_i(\theta,\delta) - e^{\kappa\delta} \right\}, \\
B_4(\theta,\delta) &= \tfrac{1}{2}(\theta_\alpha - 1)^{-1}(1-e^{-\kappa\delta})\sigma^2, \\
V_4(\theta,\delta) &= \delta^{-1}(e^{2\kappa\delta}-1), V_5(\theta,\delta) = \theta_\beta^{-2}\theta_\alpha \left\{ 2(e^{\kappa\delta}-1)^{-1} + 1 \right\}, \\
V_6(\theta,\delta) &= A_1(\theta,\delta)^2 Z_1(\theta,\delta) + A_2(\theta,\delta)^2 Z_2(\theta,\delta), \\
A_1(\theta,\delta) &= \frac{\sigma^2\delta^{-1}}{\kappa e^{-\kappa\delta}} - \frac{2\delta\sigma^2}{(1-e^{-2\kappa\delta})}, A_2(\theta,\delta) = \frac{-2\delta^{-2}\kappa}{1-e^{-2\kappa\delta}}, \\
Z_1(\theta,\delta) &= \frac{\theta_\alpha - 1}{\theta_\alpha}(1-e^{-\kappa\delta}) \quad \text{and} \\
Z_2(\theta,\delta) &= \frac{1}{4\beta_3^2} \left[1 + \frac{1}{1+e^{-\kappa\delta}} \left\{ 12e^{-2\kappa\delta} + (12\nu+48)c(\theta)^{-1} \frac{e^{-\kappa\delta}\theta_\beta}{\theta_\alpha - 1} + \right. \right. \\
&\quad \left. \left. (3\nu^2+12\nu)c(\theta)^{-2} \frac{\theta_\beta^2}{(\theta_\alpha-1)(\theta_\alpha-2)} - \frac{2(\theta_\alpha + \theta_\alpha e^{-\kappa\delta} - 2e^{-\kappa\delta}}{(1+e^{-\kappa\delta})(\theta_\alpha-1)} \right\} \right],
\end{aligned}$$

where $c(\theta,\delta) = 2\theta_\beta(1-e^{-\kappa\delta})^{-1}$ and $\nu = 2\theta_\alpha$.

The results on the pseudo-maximum likelihood estimators for the CIR process are summarized below.

Theorem 2.3 *For a stationary CIR process with* $\theta_\alpha \geq 2$,

$$E(\hat{\kappa}) = \kappa + (n\delta)^{-1}B_3(\theta,\delta) + O(n^{-2}), \quad Var(\hat{\kappa}) = (n\delta)^{-1}V_4(\theta,\delta) + O(n^{-2}),$$
$$E(\hat{\alpha}) = \alpha + O(n^{-2}), \quad Var(\hat{\alpha}) = n^{-1}V_5(\theta,\delta) + O(n^{-2}),$$
$$E(\hat{\sigma}^2) = \sigma^2 + B_4(\theta,\delta) + O(n^{-1}) \quad \textit{and} \quad Var(\hat{\sigma}^2) = n^{-1}V_6(\theta,\delta) + O(n^{-2}).$$

as $n \to \infty$.

Theorem 2.3 conveys similar features to those given in Theorem 2.1 for Vasicek processes. This is particularly the case with respect to the orders for the bias and variance of the estimators. Some of the coefficient functions for the bias and variance become more involved in the case of the CIR process. A major difference appears in the bias of $\hat{\sigma}^2$. The $B_4(\theta)$ appears in the bias of $\hat{\sigma}^2$ is a constant and does not converge to 0 when δ is fixed; and hence $\hat{\sigma}^2$ is not a consistent estimator. This is due to the discretization of the diffusion function used when we construct the pseudo-likelihood. In general, any estimation method based on discretization are likely to encounter this type of systematic bias (Lo, 1988). Therefore, $\delta \to 0$ is actually required for those estimators to be consistent. It should be noted that the inconsistency is limited to the diffusion parameter estimation as the drift parameter estimators are asymptotically unbiased and consistent. This is because the discretization of the process is confined to the diffusion part only.

We also have the following asymptotic normality for the CIR parameter estimators.

Theorem 2.4 *For a stationary CIR process with* $\theta_\alpha \geq 2$, *let* $\hat{\theta} = (\hat{\kappa}, \hat{\alpha}, \hat{\sigma}^2)^T$ *and* $\tilde{\theta} = (\kappa, \alpha, \sigma^2 + B_4(\theta,\delta))^T$, *then*

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) \xrightarrow{d} N(0, \Omega_2) \quad \textit{where}$$

$$\Omega_2 = \begin{pmatrix} \delta^{-1}V_4(\theta,\delta) & \delta^{-1}(e^{\kappa\delta}+1) & -A_1(\theta,\delta)\delta^{-1}(e^{\kappa\delta}-e^{-\kappa\delta}) \\ \delta^{-1}(e^{\kappa\delta}+1) & V_5(\theta,\delta) & A_1(\theta,\delta)\delta^{-1}(e^{\kappa\delta}+1) \\ -A_1(\theta,\delta)\delta^{-1}(e^{\kappa\delta}-e^{-\kappa\delta}) & A_1(\theta,\delta)\delta^{-1}(e^{\kappa\delta}+1) & V_6(\theta,\delta) \end{pmatrix}.$$

We note that to take into account of the inconsistency of $\hat{\sigma}^2$, we have adjusted θ to $\tilde{\theta}$ in the above asymptotic normality. An essential distinction between Theorem 2.4 and Theorem 2.2 is that the pseudo-likelihood estimates of CIR processes are no longer asymptotically uncorrelated.

For fixed δ, the estimation of the CIR process is within the framework of ARCH time series, due to the conditional heteroscedasity given by (2.10). Existing literature, for instance (Gouriérous, 1997), has the asymptotic normality result for general pseudo-likelihood estimators. Our Theorem 2.4 has explicit form under a special circumstance where the solutions of the pseudo-likelihood equations are available.

2.2.2 Diminishing δ Analysis

We present results by allowing $\delta \to 0$ so that (2.1) is satisfied. As we will demonstrate shortly, letting $\delta \to 0$ largely simplifies the coefficient functions appear in the bias and variance of the fixed-δ results given in the previous subsection.

The following two theorems are counterparts of Theorems 2.1 and 2.2 respectively.

Theorem 2.5 *For a stationary Vasicek process and under Condition (2.1),*

$$\begin{aligned} &E(\hat{\kappa}) = \kappa + 4/T - \left\{4\kappa n^{-1} + 7/(\kappa T^2)\right\} + o(n^{-1}+T^{-2}), \\ &Var(\hat{\kappa}) = 2\kappa/T + o(T^{-1}), \\ &E(\hat{\alpha}) = \alpha + o(T^{-2}), \quad Var(\hat{\alpha}) = \sigma^2\kappa^{-2}/T + o(T^{-1}), \\ &E(\hat{\sigma}^2) = \sigma^2 + O(n^{-1}) \quad \textit{and} \quad Var(\hat{\sigma}^2) = 2\sigma^4 n^{-1} + o(n^{-1}). \end{aligned}$$

Theorem 2.6 *Let* $\hat{\theta} = (\hat{\kappa}, \hat{\alpha}, \hat{\sigma}^2)^T$ *and* $\theta = (\kappa, \alpha, \sigma^2)^T$*, and under the same conditions of Theorem 2.5 as* $n \to \infty$

$$R_{n,\delta}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Omega_3),$$

where $R_{n,\delta} = diag(T^{1/2}, T^{1/2}, n^{1/2})$ *and* $\Omega_3 = diag(2\kappa, \sigma^2\kappa^{-2}, 2\sigma^4)$.

Theorem 2.5 reveals, first of all, that the leading order bias of $\hat{\kappa}$ is $4/T$, and the leading order relative bias is $4/(\kappa T)$, which gets larger as κ gets smaller (weaker mean-reverting). Secondly, the leading order variance of $\hat{\kappa}$ and $\hat{\alpha}$ are both of $1/T$, which are larger order than $1/n$, the order of $Var(\hat{\sigma}^2)$. Hence, estimation for the two drift parameters are much more variable than $\hat{\sigma}^2$. These confirm the commonly observed empirical bias behavior in κ-estimation as well as larger variability in the drift parameter estimation. The theorem also reveals that despite $\hat{\alpha}$ having a larger order variance, it is almost unbiased. At the same time, contrasting the difficulties in estimating the drift parameters, estimation of σ^2 enjoys both smaller bias and less variability as having been observed in various empirical studies. The weak convergent result in Theorem 2.6 reveals that $\hat{\kappa}$ and $\hat{\alpha}$ converge in a different rate ($T^{-1/2}$) from that of $\hat{\sigma}^2$ ($n^{-1/2}$). And it also indicates that $\hat{\kappa}, \hat{\alpha}$ and $\hat{\sigma}^2$ are asymptotically uncorrelated.

The following are theorems on the estimators of the CIR process, which are similar to Theorems 2.5 and 2.6.

Theorem 2.7 *For a stationary CIR process, and under Condition (2.1) and* $\theta_\alpha \geq 2$*,*

$$\begin{aligned}
&E(\hat{\kappa}) = \kappa + \left(4 + \frac{2}{\theta_\alpha - 1}\right) T^{-1} + o(T^{-1}), Var(\hat{\kappa}) = 2\kappa T^{-1} + o(T^{-1});\\
&E(\hat{\alpha}) = \alpha + o(n^{-1}), \quad Var(\hat{\alpha}) = 2\alpha\theta_\beta^{-1}\kappa^{-1}T^{-1} + o(T^{-1});\\
&E(\hat{\sigma}^2) = \sigma^2 - \frac{\sigma^2\kappa\delta}{2(\theta_\alpha - 1)} + O(n^{-1}), \quad Var(\hat{\sigma}^2) = \sigma^4\left(2 - \frac{1}{\theta_\alpha - 1}\right) n^{-1} + o(n^{-1}).
\end{aligned}$$

where $\theta_\alpha = 2\kappa\alpha/\sigma^2$ *and* $\theta_\beta = 2\kappa/\sigma^2$.

Theorem 2.8 *Let* $\hat{\theta} = (\hat{\kappa}, \hat{\alpha}, \hat{\sigma}^2)^T$ *and* $\theta = (\kappa, \alpha, \sigma^2)^T$, *under the same conditions of Theorem 2.7,*

$$R_{n,\delta}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Omega_4),$$

where $R_{n,\delta} = diag(T^{1/2}, T^{1/2}, n^{1/2})$ *and* $\Omega_4 = \begin{pmatrix} 2\kappa & 2 & 0 \\ 2 & 2\alpha\theta_\beta^{-1}\kappa^{-1} & 0 \\ 0 & 0 & \sigma^4\left(2 - \frac{1}{\theta_\alpha - 1}\right) \end{pmatrix}$.

Theorems 2.7 and 2.8 reveal similar features to those by Theorems 2.5 and 2.6 for the Vasicek process. These include (i) the leading order bias of $\hat{\kappa}$ is still T^{-1}; (ii) estimation of κ and α still incurs a larger order variance as compared to the estimation of σ^2. A difference is in the bias of $\hat{\sigma}^2$, which is at the order of δ^{-1}. This can be understood as a result of the piece-wise discretization of the diffusion function used in (2.11). Again, identical with what is revealed in Theorem 2.4, the estimations of CIR processes using pseudo-likelihood are not asymptotically independent. The difference is that the covariance between $\hat{\kappa}$ and $\hat{\sigma}^2$ and that between $\hat{\alpha}$ and $\hat{\sigma}^2$ are of smaller magnitude order when $\delta \to 0$.

An important message from Theorems 2.5 to 2.8 is that it is T, the total observation time, rather than the sample size n, that controls the bias and/or variance in estimation of κ and α. Our analysis is entirely based on each continuous-time process, and improves the heuristic justification used in Phillips and Yu (2005) which are based on results like (2.8) from discrete-time series. And most importantly, the results in these theorems nicely explain various empirical results reported in the literature.

2.3 Bootstrap Bias Correction

Given the explicit bias expansion in Theorem 2.5, a simple bias correction for $\hat{\kappa}$ for the Vasicek process is $\hat{\kappa}_1 = \hat{\kappa} - 4/T$. This will remove the leading order bias without

altering the variance. The same may be applied to the CIR process by constructing

$$\hat{\kappa}_1 = \hat{\kappa} - \left(4 + \frac{2}{\hat{\theta}_\alpha - 1}\right) T^{-1}$$

where $\hat{\theta}_\alpha = 2\hat{\kappa}\hat{\alpha}/\hat{\sigma}^2$. The limitation of this approach is that it would not be applicable to other processes unless similar bias expansions are established.

In a significant development, Phillips and Yu (2005) propose a jackknife method to correct bias in parameter estimation of diffusion processes. Their proposal was motivated by the bias expansions (2.8) established for discrete time series. It consists of first dividing the entire sample of n observations into m consecutive non-overlapping blocks of observations of size l such that $n = ml$; and then construct parameter estimators based on each block of observations, say $\hat{\theta}_i$ for the i-th block. The jackknife estimator that corrects bias in an original estimator $\hat{\theta}$ is

$$\hat{\theta}_J = \frac{m}{m-1}\hat{\theta} - \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m^2 - m}.$$

They suggested using $m = 4$ which was shown numerically to produce the best trade-off between bias reduction and variance inflation.

In conventional statistical settings, it is understood (Shao and Tu, 1995) that the jackknife tends to inflate variance more than the bootstrap when both are used for bias correction. Indeed, as shown in our simulations, although using $m = 4$ has reduced the variance of the jackknife estimator as opposed to using $m = 2$, the variance can still be much larger than the original estimator. This may be due to that dividing the data into shorter blocks reduces the observation time which has been shown in Theorems 2.5 and 2.7 to be the key force in influencing the variability in the estimation of the drift parameters.

We propose a parametric bootstrap procedure for bias correction. The bootstrap (Efron, 1979) has been shown to be an effective method for bias correction and variance estimation for both independent and dependent observations as summarized in

Hall (1992) and Lahiri (2003). Although our analysis was confined to the two specific processes in the previous section, the proposed bootstrap bias correction is applicable to the general multivariate diffusion process (1.1).

Let $\hat{\theta}$ be a mean square consistent estimator of θ. The parametric bootstrap procedure consists of the following steps:

Step 1. Generate a bootstrap sample path $\{X_t^*\}_{t=1}^n$ with the same sampling interval δ from $dX_t = \mu(X_t;\hat{\theta})dt + \sigma(X_t;\hat{\theta})dB_t$;

Step 2. Obtain a new estimator $\hat{\theta}^*$ from the bootstrap sample path by applying the same estimation procedure as $\hat{\theta}$;

Step 3. Repeat Steps 1 and 2 N_B number of times and obtain a set of bootstrap estimates $\hat{\theta}^{*,1}, \cdots, \hat{\theta}^{*,N_B}$.

Let $\bar{\hat{\theta}}^* = N_B^{-1}\sum_{b=1}^{N_B}\hat{\theta}^{*,b}$, the bootstrap bias-corrected estimator is $\hat{\theta}_B = 2\hat{\theta} - \bar{\hat{\theta}}^*$ and the bootstrap estimates for the variance of $\hat{\theta}$ is

$$\widehat{Var}(\hat{\theta}) = N_B^{-1}\sum_{b=1}^{N_B}\left(\hat{\theta}^{*,b} - \bar{\hat{\theta}}^*\right)\left(\hat{\theta}^{*,b} - \bar{\hat{\theta}}^*\right)^T.$$

Here A^T denotes matrix transpose.

In the above Step 1, we first generate an initial value of X_0^* from the stationary marginal distribution. For a univariate process, the stationary density is known to be

$$\pi_\theta(x) = \frac{\xi(\theta)}{\sigma^2(x,\theta)}\exp\left\{\int_{x_0}^x \frac{2\mu(t,\theta)}{\sigma^2(t,\theta)}dt\right\}.$$

If the transitional distribution of $X_{t\delta}$ given $X_{(t-1)\delta}$ is known, we can generate $X_{t\delta}^*$ given $X_{(t-1)\delta}^*$ from that distribution. If the transitional distribution is unknown, we can use the approximate transitional density of Aït-Sahalia(1999). We may also apply the Milstein scheme (Kloeden and Platen, 2000) which is more accurate than the first-order Euler scheme.

The bootstrap bias correction method shares some key features of the jackknife method, for instance it can be applied to a general diffusion process (univariate or

multivariate), and for a range of estimators including the MLE, the pseudo-MLE and discretization based estimators. The bootstrap bias correction is justified in the following theorem. Before that, let us introduce some notations.

Let $\theta = (\theta_1, \cdots, \theta_p)^T$ be a vector of parameters of the general diffusion process (1.1), and $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_p)^T$ be a consistent estimator of θ. Write the bootstrap bias corrected estimator $\hat{\theta}_B = (\hat{\theta}_{B1}, \cdots, \hat{\theta}_{Bp})^T$. For $l = 1, \cdots, p$, let $b_{nl}(\theta) = E(\hat{\theta}_{nl}) - \theta_l$ and $v_{nl}(\theta) = Var(\hat{\theta}_{nl})$ be the bias and variance components of $\hat{\theta}_l$ respectively, and

$$b_{nl}(\theta) = \beta_{nl} b_{nl}^{(0)}(\theta) \quad \text{and} \quad v_{nl}(\theta) = \nu_{nl} v_{nl}^{(0)}(\theta)$$

so that both $|b_{nl}^{(0)}(\theta)|$ and $|v_{nl}^{(0)}(\theta)|$ are uniformly bounded away from ∞ and zero with respect to n and δ. Hence, both β_{nl} and ν_{nl} are the exact orders of magnitude for the bias and variance of $\hat{\theta}_l$ respectively.

Theorem 2.9 *Suppose that for each $l = 1, \cdots, p$, (i) $\beta_{nl}^2 + \nu_{nl} \to 0$ as both n and $T \to \infty$ and (ii) $b_{nl}^{(0)}(\theta)$ and $v_{nl}^{(0)}(\theta)$, as functions of θ, are twice continuously differentiable within a hypersphere S in R^p that contains the real parameter θ; and (iii) $E\{b_{nl}^{(0)}(\hat{\theta})\}^2 = O(1)$. Then,*

$$E(\hat{\theta}_{Bl}) = \theta_l + o\{b_{nl}(\theta)\} \quad \textit{and} \quad Var(\hat{\theta}_{Bl}) = v_{nl}(\theta) + o\{v_{nl}(\theta)\}. \tag{2.17}$$

The theorem shows that the proposed bootstrap estimator $\hat{\theta}_B$ reduces the bias of the original estimator $\hat{\theta}$ while having the same leading order variance as $\hat{\theta}$. W It can be seen from (2.65) in the appendix that the bias in the bootstrap bias corrected estimator is in fact $O\{b_{nl}(v_{nl} + b_{nl}^2)^{1/2}\}$, indicating meaningful bias reduction. The processes to which the bootstrap technology can be applied are general processes in (2.3), that include both univariate and multivariate processes. This indeed makes the proposed bootstrap estimator generally applicable. We would like to highlight that the theorem applies to both fixed and diminishing δ situations. We note also that the conditions of the theorem

are quite weak, which are no more than the mean square consistent and differentiability of the bias and variance functions near θ. These are satisfied by most of the commonly used estimators, for instance those evaluated in Theorems 2.5 and 2.7.

2.4 Simulation Studies

We report in this section results from simulation studies which were designed to (i) confirm the theoretical findings of Theorems 2.5 and 2.7, (ii) evaluate the performance of the proposed bootstrap bias correction, and (iii) compare the bootstrap proposal with the jackknife bias correction proposed by Phillips and Yu (2005). In the simulation studies, both univariate (Vasicek and CIR processes) and bivariate processes were considered as well as a range of parameter estimators. All the simulation results reported in this section were all based on 5000 simulations and 1000 bootstrap resamples.

2.4.1 Univariate Processes

To confirm the theoretical results given in Theorems 2.5 and 2.7, we simulated two sets of models for both Vasicek and CIR processes. The parameter values used for the Vasicek process were $\theta = (\kappa, \alpha, \sigma^2) = (0.858, 0.0891, 0.00219)$ (Vasicek Model 1) and $(0.215,\ 0.0891, 0.0005)$ (Vasicek Model 2). For the CIR process, the parameter $\theta = (\kappa, \alpha, \sigma^2) = (0.892, 0.09, 0.033)$ (CIR Model 1) and $(0.223, 0.09, 0.008)$ (CIR Model 2) respectively. Both Vasicek Model 2 and CIR Model 2 have only a quarter of the mean-reverting force of Vasicek Model 1 and CIR 1 respectively. We chose $\delta = 1/12$ that corresponds to monthly observations in annualized term. The sample size n was 120, 300, 500 and 2000. The purpose of trying $n = 2000$ was to confirm the asymptotic bias and variance developed in Theorems 2.5 and 2.7. As the transitional distribution of these two processes are known, the simulated sample paths were generated from the known transitional distribution with the initial value X_0 from their known stationary

distributions respectively.

Table 2.1 reports the average bias, relative bias (R. Bias), standard deviation (SD) and root mean square error (RMSE) for the two Vasicek models, while Table 2.2 reports the results with the comparable setting for two CIR models. We also report in parentheses the asymptotic bias and standard deviation prescribed by expansions in Theorem 2.7. We observe that the severe bias in κ estimation was very clear, especially when the amount of the mean reverting was weak (Table 2.1(b) and Table 2.2(b)). At the same time, there was little bias in the estimation of α and the overall quality in estimating σ^2 was very high even for sample size as small as 120. These all confirmed our theoretical findings. We find the difference between the simulated bias and SD and those predicted by the theoretical expansions decreased as n and T were increased, and was very small at $n = 2000$, which was reassuring.

We then applied the bootstrap bias correction to estimation of κ for the Vasicek and CIR models. The jackknife approach proposed by Phillips and Yu (2005) was also performed with $m = 4$. The simulation results are summarized in Table 2.3. We see that the bootstrap bias correction effectively reduced the bias without increasing the variance of the estimation much. However for the jackknife bias correction, there was some non-ignorable variance inflation. The bootstrap bias correction had less RMSE than the jackknife bias correction as well as the original estimator.

We also carried out estimation and bootstrap bias correction based on the approximated likelihood estimation of Aït-Sahalia(2002) for the CIR Model 2. This was designed to see if there were significant difference between the approximated MLEs and the pseudo-likelihood estimators of Nowman (1997) which we have analyzed in Theorem 2.7. The results are reported in Table 2.4, which were similar to the pseudo-likelihood estimators in Table 2.2(b). However, we did see that the use of the approximate likelihood did produce estimates which had slightly smaller bias and standard deviation. Most importantly, the bootstrap bias correction worked well for the approximated likelihood

in reducing both the bias and mean square error in κ-estimation.

2.4.2 Multivariate Processes

To evaluate the general applicability of the proposed bootstrap procedure, we carry out simulations for the following bivariate processes:

$$dX_t = \kappa(\alpha - X_t)dt + \sigma(X_t)dB_t \tag{2.18}$$

where

$$X_t = \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}, \kappa = \begin{pmatrix} \kappa_{11} & 0 \\ \kappa_{21} & \kappa_{22} \end{pmatrix}, \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad \text{and} \quad \sigma(X_t) = \begin{pmatrix} \sigma_{11} X_{1t}^{\rho} & 0 \\ 0 & \sigma_{22} X_{2t}^{\rho} \end{pmatrix}$$

with $\rho = 0$ and $1/2$ respectively. Here $\rho = 0$ corresponds a bivariate Ornstein-Uhlenbeck (OU) process whose exact transitional density is known to be bivariate Gaussian, whereas $\rho = 1/2$ corresponds to a bivariate extension of the Feller's process. We will first report simulation results for the parameter estimation and bias correction for two bivaiate diffusion processes.

For a bivariate OU process X_t, the stationary distribution is $N(\alpha, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \frac{\sigma_{11}^2}{2\kappa_{11}} & -\frac{\kappa_{21}}{\kappa_{11}+\kappa_{22}} \frac{\sigma_{11}^2}{2\kappa_{11}} \\ -\frac{\kappa_{21}}{\kappa_{11}+\kappa_{22}} \frac{\sigma_{11}^2}{2\kappa_{11}} & \frac{\sigma_{22}^2}{2\kappa_{22}} + E \end{pmatrix},$$

where $E = \frac{\kappa_{21}^2}{\kappa_{22}(\kappa_{11}+\kappa_{22})} \frac{\sigma_{11}^2}{2\kappa_{11}}$. The transitional distribution $X_t | X_{t-1}$ is given by $N\{\mu(X_{t-1}), \Sigma_0\}$, where

$$\begin{aligned} \mu(X_{t-1}) &= \alpha + \exp(-\kappa\delta)(X_{t-1} - \alpha) \\ \Sigma_0 &= \sigma - \exp(-\kappa\delta)\Sigma \exp(-\kappa^T\delta), \end{aligned}$$

where $\exp(\cdot)$ is the matrix exponential. Therefore, the bivariate OU can be generated by the transitional distribution and the parameter estimation can be carried out by maximizing the likelihood function.

Unless $\kappa_{21} = 0$, the transitional density of the bivariate Feller's process does not admit an explicit form (Aït-Sahalia and Kimmel, 2002). We consider estimation based on the Euler discretization:

$$X_t - X_{t-1} = \kappa(\alpha - X_{t-1})\delta + \sigma(X_{t-1})\Delta B_t, \tag{2.19}$$

where $\Delta B_t = (B_{1t} - B_{1t-1}, B_{2t} - B_{2t-1})^T$ is a discretization of the bivariate Brownian motion. A pseudo-likelihood can be constructed similar to that of (2.13) from the conditional mean and variance structures:

$$E(X_t|X_{t-1}) = \kappa(\alpha - X_{t-1})\delta, \quad \text{and} \quad Var(X_t|X_{t-1}) = \sigma(X_{t-1})\text{diag}(\delta, \delta)\sigma^T(X_{t-1}).$$

The approximation (2.19) is subject to a discretization error. However, the pseudo-likelihood estimator is consistent as long as $\delta \to 0$.

We chose $(\kappa_{11}, \kappa_{21}, \kappa_{22}, \alpha_1, \alpha_2, \sigma_{11}, \sigma_{22}) = (0.223, 0.4, 0.9, 0.09, 0.08, 0.008, 0.03)$. The generation of the bivariate diffusion process was via the Milstein's scheme (Kloeden and Platen, 2000). We also pre-runned the process 1000 times before starting the real simulation to make the simulated sample path stationary.

Table 2.5 and Table 2.6 summarize the simulation performance of the parameter estimation and and the bootstrap bias corrected parameter estimation. We observe that similar to the univariate case as reported in Tables 2.2 and 2.3, the estimators of the drift parameters in κ and α had worse performance than the diffusion parameters in σ. It is encouraging to see that the bootstrap worked effectively in reducing both the bias and the mean square errors. In Table 2.6, the bootstrap bias correction for κ_{21} did not working as effectively as those for κ_{11} and κ_{22}. However, our simulation results in 2.5 for the bivariate Ornstein-Uhlenbeck process showed that the bootstrap work well for all κ coefficient including κ_{21}. Hence, this suggests that it might be due to the use of the pseudo-likelihood that only uses the conditional variance that ignores the dependence between the two marginal processes. We note that there was no need

to carry out the bootstrap bias correction for (α_1, α_2) and $(\sigma_{11}, \sigma_{22})$. This is consistent with the recommendations from Theorems 2.5 and 2.7.

2.5 A Case Study and Option Pricing

We analyze a Fed fund interest rate dataset consisting 432 monthly observations from January 1963 to December 1998. This dataset has been analyzed in (Aït-Sahalia, 1999) to demonstrate his approximate likelihood estimation.

In additional to estimate the Vasicek and CIR processes, we computed two option prices driven by these two processes: $P_{t,T}(\theta)$, the price of a zero-coupon-bond at time t that pays \$1 at a maturity time T; and $C_{t,T,S,K}(\theta)$, the price at time t of an European call option with maturity T and a strike price K on a zero-coupon bond maturing at $S > T$. See Vasicek (1977) and Cox et al. (1985) for detailed expressions of these two option prices as functions of parameters of an underlying interest rate process.

We first estimated the parameters of the underlying diffusion processes (Vasicek and CIR) by the maximum likelihood method and carried out the bootstrap bias correction. Then, we calculated the option prices $P_{t,T}(\theta)$ and $C_{t,T,S,K}(\theta)$ based on the estimated parameters of the Vasicek or CIR process with $t = 0$, $T = 1$, $S = 3$ and the initial interest rate at 5%. The face value of the European Call option on a three year discount bond was \$100 with a strike price $K = \$90$.

The parametric bootstrap was used to estimate both the bias of the parameter estimates of the process and the option prices as well as their standard deviations based on 1000 resamples. The bootstrap implementation for the option prices were readily made by extending Steps 2 and 3 in the procedure outlined in Section 4 to include computation of the option prices in each resample. The empirical results are reported in Table 2.7. It is observed that the bootstrap bias estimates (Estimated Bias) were rather substantial in both $\hat{\kappa}$ and the option price $\hat{C}(0, 1, 3, 90)$. While the large bias in $\hat{\kappa}$ was expected

from Theorems 2.5 and 2.7, it was rather alarming to see a large under-estimation (more than 10%) in $\hat{C}(0, 1, 3, 90)$. Also, the bootstrap estimate of the standard deviation for both κ and $C(0, 1, 3, 90)$ were quite large too. The large variability in the option price should be taken into consideration and indicates the difficulties in producing accurate estimated prices. The empirical analysis also indicated that the European call option is more affected by the biased parameter estimates for the underlying interest rate process than the zero-coupon bond, which can be understood due to different transformations of the underlying diffusion parameters. We also supplied in parentheses the estimated standard deviation based on the leading order variance terms prescribed by Theorems 2.5 and 2.7, which were all comparable with the bootstrap estimates.

2.6 Discussion

The estimation of the drift parameters in diffusion processes has been known to be challenging when the process is lack of dynamics. Our analysis reported in Theorems 2.5 to 2.8 quantify the underlying sources of the challenge for two commonly used interest rate processes. One source of the challenge, apart from being lack of dynamics, is that the accuracy in the estimation of the drift parameters is governed by T, the total amount of time a process is observed, rather than the sample size n. This is different from estimation for discrete time series models where n is the driving force for accuracy. Although Theorems 2.5 and 2.7 are on the two specific interest rate processes, their results could be used to understand parameter estimation of a general process.

The proposed bootstrap bias correction whose justification is contained in Theorem 2.9 is for general processes. A reason for the proposed parametric bootstrap method working more effectively than the jackknife method is that its re-creation of the full observation length in each resampling that fully utilizes the amount of observation data available and the assumed model for the process. While we have gained quite complete

understanding on parameter estimation for the two popular processes in Theorems 2.1 to 2.8, there is a need to understand more on estimation for multivariate processes, in particularly estimation of parameters that control the correlation between components of the process. Another important issue is how to reduce the variability in estimation of the drift parameters. We hope future research will address these issues.

(a) Vasicek Model 1

		κ	α	σ^2
	True Value	0.858	0.0891	0.00219
$n = 120$	Bias (A. Bias)	0.481(0.4)	$-2.4 \cdot 10^{-4}(0.0)$	$3.7 \cdot 10^{-5}(0.0)$
	R Bias (%)	56.039	0.284	1.700
	SD (Asy. SD)	0.659(0.414)	0.017(0.017)	0.0003(0.0003)
	RMSE	0.816	0.017	0.0003
$n = 300$	Bias (A. Bias)	0.181(0.16)	$-5.3 \cdot 10^{-5}(0.0)$	$1.3 \cdot 10^{-5}(0.0)$
	R Bias (%)	21.082	0.059	0.579
	SD (Asy. SD)	0.329(0.261)	0.011(0.010)	$1.8 \cdot 10^{-4}(1.8 \cdot 10^{-4})$
	RMSE	0.375	0.011	$1.9 \cdot 10^{-4}$
$n = 500$	Bias (A. Bias)	0.111(0.096)	$-6.8 \cdot 10^{-5}(0.0)$	$9.6 \cdot 10^{-6}(0.0)$
	R Bias (%)	12.880	0.076	0.438
	SD (Asy.SD)	0.240(0.202)	0.008(0.008)	$1.4 \cdot 10^{-4}(1.4 \cdot 10^{-4})$
	RMSE	0.265	0.008	$1.4 \cdot 10^{-4}$
$n = 2000$	Bias (A. Bias)	0.024(0.024)	$-1.1 \cdot 10^{-4}(0.0)$	$2.33 \cdot 10^{-7}(0.0)$
	R Bias	2.777	0.119	0.011
	SD (Asy. SD)	0.111(0.101)	0.004(0.004)	$7.3 \cdot 10^{-5}(6.9 \cdot 10^{-5})$
	RMSE	0.113	0.004	$7.3 \cdot 10^{-5}$

(b) Vasicek Model 2

		κ	α	σ^2
	True Value	0.215	0.0891	0.0005
$n = 120$	Bias (A. Bias)	0.507(0.4)	$9.2 \cdot 10^{-4}(0.0)$	$7.8 \cdot 10^{-6}(0.0)$
	R Bias(%)	236.344	1.031	1.428
	SD (Asy. SD)	0.519(0.210)	0.032(0.033)	$7.5 \cdot 10^{-5}(6.5 \cdot 10^{-5})$
	RMSE	0.726	0.032	$7.5 \cdot 10^{-5}$
$n = 300$	Bias (A. Bias)	0.191(0.16)	$-8.5 \cdot 10^{-6}(0.0)$	$3.0 \cdot 10^{-6}(0.0)$
	R Bias (%)	88.985	0.010	0.541
	SD (Asy. SD)	0.221(0.131)	0.022(0.021)	$4.5 \cdot 10^{-5}(4.1 \cdot 10^{-5})$
	RMSE	0.292	0.022	$4.5 \cdot 10^{-5}$
$n = 500$	Bias (A. Bias)	0.114(0.096)	$-1.5 \cdot 10^{-4}(0.0)$	$2.3 \cdot 10^{-6}(0.0)$
	R Bias (%)	53.033	0.174	0.413
	SD (Asy. SD)	0.150(0.110)	0.017(0.016)	$3.4 \cdot 10^{-5}(3.2 \cdot 10^{-5})$
	RMSE	0.189	0.017	$3.4 \cdot 10^{-5}$
$n = 2000$	Bias (A. Bias)	0.025(0.024)	$-2.2 \cdot 10^{-4}(0.0)$	$7.3 \cdot 10^{-8}(0.0)$
	R Bias (%)	11.602	0.244	0.013
	SD (Asy. SD)	0.059(0.051)	0.009(0.009)	$1.8 \cdot 10^{-5}(1.6 \cdot 10^{-5})$
	RMSE	0.064	0.009	$1.8 \cdot 10^{-5}$

Table 2.1 Bias, Relative bias(R.bias), standard deviation(SD) and the root mean squared error(RMSE) of the pseudo-likelihood estimator for the Vasicek models; figures inside the parentheses are those predicted by the theoretical expansions in Theorem 2.5.

(a) CIR Model 1

		κ	α	σ^2
	True Value	0.892	0.09	0.033
$n = 120$	Bias (A. Bias)	0.464(0.452)	$2.4 \cdot 10^{-4}(4.3 \cdot 10^{-4})$	$8.7 \cdot 10^{-4}(3.5 \cdot 10^{-4})$
	R Bias (%)	52.005	0.270	2.661
	SD (Asy. SD)	0.627(0.431)	0.020(0.021)	0.005(0.004)
	RMSE	0.780	0.020	0.005
$n = 300$	Bias (A. Bias)	0.179(0.180)	$2.2 \cdot 10^{-4}(1.7 \cdot 10^{-4})$	$5.8 \cdot 10^{-4}(3.5 \cdot 10^{-4})$
	R Bias (%)	20.107	0.250	1.778
	SD (Asy. SD)	0.334(0.273)	0.012(0.013)	0.003(0.003)
	RMSE	0.380	0.012	0.003
$n = 500$	Bias (A. Bias)	0.107(0.108)	$6.4 \cdot 10^{-5}(1.0 \cdot 10^{-4})$	$4.9 \cdot 10^{-4}(3.5 \cdot 10^{-4})$
	R Bias (%)	12.037	0.070	1.510
	SD (Asy. SD)	0.247(0.211)	0.009(0.01)	0.002(0.002)
	RMSE	0.269	0.009	0.002
$n = 2000$	Bias (A. Bias)	0.025(0.027)	$3.5 \cdot 10^{-5}(3.0 \cdot 10^{-5})$	$3.5 \cdot 10^{-4}(3.5 \cdot 10^{-4})$
	R Bias (%)	2.805	0.039	1.061
	SD (Asy. SD)	0.112(0.106)	0.005(0.005)	0.001(0.001)
	RMSE	0.115	0.005	0.001

(b) CIR Model 2

		κ	α	σ^2
	True Value	0.223	0.09	0.008
$n = 120$	Bias (A. Bias)	0.509(0.452)	$1.2 \cdot 10^{-3}(1.7 \cdot 10^{-3})$	$1.5 \cdot 10^{-4}(2.9 \cdot 10^{-5})$
	R Bias (%)	228.251	1.343	1.796
	SD (Asy. SD)	0.507(0.216)	0.036(0.042)	0.001(0.001)
	RMSE	0.719	0.036	0.001
$n = 300$	Bias (A. Bias)	0.185(0.180)	$9.2 \cdot 10^{-4}(7.0 \cdot 10^{-4})$	$8.7 \cdot 10^{-5}(2.9 \cdot 10^{-5})$
	R Bias (%)	82.836	1.018	1.062
	SD (Asy. SD)	0.222(0.136)	0.025(0.026)	0.0007(0.0006)
	RMSE	0.289	0.025	0.001
$n = 500$	Bias (A. Bias)	0.108(0.108)	$3.7 \cdot 10^{-4}(4.1 \cdot 10^{-4})$	$5.5 \cdot 10^{-5}(2.9 \cdot 10^{-5})$
	R Bias (%)	48.612	0.408	0.669
	SD (Asy. SD)	0.148(0.106)	0.019(0.02)	0.0005(0.0005)
	RMSE	0.183	0.019	0.001
$n = 2000$	Bias (A. Bias)	0.025(0.027)	$9.2 \cdot 10^{-5}(1.0 \cdot 10^{-4})$	$2.9 \cdot 10^{-5}(2.9 \cdot 10^{-5})$
	R Bias (%)	11.145	0.102 0.346	
	SD (Asy. SD)	0.058(0.053)	0.009(0.01)	$2.6 \cdot 10^{-4}(2.5 \cdot 10^{-4})$
	RMSE	0.063	0.009	$2.6 \cdot 10^{-4}$

Table 2.2 Bias, Relative bias(R.bias), standard deviation(SD) and the root mean squared error(RMSE) of the pseudo-likelihood estimator for the CIR models; figures inside the parentheses are those predicted by the theoretical expansions in Theorem 2.7.

(a) Vasicek Model 1 and CIR Model 1

		$\hat{\kappa}$	$\hat{\kappa}_J$	$\hat{\kappa}_B$	$\hat{\kappa}$	$\hat{\kappa}_J$	$\hat{\kappa}_B$
$n = 120$	Bias	0.481	−0.120	0.001	0.464	−0.122	0.002
	R Bias (%)	56.039	14.941	0.118	52.005	13.650	0.178
	SD	0.659	0.767	0.623	0.627	0.730	0.651
	RMSE	0.816	0.778	0.623	0.780	0.739	0.651
$n = 300$	Bias	0.181	−0.026	−0.003	0.179	−0.027	−0.004
	R Bias (%)	21.082	3.070	0.406	20.107	3.094	0.447
	SD	0.329	0.353	0.321	0.334	0.365	0.326
	RMSE	0.375	0.354	0.321	0.380	0.366	0.326
$n = 500$	Bias	0.111	0.005	0.001	0.107	−0.008	0.007
	R Bias (%)	12.880	0.586	0.073	12.037	0.842	0.826
	SD	0.240	0.250	0.235	0.247	0.257	0.245
	RMSE	0.265	0.250	0.235	0.269	0.257	0.245

(b) Vasicek Model 2 and CIR Model 2

		$\hat{\kappa}$	$\hat{\kappa}_J$	$\hat{\kappa}_B$	$\hat{\kappa}$	$\hat{\kappa}_J$	$\hat{\kappa}_B$
$n = 120$	Bias	0.507	−0.112	0.032	0.509	−0.088	0.030
	R Bias (%)	236.344	51.974	14.774	228.251	39.283	13.579
	SD	0.519	0.645	0.510	0.507	0.623	0.501
	RMSE	0.726	0.655	0.511	0.719	0.630	0.502
$n = 300$	Bias	0.191	−0.029	0.002	0.185	−0.032	0.008
	R Bias (%)	88.985	13.465	0.829	82.836	14.428	3.461
	SD	0.221	0.261	0.219	0.222	0.265	0.226
	RMSE	0.292	0.262	0.219	0.289	0.267	0.226
$n = 500$	Bias	0.114	−0.011	0.002	0.108	−0.0161	0.003
	R Bias (%)	53.033	5.230	0.861	48.612	7.209	1.325
	SD	0.150	0.170	0.147	0.148	0.167	0.150
	RMSE	0.189	0.171	0.147	0.183	0.168	0.150

Table 2.3 Comparisons of bias corrections for the Vasicek and CIR Models, $\hat{\kappa}_J$ and $\hat{\kappa}_B$ are, respectively, the jackknife and bootstrap bias corrected estimators for κ.

		CIR model 2				
		$\hat{\kappa}$	$\hat{\alpha}$	$\hat{\sigma}^2$	$\hat{\kappa}_J$	$\hat{\kappa}_B$
	True Value	0.223	0.09	0.008	0.223	0.223
$n = 120$	Bias	0.494	0.004	$1 \cdot 10^{-4}$	-0.072	0.035
	R Bias (%)	221.684	4.778	1.507	32.412	15.559
	SD	0.490	0.058	0.001	0.596	0.514
	RMSE	0.696	0.058	0.001	0.601	0.516
$n = 300$	Bias	0.180	0.001	$6 \cdot 10^{-5}$	-0.035	0.013
	R Bias (%)	80.559	1.349	0.700	15.803	5.618
	SD	0.223	0.0262	0.001	0.262	0.234
	RMSE	0.286	0.0262	0.001	0.265	0.234
$n = 500$	Bias	0.1001	$7 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	-0.022	-0.003
	R Bias (%)	45.279	0.834	0.478	9.806	1.493
	SD	0.147	0.019	0.001	0.166	0.151
	RMSE	0.178	0.019	0.001	0.167	0.151

Table 2.4 Parameters estimation and bias correction for CIR Model 2 based on the approximated likelihood method of Aït-Sahalia (1999).

	Bivariate Ornstein-Uhlenbeck Process						
$n = 120$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
θ	0.215	0.4	0.5	0.0891	0.09	0.0005	0.002
Bias	0.475	0.546	0.690	0.0017	0.0009	$5 \cdot 10^{-6}$	0.0002
	(0.106)	(0.131)	(-0.216)	(0.0015)	(0.0015)	$(-1 \cdot 10^{-6})$	(0.00051)
Rbias	221.57	136.465	137.9	1.878	0.955	0.942	12.272 (%)
	(49.47)	(32.767)	(43.289)	(1.635)	(1.692)	(0.184)	(25.494)
SD	0.522	0.758	0.663	0.032	0.038	$7.2 \cdot 10^{-5}$	0.001
	(0.493)	(0.56)	(0.784)	(0.041)	(0.037)	$(7.1 \cdot 10^{-5})$	(0.0011)
RMSE	0.706	0.934	0.956	0.032	0.038	$7.2 \cdot 10^{-5}$	0.001
	(0.504)	(0.575)	(0.813)	(0.041)	(0.037)	$(7.1 \cdot 10^{-5})$	(0.0011)
$n = 300$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
Bias	0.187	0.167	0.220	-0.0013	0.0009	$3 \cdot 10^{-6}$	$-2.1 \cdot 10^{-5}$
	(0.031)	(0.087)	(0.013)	(-0.0011)	(0.0009)	$(1 \cdot 10^{-6})$	(0.0001)
Rbias	87.152	41.863	44.02	1.454	1.006	0.553	1.04
(%)	(14.574)	(21.766)	(2.627)	(1.276)	(0.951)	(0.2)	(5.14)
SD	0.221	0.361	0.283	0.023	0.026	$4.2 \cdot 10^{-5}$	0.0006
	(0.221)	(0.312)	(0.285)	(0.023)	(0.026)	$(4.2 \cdot 10^{-5})$	(0.0007)
RMSE	0.290	0.398	0.358	0.023	0.026	$4.2 \cdot 10^{-5}$	0.0006
	(0.223)	(0.324)	(0.286)	(0.023)	(0.026)	$(4.2 \cdot 10^{-5})$	(0.0007)
$n = 500$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
Bias	0.106	0.084	0.147	0.0001	-0.0002	$5 \cdot 10^{-7}$	$-3.5 \cdot 10^{-5}$
	(0.016)	(0.033)	(0.017)	(0.0003)	(-0.0003)	$(1 \cdot 10^{-7})$	$(3.5 \cdot 10^{-5})$
Rbias	49.633	20.998	29.459	0.156	0.172	0.088	1.761 (%)
	(7.337)	(8.286)	(3.498)	(0.357)	(0.286)	(0.02)	(1.738)
SD	0.150	0.268	0.196	0.017	0.020	$3.3 \cdot 10^{-5}$	0.0006
	(0.148)	(0.237)	(0.197)	(0.017)	(0.021)	$(3.3 \cdot 10^{-5})$	(0.0006)
RMSE	0.184	0.281	0.245	0.017	0.020	$3.3 \cdot 10^{-5}$	0.0006
	(0.149)	(0.239)	(0.197)	(0.017)	(0.021)	$(3.3 \cdot 10^{-5})$	(0.0006)

Table 2.5 Bias, Relative bias(R.bias), standard deviation(SD) and the root mean squared error(RMSE) of the pseudo-likelihood estimator for a bivariate Ornstein-Uhlenbeck Process; figures in parentheses are those for the bootstrap bias corrected estimators.

	Bivariate Feller Process						
$n = 120$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
θ	0.223	0.4	0.9	0.09	0.08	0.008	0.03
Bias	0.478	0.396	0.531	0.001	-0.0001	-0.0002	-0.0006
	(0.101)	(0.168)	(0.187)	(0.0001)	(-0.001)	(-0.0001)	(-0.0005)
Rbias	214.48	98.948	88.442	1.537	0.147	2.68	2.141 (%)
	(45.227)	(41.948)	(31.086)	(0.135)	(1.273)	(1.245)	(1.672)
SD	0.468	0.584	0.561	0.037	0.036	0.001	0.0041
	(0.543)	(0.846)	(0.696)	(0.037)	(0.037)	(0.001)	(0.004)
RMSE	0.669	0.705	0.772	0.037	0.036	0.001	0.0042
	(0.553)	(0.863)	(0.721)	(0.037)	(0.037)	(0.001)	(0.004)
$n = 300$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
Bias	0.174	0.08	0.206	0.002	-0.0009	$-9 \cdot 10^{-5}$	$-7 \cdot 10^{-5}$
	(-0.003)	(0.064)	(-0.006)	(0.002)	(-0.0012)	$(-1 \cdot 10^{-5})$	(-0.0002)
Rbias	78.048	20.048	34.368	2.131	1.12	1.06	0.219 (%)
	(1.266)	(15.890)	(1.006)	(1.710)	(1.499)	(0.164)	(0.776)
SD	0.212	0.304	0.303	0.026	0.023	0.0006	0.0027
	(0.208)	(0.297)	(0.288)	(0.026)	(0.023)	(0.0007)	(0.0027)
RMSE	0.275	0.314	0.366	0.026	0.023	0.0007	0.0027
	(0.208)	(0.303)	(0.288)	(0.026)	(0.023)	(0.0007)	(0.0027)
$n = 500$	κ_{11}	κ_{21}	κ_{22}	α_1	α_2	σ_1^2	σ_2^2
Bias	0.102	0.017	0.115	$3 \cdot 10^{-5}$	0.000423	$-5 \cdot 10^{-5}$	0.0002
	(-0.003)	(0.024)	(-0.013)	$(1 \cdot 10^{-5})$	$(-8 \cdot 10^{-5})$	$(-6 \cdot 10^{-6})$	(-0.0001)
Rbias	45.52	4.271	19.087	0.035	0.528	0.656	0.64 (%)
	(1.227)	(6.102)	(2.163)	(0.012)	(0.099)	(0.077)	(0.359)
SD	0.148	0.22	0.214	0.021	0.018	0.0005	0.0024
	(0.147)	(0.227)	(0.206)	(0.021)	(0.019)	(0.0005)	(0.0021)
RMSE	0.179	0.22	0.243	0.021	0.018	0.00048	0.0024
	(0.147)	(0.228)	(0.206)	(0.021)	(0.018)	(0.00048)	(0.0023)

Table 2.6 Bias, Relative bias(R.bias), standard deviation(SD) and the root mean squared error(RMSE) of the pseudo-likelihood estimator for a Bivariate Feller's Process; figures in parentheses are those for the bootstrap bias corrected estimators.

(a) Under Vasicek Process

	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{\sigma}^2$	$\hat{P}$	$\hat{C}$
Estimates	0.261	0.07	0.0005	0.846	3.03
Estimated Bias	0.125	$2 \cdot 10^{-5}$	$2 \cdot 10^{-6}$	-0.004	-0.313
BP Estimates	0.136	0.07	0.0005	0.852	3.67
$\widehat{SD}(Asy.SD)$	0.17(0.12)	0.015(0.014)	$3.5 \cdot 10^{-5}(3.4 \cdot 10^{-5})$	0.015	1.146

(b) Under CIR Process

	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{\sigma}^2$	$\hat{P}$	$\hat{C}$
Estimates	0.146	0.07	0.0043	0.852	2.64
Estimated Bias	0.127	$8 \cdot 10^{-4}$	$3 \cdot 10^{-5}$	-0.004	-0.294
BP Estimates	0.018	0.069	0.0043	0.860	3.39
$\widehat{SD}(Asy.SD)$	0.152(0.11)	0.02(0.02)	$3.0 \cdot 10^{-4}(3.0 \cdot 10^{-4})$	0.014	0.996

Table 2.7 Results for a case study: $\hat{P}$ and $\hat{C}$ are the estimated prices for the discount bond and European call option respectively; Estimated Bias, Bootstrap Estimates and $\widehat{SD}$ are respectively the bootstrap estimate of the bias, the bootstrap bias corrected estimate and the bootstrap estimation of the standard deviation; figures in parentheses are the asymptotic standard deviation (Asy.SD) based on the leading order variance given Theorems 2.5 and 2.7.

2.7 Technical Proofs

Proof of Theorem 2.1

Define $\beta_1 = e^{-\kappa\delta}$, $\beta_2 = \alpha$ and $\beta_3 = \frac{1}{2}\sigma^2\kappa^{-1}(1-e^{-2\kappa\delta})$. We note that the discretized sample $(X_0, \ldots, X_n)$ from the stationary Vasicek process is $N(\alpha\mathbf{1}_{n+1}, \Sigma)$ distributed where $\mathbf{1}_{n+1} = (1, \ldots, 1)^T$ and $\Sigma = (\sigma_{ij})_{(n+1)\times(n+1)}$ with $\sigma_{ij} = \frac{1}{2}\sigma^2\kappa^{-1}e^{-|j-i|\kappa\delta}$.

Let $X_i' = X_i - \alpha$, $a_1 = n^{-1}\sum_{i=1}^n X_i'X_{i-1}'$ and $a_2 = n^{-2}\sum_{i=1}^n X_i'\sum_{j=1}^n X_{j-1}'$, $b_1 = n^{-1}\sum_{i=1}^n X_{i-1}'^2$ and $b_2 = n^{-2}(\sum_{i=1}^n X_{i-1}')^2$. Then by rewriting equation (2.6), we have

$$\hat{\beta}_1 = \frac{\mu_{a_1} + (a_1 - \mu_{a_1}) - a_2}{\mu_{b_1} + (b_1 - \mu_{b_1}) - b_2}, \tag{2.20}$$

where $\mu_{a_1} = E(a_1)$ and $\mu_{b_1} = E(b_1)$. Note that $(X_0', \ldots, X_n')$ follows a multivariate normal distribution with zero mean and the same covariance matrix Σ as $(X_0, \ldots, X_n)$, and $\frac{\mu_{a_1}}{\mu_{b_1}} = e^{-\kappa\delta} = \beta_1$, as $\mu_{a_1} = \frac{1}{2}\sigma^2\kappa^{-1}e^{-\kappa\delta}$ and $\mu_{b_1} = \frac{1}{2}\sigma^2\kappa^{-1}$. Further, standard derivations can show that $Var(a_1) = O(n^{-1}\delta^{-1})$ and $Var(n^{-1}\sum_{i=1}^n X_i') = O(n^{-1}\delta^{-1})$. Hence $(a_1 - \mu_{a_1}) = O_p(n^{-1/2})$, $a_2 = O_p(n^{-1/2})$, $(b_1 - \mu_{b_1}) = O_p(n^{-1/2})$ and $b_2 = O_p(n^{-1/2})$.

Define $T_1 = (b_1 - \mu_{b_1}) - b_2$ and $T_2 = (a_1 - \mu_{a_1}) - a_2$, then

$$\hat{\beta}_1 = \frac{\mu_{a_1} + T_2}{\mu_{b_1} + T_1} = \frac{\mu_{a_1} + T_2}{\mu_{b_1}}\left\{1 - \frac{T_1}{\mu_{b_1}} + \frac{T_1^2}{\mu_{b_1}^2} + o_p(n^{-1})\right\}. \tag{2.21}$$

Let

$$\begin{aligned}
A_1 &= \frac{1}{\mu_{b_1}}(a_1 - \mu_{a_1}) - \frac{\mu_{a_1}}{\mu_{b_1}^2}(b_1 - \mu_{b_1}), \\
A_2 &= -\frac{1}{\mu_{b_1}}a_2 + \frac{\mu_{a_1}}{\mu_{b_2}^2}b_2 - \frac{1}{\mu_{b_2}^2}(a_1 - \mu_{a_1})(b_1 - \mu_{b_1}) + \frac{\mu_{a_1}}{\mu_{b_1}^3}(b_1 - \mu_{b_1})^2, \\
A_3 &= \frac{1}{\mu_{b_1}^2}\{b_2(a_1 - \mu_{a_1}) + a_2(b_1 - \mu_{b_1})\} - \frac{\mu_{a_1}}{\mu_{b_1}^3}\{2b_2(b_1 - \mu_{b_1})\} \\
&+ \frac{1}{\mu_{b_1}^3}\left\{(b_1 - \mu_{b_1})^2(a_1 - \mu_{a_1})\right\} - \frac{\mu_{a_1}}{\mu_{b_1}^4}\left\{(b_1 - \mu_{b_1})^3\right\} \quad \text{and}
\end{aligned}$$

$$\begin{aligned} A_4 &= -\frac{1}{\mu_{b_1}^2}a_2b_2 + \frac{\mu_{a_1}}{\mu_{b_1}^3}b_2^2 \\ &+ \frac{1}{\mu_{b_1}^3}\left\{-(b_1-\mu_{b_1})^2a_2 - 2b_2(b_1-\mu_{b_1})(a_1-\mu_{a_1})\right\} + \frac{\mu_{a_1}}{\mu_{b_1}^4}3(b_1-\mu_{b_1})^2b_2 \\ &+ \frac{1}{\mu_{b_1}^4}\left\{-(b_1-\mu_{b_1})^3(a_1-\mu_{a_1})\right\} + \frac{\mu_{a_1}}{\mu_{b_1}^5}(b_1-\mu_{b_1})^4, \end{aligned}$$

then (2.21) implies that

$$\hat{\beta}_1 = A_1 + A_2 + A_3 + A_4 + o_p(n^{-2}). \tag{2.22}$$

Clearly, $E(A_1) = 0$. The moment generating function(MGF) will be utilized in order to derive $E(A_2)$. Note the MGF of $X = (X_0', \ldots, X_n')^T$ is

$$M_X(t) = \exp\left(\frac{1}{2}t^T\Sigma t\right) = \sum_{i=0}^{\infty}\frac{(\frac{1}{2}t^T\Sigma t)^i}{i!},$$

for $t = (t_0, \ldots, t_n)^T$. Therefore

$$E(X_i'X_{j-1}') = \left.\frac{\partial^2 M_X(t)}{\partial t_i \partial t_{j-1}}\right|_{t=0} = \left.\frac{\partial^2(\frac{1}{2}t'\Sigma t)}{\partial t_i t_{j-1}}\right|_{t=0} = Ve^{-|j-i-1|\kappa\delta}, \tag{2.23}$$

where $V = \frac{1}{2}\sigma^2\kappa^{-1}$. Hence

$$\begin{aligned} E(A_2) &= n^{-2}(e^{-\kappa\delta} - e^{\kappa\delta})\sum_{j>i}\left\{(e^{-(j-i)\kappa\delta} + 2e^{-2(j-i)\kappa\delta})\right\} \\ &= n^{-2}(e^{-\kappa\delta} - e^{\kappa\delta})\left\{f_n(\kappa\delta) + 2f_n(2\kappa\delta)\right\}, \end{aligned} \tag{2.24}$$

where

$$f_n(\kappa\delta) = \sum_{j>i}e^{-(j-i)\kappa\delta} = \frac{n-1}{e^{\kappa\delta}-1} - \frac{1-e^{-(n-1)\kappa\delta}}{(e^{\kappa\delta}-1)^2} = n(e^{\kappa\delta}-1)^{-1} + O(1). \tag{2.25}$$

Equations (2.24) and (2.25) implies that

$$E(A_2) = -n^{-1}\{e^{-\kappa\delta}(3 + e^{\kappa\delta})\} + O(n^{-2}). \tag{2.26}$$

Utilizing the moment bounds for weakly dependent sequence (Yokoyama, 1980), it is straightforward to show that $E(A_3) = O(n^{-2})$ and $E(A_4) = O(n^{-2})$. Therefore,

$$E(\hat{\beta}_1) = \beta_1 - n^{-1}\{e^{-\kappa\delta}(3 + e^{\kappa\delta})\} + O(n^{-2}). \tag{2.27}$$

Now write $A_1 = A_{11} - A_{22}$, where $A_{11} = {\mu_{b_1}}^{-1}(a_1 - \mu_{a_1})$ and $A_{22} = \mu_{a_1}\mu_{b_1}^{-2}(b_1 - \mu_{b_1})$. As $Var(A_i) = o\{Var(A_1)\}$ and $Cov(A_1, A_i) = o\{Var(A_1)\}$ for $i = 2, 3, 4$,

$$Var(\hat{\beta}_1) = Var(A_1)\{1 + o(1)\} = \{Var(A_{11}) + Var(A_{22}) - 2Cov(A_{11}, A_{22})\}\{1 + o(1)\}. \tag{2.28}$$

By replicatedly using (2.23), (2.28) implies that

$$Var(\hat{\beta}_1) = n^{-1}(1 - e^{-2\kappa\delta}) + O(n^{-2}). \tag{2.29}$$

Note that an expansion of $\hat{\kappa}$ is given by

$$\hat{\kappa} = -\frac{1}{\delta}\left\{\log(\beta_1) + \frac{(\hat{\beta}_1 - \beta_1)}{\beta_1} - \frac{(\hat{\beta}_1 - \beta_1)^2}{2\beta_1^2} + O_p(n^{-3/2})\right\}. \tag{2.30}$$

Then (2.27), (2.29) and (2.30) together establish the first two equations in Theorem 3.1.1.

The following is the proof of the second two equations on $\hat{\alpha}$. From (2.6), we have

$$\begin{aligned}
\hat{\alpha} &= \hat{\beta}_2 = \frac{n^{-1}\sum_{i=1}^{n}(X_i - \hat{\beta}_1 X_{i-1})}{1 - \hat{\beta}_1} \\
&= n^{-1}\sum_{i=1}^{n} X_i + \frac{\hat{\beta}_1}{1 - \hat{\beta}_1} n^{-1}\sum_{i=1}^{n}(X_i - X_{i-1}) \\
&= \bar{X} + \frac{\beta_1}{1 - \beta_1} n^{-1}(X_n - X_0)\{1 + O_p(n^{-1/2})\}.
\end{aligned} \tag{2.31}$$

The direct result from (2.31) is that $\hat{\alpha} = \bar{X} + o_p(n^{-1})$. Then the second two equations of Theorem 3.1.1 follow by noting $E(\bar{X}) = \alpha$ and $Var(\bar{X}) = n^{-1}(e^{\kappa\delta} + 1)(e^{\kappa\delta} - 1)\sigma^2(2\kappa)^{-1}$.

To prove the last two equations of Theorem 2.1 on $\hat{\sigma}^2$, we first work on $\hat{\beta}_3$ given by (2.6). By using the previous results on $\hat{\beta}_1$ (2.27 and 2.29) and $\hat{\beta}_2$, we have

$$\hat{\beta}_3 = n^{-1}\sum_{i=1}^{n}\{X_i - \beta_1 X_{i-1} - \beta_2(1 - \beta_1)\}^2 + o_p(n^{-1}).$$

Then it can be shown that

$$E(\hat{\beta}_3) = \beta_3 + o(n^{-1}) \quad \text{and} \quad Var(\hat{\beta}_3) = 2n^{-1}\beta_3^2 + o(n^{-1}). \tag{2.32}$$

Note from (2.5),

$$\hat{\sigma}^2 = \frac{2\hat{\kappa}}{1-\hat{\beta}_1^2}\hat{\beta}_3 = \frac{-2\delta^{-1}\log(\hat{\beta}_1)}{1-\hat{\beta}_1^2}\hat{\beta}_3. \tag{2.33}$$

Let $S(x,y) = \log(x)(1-x^2)^{-1}y$. Then $\hat{\sigma}^2 = -2\delta S(\hat{\beta}_1, \hat{\beta}_3)$. It can be established that

$$\begin{aligned} -\tfrac{1}{2}\delta\hat{\sigma}^2 &= S(\beta_1,\beta_3) + (\hat{\beta}_1-\beta_1)\frac{\partial S}{\partial\beta_1} + (\hat{\beta}_3-\beta_3)\frac{\partial S}{\partial\beta_3} + (\hat{\beta}_1-\beta_1)^2\frac{\partial^2 S}{\partial\beta_1^2} \\ &+ (\hat{\beta}_3-\beta_3)^2\frac{\partial^2 S}{\partial\beta_3^2} + (\hat{\beta}_1-\beta_1)(\hat{\beta}_3-\beta_3)\frac{\partial^2 S}{\partial\beta_1\partial\beta_3} + o_p(n^{-1}). \end{aligned} \tag{2.34}$$

Then the proof of Theorem 2.1 is completed by taking expectation and variance operations on both sides of (2.34).

Proof of Theorem 2.2

Let $\beta = (\beta_1, \beta_2, \beta_3)$ be the vector of the true parameter. Applying Taylor's expansion to the likelihood score equations, we have

$$0 = \frac{\partial^T \ell(\hat{\beta})}{\partial\beta} = \frac{\partial^T \ell(\beta)}{\partial\beta} + \frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}(\hat{\beta}-\beta) + \xi_n,$$

where ξ_n is the remainder term. This implies that

$$\sqrt{n}(\hat{\beta}-\beta) = \left\{-\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\}^{-1}\left\{\sqrt{n}\left(\frac{\partial^T\ell(\beta)}{\partial\beta} + \xi_n\right)\right\}.$$

Utilizing the central limiting theorem for mixing sequences (Bosq, 1998), it can be shown that

$$\frac{1}{\sqrt{n}}\frac{\partial^T\ell(\beta)}{\partial\beta} \xrightarrow{d} N(0, \Sigma^{-1}),$$

where $\Sigma = diag\{2\beta_3\kappa\sigma^{-2}, \beta_3(1-\beta_1)^{-2}, 2\beta_3^2\}$. As $\left\{-n^{-1}\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\} \xrightarrow{p} \Sigma^{-1}$,

$$\left\{-\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\}^{-1}\left(\sqrt{n}\frac{\partial^T\ell(\beta)}{\partial\beta}\right) \xrightarrow{d} N(0,\Sigma).$$

As a result of Theorem 2.1, $\frac{1}{\sqrt{n}}\xi_n \xrightarrow{p} 0$. Then by Slutsky's Theorem, $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\Sigma)$ and Theorem 2.2 holds by transforming back to $\hat{\theta}$ as a function of the asymptotically normal vector $\hat{\beta}$ (Serfling, 1980).

Proof of Theorem 2.3

We first note two basic facts regarding a sample $\{X_i\}_{i=1}^n$ from a stationary CIR process: (i) for any $i, j \geq 0$, $c \cdot X_i | X_j \sim \chi^2_\nu(\lambda)$ distribution where $\nu = \frac{4\kappa\alpha}{\sigma^2}$, $\lambda = cX_j e^{-|j-i|\kappa\delta}$ and $c = \frac{4\kappa}{\sigma^2(1-e^{-|j-i|\kappa\delta})}$; and (ii) $X_t \sim \Gamma(\theta_\alpha, \theta_\beta)$ where $\theta_\alpha = \frac{2\kappa\alpha}{\sigma^2}$ and $\theta_\beta = \frac{2\kappa}{\sigma^2}$.

Let $F(a_1, a_2; b; Z) = \frac{\Gamma(b)}{\Gamma(a_1)\Gamma(a_2)} \sum_{k=0}^\infty \frac{Z^k}{k!} \frac{\Gamma(a_1+k)\Gamma(a_2+k)}{\Gamma(b+k)}$ be the hypergeometric function. It can be shown using the above facts

$$E(X_i^{-1}) = \sum_{k=0}^\infty \frac{(\lambda/2)^k}{k!} e^{-\lambda/2} \frac{1/2}{\nu/2+k-1} \quad \text{and} \tag{2.35}$$

$$E\left(X_i^{-1}X_j^{-1}\right) = E\left\{X_i^{-1} E\left(X_j^{-1}|X_i\right)\right\} = \frac{\theta_\beta^2}{(\theta_\alpha-1)^2} \cdot F(1,1;\theta_\alpha, e^{-|j-i|\kappa\delta}). \tag{2.36}$$

The function $F(a_1, a_2, b; Z)$ converges absolutely if $b - a_1 - a_2 \geq 0$, therefore

$$E(X_i^{-1}X_j^{-1}) < \infty$$

if $\theta_\alpha \geq 2$. This is the reason behind assuming $\theta_\alpha \geq 2$. Similarly, it can be concluded that

$$E\left(X_{i-1}^{-1}X_i X_{j-1}^{-1}\right) = C_\theta(i,j) S_{i,j} \quad \text{for} \quad i < j \quad \text{and} \tag{2.37}$$

$$E\left(X_{j-1}^{-1}X_{i-1}^{-1}X_i\right) = \frac{\theta_\beta e^{-\kappa\delta}}{\theta_\alpha - 1} + \frac{\theta_\beta^2 \alpha(1-e^{-\kappa\delta})}{(\theta_\alpha-1)^2} F(1,1,\theta_\alpha, e^{-|i-j|\kappa\delta}) \quad \text{for} \quad j < i-1,$$

where $C_\theta(i,j) = \frac{\theta_\beta \Gamma^2(\theta_\alpha-1)}{(1-e^{-\kappa\delta})(1-e^{-(j-i-1)\kappa\delta})\Gamma(\theta_\alpha)}$, $S_{i,j} = \sum_{k=0}^\infty \sum_{l=0}^\infty D_1^k D_2^l \frac{\Gamma(\theta_\alpha+l+k+1)}{\Gamma(\theta_\alpha+l)\Gamma(\theta_\alpha+k)}$, $D_1 = e^{-\kappa\delta}/(e^{-\kappa\delta}-1)$ and $D_2 = e^{-(j-i-1)\kappa\delta}/(e^{-(j-i-1)\kappa\delta}-1)$;

$$E\left(X_{j-1}X_{i-1}^{-1}X_i\right) = -\frac{\alpha e^{-(i-j)\kappa\delta}}{\theta_\alpha - 1}(1-e^{-\kappa\delta}) + \mu \quad \text{for} \quad j < i-1 \quad \text{and} \tag{2.38}$$

$$E\left(X_{i-1}^{-1}X_i X_{j-1}\right) = e^{-(j-i-1)\kappa\delta} C(1-e^{-\kappa\delta}) + \mu \quad \text{for} \quad j > i,$$

where $\mu = \frac{(\theta_\alpha - e^{-\kappa\delta})\alpha}{\theta_\alpha - 1}$ and $C = \frac{\theta_\alpha}{\theta_\alpha-1}\left(\frac{\sigma^2}{2\kappa}+\alpha\right)(1-e^{-\kappa\delta}) + \frac{\sigma^2}{\kappa}e^{-\kappa\delta} + 2\alpha e^{-\kappa\delta} - \alpha(1+e^{-\kappa\delta}) - \frac{\alpha}{\theta_\alpha-1}$.

Let $\mu_1 = E(X_t)$, $\mu_2 = E(X_t^{-1})$, $\mu_3 = E(X_t X_{t-1}^{-1})$, $\mu_1' = \mu_1\mu_2 - \mu_3$ and $\mu_2' = \mu_1\mu_2 - 1$. It can be shown that $\mu_1 = \alpha$ and $\mu_2 = \frac{\theta_\beta}{\theta_\alpha-1}$, $\mu_3 = E\left\{X_{t-1}^{-1} E\left(X_t|X_{t-1}\right)\right\} = \frac{\theta_\alpha - e^{-\kappa\delta}}{\theta_\alpha-1}$.

Then by the definition in (2.15)

$$\hat{\beta}_1 = \frac{\mu_1'}{\mu_2'} + \frac{T_{11}}{\mu_2'} - \frac{\mu_1' T_{21}}{\mu_2'^2} + \frac{T_{12}}{\mu_2'} - \frac{\mu_1' T_{22}}{\mu_2'^2} - \left(\frac{T_{11} T_{21}}{\mu_2'^2} - \frac{\mu_1'^2 T_{21}^2}{\mu_2'^3} \right) \{1 + o_p(1)\}, \quad (2.39)$$

where

$$\begin{aligned}
T_{11} &= \mu_2 n^{-1} \sum_{i=1}^{n} (X_i - \mu_1) + \mu_1 n^{-1} \sum_{j=1}^{n} (X_{j-1}^{-1} - \mu_2) - n^{-1} \sum_{i=1}^{n} (X_i X_{i-1}^{-1} - \mu_3), \\
T_{12} &= n^{-2} \sum_{i=1}^{n} (X_i - \mu_1) \sum_{j=1}^{n} (X_{j-1}^{-1} - \mu_2), \\
T_{21} &= \mu_2 n^{-1} \sum_{i=1}^{n} (X_{i-1} - \mu_1) + \mu_1 n^{-1} \sum_{j=1}^{n} (X_{j-1}^{-1} - \mu_2) \quad \text{and} \\
T_{22} &= n^{-2} \sum_{i=1}^{n} (X_{i-1} - \mu_1) \sum_{j=1}^{n} (X_{j-1}^{-1} - \mu_2).
\end{aligned}$$

It is clear that $E(T_{11}) = E(T_{21}) = 0$. And it can be shown that

$$E\left\{ (X_i - \mu_1)(X_j^{-1} - \mu_2) \right\} = -(\theta_\alpha - 1)^{-1} e^{-|j-i|\kappa\delta}.$$

Then

$$E\left(\frac{T_{12}}{\mu_2'} - \frac{\mu_1' T_{22}}{\mu_2'^2} \right) = -n^{-2} (e^{\kappa\delta} - e^{-\kappa\delta}) f_n(\kappa\delta) \quad (2.40)$$

where

$$\begin{aligned}
f_n(\kappa\delta) &= \sum_{j>i} e^{-(j-i)\kappa\delta} = \frac{n-1}{e^{\kappa\delta} - 1} - \frac{1 - e^{-(n-1)\kappa\delta}}{(e^{\kappa\delta} - 1)^2} \\
&= \frac{n}{e^{\kappa\delta} - 1} + o(n). \quad (2.41)
\end{aligned}$$

Therefore,

$$E\left(\frac{T_{12}}{\mu_2'} - \frac{\mu_1' T_{22}}{\mu_2'^2} \right) = -n^{-1}(1 + e^{-\kappa\delta}) + o(n^{-1}). \quad (2.42)$$

The derivations of $E(T_{11}T_{21})$ and $E(T_{21}^2)$ need the following results which can be

obtained from (2.35) to (2.38),

$$n^{-2}\frac{\mu_2^2}{\mu_2'^2}E\left\{-\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu_1)(X_{j-1}-\mu_1)+e^{-\kappa\delta}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}-\mu_1)(X_{j-1}-\mu_1)\right\}$$
$$=-n^{-1}\theta_\alpha(1+e^{-\kappa\delta})+o(n^{-1}), \tag{2.43}$$
$$n^{-2}\frac{\mu_1^2}{\mu_2'^2}E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}^{-1}-\mu_2)(X_{j-1}^{-1}-\mu_2)\right\}$$
$$=n^{-2}\theta_\alpha^2\sum_{i=1}^{n}\sum_{j=1}^{n}\{F(1,1,\theta_\alpha,e^{-|j-i|\kappa\delta})-1\}, \tag{2.44}$$

$$n^{-2}\frac{\mu_1}{\mu_2'^2}E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}^{-1}-\mu_2)\right\}=n^{-1}(1-e^{-\kappa\delta})+o(n^{-1}) \tag{2.45}$$
$$\text{and}\quad n^{-2}\frac{\mu_2}{\mu_2'^2}E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}-\mu)\right\}$$
$$=n^{-2}\theta_\alpha^2(1-e^{\kappa\delta})\left\{S_1(\theta,\delta)+S_2(\theta,\delta)+S_3(\theta,\delta)+S_4(\theta,\delta)\right\}+o(n^{-1}), \tag{2.46}$$

where $S_i(\theta,\delta)$, $i=1,2,3,4$ have been defined in (2.16). We note from (2.39) that

$$T_{11}T_{21}=\mu_2^2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu_1)(X_{j-1}-\mu_1)+\mu_1^2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}^{-1}-\mu_2)(X_{j-1}^{-1}-\mu_2)$$
$$+\quad\mu_1\mu_2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu_1)(X_{j-1}^{-1}-\mu_2)+\mu_1\mu_2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}-\mu_1)(X_{j-1}^{-1}-\mu_2)$$
$$-\quad\mu_2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}-\mu_1)-\mu_1n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}^{-1}-\mu_2)$$

and

$$T_{21}^2=\mu_2^2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}-\mu_1)(X_{j-1}-\mu_1)+\mu_1^2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}^{-1}-\mu_2)(X_{j-1}^{-1}-\mu_2)$$
$$+\quad2\mu_1\mu_2n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{i-1}-\mu_1)(X_{j-1}^{-1}-\mu_2).$$

Applying the results in (2.43) to (2.46), we have

$$E\left(\frac{T_{11}T_{21}}{\mu_2'^2}-\frac{\mu_1'^2T_{21}^2}{\mu_2'^3}\right)=n^{-1}\left[-\theta_\alpha(1+e^{-\kappa\delta})+2\theta_\alpha^2S_1(\theta,\delta)(1-e^{-\kappa\delta})\right.$$
$$\left.+\theta_\alpha^2(1-e^{-\kappa\delta})\left\{\sum_{i=1}^{4}S_i(\theta,\delta)+\theta_\alpha^{-2}\right\}\right]+o(n^{-1}).$$

This together with (2.39) and (2.40) lead to

$$E(\hat{\beta}_1) = \beta_1 \quad - \quad n^{-1}\left[(1+e^{-\kappa\delta})(1-\theta_\alpha) + 2\theta_\alpha^2 S_1(\theta,\delta)(1-e^{\kappa\delta}) \right. \\ \left. +\theta_\alpha^2(1-e^{-\kappa\delta})\left\{\sum_{i=1}^{4} S_i(\theta,\delta) + \theta_\alpha^{-2}\right\}\right] + o(n^{-1}). \tag{2.47}$$

To derive $Var(\hat{\beta}_1)$, we take variance operation on both sides of (2.39) so that

$$Var(\hat{\beta}_1) = \left\{\frac{Var(T_{11})}{\mu_2'^2} + \frac{\mu_1'^2 Var(T_{21})}{\mu_2'^4} - \frac{2\mu_1'}{\mu_2'^3} Cov(T_{11}, T_{21})\right\}\{1+o(1)\}.$$

Then we apply (2.43) to (2.46) to yield

$$Var(\hat{\beta}_1) = n^{-1}(1-e^{-2\kappa\delta}) + o(n^{-1}). \tag{2.48}$$

To establish the bias and variance expansions for $\hat{\kappa}$, we note

$$\hat{\kappa} = -\frac{1}{\delta}\left[\log(\beta_1) + \frac{(\hat{\beta}_1-\beta_1)}{\beta_1} - \frac{(\hat{\beta}_1-\beta_1)^2}{2\beta_1^2} + O_p\{(\hat{\beta}_1-\beta_1)^3\}\right].$$

Applying the delta-method, we have from (2.47) and (2.48)

$$\begin{aligned} E(\hat{\kappa}) &= \kappa - \delta^{-1}E\left[\frac{\hat{\beta}_1-\beta_1}{\beta_1} - \frac{(\hat{\beta}_1-\beta_1)^2}{2\beta_1^2}\right] + o\left\{E(\hat{\beta}_1-\beta_1)^2\right\} \\ &= \kappa + (n\delta)^{-1}B_3(\theta,\delta) + o(n^{-1}) \quad \text{and} \\ Var(\hat{\kappa}) &= (n\delta)^{-1}V_4(\theta,\delta) + o(n^{-1}), \end{aligned}$$

where

$$B_3(\theta,\delta) = (1+e^{-\kappa\delta})(1-\theta_\alpha) + 2\theta_\alpha^2 S_1(\theta,\delta)(1-e^{\kappa\delta}) + (1-e^{-\kappa\delta})\left\{\theta_\alpha^2\sum_{i=1}^{4} S_i(\theta,\delta) - e^{\kappa\delta}\right\}$$

and $V_4(\theta,\delta) = \delta^{-1}(e^{2\kappa\delta}-1)$. These complete proving the first two equations of Theorem 2.3 regarding $\hat{\kappa}$.

Similar with (2.31) in Vasicek case, from (2.15) it can be shown that

$$\hat{\alpha} = \bar{X} + n^{-1}\frac{X_n - X_0}{1-\hat{\beta}_1} = \bar{X} + O_p(n^{-1}). \tag{2.49}$$

Then the second two equations of Theorem 3.1.3 regargding $\hat{\alpha}$ can be established by noting that $E(\bar{X}) = \alpha$ and $Var(\bar{X}) = n^{-1}\theta_\beta^{-2}\theta_\alpha \left\{2(e^{\kappa\delta} - 1)^{-1} + 1\right\}$.

To establish the last two equations of Theorem 2.3, firstly note that

$$\hat{\beta}_3 = n^{-1} \sum_{i=1}^{n} \left[\frac{\{X_i - X_{i-1}\beta_1 - \beta_2(1 - \beta_1)\}^2}{X_{i-1}} \right] + O_p(n^{-1}),$$

which implies $E(\hat{\beta}_3) = \beta_3 + \frac{\sigma^2}{2\kappa(\theta_\alpha - 1)}(1 - e^{-\kappa\delta})^2 + O(n^{-1})$. This result indicates that the piece-wise diffusion approximation used in the pseudo-likelihood estimation introduces a constant bias term in the estimation of β_3. In deriving the variance of $\hat{\beta}_3$, the fourth central moment of a non-central chi-square distribution is encountered. As $c \cdot X_i | X_{i-1}$ follows a noncentral Chi-square distribution random variable $\chi^2_\nu(\lambda)$, the fourth conditional central moment is given by

$$E[\{X_i - X_{i-1}\beta_1 - \beta_2(1 - \beta_1)\}^4 | X_{i-1}] = c^{-4} \left\{12(\nu + 2\lambda)^2 + 48(\nu + 4\lambda)\right\},$$

where λ depends on X_{i-1}. Then it can be shown that

$$Var(\hat{\beta}_3) = n^{-1} \left(2\beta_3^2 - \frac{(1 - e^{-\kappa\delta})^2}{\theta_\alpha - 1} \right) + o(n^{-1}). \tag{2.50}$$

Noting that the expansion given by (2.34) is also valid for CIR case, the proof of Theorem 2.3 is completed by utilizing the established results and taking expectation and variance operations on (2.34).

Proof of Theorem 2.4

Let $\beta_1 = e^{-\kappa\delta}$, $\beta_2 = \alpha$, $\beta_3 = \sigma^2(2\kappa)^{-1}(1 - e^{-2\kappa\delta})$ and $\beta = (\beta_1, \beta_2, \beta_3)$ be the $1 - 1$ mapping from $\theta = (\kappa, \alpha, \sigma^2)$. The $\ell(\theta)$ defined by (2.13) can be regarded as $\ell(\beta)$ after the reparametrization. Then the pseudo-MLE $\hat{\beta}$ is the root of $\frac{\partial^T \ell(\beta)}{\partial \beta} = 0$. It can be shown that $E\hat{\beta}_3 = \beta_3 + b(\theta, \delta) + O(n^{-1})$ where $b(\theta, \delta) = (\theta_\alpha - 1)^{-1}(1 - e^{-2\kappa\delta})^2\sigma^2$ is a bias term which does not converge to 0 unless $\delta \to 0$. Let $\tilde{\beta} = (\beta_1, \beta_2, \beta_3 + b(\theta, \delta))^T$ and applying Taylor's expansion to the pseudo-likelihood score equations, we have

$$0 = \frac{\partial^T \ell(\hat{\beta})}{\partial \beta} = \frac{\partial^T \ell(\tilde{\beta})}{\partial \beta} + \frac{\partial^2 \ell(\tilde{\beta})}{\partial \beta \partial \beta^T}(\hat{\beta} - \tilde{\beta}) + \xi_n,$$

where ξ_n is the remainder term. This implies that

$$\sqrt{n}(\hat{\beta}-\tilde{\beta})=\left\{-\frac{\partial^2\ell(\tilde{\beta})}{\partial\beta\partial\beta^T}\right\}^{-1}\left\{\sqrt{n}\left(\frac{\partial^T\ell(\tilde{\beta})}{\partial\beta}+\xi_n\right)\right\}.$$

Utilizing the central limiting theorem for mixing sequences (Bosq, 1998), it can be shown that

$$\frac{1}{\sqrt{n}}\frac{\partial^T\ell(\tilde{\beta})}{\partial\beta}\xrightarrow{d}N(0,\Sigma^{-1}),$$

where $\Sigma=\begin{pmatrix}1-e^{-2\kappa\delta} & -(1+e^{-\kappa\delta}) & 0\\ -(1+e^{\kappa\delta}) & 2\alpha\theta_\beta^{-1}(1-e^{-\kappa\delta})^{-1} & 0\\ 0 & 0 & Z_2(\theta,\delta)\end{pmatrix}$,

$$\begin{aligned}Z_2(\theta,\delta) &= \frac{1}{4\beta_3^2}\left[1+\frac{1}{1+e^{-\kappa\delta}}\left\{12e^{-2\kappa\delta}+(12\nu+48)c(\theta,\delta)^{-1}\frac{e^{-\kappa\delta}\theta_\beta}{\theta_\alpha-1}+\right.\right.\\ &\left.\left.(3\nu^2+12\nu)c(\theta,\delta)^{-2}\frac{\theta_\beta^2}{(\theta_\alpha-1)(\theta_\alpha-2)}-\frac{2(\theta_\alpha+\theta_\alpha e^{-\kappa\delta}-2e^{-\kappa\delta}}{(1+e^{-\kappa\delta})(\theta_\alpha-1)}\right\}\right],\end{aligned}$$

$c(\theta,\delta)=2\theta_\beta(1-e^{-\kappa\delta})^{-1}$and $\nu=2\theta_\alpha$. As $\left\{-n^{-1}\frac{\partial^2\ell(\tilde{\beta})}{\partial\beta\partial\beta^T}\right\}\xrightarrow{p}\Sigma^{-1}$,

$$\left\{-\frac{\partial^2\ell(\tilde{\beta})}{\partial\beta\partial\beta^T}\right\}^{-1}\left(\sqrt{n}\frac{\partial^T\ell(\tilde{\beta})}{\partial\beta}\right)\xrightarrow{d}N(0,\Sigma).$$

As a result of Theorem 2.3, $\frac{1}{\sqrt{n}}\xi_n\xrightarrow{p}0$. Then by Slutsky's Theorem, $\sqrt{n}(\hat{\beta}-\tilde{\beta})\xrightarrow{d}N(0,\Sigma)$ and Theorem 2.4 holds by transforming back to $\hat{\theta}$ as a function of the asymptotically normal vector $\hat{\beta}$ (Serfling, 1980).

Proof of Theorem 2.5

Let us start the proof of Theorem 2.5 from the proof of Theorem 2.1. By using the same notation, similar with (2.21), we may establish an expansion for $\hat{\beta}_1$ when $\delta\to0$,

$$\hat{\beta}_1 = \frac{\mu_{a_1}+T_2}{\mu_{b_1}+T_1}=\frac{\mu_{a_1}+T_2}{\mu_{b_1}}\left(1-\frac{T_1}{\mu_{b_1}}+\frac{T_1^2}{\mu_{b_1}^2}+\cdots-\frac{T_1^{2k-1}}{(1+\xi_1)^{2k}\mu_{b_1}^{2k-1}}\right) \quad (2.51)$$

for some $|\xi_1|<|T_1/\mu_{b_1}|$. Since $T_1\to0$ almost surely as $n\to\infty$ by the strong ergodic theorem (Theorem 5.5 of Karlin and Taylor (1975)), $\xi_1\to0$ almost surely.

We claim that $E|T_2T_1^{2k-1}| < Mn^{-k}\delta^{-k}$ for some $M < \infty$ and large n. The claim follows from Theorem 4 of Yokoyama (1980) which gave moment bounds for sums of α-mixing random variables. That a diffusion process is α-mixing with the mixing coefficient $\alpha(t) = e^{-\varsigma t}$ for some $\varsigma > 0$ is established in (Genon-catalot et al., 2000) under certain conditions which are satisfied for both the Vasicek and CIR processes. Then, $E|T_2T_1^{2k-1}(1+\xi_1)^{-2k}\mu_{b_1}^{-2k}| = o(n^{-2}\delta)$ if (2.1) (iii) holds.

As a result,

$$E(\hat{\beta}_1) = \frac{\mu_{a_1}}{\mu_{b_1}} + E\left(\eta_1 + \cdots + \eta_{2k-1}\right) + o(n^{-2}\delta^{-1}) \tag{2.52}$$

where $\eta_j = (-1)^{j-1}\mu_{b_1}^{-j}(T_2T_1^{j-1} - \beta_1 T_1^j)$. By rearranging terms in $\sum_{j=1}^{2k-1}\eta_j$ in an increasing order of magnitudes,

$$\sum_{j=1}^{2k-1}\eta_j = A_1 + A_2 + A_3 + A_4 + o_p(n^{-2}\delta^{-1})$$

where A_i, $i = 1, \ldots, 4$ are samely defined as those in (2.22). It is noted that in establishing the above rearrangement, $T_2T_1^{j-1} = T_1^j\{1 + O_p(\delta)\}$ and $\beta_1 = 1 + o(1)$ when $\delta \to 0$. This is the reason that $A_i = O_p(n^{-2}\delta^{-1})$ rather than $O_p(n^{-2}\delta^{-2})$ for $i = 3, 4$. Following the same approach in deriving Theorem 2.1, and expand (2.25) as $\delta \to 0$, it is true that

$$f_n(\kappa\delta) = n\kappa^{-1}\delta^{-1} - \tfrac{1}{2}n - \kappa^{-2}\delta^{-2} + o(\delta^{-2} + n). \tag{2.53}$$

Then

$$E(A_2) = -\frac{4}{n} + \frac{3\kappa\delta}{n} + \frac{3}{n^2\kappa\delta} + o(n^{-2}\delta^{-1} + n^{-1}\delta), \tag{2.54}$$

By obtaining higher order moments from the moment generation function, it can be shown that

$$E(A_3) = \frac{28}{n^2\kappa\delta}\left\{1 + o(1)\right\} \quad \text{and} \quad E(A_4) = -\frac{24}{n^2\kappa\delta}\left\{1 + o(1)\right\}. \tag{2.55}$$

Therefore

$$E(\hat{\beta}_1) = \beta_1 - \frac{4}{n} + \frac{3\kappa\delta}{n} + \frac{7}{n^2\kappa\delta} + o(n^{-2}\delta^{-1} + n^{-1}\delta). \tag{2.56}$$

And similarly from (2.28)

$$Var(\hat{\beta}_1) = n^{-1}(1 - e^{-2\kappa\delta})\left\{1 + o(1)\right\}. \tag{2.57}$$

An expansion of $\hat{\kappa}$ similar to (2.30) can be established as the following

$$\hat{\kappa} = -\frac{1}{\delta}\left\{\log(\beta_1) + \frac{(\hat{\beta}_1 - \beta_1)}{\beta_1} - \frac{(\hat{\beta}_1 - \beta_1)^2}{2\beta_1^2} + O_p\{(n\delta)^{-3/2}\}\right\}. \tag{2.58}$$

Hence, the first two equations of Theorem 2.5 is established by combining (2.56), (2.57) and (2.58).

Noting that by using the $\delta \to 0$ asymptotic, we have $Var(\bar{X}) = (n\delta)^{-1}\sigma\kappa^{-1}$. The results of $\hat{\alpha}$ can be established by utilizing an expansion similar to (2.31),

$$\hat{\alpha} = \bar{X} + \frac{\beta_1}{1 - \beta_1} n^{-1}(X_n - X_0)\{1 + O_p(n^{-1/2}\delta^{1/2})\}. \tag{2.59}$$

Substitute the above results into the expansion (2.34) for $\hat{\sigma}^2$, the proof of Theorem 2.5 is completed.

Proof of Theorem 2.6

The proof of Theorem 2.6 is almost the same as that of Theorem 2.2. Applying Taylor's expansion to $\ell(\beta)$, we have

$$R_{1n,\delta}(\hat{\beta} - \beta) = \left\{-\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\}^{-1}\left\{R_{1n,\delta}\left(\frac{\partial^T\ell(\beta)}{\partial\beta} + \xi_{n,\delta}\right)\right\},$$

where $R_{1n,\delta} = \text{diag}(n^{1/2}\delta^{-1/2}, n^{1/2}\delta^{1/2}, n^{1/2}\delta^{-1})$ and $\xi_{n,\delta}$ is the remainder term. Using the results in the proof of Theorem 2.5 and Theorem 2.1, it can be shown that

$$(R_{1n,\delta})^{-1}\frac{\partial^T\ell(\beta)}{\partial\beta} \xrightarrow{d} N(0, \Sigma^{-1}),$$

where $\Sigma = diag\{2\kappa, \sigma^2\kappa^{-2}, 2\sigma^4\}$. Noting that $\left\{-(R_{1n,\delta})^{-2}\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\} \xrightarrow{p} \Sigma^{-1}$, therefore

$$\left\{-\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T}\right\}^{-1}\left(R_{1n,\delta}\frac{\partial^T\ell(\beta)}{\partial\beta}\right) \xrightarrow{d} N(0,\Sigma).$$

From Theorem 2.5, $(R_{1n,\delta})^{-1}\xi_{n,\delta} \xrightarrow{p} 0$. Then by Slutsky's Theorem, $R_{1n,\delta}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\Sigma)$. And Theorem 2.6 follows by transforming back to $\hat{\theta}$.

Proof of Theorem 2.7

Since $\delta \to 0$ as specified in (2.1), the bias and variance in Theorem 2.3 can be largely simplified.

Note that $1-e^{\kappa\delta} = \kappa\delta + O(\delta^2)$, and (2.41) implies that $f_n(\kappa\delta) = n\kappa^{-1}\delta^{-1} - \frac{1}{2}n - \kappa^{-2}\delta^{-2} + o(\delta^{-2}+n)$. Then

$$E\left(\frac{T_{12}}{\mu_2'} - \frac{\mu_1'T_{22}}{\mu_2'^2}\right) = -2n^{-1}\{1+o(1)\}.$$

Similarly we can obtain the following results which are parallel to (2.43) to (2.46) in the proof of Theorem 2.3:

$$n^{-2}\frac{\mu_2^2}{\mu_2'^2}E\left\{-\sum_{i=1}^n\sum_{j=1}^n(X_i-\mu_1)(X_{j-1}-\mu_1) + e^{-\kappa\delta}\sum_{i=1}^n\sum_{j=1}^n(X_{i-1}-\mu_1)(X_{j-1}-\mu_1)\right\}$$
$$= -2n^{-1}\theta_\alpha + O(n^{-1}\delta), \tag{2.60}$$
$$n^{-2}\frac{\mu_1^2}{\mu_2'^2}E\left\{\sum_{i=1}^n\sum_{j=1}^n(X_{i-1}^{-1}-\mu_2)(X_{j-1}^{-1}-\mu_2)\right\} = n^{-1}\frac{2\theta_\alpha^2}{(\theta_\alpha-1)\kappa\delta}\{1+o(1)\}, \tag{2.61}$$
$$n^{-2}\frac{\mu_1}{\mu_2'^2}E\left\{\sum_{i=1}^n\sum_{j=1}^n(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}^{-1}-\mu_2)\right\} = o(n^{-1}) \quad \text{and} \tag{2.62}$$
$$n^{-2}\frac{\mu_2}{\mu_2'^2}E\left\{\sum_{i=1}^n\sum_{j=1}^n(X_iX_{i-1}^{-1}-\mu_3)(X_{j-1}-\mu)\right\} = o(n^{-1}). \tag{2.63}$$

Then following the same steps in the proof of Theorem 2.3, (2.60) to (2.63) imply that

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1 - n^{-1}\left(4+\frac{2}{\theta_\alpha-1}\right)\{1+o(1)\} \quad \text{and} \\ Var(\hat{\beta}_1) &= n^{-1}\delta 2\kappa + o(n^{-1}\delta). \end{aligned}$$

These readily imply that $E(\hat{\kappa}) = \kappa + \left(4 + \frac{2}{\theta_\alpha - 1}\right) T^{-1} + o(T^{-1})$ and $Var(\hat{\kappa}) = 2\kappa T^{-1} + o(T^{-1})$. The rest proof of Theorem 2.7 are replicated applications of Taylor's expansion and results from (2.35) to (2.38) and (2.60) to (2.63), and also by the established expansion of $\hat{\alpha}$ and $\hat{\sigma}^2$ in (2.49) and (2.34).

Proof of Theorem 2.8

The proof of Theorem 2.8 is similar to the proof of Theorem 2.4. Applying Taylor's expansion to $\ell(\beta)$, we have

$$R_{1n,\delta}(\hat{\beta} - \beta) = \left\{ -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T} \right\}^{-1} \left\{ R_{1n,\delta} \left(\frac{\partial^T \ell(\beta)}{\partial\beta} + \xi_{n,\delta} \right) \right\},$$

where $R_{1n,\delta} = \text{diag}(n^{1/2}\delta^{-1/2}, n^{1/2}\delta^{1/2}, n^{1/2}\delta^{-1})$ and $\xi_{n,\delta}$ is the remainder term. Using the results in the proof of Theorem 2.7, it can be shown that

$$(R_{1n,\delta})^{-1} \frac{\partial^T \ell(\beta)}{\partial\beta} \xrightarrow{d} N(0, \Sigma^{-1}),$$

where $\Sigma = \begin{pmatrix} 2\kappa & -2 & 0 \\ -2 & 2\alpha\theta_\beta^{-1}\kappa^{-1} & 0 \\ 0 & 0 & \sigma^4(2 - \frac{1}{\theta_\alpha - 1}) \end{pmatrix}$. Noting that $\left\{ -(R_{1n,\delta})^{-2} \frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T} \right\} \xrightarrow{p} \Sigma^{-1}$, therefore

$$\left\{ -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T} \right\}^{-1} \left(R_{1n,\delta} \frac{\partial^T \ell(\beta)}{\partial\beta} \right) \xrightarrow{d} N(0, \Sigma).$$

From Theorem 2.7, $(R_{1n,\delta})^{-1}\xi_{n,\delta} \xrightarrow{p} 0$. Then by Slutsky's Theorem, $R_{1n,\delta}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$. And Theorem 2.8 follows by transforming back to $\hat{\theta}$.

Proof of Theorem 2.9

The proof of Theorem 2.9 needs the following lemma.

Lemma 1 *Let $\hat{\theta}_n$ be an estimator of θ based on n observations, $b_n(\theta) = E(\hat{\theta}_n) - \theta$ and*

(i) For some integer $N \geq 2$, $E||\hat{\theta}_n - \theta||^N = O(\eta_{n,N})$ where $\eta_{n,N} \to 0$ as $n \to \infty$.

(ii) For some $K \geq 1$, $E\{\phi_n(\hat{\theta})\}^K = O(\xi_{n,K})$ for a sequence of constants $\{\xi_{n,K}\}_{n\geq 1}$. Then, $E\{\phi_n^k(\hat{\theta}_n)\} - E\{\phi_{nr}^k(\hat{\theta}_n)\} = O(\eta_{n,N}^{r/N} + \xi_{n,K}^{k/K}\eta_{n,N}^{(K-k)/K})$.

Proof: It can be obtained by modifying the proof of Theorem A.2 of Sargan (1976). Noticeably we use $\eta_{n,N}$ and $\xi_{n,K}$ to replace T^{-rR} and T^{λ} respectively in Sargan (1976).

Proof of Theorem 2.9: Recall $\hat{\theta}_B = \hat{\theta} - (\bar{\hat{\theta}}^* - \hat{\theta})$ where $\bar{\hat{\theta}}^* = N_B^{-1}\sum_{i=1}^n \hat{\theta}_i^*$ and N_B is the replication number of bootstrap resamples. Let χ_n be the σ-field generated by $X_1, \ldots, X_n$. As the bootstrap generates the resamples for the parametric diffusion process, where $\hat{\theta}^*$ are estimations based on the resampled path in the same way as $\hat{\theta}$ based on the original sample, we have

$$E(\hat{\theta}^*|\chi_n) = b_n(\hat{\theta}) \quad \text{and} \quad Var(\hat{\theta}^*|\chi_n) = v_n(\hat{\theta}).$$

First consider the bias of the bootstrap estimator $\hat{\theta}_B$ and note that

$$E(\hat{\theta}_B) \;=\; E\left\{E(\hat{\theta}_B|\chi_n)\right\} = E\left[2\hat{\theta}_n - \left\{\hat{\theta}_n + b_n(\hat{\theta})\right\}\right] = \theta + b_n(\theta) - E\{b_n(\hat{\theta})\}.$$

We need to show

$$E\{b_{nl}(\hat{\theta})\} - b_{nl}(\theta) = o\{b_{nl}(\theta)\}. \tag{2.64}$$

Choose $\phi_n(x) = b_{nl}(x)$, $r = 1$, $N = 2$, $k = 1$, $K = 2$, $\eta_{n,N} = O(\nu_{nl})$ and $\xi_{n,K} = 1$ in Lemma 1. Then

$$E\{b_{nl}^{(0)}(\hat{\theta})\} - b_{nl}^{(0)}(\theta) = O\{(\nu_{nl} + b_{nl}^2)^{1/2}\} \tag{2.65}$$

which readily leads to (2.64) and the first conclusion of the theorem.

Applying the Lemma in a similar fashion, we have

$$E\{b_{nl}^2(\hat{\theta})\} = b_{nl}^2(\theta) + o(\beta_{nl}^2), \tag{2.66}$$

$$E\{v_{nl}(\hat{\theta})\} = v_{nl}(\theta) + o\{v_{nl}(\theta)\} \tag{2.67}$$

Let us now consider the variance of $\hat{\theta}_B$. Note that

$$\begin{aligned} Var(\hat{\theta}_B) &= Var\left\{E(\hat{\theta}_B|\chi_n)\right\} + E\left\{Var(\hat{\theta}_B|\chi_n)\right\} \\ &= Var\left\{\hat{\theta} - b_n(\hat{\theta})\right\} + E\left\{\frac{1}{N_B}Var(\hat{\theta}^{*,1})\right\} \\ &= Var\left\{\hat{\theta} - b_n(\hat{\theta})\right\} + \frac{1}{N_B}E\left\{v_n(\hat{\theta})\right\} \end{aligned}$$

From (2.67) and by choosing N_B large enough, $N_B^{-1}E\left\{v_n(\hat{\theta})\right\} = o\{v_n(\theta)\}$. Note that (2.64) and (2.66) mean that

$$Var\{b_{nl}(\hat{\theta})\} = Eb_{nl}^2(\hat{\theta}) - E^2\{b_{nl}(\theta)\} = O\{b_{nl}^2(\theta)\} + o\{v_{nl}(\theta)\} = o\{v_{nl}(\theta)\}.$$

This and the Cauthy-Schwartz inequality lead to $|Cov\{\hat{\theta}_{nl}, b_{nl}(\hat{\theta})\}| = o\{v_{nl}(\theta)\}$. Hence $Var\left\{\hat{\theta}_{nl} - b_{nl}(\hat{\theta})\right\} = Var(\hat{\theta}_{nl}) + o\{v_{nl}(\theta)\}$. This establishes the second part of the theorem.

CHAPTER 3. Nonparametric Estimation of Enumeration Functions in Census

This Chapter elaborates on the estimations of enumeration functions in the Census, which is a key quantity resulting in population size estimation.

3.1 Overview

Dual system capture-recapture surveys are essential components of the studies in evaluating the census count, for instance those have been conducted in 1990 and 2000 in conjunction with the last two decennial US Census Accuracy and Coverage Evaluation (ACE). Its main objective is to obtain information on the census net coverage and enumeration errors for both the whole population and subpopulation groups, for instance those defined by race, age, geographic regions, social-economic and other demographical variables. In the US Census, the dual system surveys consist of two components. The first survey verifies the census enumerations on a selected sample block clusters of the census; and the data collected are called the E-sample. The aim of the E-sample is to measure erroneous enumerations in the census, which are invalid records. The second survey is an independent survey of the first one, conducted soon after the census, basically on the same sample block clusters covered by the E-sample. The data collected are called the P-sample. The purpose of the P-sample is to identify "matches"(recaptures) to census records; so as to estimate the probability of enumeration by the E-sample. Comprehensive introductions of the US Census and discussions are in Hogan (1993,

2000, 2000), Haberman et al. (1998) and Bell (1999), as well as details and issues with the dual system surveys for the US Census. The methods of estimating the population size are introduced in Wolter (1986), Pollock (1991) and Chao and Tsay (1998) on capture-recapture based approaches.

It is a well known fact that different individuals may have sharply different probability to be counted in the Census (Hogan, 1993 and 2000). A group of variables termed ROAST is known to contribute to much of the heterogeneity in the enumeration. Here ROAST stands for Race/(Hispanic) Origin, Age, Sex, and (housing) Tenure. Other variables may contribute to the heterogeneity as well, for instance geographical region, the tract level mail return rate and the census local office effect.

In general, let $X = (X_1, \cdots, X_d)$ be a vector of covariates that influences the enumerations of individuals, and $I_{i\in\mathcal{E}}$ be a binary indicator that takes value 1 for enumeration and 0 otherwise for an individual i in the population. The enumeration probability of an individual with covariate $X_i = x$ in the E-sample is

$$p(x) = P(I_{i\in\mathcal{E}} = 1|X_i = x). \tag{3.1}$$

It is apparent that without the P-sample, only individuals with $I_{i\in\mathcal{E}} = 1$ are observed, which are insufficient for estimating of $p(x)$. The P-sample makes estimation of $p(x)$ feasible by providing enumerations with both $I_{i\in\mathcal{E}} = 1$ and 0. Here, in the P-sample $I_{i\in\mathcal{E}} = 1$ if it is a match to a census record within the E-sample (enumerated in the Census). In US Census practice, a P-sample person is matched to the corresponding E-sample surrounding blocks(the search area) in the census records. And $I_{i\in\mathcal{E}} = 0$ if a match cannot be made between the P-sample and the census within the search area. This is the so-called one way approach, i.e, matching the cases in the P-sample to all the E-sample records. As a result, $p(x)$ is the enumeration probability function of the E-sample. This Chapter focuses on the estimation of $p(x)$ based on the E- and P-samples.

Figure 3.1 displays the kernel estimates (defined in Sections 2 and 3) of the enumera-

tion probability as (i) an overall function of age, (ii) a function of age whereas Region and the other ROAST variables are set at (Midwest, Hispanic, female, owner), (Northeast, White, male, owner) and (Midwest, Non-Hispanic Black, male, renter) respectively. The estimates are based on the 10% research data of ACE revision II (US Census Bureau, 2004) with a proposed kernel based imputations for missing values. The estimated $p(x)$ indicates strong heterogeneity with respect to the age, region and other ROAST variables. In particular, each plot displays a V-shape for the enumeration probability within the age range [18, 29], an age interval that is known to experience volatile enumeration. However, the detail aspects of the V-shape vary substantially among (b), (c) and (d) of Figure 3.1 when region and the other ROAST variables change values.

Post-stratification has been employed in the US Census to counter heterogeneous enumerations by sub-dividing(stratifying) the sample space of the covariates into post-strata. Although it can reduce some of the heterogeneity, still substantial amount of heterogeneity remains as illustrated in Figure 3.1. One limitation of the post-stratification is that continuously-valued covariates like age are grouped into discrete categories, implying that $p(x)$ is piecewise constant with respect to the age strata. However, as revealed by Figure 3.1, this may not be the case. This was noted in Hogan (1992) who outlined problems with the age post-strata and certain heterogeneous effects that had been unaccounted for. Any unaccounted heterogeneity may result in the so-called "correlation bias" (Wolter, 1986; Chen and Lloyd, 2000) in population size estimation. More on correlation bias is discussed in Chapter 4. The other limitation with the post-stratification is that some strata can have small sample sizes which inflates the variance in both $p(x)$ and the population size estimation. To control the variance, several small size strata are usually combined. However, doing so generally counterparts the effort to reduce the heterogeneity.

The aim of the Chapter is to propose a nonparametric estimation of $p(x)$ for the dual system estimation. The nonparametric estimation is made through kernel regression

estimators to the probabilities of enumeration. The kernel estimators effectively produce a local stratum around the value of interest, say x. The size of the local stratum shrinks as the number of observations increases. This leads to the removal of the "correlation bias". At the same time, the number of observations contained in the local stratum is managed to be increasing as the sample size increases, which controls/reduces the variances. The proposed estimators can accommodate categorical covariates which suits the human census as large number of covariates are categorical variables. The local strata constructed with respect to the categorical covariates combines data within a ring of neighboring strata, which leads to more efficiently utilizing of information as compared with a post-stratification based estimation which only utilizes information within each stratum.

Like many statistical applications, the census records encounters missing values which are inevitable in such a large scale data collection. For the dual system surveys, one important type of missing values arises when matching a P-sample individual with a census record. In US Census, A portion of enumerations in the P-sample can be neither match nor non-match to a census record, resulting in unresolved matches and hence missing values. The percentage of unresolved matches was 3.7% in the P-sample and the unresolved correct enumeration status in the E-sample was 6.6% for the 10% ACE revision II research data files (missing values in the entire US Census P- and E- samples are around 1% and 3% respectively). This is high comparing with the overall estimated level of undercounts in the 2000 Census (US Census Bureau, 2004). Therefore, fully utilizing missing data information is necessary in order to improve the dual system estimation.

In this Chapter, we will study a general nonparametric regression estimation that accommodates missing values. An imputation based estimator is proposed which is shown to be more efficient than an estimator that ignores missing observations. The benefits of the imputation and kernel smoothing are quantified theoretically and confirmed by

simulation studies.

The Chapter is organized as follows. Section 3.2 outlines nonparametric estimation of $p(x)$ in the dual system surveys. And the issues of erroneous enumerations and missing values are presented in Section 3.3, as well as the nonparametric kernel estimators of $p(x)$ and the correct enumeration probability function. A theoretical analysis on the nonparametric regression in the presence of missing values is carried out in Section 3.4. Section 3.5 analyzes the 10% research data of the ACE Revision II, including an empirical goodness-of-fit test for the model of post-stratification scheme used in the US Census estimation. Some simulation results based on models that mimic the US Census ACE data are reported in Section 3.6. All the technical proofs are deferred to Section 3.7.

3.2 Nonparametric Estimation of Enumeration Functions

Let U be the collection of the whole population of size N, $\mathcal{E}$ and $\mathcal{P}$ be the sets of individuals enumerated by the E and P-samples respectively. Let $I_{i\in\mathcal{S}}$ be the indicator of the individual $i \in U$ being included in the set $\mathcal{S}$, then $I_{i\in\mathcal{E}}$ and $I_{i\in\mathcal{P}}$ are two independent binary random variables due to the independence between the E- and P-samples. We denote the available covariate associated with the individual i by X_i and the support of X_i is $\mathcal{X}$. In this dissertation, we will assume that the individuals in the population U are independent and from some super population distribution, whose covariate X follows some probability density function $f(x)$. Assume that the enumeration probability functions of the E- and P-samples are explained by the available covariate, i.e. $P(I_{i\in\mathcal{E}} = 1|X_i = x) = p(x)$ and $P(I_{i\in\mathcal{P}} = 1|X_i = x) = g(x)$ for some $p(\cdot)$ and $g(\cdot)$ mapping $\mathcal{X}$ to $[0, 1]$.

As $E(I_{i\in\mathcal{E}}|X_i = x) = p(x)$ and the enumerations of the E- and P- samples are independent, the E-sample enumeration probability $p(x)$ can be estimated based on

$\{(X_i, I_{i\in\mathcal{E}})\}_{i\in\mathcal{P}}$ in the P-sample via a binary parametric or nonparametric regression.

If there is no erroneous enumerations and missing values, and if $\hat{p}(x)$ is a consistent estimator of $p(x)$, an estimator for the population size N of Horvitz-Thompson type is

$$\hat{N} = \sum_{i\in\mathcal{E}} \frac{1}{\hat{p}(X_i)}. \tag{3.2}$$

Parametric approach for modeling and estimating $p(x)$ has been proposed in the dual system estimation, for instance those studied in Pollock (1976, 1991), the logistic model of Huggins (1989), Alho (1990) and Alho et al. (1993) and the post-stratification being used by the US Census Bureau (US Census Bureau, 2004). However, if the parametric model is mis-specified, which is possible in practice, it may produce biased estimation. It is quite challenging to specify reasonable parametric models for the enumeration probability function $p(x)$. Indeed, Figures 3.1 and 3.2 show diverse forms for the two functions with respect to the age and other categorical variables in ROAST and region, which makes proposing a reasonable parametric model not an easy task. Our focus in this Chapter is to demonstrate that a nonparametric approach based on kernel estimators of the enumeration probability function $p(x)$ can be used as an alternative in the dual system estimation.

An aspect that encourages the proposed nonparametric estimation is that there is a good amount of data collected in both the E- and P-samples. The 10% ACE Research Data files have about 70,000 records each. With this amount of data, we can just let the data to speak for themselves. Hence the need to specify a parametric model for $p(x)$ is reduced.

The covariate X in the US Census is a combination of continuous and categorical variables. As early mentioned, the age is an important variable in both $p(x)$. Although rounding to the nearest integer makes it ordered categorical, we will treat it as a continuous variable as smoothing for the ordered categorical is the same as continuous variables (Simonnoff, 1995). Another example of continuous variable is the mail return rate at

the tract (consists of census blocks). Most covariates encountered in the census are unordered categorical. Without loss of generality, we write $X_i = (X_i^c, X_i^u)$ where X_i^c is a d_c-dimensional continuous covariate and X_i^u is a d_u-dimensional unordered categorical covariate with $d_c + d_u = d$. Here X_i^c may include the age and the tract-level mail return rate; and X_i^u can contain the other categorical for instance region and ROAST variables.

To smooth the continuous covariates, we employ a d_c-dimensional kernel K which is a radially symmetric probability density function in R^{d_c}. We define $K_h(x) = h^{-d_c}K(x/h)$. Here h is the smoothing bandwidth than controls the amount of smoothness of the kernel estimate. The conventional kernel estimation is for continuous variables based on a kernel like K; see Härdle (1990) and Fan and Gijbels (1996) for comprehensive discussions. Without loss of generality, we consider the product kernel in our studies, i.e. $K(u) = \prod_{i=1}^{d_c} K_1(u_i)$, where $K_1(\cdot)$ is some symmetric univariate density function.

To smooth unordered categorical covariates, we employ the discrete kernel originally proposed by Aitchison and Aitken (1976); see also Hall (1981), Racine and Li (2004) and Hall et al. (2004) for recent studies. Smoothing categorical variables is designed to utilize data information in the neighboring strata to improve estimation efficiency as they share similar characteristics and information with the stratum where the estimation is carried out. This is ideally suited for the Census as small size post-strata are commonly encountered in the existing post-stratification.

Suppose X_{ij}^u, the j-th component of X_i^u, takes c_j discrete values in $\{0, 1, ..., c_j - 1\}$. The bandwidth for smoothing X_{ij}^u is λ_j and the kernel weight at x_j^u is

$$\lambda_j I(X_{ij}^u = x_j^u) + \frac{1 - \lambda_j}{c_j - 1} I(X_{ij}^u \neq x_j^u)$$

where $I(\cdot)$ is the indicator function. The bandwidth λ_j takes values within $[c_j^{-1}, 1]$. Assigning $\lambda_j = c_j^{-1}$ leads to a uniform weight irrespective to the difference between X_{ij}^u and x_j^u, whereas $\lambda_j = 1$ gives a kernel weight of 1 if $X_{ij}^u = x_j^u$ and zero otherwise which coincides with the standard frequency weight. The other λ_j values between c_j^{-1} and 1

offer a range of choices for information combining for efficiency improvement.

The kernel used to smooth the entire categorical component $X_i^u = (X_{i1}^u, \cdots, X_{id_u}^u)$ at $x^u = (x_1^u, \cdots, x_{d_u}^u)$ is

$$L(x^u, X_i^u; \vec{\lambda}) = \prod_{j=1}^{d_u} \{\lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \neq x_j^u)\}, \tag{3.3}$$

where $\vec{\lambda} = (\lambda_1, \cdots, \lambda_{d_u})$ is the bandwidth vector. The overall kernel weight drawn from $X_i = (X_i^c, X_i^u)$ for local estimation at $x = (x^c, x^u)$ is $K_h(x^c - X_i^c)L(x^u, X_i^u; \vec{\lambda})$.

When some categorical variables in X_i have natural order, for instance $X_i^o = (X_{i1}^o, \ldots, X_{im}^o)$ where $X_{ik}^o \in \{0, \ldots, c_k - 1\}$. Then, the kernel used to smooth the k-th component is

$$H_k(x_k^o, X_{ik}^o, \lambda_k) = \binom{c_k - 1}{v_k} \lambda_k^{c_k - v_k - 1} (1 - \lambda_k)^{v_k}, \text{where } v_k = |X_k^o - X_{ik}^o|,$$

and $H(x^o, X_i^o, \vec{\lambda}) = \prod_{k=1}^m H_k(x_k^o, X_{ik}^o, \lambda_k)$ is the kernel for smoothing the entire ordered categorical covariates. To simplify our analysis and without loss if generality, we will only consider continuous and unordered categorical variables in this Chapter.

Let

$$\mathcal{K}_{h,\vec{\lambda}}(x, y) = K_h(x^c - y^c)L(x^u, y^u, \vec{\lambda}).$$

The kernel estimator of $p(x)$ in the ideal case is

$$\hat{p}_0(x) = \frac{\sum_{i \in \mathcal{P}} \mathcal{K}_{h,\lambda}(x, X_i) I_{i \in \mathcal{E}}}{\sum_{i \in \mathcal{P}} \mathcal{K}_{h,\lambda}(x, X_i)} = \frac{\sum_{i \in U} \mathcal{K}_{h,\lambda}(x, X_i) I_{i \in \mathcal{E}} I_{i \in \mathcal{P}}}{\sum_{i \in U} \mathcal{K}_{h,\lambda}(x, X_i) I_{i \in \mathcal{P}}}. \tag{3.4}$$

This is a Nadaraya-Watson type estimator by carrying out weighted average of the binary responses $I_{i \in \mathcal{E}}$ locally where the "closer" a X_i is to x, the higher the kernel weight. This local averaging is the key that allows us to estimate $p(x)$ without assuming a parametric model.

Let $\mathcal{C}_{x^u} = \{y^u : \sum_{j=1}^{d_u} I(x_j^u = y_j^u) = 1\}$ be the nearest strata of x^u which differs from x^u only in one component. For a $y^u \in \mathcal{C}_{x^u}$, define

$$\alpha(x^u, y^u) = \sum_{k=1}^{d_u} c_k I(x_k^u = y_k^u) \quad \text{and} \quad \beta_\lambda(x^u, y^u) = \sum_{k=1}^{d_u} \lambda_k I(x_k^u = y_k^u)$$

where c_k is the levels of X_{ik}^u. The bias induced by smoothing the discrete variables is quantified by

$$b_{0,u}(x;\vec{\lambda}) = \sum_{y^u \in \mathcal{C}_x^u} \left(\frac{1-\beta_\lambda(x^u,y^u)}{\alpha(x^u,y^u)-1} \left[\frac{g(x^c,y^u)f(x^c,y^u)}{g(x)f(x)} \{p(x^c,y^u)-p(x)\} \right] \right).$$

And let

$$b_{0,c}(x;h) = \frac{1}{2}h^2\sigma_K^2 \frac{\mathrm{tr}[\nabla^2\{p(x)g(x)f(x)\}] - p(x)\mathrm{tr}[\nabla^2\{g(x)f(x)\}]}{g(x)f(x)}$$

be the bias from smoothing the continuous variables, where ∇ is the differential operator with respect to x^c, tr is the the matrix trace and $\sigma_K^2 = \int u^2 K(u)du$. The following theorem shows the mean square consistency of $\hat{p}_0(x)$.

Theorem 3.1 *Under the regularity conditions given in Section 3.7, let* $\lambda = \min_{1\le l\le d_u}(\lambda_l)$ *and* $\lambda^{(a)} = \prod_{k=1}^{d_u} \lambda_k^a$,

$$\begin{aligned} E\{\hat{p}_0(x)\} &= p(x) + b_{0,u}(x,\vec{\lambda}) + b_{0,c}(x,h) + O\{h^2(1-\lambda)\}, \\ var\{\hat{p}_0(x)\} &= \frac{\lambda^{(2)}R(K)}{Nh^{d_c}} \frac{p(x)\{1-p(x)\}}{g(x)f(x)} + o(N^{-1}h^{-d_c}), \end{aligned}$$

as $N\to\infty$, $h\to 0$, $\lambda\to 1$ *and* $Nh^{d_c}\to\infty$, *where* $R(K) = \int K^2(t)dt$.

Note that $(1-\vec{\lambda}) \to 0$ and $h \to 0$ as $N\to\infty$, the bias of $\hat{p}_0(x)$ converges to 0 asymptotically. In finite sample implementation of the nonparametric kernel methods, the choice of smoothing bandwidth is an important practical issue. The bandwidth selection through minimizing the Cross-Validation (CV) function is discussed in the case study in Section 3.5.

3.3 Erroneous Enumerations and Missing Values

Given a consistent estimator of $p(x)$, the population size N can be estimated by (3.2). However, an estimator of form (3.2) may not be applied to the census dual system estimation of the population size due to (i) erroneous enumerations (EEs) and (ii) missing values. The EEs and missing values will be discussed in this Section.

3.3.1 Erroneous Enumerations

Erroneous Enumerations (EEs) are invalid records in the census and typically lead to over-estimation of the population size. There are two main sources of EEs as described in Hogan (1993) and Haberman et al. (1998). One is caused by persons enumerated that should not have been, which includes duplicated or fictitious records, and people born after or died before the Census. Another source of EEs is due to enumerations at wrong locations, for instance those enumerations that should be included in the Census but not at the location they were counted. In the US Census ACE, studies on the E-sample were carried out to identify the EEs which consist of computer identification, clerical investigation and in field follow-up work (US Census Bureau, 2004).

In the following parts of the dissertation, let $\tilde{\mathcal{E}} \subset \mathcal{E}$ be the set of correct enumerations and $I_{i\in\tilde{\mathcal{E}}}$ be the indicator of the i^{th} individual in the E-sample being a correct enumeration. Correspondingly, let U be the hypothetical collection of population where $\mathcal{E}$ is sampled from and $\tilde{U}$ be the correctly enumerated part of U. Therefore, the size of $\tilde{U}$ is the true population size. We assume that $P(I_{i\in\tilde{\mathcal{E}}} = 1 | I_{i\in\mathcal{E}} = 1, X_i = x) = e(x)$ and the expected size of $\tilde{U}$ is $N\left\{\int_{\mathcal{X}} e(x)f(x)dx\right\}$.

Similar to (3.4), a nonparametric estimator of $e(x)$ is given by

$$\hat{e}_0(x) = \frac{\sum_{i\in\mathcal{E}} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\tilde{\mathcal{E}}}}{\sum_{i\in\mathcal{E}} \mathcal{K}_{h,\vec{\lambda}}(x, X_i)} = \frac{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\tilde{\mathcal{E}}} I_{i\in\mathcal{E}}}{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{E}}}. \tag{3.5}$$

The consistency of $\hat{e}_0(x)$ is in exactly the same form of that given by Theorem 3.1.

Given consistent estimators $\hat{e}(x)$ and $\hat{p}(x)$, a general estimator for the population size, that taking into account of both EEs and missing values, is

$$\hat{N} = \sum_{i\in\mathcal{E}} \frac{\hat{e}(X_i)}{\hat{p}(X_i)}. \tag{3.6}$$

Next, we will propose nonparametric kernel estimators for both $p(x)$ and $e(x)$ that account for missing values. A study on the above estimator $\hat{N}$ with the proposed kernel estimators for $p(x)$ and $e(x)$ will be considered in Chapter 4.

3.3.2 Missing Values and Estimation of Enumeration Functions

As illustrated in the introduction, both the matching status ($I_{i\in\mathcal{E}}$) in the P-sample and correct enumeration status ($I_{i\in\tilde{\mathcal{E}}}$) in the E-sample may be missing when the enumeration status are unresolved. Let δ_i and η_i be the missing indicators of $I_{i\in\mathcal{E}}$ and $I_{i\in\tilde{\mathcal{E}}}$ respectively, namely $\delta_i/\eta_i = 0(1)$ for unresolved (resolved) $I_{i\in\mathcal{E}}/I_{i\in\tilde{\mathcal{E}}}$. In this section, we will treat only missing values for un-resolved enumeration status. The missingness in X_i is beyond the scope of this dissertation.

Missing at random (MAR) (Rosenbaum and Rubin, 1983) is an important notion in missing data analysis. Under a standard circumstance, MAR means that conditioning on the covariate X_i the missingness of $I_{i\in\mathcal{E}}$ in the P-sample is independent of $I_{i\in\mathcal{E}}$. In other words, $\delta_i/\eta_i/$ and $I_{i\in\mathcal{E}}/I_{i\in\tilde{\mathcal{E}}}$) are conditionally independent given X_i, namely

$$\begin{aligned} P(\eta_i = 1|I_{i\in\tilde{\mathcal{E}}}, X_i) &= P(\eta_i = 1|X_i) =: w_e(X_i), \\ P(\delta_i = 1|I_{i\in\mathcal{E}}, I_{i\in\mathcal{P}}, X_i) &= P(\delta_i = 1|X_i) =: w_p(X_i). \end{aligned} \tag{3.7}$$

Here, w_p/w_e is called the missing propensity of $I_{i\in\mathcal{E}}/I_{i\in\tilde{\mathcal{E}}}$. MAR is a weaker assumption than missing completely at random since the later implies that the propensity $w_p(x)$ is a constant function.

Figure 3.3 displays the kernel estimates for the missing propensity score $w_e(x)$, which is as interesting as Figures 3.1 and 3.2. For instance, the White Male owners had very small chance of being missing as the estimate of $w_e(x)$ were very close to 1. In contrast, the Hispanic Female owners endured more missingness while the Black Male renters experienced the highest missing values among the three.

The MAR in the form of (3.7) may not be realistic for the census. Analyses on the census data indicate that MAR in the form of (3.7) may not be valid as there are extra characteristics in additional to those listed in X_i that contributes to the missingness. For instance, the before-follow-up coding status has been shown to be influential in Belin et al. (1993). To reflect this reality of the census, we assume that in additional

to X_i, there are extra covariate Z_i that contributes to the missingness of $I_{i\in\mathcal{E}}$ or $I_{i\in\tilde{\mathcal{E}}}$. Specifically, we assume that $I_{i\in\mathcal{E}}$ and $I_{i\in\tilde{\mathcal{E}}}$ are missing at random given (X_i, Z_i), namely

$$P(\delta_i = 1 \mid I_{i\in\mathcal{E}}, X_i, Z_i) = P(\delta_i = 1 \mid X_i, Z_i) =: w_p(X_i, Z_i) \quad \text{and}$$
$$P(\eta_i = 1 \mid I_{i\in\tilde{\mathcal{E}}}, X_i, Z_i) = P(\eta_i = 1 \mid X_i, Z_i) =: w_e(X_i, Z_i)$$

where $w_p(x, z)$ and $w_e(x, z)$ are the unknown missing propensity scores.

At the same time, we assume

$$\begin{aligned} P(I_{i\in\mathcal{E}} = 1|X_i = x, Z_i = z) &= p(x), \quad P(I_{i\in\mathcal{P}} = 1|X_i = x, Z_i = z) = g(x) \text{ and} \\ P(I_{i\in\tilde{\mathcal{E}}} = 1|X_i = x, Z_i = z) &= e(x). \end{aligned} \tag{3.8}$$

which means that the extra covariates Z that contributes to the pattern of missingness do not have any predicting power on the enumeration or correct enumeration of individuals. If part of Z possesses such power, the part should be included as part of X.

We will concentrate on the estimation of $p(x)$, as that for $e(x)$ can be readily extended. The first estimator of $p(x)$ is a version of (3.4) but ignores data with missing values

$$\hat{p}_1(x) = \frac{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{E}} I_{i\in\mathcal{P}} \delta_i}{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} \delta_i}. \tag{3.9}$$

We call it the complete case estimator. Despite there is a selection bias in the missingness as specified by the propensity $w_p(x, z)$, as (3.9) is a ratio estimator, the biases in the numerator and denominator arise from the missingness cancel each other. And as a result, the estimator is still consistent. However, it has not fully utilized data information in (X_i, Z_i) with missing Y_i.

To improve estimation efficiency, we impute each missing $I_{i\in\mathcal{E}}$ by $\hat{p}_1(X_i)$, which leads to the proposed imputation based estimator

$$\hat{p}_2(x) = \frac{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} \left\{ I_{i\in\mathcal{E}} \delta_i + \hat{p}_1(X_i)(1 - \delta_i) \right\}}{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}}}. \tag{3.10}$$

The properties of (3.9) and (3.10) be be discussed in the next Section.

3.4 Effects of the Imputation

Before we apply the proposed estimator $\hat{p}_2(x)$ for the US Census ACE data, we would like to quantify the effects of smoothing the discrete covariates and the missing value imputation in this section. We note that our proposal can be used well beyond the census in general nonparametric regression.

Let $\tilde{f}(x,z)$ be the probability density function of (X_i, Z_i),

$$\tilde{w}_p(x) = \int w_p(x,z)\tilde{f}_p(x,z)dz$$

be the marginal propensity function.

The following definition are the same as those in Theorem 3.1, in particular, let $\mathcal{C}_{x^u} = \{y^u : \sum_{j=1}^{d_u} I(x_j^u = y_j^u) = 1\}$ be the nearest strata of x^u which differs from x^u only in one component. For a $y^u \in \mathcal{C}_{x^u}$, recall

$$\alpha(x^u, y^u) = \sum_{k=1}^{d_u} c_k I(x_k^u = y_k^u) \quad \text{and} \quad \beta_\lambda(x^u, y^u) = \sum_{k=1}^{d_u} \lambda_k I(x_k^u = y_k^u)$$

where c_k is the levels of X_{ik}^u. For $\hat{p}_1(x)$, the bias induced by the smoothing of the discrete variables is quantified by

$$b_{1,u}(x;\vec{\lambda}) = \sum_{y^u \in \mathcal{C}_x^u} \left(\frac{1-\beta_\lambda(x^u,y^u)}{\alpha(x^u,y^u)-1} \left[\frac{g(x^c,y^u)\tilde{w}_p(x^c,y^u)}{g(x)\tilde{w}_p(x)} \{p(x^c,y^u) - p(x)\} \right] \right),$$

and the bias by smoothing the continuous variables is

$$b_{1,c}(x;h) = \frac{1}{2}h^2\sigma_K^2 \frac{\mathrm{tr}[\nabla^2\{p(x)g(x)\tilde{w}_p(x)\}] - p(x)\mathrm{tr}[\nabla^2\{g(x)\tilde{w}_p(x)\}]}{g(x)\tilde{w}_p(x)},$$

The following theorem shows the property of $\hat{p}_1(x)$.

Theorem 3.2 *Under the regularity conditions given in Section 3.7, let* $\lambda = \min_{1\le l\le d_u}(\lambda_l)$ *and* $\lambda^{(a)} = \prod_{k=1}^{d_u} \lambda_k^a$,

$$\begin{aligned} E\{\hat{p}_1(x)\} &= p(x) + b_{1,u}(x,\vec{\lambda}) + b_{1,c}(x,h) + O\{h^2(1-\lambda)\}, \\ var\{\hat{p}_1(x)\} &= \frac{\lambda^{(2)}R(K)}{Nh^{d_c}} \frac{p(x)\{1-p(x)\}}{g(x)\tilde{w}_p(x)} + o(N^{-1}h^{-d_c}), \end{aligned}$$

as $N \to \infty$, $h \to 0$, $\lambda \to 1$ *and* $Nh^{d_c} \to \infty$.

Theorem 3.2 is with very similar form to that of Theorem 3.1. The effect on the variance from the un-resolved enumeration status and ignoring the missing values is the inflation term $\tilde{w}_p^{-1}(x)$. As $w_p(x,z) \leq 1$, $\int \tilde{w}_p(x,z)\tilde{f}(x,z)dz \leq f(x)$. Therefore, $var\{\hat{p}_1(x)\} \geq var\{\hat{p}_0(x)\}$. This represents the loss of information due to missing values.

We may similarly quantify the biases of $\hat{p}_2(x)$ from smoothing discrete and continuous variables. Let

$$\begin{aligned} b_{2,u}(x;\vec{\lambda}) &= b_{0,u}(x;\vec{\lambda}) + \frac{f(x)-\tilde{w}_p(x)}{f(x)} b_{1,u}(x,\vec{\lambda}) \text{ and} \\ b_{2,c}(x;h) &= b_{0,c}(x;h) + \frac{f(x)-\tilde{w}_p(x)}{f(x)} b_{1,c}(x;h). \end{aligned}$$

The following Theorem shows the property of $\hat{p}_2(x)$.

Theorem 3.3 *Under the regularity conditions given in Section 3.7, let* $\lambda = \min_{1\leq l\leq d_u}(\lambda_l)$ *and* $\lambda^{(a)} = \prod_{k=1}^{d_u} \lambda_k^a$,

$$\begin{aligned} E\{\hat{p}_2(x)\} &= p(x) + b_{2,u}(x,\vec{\lambda}) + b_{2,c}(x,h) + O\{h^2(1-\lambda)\}, \\ var\{\hat{p}_2(x)\} &= \frac{1}{Nh^{d_c}} \frac{p(x)\{1-p(x)\}}{f^2(x)} \left[\lambda^{(2)} R(K)\tilde{w}_p(x) + 2\lambda^{(3)} R_2(K)\{f(x)-\tilde{w}_p(x)\}\right. \\ &+ \left. \frac{\lambda^{(4)} R_3(K)\{f(x)-\tilde{w}_p(x)\}^2}{\tilde{w}_p(x)}\right] + o(N^{-1}h^{-d_c}). \end{aligned}$$

as $N \to \infty$, $h \to 0$, $\lambda \to 1$ *and* $Nh^{d_c} \to \infty$, *where* $R(K) = \int K^2(t)dt$, $R_2(K) = \int K^{(2)}(t)K(t)dt$, $R_3(K) = \int K^{(3)}(t)K(t)dt$ *and* $K^{(3)}(t)$ *is the third convolution of* $K(t)$.

Remark 1. The first implication of the above results is that the imputation based estimator $\hat{p}_2(x)$ is more efficient than $\hat{p}_1(x)$. This can be appreciated by comparing the leading variance term in Theorem 3.3 and Theorem 3.2. As $R_2(K) < R(K)$ and $R_3(K) < R(K)$ for all K with 0 being the unique maxima, which is fulfilled by all commonly used symmetric kernels except the uniform kernel, then by noting $\lambda^{(k)} = 1 + o(1)$, the variance of $\hat{p}_2(x)$ is less than

$$(N^{-1}h^{-d_c})\frac{R(K)p(x)\{1-p(x)\}}{g(x)f^2(x)\tilde{w}_p(x)}\{\tilde{w}_p(x) + \tilde{f}_p(x) - \tilde{w}_p(x)\}^2 = var\{\hat{p}_1(x)\}.$$

When there is no missing value, i.e $w_p(x,z) \equiv 1$, the theorem implies that the leading variance of the Oracle (who knows all missing $I_{i\in\mathcal{E}} I_{i\in\mathcal{P}}$) estimator $\hat{p}_0(x)$ in (3.4) is

$$(N^{-1}h^{-d_c})R(K)p(x)\{1-p(x)\}/\{g(x)f(x)\},$$

as given in Theorem 3.1. Thus, $\hat{p}_1(x)$ which ignores the missing values endures a variance inflation by a factor $f(x)/\tilde{w}_p(x)$. The proposed $\hat{p}_2(x)$ removes part of the variance inflation by utilizing missing value information.

Remark 2. Both $\hat{p}_1(x)$ and $\hat{p}_2(x)$ enjoy variance reductions by smoothing the categorical variables as shown by the terms involving $\lambda^{(j)}$, as $\lambda^{(j)} = 1 - j\sum_{k=1}^{d_u}(1-\lambda_k) + O\{(1-\lambda)^2\}$. This is a result of combining data information within neighboring strata defined by the categorical variables. Although the variance reductions are at the second order $(1-\lambda)/(Nh^{d_c})$, the realized reduction in finite samples can be substantial. This is especially the case when there are a large number of categorical variables and some strata have low sample size as is the case for the census. Our simulation study reported in Section 3.6 confirms this point.

Remark 3. The Theorem also contains results on the bias of the estimators. The general message there is that the two estimators have comparable bias behaviors and both are at the order of $O\{(h^2 + (1-\lambda)\}$. The optimal bandwidths $(h, 1-\vec{\lambda})$ that minimize the mean square error (MSE) or the mean integrated square error (MISE) satisfy $h \sim N^{-1/(4+d_c)}$ and $(1-\vec{\lambda}) \sim N^{-2/(4+d_c)}$, which means that $\vec{A}(\lambda) \sim N^{-2/(4+d_c)}$ as well. These rates coincide with the rates obtained in Hall et al. (2004) when there is no missing values. The smoothing bandwidths can be chosen by the cross-validation method, which will be demonstrated in the next section when we analyze the US Census data.

3.5 Analyzing Census Data

3.5.1 Estimation of Enumeration Probabilities

In this section, we apply the proposed kernel estimators to the 10% ACE revision II research data files. In studying the enumeration probability function, ACE revision II data files are preferred as the revision studies corrected various data coding errors in the US Census ACE samples, including some error enumerations coding, duplicated enumerations and other measurement errors (US Census Bureau, 2004). The error from data coding may have substantial effect on the final estimate (US Census Bureau, 2004). The data files contain about $60,000$ P-sample cases and $70,000$ E-sample cases. The individuals in the samples are properly weighted representing the sampling procedure constructing the final samples on files.

In our analysis, the covariates are the ROAST variables which include age, sex (2 levels), housing tenure (2 levels: owner and renter), and racial origins (7 levels: American Indian or Alaska Natives on Reservation, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian and Non-Hispanic white or other races). Geographical region (4 levels: Northeast, Midwest, South and West) is also included in our studies. Additional covariates may be included without changing the tune of the analysis. And the responses are the match status $I_{i\in\mathcal{E}}$ in the P-sample and correct enumeration status $I_{i\in\tilde{\mathcal{E}}}$ in the E-sample.

The ROAST covariates have been demonstrated by the existing US Census research to be significant in identifying relatively homogeneous subgroups in enumeration and correct enumeration. They are the variables used in the existing post-stratification in the census dual system estimation. The post-stratification discretizes the age in conjunction with the sex variable to form 7 sub-groups: Under 18, 18-29 male, 18-29 female, 30-49 male, 30-49 female, 50+ male and 50+ female. In the ACE revision II studies, two sub groups of age under 18 were created as $0-9$ and $10-17$ regardless of sex.

Missing values exist in both the P-sample and E-sample and in both the covariates and responses (US Census Bureau, 2004). The percentages of missing $I_{i\in\mathcal{E}}$ and $I_{i\in\tilde{\mathcal{E}}}$ are 3.7% and 6.6% respectively in the P- and E-samples, which are high considering the overall level of undercounts in the Census. In our analysis, we consider only missing responses. For missing covariates, we use the existing imputed values assigned in the research files. The missing $I_{i\in\mathcal{E}}$ and $I_{i\in\tilde{\mathcal{E}}}$ are imputed using our proposed method.

We chose the biweight kernel $K(x) = 15/16(1-x^2)^2 I(|x| \leq 1)$ to smooth the age and the discrete kernel (3.3) to smooth the other categorical covariates. The smoothing bandwidths were chosen by the Cross-Validation (CV) method. Let $\hat{p}^{(-i)}_{h,\lambda}(x)$ and $\hat{e}^{(-i)}_{h,\lambda}(x)$ be the estimators of $p(x)$ and $e(x)$ after excluding the i^{th} data pair $(X_i, I_{i\in\mathcal{E}})$.

For estimation of $p(x)$, we chose $(h, \vec{\lambda})$ that minimized the Cross-validation score

$$CV_p(h,\vec{\lambda}) = n^{-1}\sum_{i\in\mathcal{P}}\{I_{i\in\mathcal{E}} - \hat{p}^{(-i)}_{h,\lambda}(X_i)\}^2\delta_i.$$

For estimation of $e(x)$, the bandwidths were chosen by minimizing

$$CV_e(h,\lambda) = n^{-1}\sum_{i\in\mathcal{E}}\{I_{i\in\tilde{\mathcal{E}}} - \hat{e}^{(-i)}_{h,\lambda}(X_i)\}^2\eta_i,$$

where $\vec{\lambda} = (\lambda_1, \cdots, \lambda_4)$ corresponding to the four discrete covariates. The bandwidths prescribed by the CV were $h = 5.5$ and $\lambda = 0.8$ for the estimation of $p(x)$, and $h = 5.0$ and $\lambda = 0.8$ for the estimation of $e(x)$. These bandwidths were used in the imputation based estimates for $p(x)$ and $e(x)$ in Figures 3.1 and 3.2.

It is observed from Figures 3.1 and 3.2 that the geographical region and ROAST variables contributed to the heterogeneity in both the enumeration and correct enumeration probabilities. The age effect was quite apparent in the estimates for $p(x)$ and $e(x)$. At the same time, the kernel estimates changes substantially with respect to the other categorical ROAST variables. Figure 3.1 indicated that Northeast White Male Owner had an overall higher enumeration probability than Northeast Hispanic Female owners and Midwest Black Male renters, which might be expected. However, Figure 3.2

showed quite different features regarding the correct enumeration. Here, the Black Male renters exhibit the V-shape around 30 years old rather than around 20 years old as the White Male Owner. The V-shape was very much muted for the Hispanic female Owners. These observations confirm the significance of the ROAST variables in explaining the heterogeneous enumerations and correct enumerations. While these confirm the effects of these covariates, they do reveal the difficulty in capture the underlying forms of the functions with respect to these discrete covariates. The wave-like pattern in both $p(x)$ and $e(x)$ estimates in some cells suggests some age-heaping in a multiple of 5 or 10 years in age beyond 30.

Figure 3.3 displays the kernel estimates for the missing propensity score $w_e(x)$, which was as interesting as Figures 3.1 and 3.2. For instance, the White Male owners had very small chance of being missing as the estimate of $w_p(x)$ were very close to 1. In contrast, the Hispanic Female owners endured larger missingness while the Black Male renters experienced the highest missing values among the three.

These figures, together with many other plots we generated from the data, also reveal challenges that one would face in proposing a reasonable parametric models. There are 112 post-strata based on the categorical ROAST variables and region. The sample size within some of these 112 post-strata can be very small, for instance the Native Hawaiian or Pacific Islander (NHPI). Getting a workable model for each stratum is quite a task. The task will only grow when more covariates are included.

At the same time, the figures show that the proposed the kernel estimation is flexible and adaptive to varying functional forms in both $p(x)$ and $e(x)$. As shown by our theoretical investigation, the kernel estimates are consistent and reflective to the underlying model structure without imposing any subjective assumptions.

3.5.2 Model Checking

An immediate application of the proposed kernel estimators is to provide a diagnostic on the goodness-of-fit of a specific model for either the enumeration and correct enumeration probabilities. Model checking is important in all statistical applications as we want to avoid using a mis-specified model. Let $p(x;\theta)$ be a parametric model indexed by a parameter θ for the enumeration probability, and $\hat{\theta}$ be an consistent estimator of θ under the proposed model $p(x;\theta)$. For instance, $p(x;\theta)$ may be the logistic regression model described in Alho et al. (1993). The existing post stratification can be regarded as a special kind of parametric model which assumes a piecewise constant structure.

To test the validity of a parametric $p(x;\theta)$, it is natural to compare it with the imputed kernel estimator $\hat{p}_2(x)$ over a set S within the domain of the covariates x. Our proposed test statistic is

$$T_n = \int_{x\in S} \{\hat{p}_2(x) - \tilde{p}(x;\hat{\theta})\}^2 dx \tag{3.11}$$

For discrete covariates, summation instead of integration over S is understood in (3.11). Here we use $\tilde{p}(x;\hat{\theta})$ instead of $p(x;\hat{\theta})$ where

$$\tilde{p}(x;\hat{\theta}) = \frac{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i) p(X_i;\hat{\theta})}{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i)}$$

is a smoothed version of $p(x,\hat{\theta})$ by the same combination of continuous and discrete kernels. This is to cancel the bias in the kernel estimation so that the bias will not get into the asymptotic distribution of the test statistics; see Härdle and Mammen (1993) for details.

Our main interest is in testing the validity of the post-stratification. The set S was $S_1 \times S_2 \times S_3 \times S_4 \times S_5$ where $S_1 = [0, 80]$ for the age, $S_2 = S_3 = \{0, 1\}$ for sex and housing tenure, $S_4 = \{0, 1, 2, 3\}$ for the region and $S_5 = \{0, 1, \cdots, 6\}$ for the racial origin. These reflect the domain of the ROAST variables plus Region.

We propose using a variation of the bootstrap procedure proposed in (Chen and Gao, 2007) and (Härdle and Mammen, 1993) to approximate l_α, the upper-α quantile (critical

values) of T_n. Both the naive bootstrap approach which resamples $\{X_i, I_{i\in\mathcal{E}}\}$ and the direct resampling of the residuals generally fail in the regression when there is conditional heteroscedasticity. Instead, the wild bootstrap (Wu, 1986; Härdle and Mammen, 1993) should be used. However, the presence of the missing values in regression brings in complexity. Bootstrap procedures have been proposed in the presence of missing values, for instance that given in Shao and Sitter (1996). We propose the following bootstrap procedure which is related to the one used in Chen and Gao (2007).

We first define the following kernel estimators of the conditional variance and the propensity:

$$\begin{aligned}\hat{\sigma}^2(x) &= \frac{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i)\delta_i\{I_{i\in\mathcal{E}} - \hat{p}_1(X_i)\}^2}{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i)\delta_i} \quad \text{and} \\ \hat{w}_p(x) &= \frac{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i)\delta_i}{\sum_{i\in\mathcal{P}} K_{h,\vec{\lambda}}(x, X_i)}.\end{aligned}$$

The bootstrap procedure consists of the following steps:

1. Generate a bootstrap resample $\{(X_i, Y_i^*, \delta_i^*)\}_{i=1}^n$ such that, for each $i = 1, \ldots, n$, let $Y_i^* = p(X_i; \hat{\theta}) + \hat{\sigma}(X_i)\epsilon_i^*$, where ϵ_i^*s are IID observations generated from a distribution satisfying $E(\epsilon_i^*) = 0$ and $E(\epsilon_i^{*2}) = 1$. Let δ_i^* be Bernoulli$\{\hat{w}_p(X_i)\}$. Both the missing value mechanism and the regression structure resample those of the original data.

2. Re-estimate θ by $\hat{\theta}^*$ and $p(x)$ by $\hat{p}_2^*(x)$ based on the resample $\{(X_i, Y_i^*, \delta_i^*)\}_{i=1}^n$, and let $T_n^* = \int_{x\in S}\{\hat{p}_2^*(x) - \tilde{p}(x; \hat{\theta}^*)\}^2 dx$.

3. Repeat the above steps B times to obtain $\{T_{n,b}^*\}_{b=1}^B$ and estimate l_α, the $1-\alpha$-quantile of T_n, by l_α^* which is the $1-\alpha$ sample quantile of $\{T_{n,b}^*\}_{b=1}^B$.

The test reject $H_0 : p(x) = p(x; \theta)$ at α level of significance if $T_n \geq l_\alpha^*$.

We implemented the above goodness-of-fit procedure to the Census data testing for the validity of the post-stratification. The test statistic given by (3.11) was $T_n = 3.546$

with bandwidths $h = 5.0$ and $\lambda = 0.8$ as prescribed by the cross-validation. The 5% critical value $l^*_{0.05}$ based on $B = 500$ bootstrap resamples was 3.228. The bootstrap approximation to the P-value of the test was 0.004. It indicated a strong evidence that the post-stratification is a mis-specified model for the enumeration probability function. This confirmed the plots given in Figure 3.1 which showed heterogeneity due to the age and other ROAST variables in the enumeration probability.

3.6 Simulation Studies

In this section we report results from simulation studies which are designed to evaluate the performance of the imputation based estimator (3.10) and the proposed goodness-of-fit test for model checking.

The setting of the simulation was motivated by those empirical studies on the Census data as revealed by Figures 1 and 2. We chose $X = (X_1, \ldots, X_5)$ with $X_1 \in (0, 50)$, a continuous variable that mimics the age, and X_2, X_3, X_4 and X_5 mimic the other ROAST variables and region. Here we took a shorter range for the age to reduce the computation burden. The covariate $(X_1, X_2, \cdots X_5)$ were independent uniform distributed. In particular, $X_1 \sim \text{Unif}(0, 50)$ and each categorical variable is uniform over the possible discrete values.

We chose

$$P(Y = 1|X = x) = p(x) = e^{b(x)}/(1 + e^{b(x)}), \tag{3.12}$$

where, instead of using a linear function for $b(x)$ as in the standard logistic model, $b(x)$ was a nonlinear function:

$$b(x) = \mu_{l(x_2,x_3,x_4,x_5)} + \beta_{1g}x_1 + \beta_{2g}\phi(\frac{x_1 - \beta_{3g}}{\beta_{4g}}).$$

Here $\phi(\cdot)$ is the density of $N(0, 1)$ distribution, $\mu_g = 2 + 0.01\log(g)$ for $g \in \{1, \ldots, 112\}$ and $l(X_2, X_3, X_4, X_5)$ is a one-to-one transformation from the domain of $X_2 \times X_3 \times$

$X_4 \times X_5$ to $\{1, \ldots, 112\}$, representing 112 strata defined by the four discrete variables. We note here that $\mu_{l(x_2,x_3,x_4,x_5)}$ defines individual stratum effect on the regression. The motivation for using $\phi(\frac{x_1-\beta_{3g}}{\beta_{4g}})$ is to re-create the "V" shape with respect to the age as observed in Figures 3.1 and 3.2.

For $g = 1, \ldots, 112$, the parameter $\beta_g = (\beta_{1g}, \beta_{2g}, \beta_{3g}, \beta_{4g})$ was randomly generated from $N(\mu_\beta, \text{diag}\{s_1^2, s_2^2, s_3^2, s_4^2\})$ and then kept fixed throughout the simulation. Here $\mu_\beta = (0.0083, -8.726, 24.292, 4.824)$ was the maximum likelihood estimates (MLEs) of the above nonlinear logistic model based on the Census data. And $(s_1^2, s_2^2, s_3^2, s_4^2)$ was set to be $(0.008^2, 1, 1, 1)$ to allow noticeable distinction among strata. The model has 112×5 parameters, and the effect of x_1 (age) on $p(x)$ varies across the 112 strata determined by $(X_2, \ldots, X_5)$.

The missing propensity function $w_p(x)$ was similarly defined as

$$w_p(x) = P(\delta = 1|X = x) = e^{c(x)}/(1 + e^{c(x)}), \tag{3.13}$$

where $c(x) = 1 + 0.02l(x_2, x_3, x_4, x_5) + \theta_{1g}x_1 + \theta_{2g}\phi(\frac{x_1-\theta_{3g}}{\theta_{4g}})$ with $\theta_g = (\theta_{1g}, \theta_{2g}, \theta_{3g}, \theta_{4g}) \sim N(\mu_\theta, \text{diag}\{c_1^2, c_2^2, c_3^2, c_4^2\})$ and kept fixed where $\mu_\theta = (-0.0018, -17.61, 24.78, 5.52)$ was the MLEs based on the Census data. Similar to the setting for (3.12), $(c_1^2, c_2^2, c_3^2, c_4^2)$ was set to be $(0.006^2, 1, 1, 1)$. This setting led to less than 10% missing values which was in line with the Census data.

In the simulation studies, h and $\vec{\lambda} = (\lambda_1, \ldots, \lambda_4)$ is chosen by minimizing the CV object function. The average bandwidths $(h_{cv}, \vec{\lambda}_{cv})$ were obtained by pre-running the simulation and are marked in Tables 3.2 and 3.3. Two other sets of bandwidths around (h_{cv}, λ_{cv}), which increasing and decreasing the optimal value by 10%, were also considered.

Two performance measures for an estimator $\tilde{p}(x)$ of $p(x)$ were considered. One was

a cumulated mean square error (CMSE) of $\tilde{p}(x)$

$$CMSE\{\tilde{p}(x)\} = \sum_{x \in \mathcal{X}} \left[\text{bias}^2\{\tilde{p}(x)\} + Var\{\tilde{p}(x)\} \right], \tag{3.14}$$

which is the summation of the cumulated square bias ($CBias^2$) and variance ($CVar^2$). The CMSE is proportional to the mean integrated square error. The other measure was the CMSE of $1/\tilde{p}(x)$ for estimation of $1/p(x)$.

In the simulation, three estimation methods were compared: the proposed imputation based estimator $\hat{p}_2(x)$, the $\hat{p}_1(x)$ that ignores missing values and the post-stratification based estimation. For the latter, the post-strata were created by $X_2, \ldots, X_5$ crossing with 3 age groups: $0-18$, $18-29$ and $29-50$. So a total 336 post-strata were used in the estimation. The mean of the post-strata was used to impute the missing response. The sample sizes considered were 3000, 5000 and $10,000$.

Tables 3.1 and 3.2 summarize the simulation results based on 1000 replications. It is observed that the $CMSE$, $Cbias^2$ and $CVar$ of both estimators $\hat{p}_1$ and $\hat{p}_2$ decreased as the sample size increased. This reflected the fact that both $\hat{p}_1$ and $\hat{p}_2$ are consistent estimators. However, the proposed imputation based estimator $\hat{p}_2$ had smaller variance and CMSE than $\hat{p}_1$, which was expected from Theorem 1. These improvements were in fact quite impressive given the fact that the percentage of missing values was less than 10%. A relative constant bias was observed for the post-stratification based estimation, which did not diminish as much as the sample size was increased. And the post-stratification incurred much larger $CVAR$ and $CMSE$. The main reason of the large variability was likely due to the small sample size in the strata. Although the variability of the post-stratification based estimates dropped as the sample size increased, its variance and the MSE were still a lot larger than those of the proposed kernel estimator.

Due to the nature of the functions used in the simulation and the sample size experimented, the cumulated variance contributed more to the cumulated MSE than the cumulated square bias. As a result, the bandwidths prescribed by minimizing the CMSE

led to bandwidth $h > 10$, as the main task was to reduce the variance.

The main benefit of local post-stratification was in smoothing discrete variables which led to a substantial reduction in both the variance and the MSE. The cumulated variance of $\hat{p}_2$ was at most 1/20 of that of the post-stratification, whereas the CMSE of $\hat{p}_2$ was at most 1/10 of the post-stratification when the performance was measured by the CMSE of $p(x)$ estimator. When we changed the measure to the estimation of $1/p(x)$, the improvement by $\hat{p}_2$ was larger.

The goodness-of-fit test proposed in the previous section was also evaluated in the simulation. The null hypothesis was the model described under the post-stratification, namely $H_0 : Y_i = \mu_g + \sigma_g \epsilon_i$, where $g \in \{1, \ldots, 336\}$ corresponding to the strata determined by $X_1, \ldots, X_5$. The alternative hypothesis was the generalized logistic regression model specified in (3.12). The test procedure based on the bootstrap as described in Section 7 was used to carry out the test for H_0 for each simulation. The goodness-of-fit test was studied in two settings: one without missing values and the test was performed on the entire observations. The other encountered missing responses and the imputation based estimation is used in the formulation of T_n. The results are reported in Table 3.3, which illustrate that the goodness-of-fit testing procedure is an effective way in conducting the model diagnostic. Although the test lost some power in the presence of the missing values, the level of power is still very good and satisfactory in detecting the discrepancy between the null and alternative models.

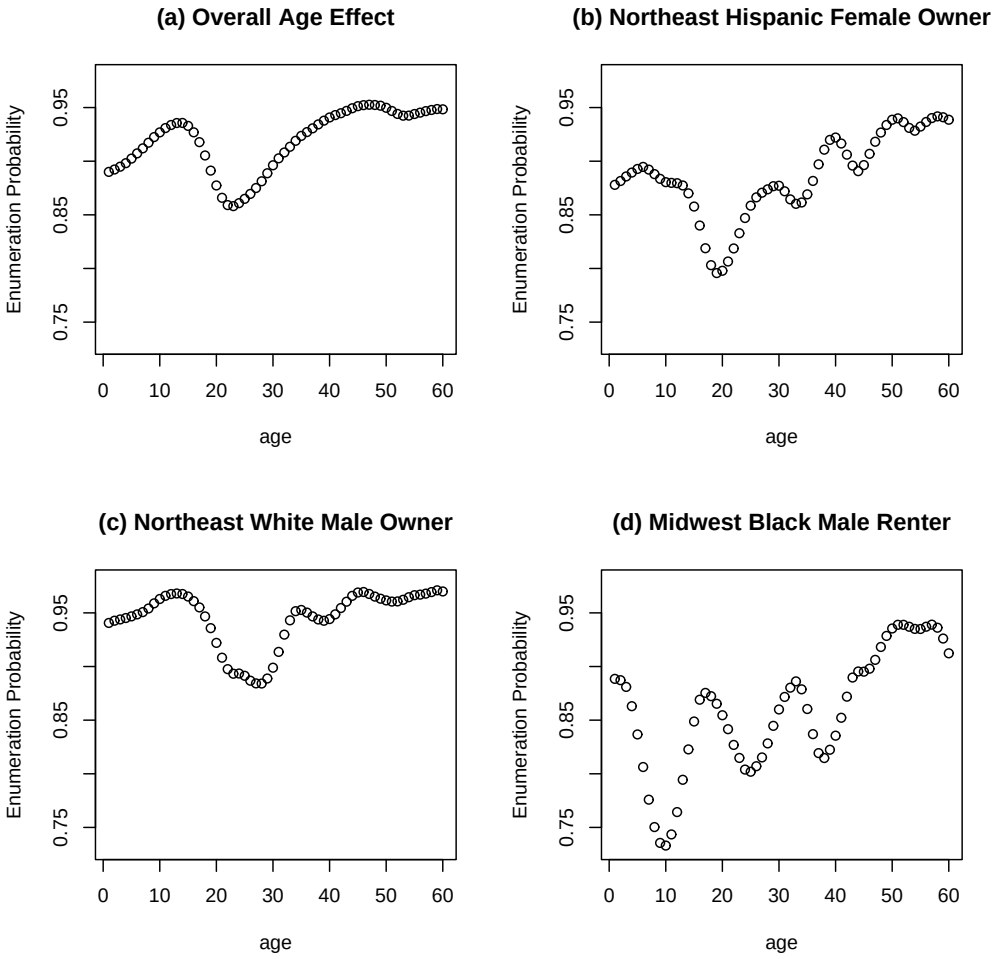

Figure 3.1 Kernel estimates of the enumeration probability $p(x)$ based on $\hat{p}_2(x)$. Bandwidths used are $h = 5.5$ and $\lambda = 0.8$.

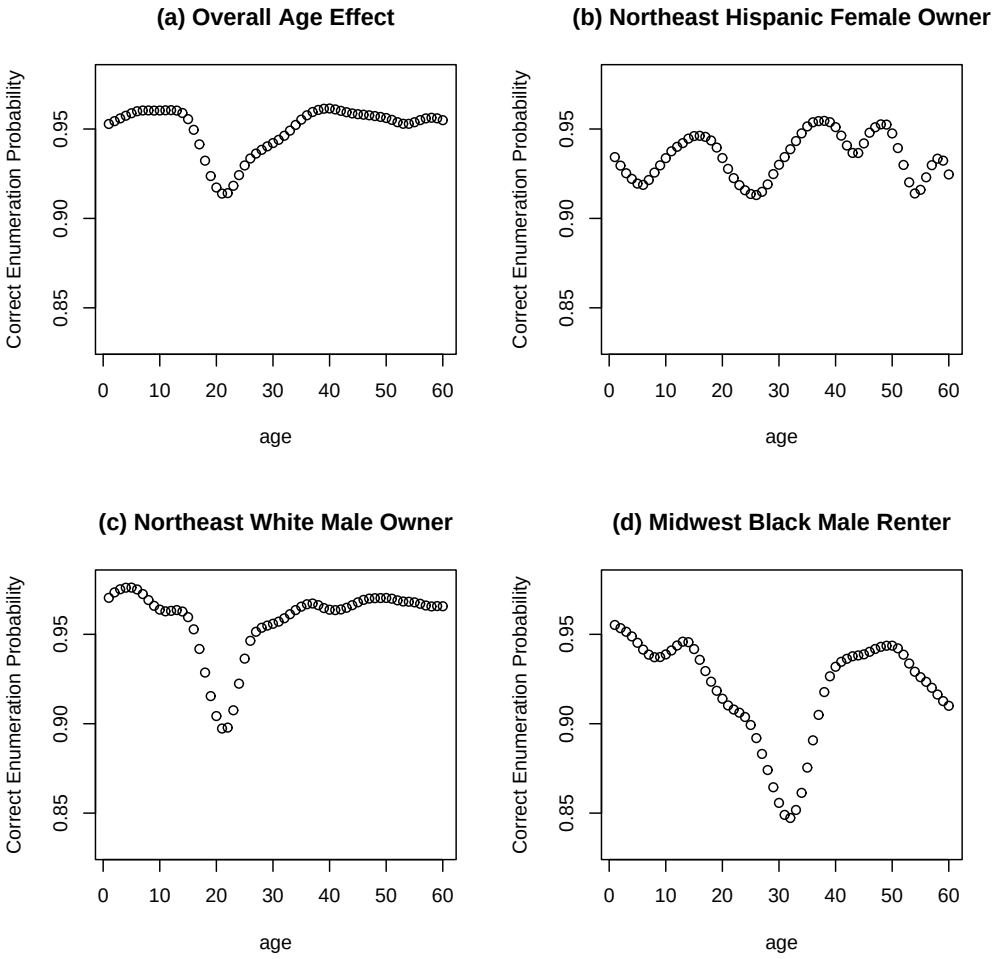

Figure 3.2 Kernel estimates of the correct enumeration probability $e(x)$ based on the nonparametric imputation. Bandwidths used are $h = 5.0$ and $\lambda = 0.8$.

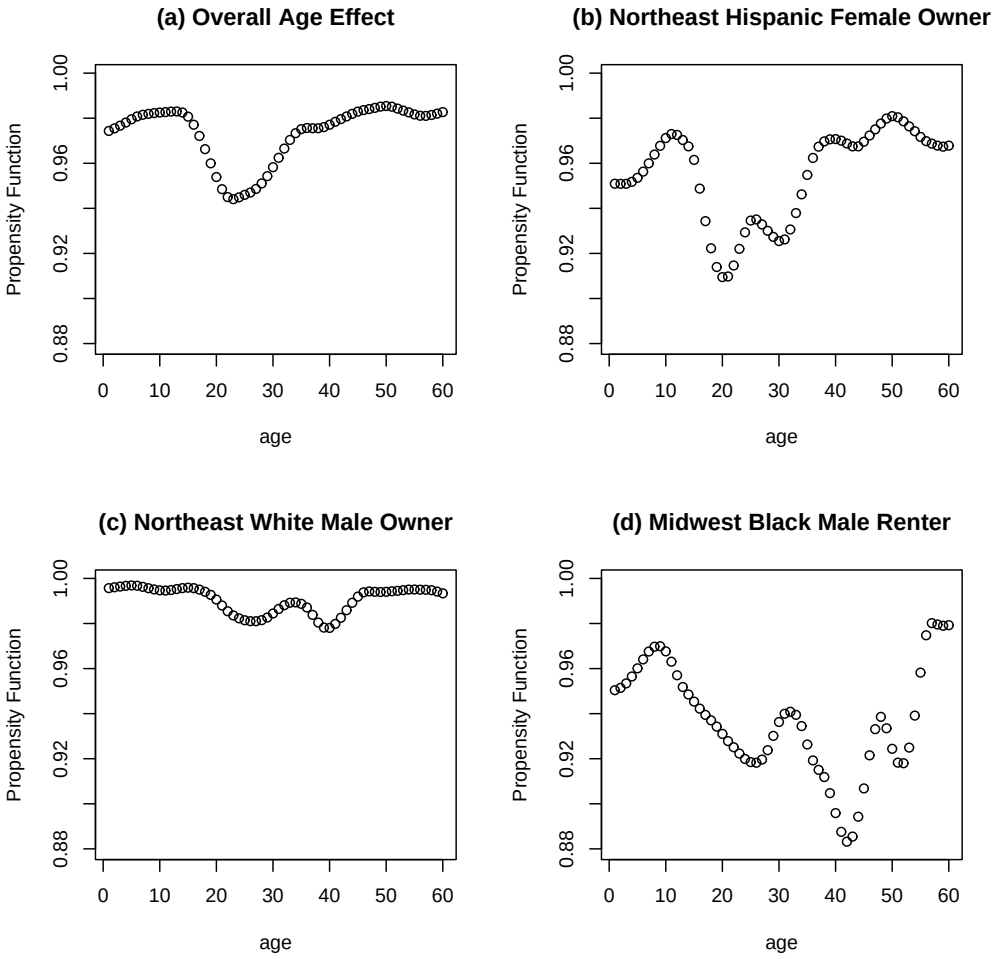

Figure 3.3 Kernel estimates of the E-sample missing propensity function $w_p(x)$ based on the proposed imputation. Bandwidths used are $h = 5.0$ and $\lambda = 0.8$.

Bandwidths		$CBias^2$		$CVar$		$CMSE$	
$n = 3000$		$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_1$	$\hat{p}_2$
$h = 12.0$	$0.9\lambda^*$	3.074	3.359	4.649	3.528	7.722	6.887
$h^* = 13.0$	λ^*	3.442	3.708	2.980	2.264	6.421	5.972
$h = 14.0$	$1.1\lambda^*$	3.773	4.017	1.956	1.501	5.729	5.517
Post-Stratification		4.037		60.70		64.74	
$n = 5000$							
$h = 11.0$	$0.9\lambda^*$	2.914	3.201	2.987	2.261	5.901	5.462
$h^* = 12.0$	λ^*	3.29	3.557	1.903	1.444	5.193	5.001
$h = 13.0$	$1.1\lambda^*$	3.627	3.872	1.245	0.9544	4.871	4.826
Post-Stratification		2.922		33.50		46.15	
$n = 10000$							
$h = 10.0$	$0.9\lambda^*$	1.561	1.836	5.08	3.906	6.641	5.742
$h^* = 11.0$	λ^*	2.152	2.445	2.800	2.120	4.952	4.565
$h = 12.0$	$1.1\lambda^*$	2.951	3.217	1.084	0.8163	4.035	4.033
Post-Stratification		2.769		22.07		24.83	

Table 3.1 Empirical cumulative square bias, variance and MSE of $\hat{p}_1$, $\hat{p}_2$ and the post-stratification for estimation of $p(x)$. Bandwidths marked with h^* and λ^* are those prescribed by the cross-validation.

Bandwidths		$CBias^2$		$CVar$		$CMSE$	
$n = 3000$		$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_1$	$\hat{p}_2$
$h = 12.0$	$0.9\lambda^*$	6.130	6.713	8.445	6.304	14.580	13.02
$h^* = 13.0$	λ^*	6.874	7.395	5.276	3.963	12.150	11.360
$h = 14.0$	$1.1\lambda^*$	7.522	7.987	3.404	2.59	10.930	10.580
Post-Stratification		10.763		201.4		212.16	
$n = 5000$							
$h = 11.0$	$0.9\lambda^*$	5.873	6.445	5.325	3.979	11.20	10.42
$h^* = 12.0$	λ^*	6.615	7.131	3.331	2.503	9.946	9.635
$h = 13.0$	$1.1\lambda^*$	7.265	7.728	2.151	1.638	9.416	9.366
Post-Stratification		8.014		164.2		172.2	
$n = 10000$							
$h = 10.0$	$0.9\lambda^*$	3.158	3.753	9.5	7.154	12.66	10.91
$h^* = 11.0$	λ^*	4.393	4.991	5.048	3.768	9.442	8.758
$h = 12.0$	$1.1\lambda^*$	5.974	6.486	1.893	1.414	7.867	7.9
Post-Stratification		5.952		64.15		70.10	

Table 3.2 Empirical cumulative square bias, variance and MSE of $1/\hat{p}_1(x)$, $1/\hat{p}_2(x)$ and the post-stratification for estimation of $1/p(x)$. Bandwidths marked with h^* and λ^* are those prescribed by the cross-validation.

Test	Size	Power
Full Data	0.048	0.74
With Missing Values	0.052	0.69

Table 3.3 Empirical Size and Power of the goodness-of-fit test for the post-stratification (H_0 and size) against the generalized logistic regression model (H_1 and power) as given in (3.12).

3.7 Technical Proofs

Technical Assumptions

Let covariate $X_i = (X_i^c, X_i^u)$ where X_i^c is a d_c-dimensional continuous covariate and X_i^u is a d_u-dimensional unordered categorical covariate, $\mathcal{X} = \{\mathcal{X}^c, \mathcal{X}^u\}$ be the support of X_i, where $\mathcal{X}^c$ and $\mathcal{X}^u$ are the supports of X_i^c and X_i^u respectively. We assume data pairs $\{(X_i, Z_i, I_{i\in\mathcal{E}}, I_{i\in\tilde{\mathcal{E}}})\}_{i=1}^N$ are independent and identically distributed. And the following conditions are assumed in the following proofs.

C.1 Let $K(\cdot)$ be a d_c variates nonnegative, bounded and symmetric probability density function with bounded second derivative. The smoothing bandwidths satisfy that $h \to 0$ and $\max_{1\le j\le d_u}\{(1-\lambda_j)\} \to 0$ and $Nh^{d_c} \to \infty$ as $N \to \infty$.

C.2 We assume missing at random in $I_{i\in\mathcal{E}}$, namely $P(\delta_i = 1|I_{i\in\mathcal{E}}, X_i = x, Z_i = z) = P(\delta_i = 1|X_i = x, Z_i = z) := w_p(x, z)$, where $w_p(x^c, x^u, z) \ge C_w$ for a constant $C_w > 0$ and $w_p(x^c, x^u, z)$ has bounded continuous second partial derivative with respect to x^c within $\mathcal{X}^c$.

C.3 For any $x^u \in \mathcal{X}^u$, $p(x^c, x^u)$, $g(x^c, x^u)$ and the probability density function $f(x^c, x^u)$ have bounded continuous second partial derivatives with respect to x^c in $\mathcal{X}^c$, and there exist $C_f > 0$ such that $f(x^c, x^u) \ge C_f$ for all $(x^c, x^u) \in \mathcal{X}$.

Proof of Theorem 3.1

Let $\theta_1(x) = p(x)g(x)f(x)$, $\theta_2(x) = g(x)f(x)$ and define

$$\hat{\theta}_1(x) = N^{-1}\sum_{i\in U}\mathcal{K}_{h,\vec{\lambda}}(x, X_i)I_{i\in\mathcal{P}}I_{i\in\mathcal{E}} \text{ and } \hat{\theta}_2(x) = N^{-1}\sum_{i\in U}\mathcal{K}_{h,\vec{\lambda}}(x, X_i)I_{i\in\mathcal{P}},$$

then (3.4) can be written as

$$\hat{p}_0(x) = \frac{\hat{\theta}_1(x)}{\hat{\theta}_2(x)} = \frac{\theta_1(x)}{\theta_2(x)} + \frac{1}{\theta_2(x)}\{\hat{\theta}_1(x) - \theta_1(x)\} - \frac{\theta_1(x)}{\theta_2^2(x)}\{\hat{\theta}_2(x) - \theta_2(x)\}\{1 + o_p(1)\}. \quad (3.15)$$

The following steps kernel smoothing are used replicatedly in the derivations. For any function $q(\cdot) : R^{d_c} \to R$ twice continuously differentiable at x, as $h \to 0$,

$$\begin{aligned}\int K_h(x-y)q(y)dy &= \int h^{-d_c}K\left(\frac{x-y}{h}\right)q(y)dy = \int K(t)q(x-ht)dt \\ &= \int K(t)\{q(x) - htq'(x) + \frac{1}{2}(ht\boldsymbol{J})^T\nabla^2\{q(x)\}(ht\boldsymbol{J}) + O(h^3)\}dt \\ &= q(x) + \frac{1}{2}h^2\sigma_K^2\mathrm{tr}[\nabla^2\{q(x)\}] + o(h^2), \qquad (3.16)\end{aligned}$$

where $\boldsymbol{J}$ is a d_c dimensional column vector whose elements are all 1 and T is the matrix transpose, and

$$\begin{aligned}&E[\{L(x^u, y^u, \vec{\lambda})q(X^c, y^u)\}|X^c = x^c] = \sum_{y^u \in \mathcal{X}^U} L(x^u, y^u, \vec{\lambda})q(x^c, y^u)f(x^c, y^u) \\ =\ & \lambda^{(1)}q(x^c, x^u)f(x^c, x^u) + \sum_{y^u \in \mathcal{C}_x^u}\left\{\frac{1-\beta_\lambda(x^u, y^u)}{\alpha(x^u, y^u) - 1}q(x^c, y^u)f(x^c, y^u)\right\} \\ +\ & O\{(1-\lambda)^2\}. \qquad (3.17)\end{aligned}$$

The bias part of Theorem 3.1 comes from

$$\begin{aligned}E\{\hat{\theta}_1(x)\} &= E\left\{\mathcal{K}_{h,\vec{\lambda}}(x, X_i)I_{i\in\mathcal{P}}I_{i\in\mathcal{E}}\right\} = \int_{\mathcal{X}}\mathcal{K}_{h,\vec{\lambda}}(x, z)p(z)g(z)f(z)dz \\ &= \int_{\mathcal{X}} K_h(x^c - z^c)L(x^u, z^u, \vec{\lambda})p(z)g(z)f(z)dz \\ &= \lambda^{(1)}\int_{\mathcal{X}} K_h(x^c - z^c)p(z^c, x^u)g(z^c, x^u)f(z^c, x^u)dz^c \\ &+ \sum_{y^u \in \mathcal{C}_x^u}\left\{\frac{1-\beta_\lambda(x^u, y^u)}{\alpha(x^u, y^u) - 1}\int_{\mathcal{X}} K_h(x^c - z^c)p(z^c, y^u)g(z^c, y^u)f(z^c, y^u)dz^c\right\} \\ &+ O\{(1-\lambda)^2\} \\ &= \theta_1(x) + \frac{1}{2}h^2\sigma_K^2\mathrm{tr}[\nabla^2\{p(x)g(x)f(x)\}] \\ &+ \sum_{y^u \in \mathcal{C}_x^u}\left\{\frac{1-\beta_\lambda(x^u, y^u)}{\alpha(x^u, y^u) - 1}p(x^c, y^u)g(x^c, y^u)f(x^c, y^u)\right\} + O\{h^2(1-\lambda)^2\}.\end{aligned}$$

And by exactly the same steps,

$$\begin{aligned}E\{\hat{\theta}_2(x)\} &= \theta_2(x) + \frac{1}{2}h^2\sigma_K^2\mathrm{tr}[\nabla^2\{g(x)f(x)\}] \\ &+ \sum_{y^u \in \mathcal{C}_x^u}\left\{\frac{1-\beta_\lambda(x^u, y^u)}{\alpha(x^u, y^u) - 1}g(x^c, y^u)f(x^c, y^u)\right\} + O\{h^2(1-\lambda)^2\}.\end{aligned}$$

Hence the bias part of Theorem 3.1 is established by taking expectation on (3.15). Further

$$\begin{aligned} var\{\hat{\theta}_1(x)\} &= N^{-1} var\{\mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{E}}\} \\ &= N^{-1}\left(E\left\{\mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{E}}\right\}^2 - \left[E\left\{\mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{E}}\right\}\right]^2\right) \\ &= N^{-1}\left\{\int_{\mathcal{X}} K^2_{h,\vec{\lambda}}(x, z) p(z) g(z) f(z) dz + O(1)\right\} \\ &= \frac{\lambda^{(2)} R(K)}{N h^{-d_c}} p(x) g(x) f(x) + o(N^{-1} h^{-d_c}). \end{aligned}$$

And similarly,

$$\begin{aligned} var\{\hat{\theta}_2(x)\} &= \frac{\lambda^{(2)} R(K)}{N h^{-d_c}} g(x) f(x) + o(N^{-1} h^{-d_c}), \text{ and} \\ cov\{\hat{\theta}_1(x), \hat{\theta}_2(x)\} &= \frac{\lambda^{(2)} R(K)}{N h^{-d_c}} p(x) g(x) f(x) + o(N^{-1} h^{-d_c}). \end{aligned}$$

Then the variance part of Theorem 3.1 is obtained by taking variance operation on (3.15),

$$\begin{aligned} var\{\hat{p}_0(x)\} &= \frac{1}{g^2(x) f^2(x)} var\{\hat{\theta}_1(x)\} + \frac{p^2(x)}{g^2(x) f^2(x)} var\{\hat{\theta}_2(x)\} \\ &+ \frac{2p(x)}{g^2(x) f^2(x)} cov\{\hat{\theta}_1(x), \hat{\theta}_2(x)\} + o(N^{-1} h^{-dc}) \\ &= \frac{\lambda^{(2)} R(K)}{N h^{d_c}} \frac{p(x)\{1 - p(x)\}}{g(x) f(x)} + o(N^{-1} h^{-d_c}). \end{aligned}$$

Proof of Theorem 3.2

The proof of Theorem 3.2 follows exactly the same steps in proof of Theorem 3.1. By letting $\theta_3(x) = p(x) g(x) \tilde{w}_p(x)$, $\theta_4(x) = g(x) \tilde{w}_p(x)$ and define

$$\hat{\theta}_3(x) = N^{-1} \sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} \delta_i \text{ and } \hat{\theta}_4(x) = N^{-1} \sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} \delta_i,$$

then (3.4) can be written as

$$\hat{p}_1(x) = \frac{\hat{\theta}_3(x)}{\hat{\theta}_4(x)} = \frac{\theta_3(x)}{\theta_4(x)} + \frac{1}{\theta_4(x)}\{\hat{\theta}_3(x) - \theta_3(x)\} - \frac{\theta_3(x)}{\theta_4^2(x)}\{\hat{\theta}_4(x) - \theta_4(x)\}\{1 + o_p(1)\}. \quad (3.18)$$

The rest of the steps are replications of those in establish Theorem 3.1. We see that in $\theta_3(x)$ and $\theta_4(x)$, which are the quantities that $\hat{\theta}_3(x)$ and $\hat{\theta}_4(x)$ consistently estimating, a bias $\tilde{w}_p(x)$ due to the missing value appears. As $\hat{p}_1(x)$ is a ratio estimator, the bias in the numerator and denominator cancel each other. And therefore, $\hat{p}_1(x)$ is still a consistent estimator of $p(x)$.

Proof of Theorem 3.3

Let $\hat{\theta}_5(x) = N^{-1}\mathcal{K}_{h,\vec{\lambda}}(x, X_i)I_{i\in\mathcal{P}}\{I_{i\in\mathcal{E}}\delta_i + \hat{p}_1(X_i)(1-\delta_i)\}$, $\hat{\theta}_5$ also estimates $\theta_1(x) = p(x)g(x)f(x)$ consistently. Then

$$\hat{p}_2(x) = \frac{\hat{\theta}_5(x)}{\hat{\theta}_2(x)} = \frac{\theta_1(x)}{\theta_2(x)} + \frac{1}{\theta_2(x)}\{\hat{\theta}_5(x) - \theta_1(x)\} - \frac{\theta_1(x)}{\theta_2^2(x)}\{\hat{\theta}_2(x) - \theta_2(x)\}\{1 + o_p(1)\}. \quad (3.19)$$

Use the result in Theorem 3.2, for any given $X_i \in \mathcal{X}$

$$E\{\hat{p}(X_i)|X_i\} = p(X_i) + b_{1,c}(X_i; h) + b_{1,u}(X_i; \vec{\lambda})\{1 + o_p(1)\}. \quad (3.20)$$

Then by firstly taking expectation conditional on X_i,

$$\begin{aligned}
E\{\hat{\theta}_5(x)\} &= E\left[\mathcal{K}_{h,\vec{\lambda}}(x, X_i)I_{i\in\mathcal{P}}\{I_{i\in\mathcal{E}}\delta_i + \hat{p}_1(X_i)(1-\delta_i)\}\right] \\
&= E[\mathcal{K}_{h,\vec{\lambda}}(x, X_i)g(X_i)\{p(X_i)w_p(X_i, Z_i) + \{p(X_i) + b_{1,c}(X_i; h) \\
&+ b_{1,u}(X_i; \vec{\lambda})\}\{1 - w_p(X_i, Z_i\}] + O\{h^2(1-\lambda)^2\} \\
&= \theta_1(x) + \frac{1}{2}h^2\sigma_K^2 \mathrm{tr}[\nabla^2\{p(x)g(x)f(x)\}] \\
&+ \sum_{y^u \in \mathcal{C}_x^u}\left\{\frac{1 - \beta_\lambda(x^u, y^u)}{\alpha(x^u, y^u) - 1}p(x^c, y^u)g(x^c, y^u)f(x^c, y^u)\right\} \\
&+ g(x)(f - \tilde{w}_p(x))\{b_{1,c}(x; h) + b_{1,u}(x; \vec{\lambda})\} + O\{h^2(1-\lambda)^2\}.
\end{aligned}$$

Therefore, from expansion (3.19) and results of $E\{\hat{\theta}_2(x)\}$ from the proof of Theorem 3.1, we have

$$E\{\hat{p}_2(x)\} = p(x) + b_{2,u}(x; \vec{\lambda}) + b_{2,c}(x; h) + O\{h^2(1-\lambda)^2\}.$$

To derive the variance of $\hat{p}_2(x)$, decompose $\hat{\theta}_5(x)$ as follows by using (3.20),

$$\hat{\theta}_5(x) = \hat{\theta}_3(x) + \hat{\theta}_{51}(x) + \hat{\theta}_{52}(x) + \hat{\theta}_{53}(x)\{1 + o_p(1)\}, \tag{3.21}$$

where

$$\begin{aligned}
\hat{\theta}_{51}(x) &= N^{-1}\mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} p(X_i)(1-\delta_i), \\
\hat{\theta}_{52}(x) &= N^{-2}\sum_{i,j\in U} a_1^{i,j}(x) \text{ and } \hat{\theta}_{53}(x) = -N^{-2}\sum_{i,j\in U} a_2^{i,j}(x).
\end{aligned}$$

The definitions of $a_1^{i,j}(x)$ and $a_2^{i,j}(x)$ are given by

$$\begin{aligned}
a_1^{i,j}(x) &= \mathcal{K}_{h,\vec{\lambda}}(x, X_i)\frac{I_{i\in\mathcal{P}}(1-\delta_i)}{g(X_i)\tilde{w}_p(X_i)}\mathcal{K}_{h,\vec{\lambda}}(X_i, X_j) I_{j\in\mathcal{E}} I_{j\in\mathcal{P}}\delta_j \text{ and} \\
a_2^{i,j}(x) &= \mathcal{K}_{h,\vec{\lambda}}(x, X_i)\frac{I_{i\in\mathcal{P}}(1-\delta_i)p(X_i)}{g(X_i)\tilde{w}_p(X_i)}\mathcal{K}_{h,\vec{\lambda}}(X_i, X_j) I_{j\in\mathcal{P}}\delta_j.
\end{aligned}$$

Then

$$\begin{aligned}
var\{\hat{\theta}_{52}(x) &+ \hat{\theta}_{53}(x)\} = N^{-4}\sum_{i,j,k,l\in U}\Big(cov[\{a_1^{i,k}(x) - a_2^{i,k}(x)\}, \{a_1^{j,l}(x) - a_2^{j,l}(x)\}]\Big) \\
&= N^{-4}\sum_{k=l \text{ or } i=l \text{ or } i=j \text{ or } k=j}\Big(cov[\{a_1^{i,k}(x) - a_2^{i,k}(x)\}, \{a_1^{j,l}(x) - a_2^{j,l}(x)\}]\Big) \\
&+ o(N^{-1}h^{-d_c}) \\
&= N^{-4}\sum_{k=l}\Big(cov[\{a_1^{i,k}(x) - a_2^{i,k}(x)\}, \{a_1^{j,l}(x) - a_2^{j,l}(x)\}]\Big) + o(N^{-1}h^{-d_c}).
\end{aligned}$$

The last equation holds as in those cases when $k \neq l$, the leading terms in $cov[\{a_1^{i,k}(x) - a_2^{i,k}(x)\}, \{a_1^{j,l}(x) - a_2^{j,l}(x)\}]$ cancel each other and result in smaller order terms. Further,

$$\begin{aligned}
cov\{a_1^{i,k}(x), a_2^{j,k}(x)\} &= E\Big\{\mathcal{K}_{h,\vec{\lambda}}(x, X_i)\mathcal{K}_{h,\vec{\lambda}}(x, X_j)\frac{\{I_{i\in\mathcal{P}}(1-\delta_i)\}\{I_{j\in\mathcal{P}}(1-\delta_j)\}}{g(X_i)\tilde{w}_p(X_i)g(X_j)\tilde{w}_p(X_j)} \\
&\times \mathcal{K}_{h,\vec{\lambda}}(X_i, X_k)\mathcal{K}_{h,\vec{\lambda}}(X_j, X_k) I_{k\in\mathcal{E}} I_{k\in\mathcal{P}}\delta_k\Big\} + O(1) \\
&= \lambda^{(4)}h^{-d_c}R_3(K)p(x)g(x)\frac{\{f(x) - \tilde{w}_p(x)\}^2}{\tilde{w}_p(x)} + O(1),
\end{aligned}$$

where $R_3(K) = \int K^{(3)}(t)K(t)dt$, $K^{(3)}(t)$ is the third convolution of $K(t)$. And similarly,

$$\begin{aligned} cov\{a_2^{i,k}(x), a_2^{j,k}(x)\} &= cov\{a_1^{i,k}(x), a_2^{j,k}(x)\} \\ &= \lambda^{(4)} h^{-d_c} R_3(K) p^2(x) g(x) \frac{\{f(x) - \tilde{w}_p(x)\}^2}{\tilde{w}_p(x)} + O(1). \end{aligned}$$

Therefore, we conclude that

$$var\{\hat{\theta}_{52}(x) + \hat{\theta}_{53}(x)\} = \frac{\lambda^{(4)}}{Nh^{d_c}} R_3(K) p(x)\{1 - p(x)\} g(x) \frac{\{f(x) - \tilde{w}_p(x)\}^2}{\tilde{w}_p(x)} + o(N^{-1}h^{-d_c}). \tag{3.22}$$

Let

$$b_1^i(x) = \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} \delta_i \text{ and } b_2^i(x) = \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} p(X_i)(1 - \delta_i),$$

then

$$\begin{aligned} cov[\hat{\theta}_3(x), \{\hat{\theta}_{52}(x) &+ \hat{\theta}_{53}(x)\}] = N^{-3} \sum_{i,j,k\in U} \Big[cov\{b_1^i(x), a_1^{j,k}(x)\} - cov\{b_1^i(x), a_2^{j,k}(x)\} \Big] \\ &= N^{-3} \sum_{i=k} \Big[cov\{b_1^i(x), a_1^{j,k}(x)\} - cov\{b_1^i(x), a_2^{j,k}(x)\} \Big] + o(N^{-1}h^{-d_c}) \\ &= \frac{\lambda^{(3)}}{Nh^{d_c}} R_2(K) p(x)\{1 - p(x)\} g(x)\{f - \tilde{w}_p(x)\} + o(N^{-1}h^{-d_c}), \end{aligned} \tag{3.23}$$

where $R_2(K) = \int K^{(2)}(t)K(t)dt$,

$$\begin{aligned} var\{\hat{\theta}_3(x)\} &= N^{-2} \sum_{i,j} \big[cov\{b_1^i(x), b_1^j(x)\} \big] \\ &= \frac{\lambda^{(2)}}{Nh^{d_c}} R(K) g(x) p(x) \tilde{w}_p(x) + o(N^{-1}h^{d_c}), \end{aligned} \tag{3.24}$$

$$\begin{aligned} var\{\hat{\theta}_{51}(x)\} &= N^{-2} \sum_{i,j} \big[cov\{b_2^i(x), b_2^j(x)\} \big] \\ &= \frac{\lambda^{(2)}}{Nh^{d_c}} R(K) g(x) p^2(x)\{1 - \tilde{w}_p(x)\} + o(N^{-1}h^{d_c}), \end{aligned} \tag{3.25}$$

$$\begin{aligned} cov\{\hat{\theta}_3(x), \hat{\theta}_{51}(x)\} &= N^{-2} \sum_{i,j} \big[cov\{b_1^i(x), b_2^j(x)\} \big] \\ &= o(N^{-1}h^{d_c}), \end{aligned} \tag{3.26}$$

$$cov\left[\hat{\theta}_{51}(x), \{\hat{\theta}_{52}(x) + \hat{\theta}_{53}(x)\}\right] = N^{-3} \sum_{i,j,k \in U} \left[cov\{b_2^i(x), a_1^{j,k}(x)\} - cov\{b_2^i(x), a_2^{j,k}(x)\}\right]$$
$$= o(N^{-1}h^{-d_c}). \quad (3.27)$$

Hence $var\{\hat{\theta}_5(x)\}$ is concluded from (3.22)-(3.27). To derive $cov\{\hat{\theta}_2(x), \hat{\theta}_5(x)\}$, we need the following. Let $b_3^i(x) = \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i \in \mathcal{P}}$, then

$$cov[\hat{\theta}_2(x), \{\hat{\theta}_3(x) + \hat{\theta}_{51}(x)\}] = N^{-2} \sum_{i,j \in U} \left[cov\{b_3^i(x), b_1^j(x)\} + cov\{b_3^i(x), b_2^j(x)\}\right]$$
$$= \frac{\lambda^{(2)}}{Nh^{d_c}} R(K) g(x) p^2(x) + o(N^{-1}h^{d_c}), \quad (3.28)$$

$$cov[\hat{\theta}_2(x), \{\hat{\theta}_{52}(x) + \hat{\theta}_{53}(x)\}] = N^{-3} \sum_{i,j,k \in U} \left[cov\{b_3^i(x), a_1^{j,k}(x)\} - cov\{b_3^i(x), a_2^{j,k}(x)\}\right]$$
$$= o(N^{-1}h^{d_c}). \quad (3.29)$$

Therefore, from expansion (3.19) and using (3.22)-(3.29),

$$var\{\hat{p}_2(x)\} = \frac{1}{g^2(x)f^2(x)} var\{\hat{\theta}_5(x)\} + \frac{p^2(x)}{g^2(x)f^2(x)} var\{\hat{\theta}_2(x)\}$$
$$+ \frac{2p(x)}{g^2(x)f^2(x)} cov\{\hat{\theta}_5(x), \hat{\theta}_2(x)\} + o(N^{-1}h^{-dc})$$
$$= \frac{1}{Nh^{d_c}} \frac{p(x)\{1-p(x)\}}{f^2(x)} \left[\lambda^{(2)} R(K) \tilde{w}_p(x) + 2\lambda^{(3)} R_2(K)\{f(x) - \tilde{w}_p(x)\}\right.$$
$$\left. + \frac{\lambda^{(4)} R_3(K)\{f(x) - \tilde{w}_p(x)\}^2}{\tilde{w}_p(x)}\right] + o(N^{-1}h^{-d_c}).$$

CHAPTER 4. A Nonparametric Approach in Population Size Estimation

This Chapter focuses on the statistical analysis in the population size estimation. Motivated by the features of the human population census, for instance the data collected from the 2000 US Census, we propose using a nonparametric kernel smoothing method in estimating the population size. The application of the nonparametric methods has the following features. In dual system estimation, in-consistent estimation of the enumeration probability function can result in systematic bias in the population size, which is so-called correlation bias. As demonstrated in Chapter 3, the proposed nonparametric estimation of the enumeration probability functions is consistent and hence is free of correlation bias. The proposed nonparametric method also includes smoothing of the available discrete variables, sex, housing tenure status and etc. Therefore it is capable of sufficiently utilizing the data information available and provides opportunity of efficiency gain. To incorporate the information from missing values, an imputation based on nonparametric smoothing is proposed for the un-resolved enumeration and correct enumeration statuses in Chapter 3. The theoretical and empirical performance of the proposed nonparametric methods of population size estimation are studied in this chapter.

Based on the capture-recapture design, the Hovitz-Thompson type estimator of population size is given by

$$\hat{N} = \sum_{i \in \mathcal{S}} \frac{1}{\hat{p}_i}, \tag{4.1}$$

where $\mathcal{S}$ is some collection of samples and $\hat{p}_i$ is the estimated enumeration probability of the i^{th} individual. Various versions of population size estimation can be formulated by using different $\hat{p}_i$.

In this chapter, we will firstly demonstrate the statistical properties of the post-stratification in a relatively simple setting without erroneous enumeration and missing values. We will quantity the correlation bias by studying the population size estimator using post-stratification. Then the proposed nonparametric methods for estimating the $p(x)$ and $e(x)$ in Chapter 3 are implemented in the population size estimation with statistical properties explored.

This Chapter is structured as follows. Section 4.1 overviews the population size estimation and the correlation bias resulting from inconsistent estimator of enumeration probability function $p(x)$. The statistical properties of the population size estimation using the nonparametric methods introduced in Chapter 3 are studied in Section 4.2. The issues of erroneous enumerations and missing values are considered. Section 4.3 gives simulation studies of the population size estimation and Section 4.4 provides a comprehensive US Census data analysis on the census count and ACE. All technical details are given in Section 4.5.

4.1 Overview

Let U be the collection of the census records of size N, $\mathcal{P}$ and $\mathcal{E}$ be the collection of P- and E- samples, $\tilde{\mathcal{E}}$ be the collection of correct enumerations in the E-sample. Let $I_{i\in\mathcal{S}} = 1$ if the i^{th} individual with covariate X_i in the population U is included in set $\mathcal{S}$ and 0 otherwise, for $\mathcal{S} = \mathcal{P}$ or $\mathcal{E}$ or $\tilde{\mathcal{E}}$. Let n_1, n_2 be the sizes of the E- and P-samples and let n_3 be the number of re-captures in the P-sample. Then under the assumption that the enumeration probability is homogeneous across the whole population, the enumeration probability of the E-sample can be estimated by $\hat{p} = n_3/n_2$. Plugging in this estimation

of enumeration probability into (4.1) results in the Petersen's estimator (Petersen, 1896),

$$\hat{N} = \frac{n_1 n_2}{n_3}. \tag{4.2}$$

Or the population size estimation can be formulated by $\hat{p} = n_3/n_1$ estimating the enumeration probability of the P sample. This ends up with the same population size estimator.

It is called symmetric when the re-capture information is available in both P- and E- samples, i.e. the enumeration probability function of the P-sample/E-sample can be estimated from the re-capture information in the E-sample/P-sample. Parametric approach based on logistic regression is proposed in the case of symmetric samples (Alho et al., 1993), which attempts to fully utilized the information from both samples. Considering only continuous auxiliary variables being available in the samples, Chen and Lloyd (2000) show that

$$\hat{N} = \sum_{i \in \mathcal{P} \cup \mathcal{E}} \frac{1}{\hat{p}(X_i)}$$

improves the efficiency of population size estimation compared to $\hat{N} = \sum_{i \in \mathcal{E}} \frac{1}{\hat{p}_e(X_i)}$, where $\hat{p}(x) = \hat{p}_p(x) + \hat{p}_e(x) - \hat{p}_p(x)\hat{p}_e(x)$, $\hat{p}_e(x)$ and $\hat{p}_p(x)$ are nonparametric estimators of the E- and P-samples enumeration probability functions. In practice, the symmetric samples are not automatically available. For instance, in the US Census ACE, match is only conducted for individuals in the P-sample, which is called the one way approach. The one way approach causes in-complete recapture information in the E-sample. Hence, the symmetric samples based formulation may not be feasible (Alho et al., 1993). In this chapter, we consider the case when matching information is only available in the P-sample.

Assume $\{X_i\}_{i=1}^N$ be iid from some super population with probability density $f(x)$, $P(I_{i\in\mathcal{E}} = 1|X_i = x) = p(x)$, $P(I_{i\in\mathcal{P}} = 1|X_i = x) = g(x)$. Let $\hat{\eta}_1 = N^{-1} \sum_{i \in U} I_{\mathcal{E}}(X_i)$, $\hat{\eta}_2 = N^{-1} \sum_{i \in U} I_{\mathcal{P}}(X_i)$ and $\hat{\eta}_3 = N^{-1} \sum_{i \in U} I_{\mathcal{P}}(X_i) I_{\mathcal{E}}(X_i)$, $\eta_1 = \int_{\mathcal{X}} p(x) f(x)$, $\eta_2 = \int_{\mathcal{X}} g(x) f(x)$

and $\eta_3 = \int_{\mathcal{X}} p(x)g(x)f(x)$, then Petersen's estimator can be written as

$$\hat{N} = N\frac{\hat{\eta}_1\hat{\eta}_2}{\hat{\eta}_3}. \tag{4.3}$$

An expansion of (4.3) is given by the following,

$$\hat{N} = N\left\{\frac{\eta_1\eta_2}{\eta_3} + \frac{\eta_1}{\eta_3}(\hat{\eta}_2 - \eta_2) - \frac{\eta_1\eta_2}{\eta_3^2}(\hat{\eta}_3 - \eta_3) + \frac{\eta_2}{\eta_3}(\hat{\eta}_1 - \eta_1) + O_p(N^{-1})\right\}. \tag{4.4}$$

Since $\hat{\eta}_j$, $j = 1, 2, 3$, are unbiased estimators of η_j, by taking expectation on (4.4), we have

$$E(\hat{N}) = N\frac{\eta_1\eta_2}{\eta_3} + O(1). \tag{4.5}$$

Noting that $var(\hat{\eta}_i) = N^{-1}\eta_i(1-\eta_i)$, $cov(\hat{\eta}_1, \hat{\eta}_2) = N^{-1}\{\eta_3 - \eta_1\eta_2\}$, $cov(\hat{\eta}_1, \hat{\eta}_3) = N^{-1}\eta_3(1-\eta_1)$ and $cov(\hat{\eta}_2, \hat{\eta}_3) = N^{-1}\eta_3(1-\eta_2)$, the variance of (4.3) is obtained by taking variance operation on (4.4),

$$var(\hat{N}) = N\left[\frac{\eta_1\eta_2}{\eta_3^2}\left\{(1-\eta_1)(1-\eta_2) + (\eta_3 - \eta_1\eta_2)\left(2 - \frac{1}{\eta_3}\right)\right\}\right] + O(1). \tag{4.6}$$

And if the enumeration probability functions $p(x)$ and $g(x)$ are constants across $\mathcal{X}$, we can show that (4.6) is actually

$$var(\hat{N}) = N\frac{(1-p)(1-g)}{pg}, \tag{4.7}$$

where p and g are constant enumeration probabilities of the E- and P- sample.

We note from (4.5) that in case either $p(x)$ or $g(x)$ is a constant function over $\mathcal{X}$, $\eta_1\eta_2/\eta_3 = 1$, which means no correlation bias exist. But in realty, the enumeration probability functions usually are heterogeneous among different groups of people. On the other hand, based on the samples, the information of the P-sample enumeration may be less understood. For instance, as mentioned earlier, the US Census Bureau only conducted the so-called one-way match, i.e. only individuals in the P-sample are matched to the E-sample ones while not from the other direction. And hence it is not

feasible to verify whether the homogeneous enumeration probability of the P-sample is the case or not.

In the 2000 US Census ACE, the post-stratification is employed to account for the heterogeneity of the enumeration probabilities. Let $X = (X^c, X^u)$ be the vector of variables affecting the enumeration and correct enumeration probability functions, and let $\mathcal{X} = (\mathcal{X}^c, \mathcal{X}^u)$ be the support of X. The post-stratification essentially partitions $\mathcal{X}$ into K non-overlapping parts, say $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K$, and assuming $p(x)$ is constant over each $\mathcal{X}_k$, $k \in \{1, \ldots, K\}$. Let n_{1k}, n_{2k} and n_{3k} be the size of E-sample, P-sample and the joint part of E- and P- samples on $\mathcal{X}_k$. Then by Petersen's estimator of the population size on each part $\mathcal{X}_k$, the population size estimator based post-stratification is

$$\hat{N}_p = \sum_{k=1}^{K} \frac{n_{1k} n_{2k}}{n_{3k}}. \tag{4.8}$$

The (4.2) is actually a special case when $K = 1$, i.e. assuming homogeneous enumeration probability of the E-sample across $\mathcal{X}$. For the post-stratification situation, define $\hat{\eta}_{1k} = N^{-1} \sum_{i \in U} I_{\mathcal{E}}(X_i) I_{\mathcal{X}_k}(X_i)$, $\hat{\eta}_{2k} = N^{-1} \sum_{i \in U} I_{\mathcal{P}}(X_i) I_{\mathcal{X}_k}(X_i)$, $\hat{\eta}_{3k} = N^{-1} \sum_{i \in U} I_{\mathcal{E}}(X_i) I_{\mathcal{P}}(X_i) I_{\mathcal{X}_k}(X_i)$ and $\hat{N}_k = N \frac{\hat{\eta}_{1k} \hat{\eta}_{2k}}{\hat{\eta}_{3k}}$. Then for each k, the following expansion

$$\hat{N}_k = N \left\{ \frac{\eta_{1k} \eta_{2k}}{\eta_{3k}} + \frac{\eta_{1k}}{\eta_{3k}} (\hat{\eta}_{2k} - \eta_{2k}) - \frac{\eta_{1k} \eta_{2k}}{\eta_{3k}^2} (\hat{\eta}_{3k} - \eta_{3k}) + \frac{\eta_{2k}}{\eta_{3k}} (\hat{\eta}_{1k} - \eta_{1k}) + O_p(N^{-1}) \right\} \tag{4.9}$$

holds. By $\hat{N}_p = \sum_{k=1}^{K} \hat{N}_k$, we have

$$\begin{aligned} E(\hat{N}_p) &= N \sum_{k=1}^{K} \alpha_k + O(1), \\ Var(\hat{N}_p) &= N \sum_{k=1}^{K} \left[\frac{\eta_{1k} \eta_{2k}}{\eta_{3k}^2} \left\{ (1 - \eta_{1k})(1 - \eta_{2k}) + (\eta_{3k} - \eta_{1k} \eta_{2k}) \left(2 - \frac{1}{\eta_{3k}} \right) \right\} \right] \\ &+ O(1) \end{aligned} \tag{4.10}$$

where $\alpha_k = \frac{\eta_{1k} \eta_{2k}}{\eta_{3k}}$,$\eta_{1k} = \int_{\mathcal{X}_k} p(x) f(x)$, $\eta_{2k} = \int_{\mathcal{X}_k} g(x) f(x)$ and $\eta_{3k} = \int_{\mathcal{X}_k} p(x) g(x) f(x)$.

The so-called correlation bias takes place when $\sum_k \alpha_k \neq 1$, which corresponds to the case that the enumeration probabilities $p(x)$ and $g(x)$ are not piecewise constant over $\mathcal{X}_k$.

And this correlation bias results from the in-consistency in estimating the enumeration probability function using post-stratification. As demonstrated in case study of Chapter 3, the construction of homogeneous strata based on age and other available variables is hard. And the left heterogeneity may still be significant. Theoretically speaking, if the total number of post-strata K increases to ∞, while the size of each stratum shrinks to 0, the enumeration probabilities functions in each stratum can be estimated consistently. In finite sample application, as the number of K increasing, the chance of observing no data in some given strata becomes high. To avoid empty strata and reduce variability, several small size strata are usually combined. This counters the effort of reducing the heterogeneity.

4.2 Nonparametric Approach: Population Size Estimation

4.2.1 Effect of Erroneous Enumerations

As mentioned in Chapter 3, the Horvitz-Thompson type population size estimator (4.1) is not applicable when erroneous enumerations and missing values present. Consistent nonparametric estimators of $p(x)$ are studied in Chapter 3 and is readily to be extended to estimate the correct enumeration probability function $e(x)$. In case when erroneous enumeration exist, we assume the true population size is

$$\tilde{N} = N \int_{\mathcal{X}} e(x) f(x) dx.$$

In the ideal case when un-resolve cases are absent, the estimators of $p(x)$ and $e(x)$ are given by

$$\begin{aligned} \hat{p}_0(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{E}} I_{i\in\mathcal{P}}}{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{P}}} \text{ and} \\ \hat{e}_0(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}}}{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}}}, \end{aligned} \tag{4.11}$$

where $\mathcal{K}_{h_k,\lambda_k}(x,y)$, $k=1,2$ is the kernel function defined by

$$\mathcal{K}_{h_k,\lambda_k}(x,y) = h_k^{-d_c} K\left(\frac{x^c - y^c}{h_k}\right) \prod_{j=1}^{d_u} \left\{ \lambda_{kj} I(x_j^u = y_j^u) + \frac{1-\lambda_{kj}}{c_j - 1} I(x_j^u \neq y_j^u) \right\},$$

where $K(\cdot)$ is a d_c-dimensional probability density function, h_k is the bandwidth, $\vec{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kd_u})$ is the bandwidth vector for smoothing categorical covariates. The nonparametric approach avoids subjectively stratifying the sample space of the available variables. Instead, the approach allows data speak for themselves.

Furthermore, by smoothing the discrete variables, the nonparametric approach utilizes information from the "neighbors". This is in contrast to the post-stratification approach which only uses the data within a given stratum. The nonparametric approach assigns weights to data according to the distances defined by the kernel function. The closer the data to the object point, which is in the sense of the smaller distance, the higher weight is assigned. And by the choice of smoothing parameters h and $\vec{\lambda}$, the nonparametric approach provides opportunity of efficiently gain.

It has been demonstrated that the nonparametric approach in estimating the $p(x)$ and $e(x)$ is consistent. We anticipate that by using the nonparametric estimator $\hat{p}(x)$ and $\hat{e}(x)$, the estimates of the population size can overcome the correlation-bias in Petersen's estimator used by the post-stratification approach.

By using the nonparametric estimates of the enumeration probability and correct enumeration probability functions, the resulting estimator of the population size is given by

$$\hat{N}_0 = \sum_{i \in \mathcal{E}} \frac{\hat{e}_0(X_i)}{\hat{p}_0(X_i)}. \tag{4.12}$$

The following theorem summarizes the statistical properties of $\hat{N}_0$, whose proof is given in Section 4.5. Define $\tilde{N} = N \int_{\mathcal{X}} e(x) f(x) dx$, which is the true population size where

erroneous enumeration exist. Let

$$\begin{aligned}
b_{0,c}(x;h_1) &= \frac{1}{2}h_1^2\sigma_K^2\frac{\mathrm{tr}\left[\nabla^2\{p(x)f(x)g(x)\}-p(x)\nabla^2\{g(x)f(x)\}\right]}{g(x)f(x)},\\
b_{0,u}(x;\vec{\lambda}_1) &= \sum_{y^u\in\mathcal{C}_x^u}\left(\frac{1-\beta_{\lambda_1}(x^u,y^u)}{\alpha(x^u,y^u)-1}\left[\frac{g(x^c,y^u)f(x^c,y^u)}{g(x)f(x)}\{p(x^c,y^u)-p(x)\}\right]\right),\\
b_{0,c}(x;h_2) &= \frac{1}{2}h_2^2\sigma_K^2\frac{\mathrm{tr}\left[\nabla^2\{e(x)p(x)f(x)\}-e(x)\nabla^2\{p(x)f(x)\}\right]}{p(x)f(x)} \text{ and}\\
b_{0,u}(x;\vec{\lambda}_2) &= \sum_{y^u\in\mathcal{C}_x^u}\left(\frac{1-\beta_{\lambda_2}(x^u,y^u)}{\alpha(x^u,y^u)-1}\left[\frac{p(x^c,y^u)f(x^c,y^u)}{p(x)f(x)}\{e(x^c,y^u)-e(x)\}\right]\right),
\end{aligned}$$

where $\mathcal{C}_{x^u}=\{y^u:\sum_{j=1}^{d_u}I(x_j^u=y_j^u)=1\}$,

$$\alpha(x^u,y^u)=\sum_{k=1}^{d_u}c_kI(x_k^u=y_k^u)\quad\text{and}\quad\beta_{\lambda_k}(x^u,y^u)=\sum_{l=1}^{d_u}\lambda_{kl}I(x_l^u=y_l^u)$$

for $k=1,2$ are similarly defined as those in Chapter 3. And let $\lambda_k^{(a)}=\prod_{k=1}^{d_u}\lambda_k^a$

$$r_0(x) = \xi_{1,K}\lambda_1^{(1)}\lambda_2^{(1)}\frac{\{1-p(x)\}\rho(x)}{g(x)f(x)p(x)}\text{ and }v_0(x)=\lambda_1^{(2)}R(K)\frac{p(x)(1-p)(x)}{g(x)f(x)},$$

where $\sigma_K^2=\int t^2K(t)dt$, $\xi_{1,K}=\int K(t)K(\frac{h_1}{h_2}t)dt$ and $R(K)=\int K^2(t)dt$.

Theorem 4.1 *Under the regularity conditions given in Section 4.5, let* $h=\min(h_1,h_2)$, $A(\vec{\lambda})=\max_{\substack{1\le j\le d_u\\ k\in\{1,2\}}}(1-\lambda_{jk})$,

$$\begin{aligned}
E(\hat{N}_0) &= \tilde{N}+N\int_{\mathcal{X}}\sum_{j=1}^{2}S_j(x)\left\{b_{0,c}(x;h_j)+b_{0,u}(x;\vec{\lambda}_j)\right\}f(x)dx-\frac{1}{h_2^{d_c}}\int_{\mathcal{X}}\frac{r_0(x)}{p(x)}f(x)dx\\
&+ \frac{1}{h_1^{d_c}}\int_{\mathcal{X}}\frac{e(x)v_0(x)}{p^2(x)}f(x)dx+o\left[N\left\{h^2+A(\vec{\lambda})\right\}+\frac{1}{h^{d_c}}\right]\\
var(\hat{N}_0) &= N\left\{\int_{\mathcal{X}}\frac{e^2f}{p}-\left(\int_{\mathcal{X}}ef\right)^2+\lambda_1^{(2)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{pg}-2\lambda_1^{(1)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{p}\right.\\
&+ \left.\lambda_2^{(2)}\int_{\mathcal{X}}\frac{e(1-e)f}{p}-2\lambda_1^{(1)}\lambda_2^{(1)}\int_{\mathcal{X}}\frac{e(1-p)\rho f}{p^2g}\right\}+o(N),
\end{aligned}$$

where $S_1(x)=-e(x)/p(x)$, $S_2(x)=1$ *and* $\rho(x)=cov\{I_{i\in\mathcal{E}}I_{i\in\mathcal{P}},I_{i\in\tilde{\mathcal{E}}}|X_i=x\}$.

In the integration expressions in the variance part of Theorem 4.1 and the following theorems to be presented, the dummy variable x is suppressed in the functions, i.e. $\int q$ represents $\int q(x)dx$.

Note that when no erroneous enumerations exist, i.e. $e(x) = 1$, and only continuous variables are available, the variance given in Theorem 4.1 reduces to

$$var(\hat{N}_0) = N \int_{\mathcal{X}} \frac{(1-p)(1-g)f}{pg} + o(N), \tag{4.13}$$

which is exactly the variance in Chen and Lloyd (2002) where a nonparametric approach is applied in estimating the population size by smoothing continuous variables with no erroneous enumerations. The form of (4.13) is with similar structure to that of (4.7), which is the variance of the Petersen's estimator of population size in the ideal case when both $p(x)$ and $g(x)$ are constant functions.

The variance of $\hat{N}_0$ is $O(N)$. The bias terms incurring in Theorem 4.1 are standard ones from smoothing, in particular, $b_{0,c}(x;h)$ from smoothing the continuous variable and $b_{0,u}(x;\vec{\lambda})$ from smoothing of the discrete covariate. Different from the systematic bias, in terms of α_k in (4.7), $b_{0,u}(x;\vec{\lambda}_j)$ and $b_{0,c}(x,h_j)$ shrink to 0 when $N \to \infty$. Another terms in the bias of $\hat{N}_0$ are of magnitude $O(h_k^{-d_c})$, $k = 1, 2$. We note that $h_k^{-d_c} = o(N)$, as we require $Nh_k^{-d_c} \to \infty$ as $N \to \infty$. This means $\hat{N}_0/\tilde{N}$ converges to 1 in probability and in L^2, while it is not the case for the post-stratification if the enumeration function $p(x)$ and correct enumeration function $e(x)$ are not piecewise constants over the strata.

The effect of smoothing the discrete variable can be distinguished by those $\lambda_k^{(a)}$ terms. To simplify th discussion, without loss of generality, we assume $\lambda_{k1} = \lambda_{k2} = \cdots = \lambda_{kd_u} = \lambda_k$ for $k = 1, 2$. Ideally, $\rho(x)$ should be 0 by operational independence of the two surveys resulting P- and E-samples. And if $\rho(x) = 0$, by smoothing discrete variables, a variance reduction at the second order is induced. This is by noting the fact that $\lambda = 1 + o(1)$, i.e, $\lambda \to 1$ as $N \to \infty$, and

$$\lambda_k^{(a)} = 1 - a \sum_{j=1}^{d_u} (1 - \lambda_{kj}) + O\{(1-\lambda)^2\}.$$

The the second order variance can be quantified by

$$-2N \left\{ d_u(1-\lambda_1) \int_{\mathcal{X}} \frac{e^2(1-p)f}{pg} - d_u(1-\lambda_1) \int_{\mathcal{X}} \frac{e^2(1-p)f}{p} + d_u(1-\lambda_2) \int_{\mathcal{X}} \frac{e(1-e)f}{p} \right\},$$

which is strictly less than 0. We see that by smoothing the discrete variables, a second order variance reduction is achieved when $\rho(x) = 0$.

4.2.2 Effect of Missing Values

In the presence of un-resolved matching and correct enumeration statuses, Chapter 3 studies the properties of the following two estimators. The first one ignores the missing values and estimate $p(x)$ and $e(x)$ based on complete data only:

$$\begin{aligned} \hat{p}_1(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{E}} I_{i\in\mathcal{P}} \delta_i}{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{P}} \delta_i} \text{ and} \\ \hat{e}_1(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} \eta_i}{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}} \eta_i}, \end{aligned} \tag{4.14}$$

where $\delta_i = 1(0)$ if i is an observation in the P-sample with resolved (un-resolved) matching status and $\eta_i = 1(0)$ if i is an observation in the E-sample with resolved (un-resolved) correct enumeration status. The straightforward approach of ignoring the un-resolve cases may result in in-consistency in the parametric inference (Little and Rubin, 2002). As the estimator (4.14) is in a form of ratio, Chapter 3 shows that the bias incurs from ignoring the un-resolved cases in the numerator and denominator cancel each other. And the $\hat{e}_1(x)$ and $\hat{p}_1(x)$ are still consistent estimators of $e(x)$ and $p(x)$. Comparing with (4.11), the variances of $\hat{e}_1(x)$ and $\hat{p}_1(x)$ are inflated as showed in Chapter 3. Recall that we assume missingness of the enumeration and correct enumeration status are given by

$$\begin{aligned} E(\delta_i | I_{i\in\mathcal{E}}, I_{i\in\mathcal{P}}, X_i = x, Z_i = z) &= w_p(x, z) \text{ and} \\ E(\eta_i | I_{i\in\mathcal{E}}, I_{i\in\tilde{\mathcal{E}}}, X_i = x, Z_i = z) &= w_e(x, z), \end{aligned} \tag{4.15}$$

i.e. missing at random (MAR) (Rosenbaum and Rubin, 1983) given the covariate (X_i, Z_i). The reason initials the MAR assumption of the form (4.15) is the fact the variables affecting the missing mechanism may be beyond those explaining the enumeration probability functions. For instance, in the US Census data, the final un-resolved

cases in both enumeration and correct enumeration status are results after several stages of follow-ups (Belin et al., 1993). We expect that the effects during the follow-up may affect the missing mechanism but may be irrelevant to the enumeration status.

The estimator of the population size is then

$$\hat{N}_1 = \sum_{i \in \mathcal{E}} \frac{\hat{e}_1(X_i)}{\hat{p}_1(X_i)}. \tag{4.16}$$

Define

$$\begin{aligned}
b_{1,c}(x;h_1) &= \frac{1}{2}h_1^2\sigma_K^2 \frac{\mathrm{tr}\left[\nabla^2\{p(x)g(x)\tilde{w}_p(x)\} - p(x)\nabla^2\{g(x)\tilde{w}_p(x)\}\right]}{g(x)\tilde{w}_p(x)}, \\
b_{1,u}(x;\vec{\lambda}_1) &= \sum_{y^u \in \mathcal{C}_x^u} \left(\frac{1-\beta_{\lambda_1}(x^u,y^u)}{\alpha(x^u,y^u)-1} \left[\frac{g(x^c,y^u)\tilde{w}_p(x^c,y^u)}{g(x)\tilde{w}_p(x)} \{p(x^c,y^u)-p(x)\} \right] \right), \\
b_{1,c}(x;h_2) &= \frac{1}{2}h_2^2\sigma_K^2 \frac{\mathrm{tr}\left[\nabla^2\{e(x)p(x)\tilde{w}_e(x)\} - e(x)\nabla^2\{p(x)\tilde{w}_e(x)\}\right]}{p(x)\tilde{w}_e(x)}, \\
b_{1,u}(x;\vec{\lambda}_2) &= \sum_{y^u \in \mathcal{C}_x^u} \left(\frac{1-\beta_{\lambda_2}(x^u,y^u)}{\alpha(x^u,y^u)-1} \left[\frac{p(x^c,y^u)\tilde{w}_e(x^c,y^u)}{p(x)\tilde{w}_e(x)} \{e(x^c,y^u)-e(x)\} \right] \right), \\
r_1(x) &= \xi_{1,K}\lambda_1^{(1)}\lambda_2^{(1)} \frac{\tilde{w}_b(x)\{1-p(x)\}\rho(x)}{g(x)\tilde{w}_p(x)p(x)\tilde{w}_e(x)} \text{ and} \\
v_1(x) &= \lambda_1^{(2)} R(K) \frac{p(x)\{1-p(x)\}}{g(x)\tilde{w}_p(x)},
\end{aligned}$$

where $\mathcal{C}_{x^u}$, $\alpha(x^u,y^u)$ and $\beta_{\lambda_k}(x^u,y^u)$ are samely defined as those in Theorem 4.1. The properties of $\hat{N}_1$ is summarized in the following theorem.

Theorem 4.2 *Under the regularity conditions given in Section 4.5, let* $h = \min(h_1,h_2)$, $A(\vec{\lambda}) = \max\limits_{\substack{1\le j\le d_u \\ k\in\{1,2\}}} (1-\lambda_{jk})$,

$$\begin{aligned}
E(\hat{N}_1) &= \tilde{N} + N\int_{\mathcal{X}} \sum_{j=1}^{2} S_j(x)\left\{ b_{1,c}(x;h_j) + b_{1,u}(x;\vec{\lambda}_j) \right\} f(x)dx - \frac{1}{h_2^{d_c}}\int_{\mathcal{X}} \frac{r_1(x)}{p(x)} f(x)dx \\
&+ \frac{1}{h_1^{d_c}}\int_{\mathcal{X}} \frac{e(x)v_1(x)}{p^2(x)} f(x)dx + o\left[N\left\{ h^2 + A(\vec{\lambda}) \right\} + \frac{1}{h^{d_c}} \right] \\
var(\hat{N}_1) &= N\left\{ \int_{\mathcal{X}} \frac{e^2 f}{p} - \left(\int_{\mathcal{X}} ef \right)^2 + \lambda_1^{(2)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{pg\tilde{w}_p} - 2\lambda_1^{(1)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{p} \right. \\
&+ \left. \lambda_2^{(2)} \int_{\mathcal{X}} \frac{e(1-e)f}{p\tilde{w}_e} - 2\lambda_1^{(1)}\lambda_2^{(1)} \int_{\mathcal{X}} \frac{\tilde{w}_b}{\tilde{w}_p\tilde{w}_e} \frac{e(1-p)\rho f^2}{p^2 g} \right\} + o(N),
\end{aligned}$$

where $S_1(x) = -e(x)/p(x)$, $S_2(x) = 1$.

The form of Theorem 4.2 is quite similar to that of Theorem 4.1. By similar arguments as that in Theorem 4.1, $\hat{N}_1/\tilde{N}$ converges to 1 in probability and in L^2 as $N \to \infty$.

We may note some effect of the missing values in $var(\hat{N}_1)$, also under the ideal case $\rho(x) = 0$. Comparing the variance in Theorem 4.2 and that in Theorem 4.1, we see a variance inflation by noting

$$\int_{\mathcal{X}} \frac{e^2(1-p)f}{pg\tilde{w}_p} \geq \int_{\mathcal{X}} \frac{e^2(1-p)f}{pg} \text{ and } \int_{\mathcal{X}} \frac{e(1-e)f}{p\tilde{w}_e} \geq \int_{\mathcal{X}} \frac{e(1-e)f}{p}.$$

Similar to estimating the $p(x)$ and $e(x)$, the missing values bring in loss of efficiency. Similar to the case in $var(\hat{N}_0)$, we note that by smoothing discrete variables, a second order variance reduction by $\lambda_k^{(a)}$ is also associated with $\hat{N}_1$.

The nonparametric estimator (4.16) ignores all un-resolved cases and hence uses no information from the available X associated with those individuals. Though the enumeration status or the correct enumeration status are un-resolved, the X may at least partially recorded. We consider the approach of imputing the un-resolved status by using the nonparametric estimates of $e(x)$ and $p(x)$ given by (4.14), which have been discussed in Chapter 3. This leads to the second approach when the missing values present.

The second estimator is based on imputing the missing values with the estimated conditional means.

$$\begin{aligned} \hat{p}_2(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{P}} \{I_{i\in\mathcal{E}}\delta_i + \hat{p}_1(X_i)(1-\delta_i)\}}{\sum_{i\in U} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_i) I_{i\in\mathcal{P}}} \text{ and} \\ \hat{e}_2(x) &= \frac{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}} \{I_{i\in\tilde{\mathcal{E}}}\eta_i + \hat{e}_1(X_i)(1-\eta_i)\}}{\sum_{i\in U} \mathcal{K}_{h_2,\vec{\lambda}_2}(x, X_i) I_{i\in\mathcal{E}}}. \end{aligned} \tag{4.17}$$

Under some mild assumptions, in Chapter 3 we show that (4.17) have smaller variance in the leading order than (4.14). By using (4.17), the resulting estimator of the population size is

$$\hat{N}_2 = \sum_{i\in\mathcal{E}} \frac{\hat{e}_2(X_i)}{\hat{p}_2(X_i)}. \tag{4.18}$$

Let

$$
\begin{aligned}
b_{2,c}(x;h_1) &= b_{0,c}(x;h_1) + \frac{b_{1,c}(x;h_1)\{f(x)-\tilde{w}_p(x)\}}{f(x)},\\
b_{2,u}(x;\vec{\lambda}_1) &= b_{0,u}(x;\vec{\lambda}_1) + \frac{b_{1,u}(x;\vec{\lambda}_1)\{f(x)-\tilde{w}_p(x)\}}{f(x)},\\
b_{2,c}(x;h_2) &= b_{0,c}(x;h_2) + \frac{b_{1,c}(x;h_2)\{f(x)-\tilde{w}_e(x)\}}{f(x)},\\
b_{2,u}(x;\vec{\lambda}_2) &= b_{0,u}(x;\vec{\lambda}_2) + \frac{b_{1,u}(x;\vec{\lambda}_2)\{f(x)-\tilde{w}_e(x)\}}{f(x)},\\
r_2(x) &= \frac{\rho(x)}{f^2(x)}\sum_{l=1}^{4}\left[\xi_{l,K}\{r_{2,l}(x)\}\right] \text{ and}\\
v_2(x) &= \frac{p(x)\{1-p(x)\}}{g(x)f^2(x)}\left\{\lambda_1^{(2)}K^{(2)}(0)\tilde{w}_p(x) + \lambda_1^{(3)}2K^{(3)}(0)\{f(x)-\tilde{w}_p(x)\}\right.\\
&+ \left.\lambda_1^{(4)}K^{(4)}(0)\frac{\{f(x)-\tilde{w}_p(x)\}^2}{\tilde{w}_p(x)}\right\},
\end{aligned}
$$

where $r_{2,1}(x) = \lambda_1^{(1)}\lambda_2^{(1)}\tilde{w}_b(x)$, $r_{2,2}(x) = \lambda_1^{(1)}\lambda_2^{(2)}\frac{\tilde{w}_b(1-\tilde{w}_e)(x)}{\tilde{w}_e(x)}$, $r_{2,3}(x) = \lambda_1^{(2)}\lambda_2^{(1)}\frac{\tilde{w}_b(1-\tilde{w}_p)(x)}{\tilde{w}_p(x)}$, $r_{2,4}(x) = \lambda_1^{(2)}\lambda_2^{(2)})\frac{\tilde{w}_b(1-\tilde{w}_p)(1-\tilde{w}_e)(x)}{\tilde{w}_p\tilde{w}_e(x)}$. And the quantities associated with the kernel K are defined by $\xi_{2,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(\frac{h_1}{h_2}t_1 - t_2)dt_1dt_2$, $\xi_{3,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1 - t_2)dt_1dt_2$, $\xi_{3,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1 - t_2)dt_1dt_2$, $\xi_{4,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1 - t_2)K(\frac{h_1}{h_2}t_1 - t_3)dt_1dt_2dt_3$and $K^{(j)}(x)$ is the j^{th} convolution of the kernel $K(x)$.

The following theorem gives the properties of $\hat{N}_2$.

Theorem 4.3 *Under the regularity conditions given in Section 4.5, let* $h = \min(h_1, h_2)$, $A(\vec{\lambda}) = \max_{\substack{1\le j\le d_u\\ k\in\{1,2\}}}(1-\lambda_{jk})$,

$$
\begin{aligned}
E(\hat{N}_2) &= \tilde{N} + N\int_{\mathcal{X}}\sum_{j=1}^{2}S_j(x)\left\{b_{2,c}(x;h_j) + b_{2,u}(x;\vec{\lambda}_j)\right\}f(x)dx - \frac{1}{h_2^{d_c}}\int_{\mathcal{X}}\frac{r_2(x)}{p(x)}f(x)dx\\
&+ \frac{1}{h_1^{d_c}}\int_{\mathcal{X}}\frac{e(x)v_2(x)}{p^2(x)}f(x)dx + o\left[N\left\{h^2 + A(\vec{\lambda})\right\} + \frac{1}{h^{d_c}}\right]
\end{aligned}
$$

$$
\begin{aligned}
var(\hat{N}_2) &= N\left[\int_{\mathcal{X}} \frac{e^2 f}{p} - \left(\int_{\mathcal{X}} ef\right)^2 - 2\lambda_1^{(2)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{p}\right. \\
&+ \int_{\mathcal{X}} \frac{e^2(1-p)}{pg}\left\{\lambda_1^{(2)}\tilde{w}_p + 2\lambda_1^{(3)}(f-\tilde{w}_p) + \lambda_1^{(4)}\frac{(f-\tilde{w}_p)^2}{\tilde{w}_p}\right\} \\
&+ \int_{\mathcal{X}} \frac{e(1-e)}{p}\left\{\lambda_2^{(2)}\tilde{w}_e + 2\lambda_2^{(3)}(f-\tilde{w}_e) + \lambda_2^{(4)}\frac{(f-\tilde{w}_e)^2}{\tilde{w}_e}\right\} \\
&- 2\int_{\mathcal{X}} \frac{e(1-p)\rho}{p^2 g}\left\{\lambda_1^{(1)}\lambda_2^{(1)}\tilde{w}_b + \lambda_1^{(2)}\lambda_2^{(1)}\frac{(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p} + \lambda_1^{(1)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)\tilde{w}_b}{\tilde{w}_e}\right. \\
&+ \left.\left.\lambda_1^{(2)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p\tilde{w}_e}\right\} + o(N),
\end{aligned}
$$

where $S_1(x) = -e(x)/p(x)$, $S_2(x) = 1$.

It is true that $\lambda_k^{(a)} = 1 + o(1)$ and

$$\tilde{w}_p + 2(f-\tilde{w}_p) + \frac{(f-\tilde{w}_p)^2}{\tilde{w}_p} = \frac{1}{\tilde{w}_p} \text{ and } \tilde{w}_e + 2(f-\tilde{w}_e) + \frac{(f-\tilde{w}_e)^2}{\tilde{w}_e} = \frac{1}{\tilde{w}_e}.$$

Under the ideal case $\rho(x) = 0$, by comparing Theorem 4.3 with Theorem 4.2, we note that $var(\hat{N}_1)$ and $var(\hat{N}_2)$ agree with each other in the leading order terms. Since we have shown in Chapter 3 that $\hat{p}_2(x)$ improves the variance of $\hat{p}_1(x)$ in the leading order, it is a little surprising that the imputation approach does not improve the estimation of the population in the leader order terms of variance.

From the proofs in Section 4.5, we note that the variances of nonparametric population size estimators $\hat{N}_k$ is determined by

$$\sum_{i,j\in U} cov\{\hat{p}_k(X_i), \hat{p}_k(X_j)\}, \sum_{i,j\in U} cov\{\hat{e}_k(X_i), \hat{e}_k(X_j)\} \text{ and } \sum_{i,j\in U} cov\{\hat{p}_k(X_i), \hat{e}_k(X_j)\}$$

for $k = 1, 2$. As for X and Y are independent and follow the pdf $f(\cdot)$, it is true that

$$
\begin{aligned}
cov\{\hat{e}_2(X), \hat{e}_2(Y)\} &= cov\{\hat{e}_1(X), \hat{e}_1(Y)\}, \\
cov\{\hat{p}_2(X), \hat{p}_2(Y)\} &= cov\{\hat{p}_1(X), \hat{p}_1(Y)\} \text{ and} \\
cov\{\hat{p}_2(X), \hat{e}_2(Y)\} &= cov\{\hat{p}_1(X), \hat{e}_1(Y)\}
\end{aligned}
$$

in the leading order of magnitude. Therefore, the $N(N-1)$ off diagonal terms of $\hat{N}_2$ agree with those of $\hat{N}_1$ in the leading order terms. The contribution of $var\{\hat{p}_2(X)$ and

$var\{\hat{e}_2(X)\}$ only have N terms and hence is of smaller order. This explains why the imputation does not improve the variance of the population size estimation in the leading order.

Though by imputation $\hat{N}_2$ has no improvement in $O(N)$, we note some second order variance reductions. Assume $\lambda_{k1} = \lambda_{k2} = \cdots = \lambda_{kd_u} = \lambda_k$ for $k = 1, 2$, we have

$$\lambda_1^{(2)} \left\{ \tilde{w}_p + 2\lambda_1^{(1)}(f - \tilde{w}_p) + \lambda_1^{(2)} \frac{(f - \tilde{w}_p)^2}{\tilde{w}_p} - \frac{1}{\tilde{w}_p} \right\} < 0 \text{ and}$$
$$\lambda_2^{(2)} \left\{ \tilde{w}_e + 2\lambda_2^{(1)}(f - \tilde{w}_e) + \lambda_1^{(2)} \frac{(f - \tilde{w}_e)^2}{\tilde{w}_e} - \frac{1}{\tilde{w}_e} \right\} < 0.$$

Therefore, the imputation in $\hat{N}_2$ still brings in variance reduction in the second order.

4.3 Simulation Studies

To demonstrate the performance of the nonparametric approach in population size estimation, the following simulations were conducted. The complete setting of simulations includes specification of X, $\mathcal{X}$, $p(x)$, $g(x)$ and $e(x)$, as well as the missing propensity function $w_p(x)$ and $w_e(x)$.

Motivated by the situation of human census, the choice of X is the following. Let $X = (X_1, \ldots, X_5)$ be the available variables, where $X_1 \in [0, 70]$ is continuous (age), $X_2, X_3 \in \{0, 1\}$ are discrete variables of two levels (sex and housing tenure status) and $X_4 \in \{0, 1, 2, 3\}$ and $X_5 \in \{0, 1, \ldots, 6\}$ are two discrete variables of 4 and 7 levels (region and race/origin domains). Then $\mathcal{X} = [0, 70] \times \{0, 1\}^2 \times \{0, 1, 2, 3\} \times \{0, 1, \ldots, 6\}$, where the set $A \times B = \{(x, y), x \in A \text{ and } y \in B\}$. Without loss of generality, each $X_i \in U$ was set to be independent and followed a uniform distribution over $\mathcal{X}$.

The setting of $p(x)$ and $e(x)$ incorporated the heterogeneity from both continuous and discrete variables. Let $P\{I_{i\in\mathcal{E}} = 1|X_i = x\} = p(x) = [1 + \exp\{-b_1(x)\}]^{-1}$, $P\{I_{i\in\mathcal{P}} = 1|X_i = x\} = g(x) = [1 + \exp\{-b_2(x)\}]^{-1}$. The second component of the simulation is the specification of $e(x)$, i.e. the probability function of correct enumeration. Let

$P\{I_{i\in\tilde{\mathcal{E}}} = 1|X_i = x\} = e(x) = [1 + \exp\{-b_3(x)\}]^{-1}$. The $b_1(x)$, $b_2(x)$ and $b_3(x)$ are nonlinear functions defined by

$$b_i(x) = \beta_{il0} + \beta_{il1}x_1 + \beta_{il2}\phi\left(\frac{x_1 - \beta_{il3}}{\beta_{il4}}\right), \tag{4.19}$$

where $i = 1, 2, 3$, $\phi(x)$ is the probability density function of standard normal distribution and $l = 56x_2 + 28x_3 + 7x_4 + x_5 + 1$ is a 1-1 onto mapping from $\{\{0, 1\}^2 \times \{0, 1, 2, 3\} \times \{0, 1, \ldots, 6\}$ to $\{1, \ldots, 112\}$. Let $\beta_{il} = (\beta_{il0}, \beta_{il1}, \beta_{il2}, \beta_{il3}, \beta_{il4})$, $i \in \{1, 2, 3\}$ and $l \in \{1, \ldots, 112\}$ be the vector of parameters in the nonlinear model (4.19) and set β_{il} follows some multivariate normal distribution with mean μ_{β_i} and variance covariance matrix Σ_i, i.e. $\beta_{il} \sim N(\mu_{\beta_i}, \Sigma_i)$. The idea of this setting is that the heterogeneity from the continuous variable age was represented by employing the normal probability density function which displays a sharp drop in enumeration probability at some age. And the heterogeneity between groups, which are defined by distinctive levels of discrete variables, is represented by the random coefficient β_{il}. Essentially, in different groups, the enumeration probability functions are different. The pattern of function $p(x)$ resulting from the nonlinear function (4.19) evaluated at μ_{β_1}, which was the setting in the simulation, is displayed in Figure 4.1.

The simulation also involved the un-resolved enumeration and correct enumeration statuses by setting the following two binary random variables δ and η. Let $P(\delta = 1|I_{\mathcal{E}}(X), I_{\mathcal{P}}(X), X = x) = w_p(x)$ and $P(\eta = 1|I_{\mathcal{E}}(X), I_{\tilde{\mathcal{E}}}(X), X = x) = w_p(x)$, i.e. missing at random MAR given the variable X. The specifications of $w_p(x)$ and $w_e(x)$ were similar to (4.19), in particular, $w_p(x) = [1 + \exp\{-c_1(x)\}]^{-1}$ and $w_e(x) = [1 + \exp\{-c_2(x)\}]^{-1}$. And for $i = 1, 2$

$$c_i(x) = \theta_{il0} + \theta_{il1}x_1 + \theta_{il2}\phi\left(\frac{x_1 - \theta_{il3}}{\theta_{il4}}\right), \tag{4.20}$$

where $\theta_{il} \sim N(\mu_{\theta_i}, \Omega_i)$.

The values of the parameters set in the simulation are given in Table 4.1. Two cases of the size, $N = 5,000$ and $N = 10,000$ were conducted. In this particular setting, the

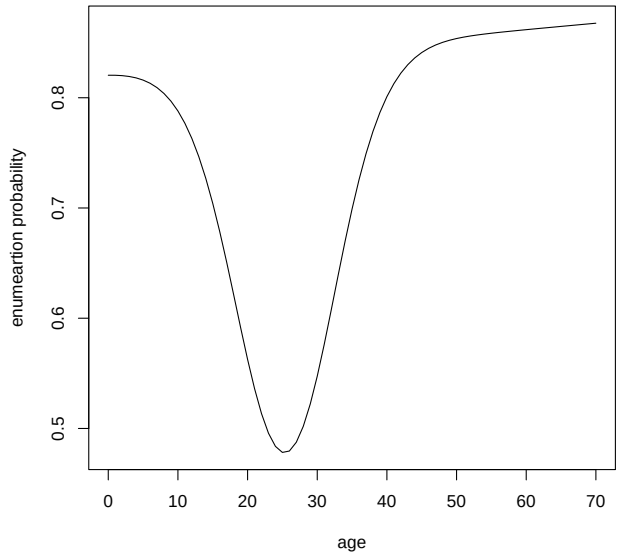

Figure 4.1 Simulation setting of the $p(x)$ at μ_{β_1}.

$p(x)$ and $g(x)$, i.e. the enumeration probability functions of the E- and P-samples, are around 0.6-0.8. The correct enumeration probability $e(x)$, is about 0.90-0.95. And the missing propensities $w_p(x)$ and $w_e(x)$ are about 0.75-0.90.

As the true population size $\tilde{N} = N \int_{\mathcal{X}} e(x)f(x)dx$, the quantity $\int_{\mathcal{X}} e(x)f(x)dx$ is needed in the numerical studies of the nonparametric methods. The numerical integration using Monte Carlo simulation was applied in evaluating this quantity. In particular, by generating large number, say $M = 10^5$, of X following $f(x)$, and evaluating $E\{e(X)\}$ numerically by $\bar{e}(X) = M^{-1} \sum e(X)$. The numerical integration $\int_{\mathcal{X}} e(x)f(x)dx = 0.918$ for the setting in Table 4.1.

The nonparametric approach was also compared to the post-stratification in the population size estimation. The post-strata were constructed as follows. Subdividing the age $[0, 70]$ into 4 parts, in particular $[0, 20)$, $[20, 35)$, $[35, 50)$ and $[50, 70]$, then the four age strata cross the 112 cells defined by the discrete variables resulted in 448 strata. By applying the Petersen's estimator on each strata and summed the estimates up, the

population size estimation was obtained. In case empty cell appears in the simulation, the empty cell merged to the neighbor cell, with the neighbor defined by the closeness of age group.

One approach obtaining the smoothing bandwidth h_i and λ_i is by minimizing the CV function.

$$CV_p(h_1, \vec{\lambda_1}) = n^{-1} \sum_{i \in \mathcal{P}} \{I_{i \in \mathcal{E}} - \hat{p}^{(-i)}_{h_1, \lambda_1}(X_i)\}^2 \delta_i.$$

For estimation of $e(x)$, the bandwidths were chosen by minimizing

$$CV_e(h_2, \lambda_2) = n^{-1} \sum_{i \in \mathcal{E}} \{I_{i \in \tilde{\mathcal{E}}} - \hat{e}^{(-i)}_{h_2, \lambda_2}(X_i)\}^2 \eta_i,$$

where $\vec{\lambda} = (\lambda_1, \cdots, \lambda_4)$ corresponding to the four discrete covariates, $\hat{p}^{(-i)}_{h_1, \lambda_1}(x)$ and $\hat{e}^{(-i)}_{h_2, \lambda_2}(x)$ are the estimators of $p(x)$ and $e(x)$ after excluding the i^{th} data pair.

In the simulation, for estimating $\hat{p}(x)$ and $\hat{e}(x)$, it was set that $\lambda_{11} = \lambda_{12} = \lambda_{13} = \lambda_{14} = \lambda_1$ and $\lambda_{21} = \lambda_{22} = \lambda_{23} = \lambda_{24} = \lambda_2$, i.e., to smooth the discrete components using one bandwidth. The pre-run of the simulation gave the average values of the bandwidths. In particular, for $N = 5,000$, $\bar{h}_{1,cv} = 6.5(1.1)$, $\bar{h}_{2,cv} = 6.4(1.0)$, $\bar{\lambda}_{1,cv} = 0.83(0.2)$ and $\bar{\lambda}_{1,cv} = 0.86(0.2)$, where the values in the brackets were the standard deviations of the averages from the pre-running of simulation. And for $N = 10,000$, $\bar{h}_{1,cv} = 5.7(1.2)$, $\bar{h}_{2,cv} = 5.9(1.0)$, $\bar{\lambda}_{1,cv} = 0.85(0.2)$ and $\bar{\lambda}_{1,cv} = 0.83(0.2)$. In the simulation, values of smoothing parameters were selected by combination of values h_k and λ_k around the mean values given by minimizing the CV functions in the pre-running.

Tables 4.2 and 4.3 display the bias, variance and mean square error(MSE) of $\hat{N}_0$, $\hat{N}_1$ and $\hat{N}_2$, corresponding to two sets of N values, $N = 5,000$ and $N = 10,000$. For each sample size, results corresponding to six sets of bandwidths combinations are reported. The results of the post-stratifications based estimator $\hat{N}_p$ are reported on the top two lines of each table.

As we can see from Tables 4.2 and 4.3, the post-stratification based estimator, was most biased, about 2% measured by percentage. And the relative bias level was not

affected when N increased. While for the nonparametric population size estimators $\hat{N}_0$, $\hat{N}_1$ and $\hat{N}_2$, we see that as the N increased from 5,000 to 10,000, the relative bias dropped from around 1% to below 1%. This is consistent with our theoretical finds in Theorems 4.1-4.3. By applying nonparametric kernel smoothing in estimating the $p(x)$ and $e(x)$, the resulting estimators are consistent. And by applying such estimators in the population size estimation, no correlation bias incurred. While for the post-stratification, the remaining hetergenity caused a bias of stable level.

We note that in the results it was the bias dominating the MSE for the nonparametric population size estimation. The MSEs of all three nonparametric population size estimators were smaller than that of $\hat{N}_p$. And in Tables 4.2 and 4.3, $\hat{N}_1$ displayed the best performance, whose biases were consistently smallest. By incorporating imputation, $\hat{N}_2$ had smaller variance than $\hat{N}_1$. But the biases of $\hat{N}_2$ were slightly larger than those of $\hat{N}_1$. Our Theorems 4.2 and 4.3 imply that $\hat{N}_1$ and $\hat{N}_2$ have the same leading order terms in the variance, while $\hat{N}_2$ has variance reduction in the second order. And this was confirmed by the simulation. We also note that the variances of $\hat{N}_0$, $\hat{N}_1$ and $\hat{N}_2$ generally became smaller for fixed h_1 and h_2 as the λ increased. This represents the second order variance reduction from smoothing the discrete variables.

4.4 Census Data Analysis

In this section, we report the procedure of implementing the nonparametric approach on the US Census 2000 ACE data. The results of the the analysis is available in (Chen, Tang, and Mule, 2008). As the main purpose of the 2010 Census is providing information for future census improvement, the identification of the errors and inaccuracy from various components is a task from analyzing the ACE data (Bell and Cohen, 2007). We will demonstrate in this section how the nonparametric approach can be utilized to provide information on census omissions, erroneous enumerations and the compound

estimation of the one number census count.

The data used in this study are those from US Census Bureau's 2000 Accuracy and Coverage Evaluation (US Census Bureau, 2004). The sample size, i.e. the number of individuals contained in the samples, are $712,900$ and $721,734$ for E- and P- samples respectively. The E- and P-samples are properly weighted representing a multiphase sampling procedure from the entire US. In this study, the variables incorporated in the nonparametric estimation of population size included the ROAST (Race/Original(7 levels), Age(continuous), Sex(2 levels), Housing Tenure(2 levels)), geographical region(4 levels)).

The ideal hypothetical estimator of the population size is given by

$$\hat{N}_{null} = \sum_{i \in \mathcal{E}} \frac{e(X_i)}{p(X_i)},$$

where $\mathcal{E}$ is the collection of census records. By plugging in the estimators of $e(x)$ and $p(x)$ based on the ACE E- and P-samples, we have

$$\hat{N}'_k = \sum_{i \in \mathcal{E}} \frac{\hat{e}_k(X_i)}{\hat{p}_k(X_i)},$$

where $k = 1, 2$ corresponding to the population size estimators introduced in Section 4.2. In practice, $\hat{N}'_k$ is not feasible, as not every census record in $\mathcal{E}$ has enough information X. In the US Census ACE, data with fewer than two characteristics recorded were removed from the E-sample. And hence those removed were therefore excluded from matching to the P-sample and correct enumeration determination. In the US Census 2000 ACE, the number of the removed records is about $8,000,000$. A record in the census with two or more recorded characteristic is called a data defined person. Let $\mathcal{D}$ be the collection of data defined person in the census. Then the applicable population size estimation is given by

$$\hat{N}_k = \sum_{i \in \mathcal{D}} \frac{\hat{e}_k(X_i)}{\hat{p}_k(X_i)}. \tag{4.21}$$

The estimator of the form (4.21) is actually what were used for population size estimation in the US Census 2000 ACE. As long as $\hat{p}_k(x)$ based on the resulting P- and E-samples is consistent to the probability that an individual being included in $\mathcal{D}$, (4.21) is still a reasonable estimator to the true population size. As $\mathcal{D} \subset \mathcal{E}$, we can write the nonparametric estimation of $p(x)$ in the following form

$$\hat{p}(x) = \frac{\sum_{i\in\mathcal{P}} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{E}\cap\mathcal{D}}}{\sum_{i\in\mathcal{P}} \mathcal{K}_{h,\vec{\lambda}}(x, X_i)} = \frac{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}} I_{i\in\mathcal{D}}}{\sum_{i\in U} \mathcal{K}_{h,\vec{\lambda}}(x, X_i) I_{i\in\mathcal{P}}}.$$

Then if no data defined problem exist in the P-sample enumeration procedure, the nonparametric estimator still estimate $q(x) = P(I_{i\in\mathcal{D}}|X_i = x)$ consistently. Therefore, (4.21) still estimate the size of the population as long as the P-sample data collection is ideal. And similarly, the nonparametric estimator $\hat{e}(x)$ still estimates the correct enumeration probability function $e(x)$ under an appropriate data situation. In the following discussion, without causing confusion keeping the notation consistent, $\mathcal{E}$ represents the set of data defined.

In this study, the discrete variables selected (Race/Original(7 levels), Sex(2 levels), Housing Tenure(2 levels), geographical region(4 levels)) effectively define 112 distinctive groups. Let $\mathcal{E}_1, \ldots, \mathcal{E}_{112}$ be the corresponding partition of the data defined individuals in the census, i.e. $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_{112}$, then the population size can be estimated for each group,

$$\hat{N}_{k,d} = \sum_{i\in\mathcal{E}_d} \frac{\hat{e}_k(X_i)}{\hat{p}_k(X_i)}. \tag{4.22}$$

Let $\{N_{C,d}\}_{d=1}^{112}$ be the census count in these 112 groups and the total census count $N_C = \sum_{d=1}^{112} N_{C,d}$, then the undercounts corresponding to the total census count and counts in each of the 112 groups are estimated by

$$u_{C,k} = \frac{\hat{N}_k - N_C}{N_C} \text{ and } u_{d,k} = \frac{\hat{N}_{k,d} - N_{C,d}}{N_{C,d}}, \tag{4.23}$$

for $k = 1, 2$ representing the two versions of estimators for $e(x)$ and $p(x)$.

In implementing the nonparametric estimators of $e(x)$, $p(x)$ and then the population size estimation. We have some further technical issues to be considered.

4.4.1 Bandwidth Selection

As the amount of data available from the ACE files differs dramatically in different Race/Origin domains, the smoothing bandwidths are needed to specified appropriately to represent the fact. In implementing the nonparametric method in estimating $p(x)$ and $e(x)$ from the US Census ACE data, a two stage data adaptive bandwidth selection procedure was conducted, i.e. the smoothing parameters depend on the abundance of data near the point of interest. The first stage specified the bandwidth h for the smoothness of continuous variable age. And the second stage chose the smoothing bandwidth $\vec{\lambda} = (\lambda_1, \ldots, \lambda_4)$. When conducting the first stage bandwidth selection, the $\vec{\lambda}$ was set to be 1, then the corresponding CV function for $\hat{e}(x)$ and $\hat{p}(x)$ are minimized, where on each Race/Origin domain the h was assigned an individual value. While in the second stage, the bandwidths h chosen from the first stage were fixed, and the $\vec{\lambda}$ were chosen by minimizing the corresponding CV function again. In particular, for estimating the enumeration probability function $p(x)$, let

$$\hat{p}^{-i}_{h,\vec{\lambda}}(x) = \frac{\sum\limits_{j\in\mathcal{P}, j\neq i} \mathcal{K}_{h,\vec{\lambda}}(x, X_j) I_{j\in\mathcal{E}} \delta_j}{\sum\limits_{j\in\mathcal{P}, j\neq i} \mathcal{K}_{h,\vec{\lambda}}(x, X_j) \delta_j}$$

be the complete cases based estimator leaving out the i^{th} individual. Let $\vec{b} = (b_1, \ldots, b_7)$ be the vector of bandwidth corresponding to the 7 Race/Origin domains and $h(x^u) = b_r$, if x^u belongs to the r^{th} Race/Origin domain, where $r \in \{1, \ldots, 7\}$. Then the first stage of bandwidth selection minimizes the following cross-validation function

$$CV_c(\vec{b}) = n_p^{-1} \sum_{i\in\mathcal{P}} \{I_{i\in\mathcal{E}} - \hat{p}^{-i}_{h(X_i^u),\vec{1}}(X_i)\}^2 \delta_i, \tag{4.24}$$

where n_p is the P-sample size. Fixe the $\vec{b}$ selected and then minimizes

$$CV_u(\vec{\lambda}) = n_p^{-1} \sum_{i \in \mathcal{P}} \{I_{i \in \mathcal{E}} - \hat{p}^{-i}_{h(X_i^u),\vec{\lambda}}(X_i)\}^2 \delta_i. \tag{4.25}$$

Denote the resulting bandwidth by $(\vec{b}_{cv}, \vec{\lambda}_{cv})$ as the smoothing parameters, when carrying out $\hat{p}(x)$, the bandwidth smoothing the age is $h(x^u) = b_{cv,r}$, if x^u belongs to the r^{th} Race/Origin domain. And the smoothing bandwidth for discrete covariate is $\vec{\lambda}_{cv}$. We may extend the procedure to the estimation of $e(x)$ automatically. Though this data adaptive bandwidth selection procedure depends on one component in the discrete covariate, we note that it could be utilized in more general data adaptive bandwidth selection.

The reason for such a bandwidth selection procedure is the following. The purpose of the nonparametric smoothing is to capture the heterogeneity in $p(x)$, $e(x)$ and the age is a most important continuous variable. The first stage of bandwidth selection essentially restricts the smoothing within each group and allows a data based optimal balance between the bias and variance over the smoothing of the age. While in the second stage, the $\vec{\lambda}$ are chosen by choosing a balance point of the bias and variance on borrowing information from neighboring groups. By such a procedure, the different data dense level in different domains are taken account into the selection and the resulting bandwidth avoids over-smooth in domain with abundant data and under-smooth in domains with sparse data.

4.4.2 Boundary Bias Correction

When smoothing the age as a continuous variable, the so-called boundary bias incurs when estimating $p(x^c, x^u)$ at $x^c < h$, where h is the smoothing bandwidth. Since the age has the lower bound 0, it is know that in the boundary region, the bias is of $O(h)$ which is a larger order of magnitude than $O(h^2)$ (Silverman, 1986).

The following bias modification approach is proposed by Rice (1984) in the boundary region. For $x < h$, let $\rho = x/h$, $R(\rho) = w_1(\rho)/w_0(\rho)$, where $w_k(\rho) = \int_{-1}^{\rho} t^k K(t)dt$, the boundary bias corrected estimator is given by

$$\hat{m}'_h(x) = \hat{m}_h(x) - \beta \left\{ \hat{m}_h(x) - \hat{m}_{\alpha,h}(x) \right\}, \tag{4.26}$$

where $\hat{m}_h(x)$ is the nonparametric estimator of the regression function $m(x)$ using bandwidth h in smoothing the continuous variable, and $\hat{m}_{\alpha,h}(x)$ is the same nonparametric estimator using bandwidth αh. The β is defined as

$$\beta = \frac{R(\rho)}{\alpha R(\rho/\alpha) - R(\rho)}.$$

Rice (1984) shows that by choosing $\alpha = 2 - \rho$, the bias of the Jackknife type estimator (4.26) is of $O(h^2)$. In the US Census ACE data analysis, we implemented (4.26) in the boundary region for $\hat{p}_k(x)$ and $\hat{e}_k(x)$, $k = 1, 2$ in estimating $p(x)$ and $e(x)$ when the estimating takes place at $x^c < h_k$.

4.4.3 Small Groups Treatment

In the US population, the population sizes of some domains are small. This is also reflected in the ACE data files. For instance, the sample sizes from the domain of Native Hawaiian or Pacific Islander(NHPI) in the ACE E- and P- samples are less than 3000. In the post-stratifications, the collapse of strata was implemented to avoid strata empty or with too few data. In particular, all regions in NHPI are combined as one and essentially only sex and housing tenure status were used in creating post-strata. The combining of regions also represented the geographical residency distribution of the NHPI group. For detail, see (US Census Bureau, 2004).

When dealing with domains with sparse data, we implemented bandwidth adjustment which has the same effect as strata combining. This is the remedy for domains with few data. In particular, when estimating $p(x)$ and $e(x)$ in domain of Native Hawaiian or

Pacific Islander, the component in bandwidth $\vec{\lambda}_k$ corresponding to region were assigned uniform value 1. While in the American Indian or Alaska Native on/off Reservations and Non-Hispanic Asian, the regions were only discriminated as whether the individual is in West or not, i.e. equivalent to a discrete variable with only 2 levels rather than 4. The discrete kernel used when estimating $p(x)$ and $e(x)$ in such domains was just the one used for smoothing binary variables, for instance sex and housing tenure.

4.4.4 Variance Estimation

The Jackknife estimation of variance (Shao and Wu, 1989) is a commonly used method in estimating variance in survey sampling related area (Wolter, 2007 and Shao and Tu, 1995). For $X_1, \ldots, X_n$ iid from some distribution F and the estimator $\hat{\theta}$ being some function of $X_1, \ldots, X_n$, the delete-1 Jackknife estimator of the variance is with the form

$$v_{J(1)} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}^{-i} - \hat{\theta})^2, \tag{4.27}$$

where $\hat{\theta}^{-i}$ is the estimator calculated based on a sample leaving the i^{th} observation out. For $\hat{\theta}$ being some smooth function, then (4.27) estimates the variance consistently. Alternative group Jackknife method (Wu, 1986) can be used for nonsmooth function $\hat{\theta}$.

The delete-1 Jackknife variance estimation is not applicable in the Census ACE data analysis. First of all, the primary sampling units (PSU) in the ACE samples are the clusters of households. Leaving one individual or even one household out of the final samples ignores the dependence from the sampling scheme and hence the resulting variance estimation is biased. By representing the multi-phase sampling procedure constructing the ACE samples, US Census Bureau (2004) conducted variance estimation through the leaving out one PSU of the first sampling phase.

Because the leave out one PSU of the first phase of sampling results in a large number of replicates (around 30,000), we implemented the following leave one group

out Jackknife estimation of variance. Based on the clusters where the individuals are sampled from, 100 pseudo-random groups are constructed according to the last two digits of the corresponding cluster numbers. Let $\hat{p}_k^{-j}(x)$ and $\hat{e}_k^{-j}(x)$ be the estimation of $p(x)$ leaving the j^{th} pseudo-random group out, for $k = 1, 2$. The ratio $r(x) = e(x)/p(x)$ can be estimated by $\hat{r}_k(x) = \hat{e}_k(x)/\hat{p}_k(x)$, whose replicate can be produced by $\hat{r}_k^{-j}(x) = \hat{e}_k^{-j}(x)/\hat{p}_k^{-j}(x)$. Then based on the replicates, the variance of $\hat{p}_k(x)$, $\hat{e}_k(x)$ and $\hat{r}_k(x)$ can be estimated by

$$\begin{aligned} v_{p,k}(x) &= \frac{99}{100}\sum_{j=1}^{100}\{\hat{p}_k^{-j}(x) - \hat{p}_k(x)\}^2, \\ v_{e,k}(x) &= \frac{99}{100}\sum_{j=1}^{100}\{\hat{e}_k^{-j}(x) - \hat{e}_k(x)\}^2 \text{ and} \\ v_{r,k}(x) &= \frac{99}{100}\sum_{j=1}^{100}\{\hat{r}_k^{-j}(x) - \hat{r}_k(x)\}^2. \end{aligned} \tag{4.28}$$

And the the replicates of the populations size estimation are given by

$$\hat{N}_k^{-j} = \sum_{i\in\mathcal{E}} \frac{\hat{e}_k^{-j}(X_i)}{\hat{p}_k^{-j}(X_i)}.$$

Therefore the variance estimation of $\hat{N}_k$ is given by

$$v_{N,k} = \frac{99}{100}\sum_{j=1}^{100}(\hat{N}_k^{-j} - \hat{N}_k)^2. \tag{4.29}$$

Replicates of the population size estimation for those 112 groups defined by the discrete variable utilized in this study can be carried out in the same way,

$$\hat{N}_{k,d}^{-j} = \sum_{i\in\mathcal{E}_d} \frac{\hat{e}_k^{-j}(X_i)}{\hat{p}_k^{-j}(X_i)}.$$

And the corresponding variance estimation is given by

$$v_{k,d} = \frac{99}{100}\sum_{j=1}^{100}(\hat{N}_{k,d}^{-j} - \hat{N}_{k,d})^2,$$

where $\hat{N}_{k,d}$ is given by (4.22).

4.5 Discussions

In Chapters 3 and 4, we propose a local post-stratification based on nonparametric kernel estimation for the enumeration and correct enumeration probabilities in the US Census dual system surveys. The local post-stratification can capture the underlying data characteristics objectively and is free of the risk of model mis-specification. Comparing with the existing post-stratification, it avoids construction of post-strata and accounts for the correlation bias in the estimation. From the empirical results in Chapters 3 and 4, we discover a potential use of the local post-stratification is in providing guidance for appropriate parametric model selection, by detecting informative feature from huge amount of available information.

An attraction of our proposal is in smoothing the categorical variables, which is the most relevant to the Census as most of its variables are categorical. Smoothing many categorical variables does not lead to the curse of dimension as encountered in the kernel smoothing of continuous only X_i. And the census does not have many continuous variables.

The US Census dual system estimation is based on well designed surveys and data of both high quality and good quantity. It provides fresh and challenging research issues for dual system capture-recapture surveys, for instance the EEs and missing values. The proposed local stratification, although having been described in close connection to the Census, is applicable for other capture-recapture surveys after minor modifications. The proposed imputation based estimation for nonparametric regression is generally applicable well beyond the dual system surveys, as it concerns improving estimation efficiency in the presence of missing values.

Parameter	Mean	Variance-Covariance Matrix
β_{1l}	$\mu_{\beta_1} = (1.53, 0.005, -35.0, 20.5, 10.0)$	$\Sigma_1 = \text{diag}(0.1, 0.001, 1.0, 2.0, 1.0)$
β_{2l}	$\mu_{\beta_2} = (1.80, 0.01, -32.0, 22.0, 9.0)$	$\Sigma_2 = \text{diag}(0.1, 0.001, 1.0, 2.0, 1.0)$
β_{3l}	$\mu_{\beta_3} = (2.56, -0.001, -7.0, 24.0, 18.0)$	$\Sigma_3 = \text{diag}(0.1, 0.001, 1.0, 2.0, 1.0)$
θ_{1l}	$\mu_{\theta_1} = (1.90, 0.005, -20.0, 25.0, 10.0)$	$\Omega_1 = \text{diag}(0.2, 0.001, 1.5, 2.5, 1.5)$
θ_{2l}	$\mu_{\theta_2} = (1.90, 0.005, -25.0, 22.0, 11.0)$	$\Omega_2 = \text{diag}(0.2, 0.001, 1.5, 2.5, 1.5)$

Table 4.1 Simulation Setting 1 of Population Size Estimation

Estimator	Bias	Relative Bias	Variance	MSE
$\hat{N}_p$	-85.26	-0.019	2494	9763.64
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.85$			
$\hat{N}_0$	-71.68	-0.016	958	6095.56
$\hat{N}_1$	-32.17	-0.007	1854	2889.18
$\hat{N}_2$	-46.38	-0.010	1684	3834.96
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.8$			
$\hat{N}_0$	-72.38	-0.016	934	6172.99
$\hat{N}_1$	-48.28	-0.011	1542	3872.70
$\hat{N}_2$	-54.27	-0.012	1498	4443.29
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.75$			
$\hat{N}_0$	-65.08	-0.014	908	5143.42
$\hat{N}_1$	-50.01	-0.010	1480	3981.12
$\hat{N}_2$	-51.70	-0.011	1430	4103.01
	$h_1 = 7.5$, $h_2 = 7.5$, $\lambda_1 = \lambda_2 = 0.85$			
$\hat{N}_0$	-68.55	-0.015	956	5655.14
$\hat{N}_1$	-36.87	-0.008	1708	3067.26
$\hat{N}_2$	-48.61	-0.011	1492	3854.68
	$h_1 = 7.5$, $h_2 = 7.5$, $\lambda_1 = \lambda_2 = 0.8$			
$\hat{N}_0$	-68.38	-0.015	926	5601.54
$\hat{N}_1$	-49.36	-0.011	1500	3936.25
$\hat{N}_2$	-54.38	-0.012	1448	4404.69
	$h_1 = 7.5$, $h_2 = 7.5$, $\lambda_1 = \lambda_2 = 0.75$			
$\hat{N}_0$	-61.56	-0.013	914	4703.70
$\hat{N}_1$	-50.04	-0.011	1430	3934.05
$\hat{N}_2$	-51.51	-0.011	1372	4025.41

Table 4.2 Simulation Results of Population Size Estimation for $N = 5000$ under the setting given by Table 4.1.

Estimator	Bias	Relative Bias	Variance	MSE
$\hat{N}_p$	-163.71	-0.017	4976	31776.96
	$h_1 = 6.0$, $h_2 = 6.0$, $\lambda_1 = \lambda_2 = 0.85$			
$\hat{N}_0$	-83.77	-0.010	2260	9277.99
$\hat{N}_1$	-44.40	-0.005	3704	5675.39
$\hat{N}_2$	-58.64	-0.006	3456	6894.95
	$h_1 = 6.0$, $h_2 = 6.0$, $\lambda_1 = \lambda_2 = 0.8$			
$\hat{N}_0$	-82.59	-0.009	2172	8993.24
$\hat{N}_1$	-58.10	-0.006	3320	6696.02
$\hat{N}_2$	-64.26	-0.007	3224	7352.82
	$h_1 = 6.0$, $h_2 = 6.0$, $\lambda_1 = \lambda_2 = 0.75$			
$\hat{N}_0$	-72.83	-0.008	2068	7372.65
$\hat{N}_1$	-57.45	-0.006	3192	6492.64
$\hat{N}_2$	-59.18	-0.006	3156	6658.47
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.85$			
$\hat{N}_0$	-82.49	-0.009	2120	8923.99
$\hat{N}_1$	-47.92	-0.005	3576	5870.92
$\hat{N}_2$	-61.01	-0.007	3224	6945.83
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.8$			
$\hat{N}_0$	-80.98	-0.009	1964	8522.15
$\hat{N}_1$	-59.77	-0.007	3352	6924.98
$\hat{N}_2$	-65.58	-0.007	3248	7548.23
	$h_1 = 6.5$, $h_2 = 6.5$, $\lambda_1 = \lambda_2 = 0.75$			
$\hat{N}_0$	-71.66	-0.008	1952	7087.88
$\hat{N}_1$	-58.70	-0.006	3108	6353.37
$\hat{N}_2$	-60.50	-0.007	3096	6756.84

Table 4.3 Simulation Results of Population Size Estimation for $N = 10000$ under the setting given by Table 4.1.

4.6 Technical Proofs

Technical Assumptions

Let covariate $X_i = (X_i^c, X_i^u)$ where X_i^c is a d_c-dimensional continuous covariate and X_i^u is a d_u-dimensional unordered categorical covariate, $\mathcal{X} = \{\mathcal{X}^c, \mathcal{X}^u\}$ be the support of X_i, where $\mathcal{X}^c$ and $\mathcal{X}^u$ are the supports of X_i^c and X_i^u respectively. We assume data pairs $\{(X_i, Z_i, I_{i\in\mathcal{E}}, I_{i\in\tilde{\mathcal{E}}})\}_{i=1}^N$ are independent and identically distributed. And the following conditions are assumed in the following proofs.

C.1 Let $K(\cdot)$ be a d_c variates nonnegative, bounded and symmetric probability density function with bounded second derivative. The smoothing bandwidths satisfy that $h \to 0$ and $\max_{1\le j\le d_u}\{(1-\lambda_j)\} \to 0$ and $Nh^{d_c} \to \infty$ as $N \to \infty$.

C.2 We assume missing at random in $I_{i\in\mathcal{E}}$ and $I_{i\in\tilde{\mathcal{E}}}$, namely $P(\delta_i = 1|I_{i\in\mathcal{E}}, X_i = x, Z_i = z) = P(\delta_i = 1|X_i = x, Z_i = z) := w_p(x,z)$, $P(\eta_i = 1|I_{i\in\tilde{\mathcal{E}}}, X_i = x, Z_i = z) = P(\eta_i = 1|X_i = x, Z_i = z) := w_e(x,z)$ where $w_p(x^c, x^u, z) \ge C_w$ and $w_e(x^c, x^u, z) \ge C_w$ for a constant $C_w > 0$ and $w_p(x^c, x^u, z)$ and $w_e(x^c, x^u, z)$ have bounded continuous second partial derivative with respect to x^c within $\mathcal{X}^c$.

C.3 For any $x^u \in \mathcal{X}^u$, $p(x^c, x^u)$, $g(x^c, x^u)$, $e(x^c, x^u)$ and the probability density function $f(x^c, x^u)$ have bounded continuous second partial derivatives with respect to x^c in $\mathcal{X}^c$, and there exist $C_f > 0$ such that $f(x^c, x^u) \ge C_f$ for all $(x^c, x^u) \in \mathcal{X}$.

We will first show the variance part of Theorem 4.1-4.3. The bias parts are postponed to the end of this section.

Proof of Theorem 4.1

$$\hat{N}_0 = \sum_{i\in\mathcal{E}} \frac{\hat{e}_0(X_i)}{\hat{p}_0(X_i)} = \sum_{i\in U} \frac{\hat{e}_0(X_i)}{\hat{p}_0(X_i)} I_{i\in\mathcal{E}}. \tag{4.30}$$

Define

$$\begin{aligned}
\hat{\theta}_1(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_1,\vec{\lambda}_1}(x,X_j)I_{j\in\mathcal{P}}I_{j\in\mathcal{E}},\\
\hat{\theta}_2(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_1,\vec{\lambda}_1}(x,X_j)I_{j\in\mathcal{P}},\\
\hat{\gamma}_1(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_2,\vec{\lambda}_2}(x,X_j)I_{j\in\mathcal{E}}I_{j\in\tilde{\mathcal{E}}} \text{ and}\\
\hat{\gamma}_2(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_2,\vec{\lambda}_2}(x,X_j)I_{j\in\mathcal{E}},
\end{aligned}$$

then the expansions of $\hat{e}_0(x)$ and $\frac{1}{\hat{p}_0(x)}$ are given by

$$\begin{aligned}
\frac{1}{\hat{p}_0(x)} &= \frac{1}{p(x)}+\frac{\{\hat{\theta}_1(x)-\theta_1(x)\}}{p(x)g(x)f(x)}-\frac{\{\hat{\theta}_2(x)-\theta_2(x)\}}{p^2(x)g(x)f(x)}\{1+o_p(1)\} \text{ and}\\
\hat{e}_0(x) &= e(x)+\frac{1}{p(x)f(x)}\{\hat{\gamma}_1(x)-\gamma_1(x)\}-\frac{e(x)}{p(x)f(x)}\{\hat{\gamma}_2(x)-\gamma_2(x)\}\{1+o_p(1)\}.
\end{aligned}$$

Let $T_0(X_i)=e(X_i)I_{i\in\mathcal{E}}/f(X_i)$, $T_1(X_i)=T_{11}(X_i)+T_{12}(X_i)$, $T_2(X_i)=T_{21}(X_i)+T_{22}(X_i)$, $T_{11}(X_i) = \frac{e(X_i)I_{i\in\mathcal{E}}\hat{\theta}_2(X_i)}{p(X_i)g(X_i)f(X_i)}$ and $T_{12}(X_i) = -\frac{e(X_i)I_{i\in\mathcal{E}}\hat{\theta}_1(X_i)}{p^2(X_i)g(X_i)f(X_i)}$, $T_{21}(X_i) = \frac{I_{i\in\mathcal{E}}\hat{\gamma}_1(X_i)}{p^2(X_i)f(X_i)}$ and $T_{22}(X_i)=-\frac{I_{i\in\mathcal{E}}\hat{\gamma}_2(X_i)}{p^2(X_i)f(X_i)}$, then we have

$$\hat{N}_0=\sum_{i\in U}\left\{T_0(X_i)+T_1(X_i)+T_2(X_i)\right\}\left\{1+o_p(1)\right\}. \tag{4.31}$$

We note that $T_1(\cdot)$ and $T_2(\cdot)$ are the terms associated with $\hat{p}(\cdot)$ and $\hat{e}(\cdot)$ respectively after linearization.

To derive $Var(\hat{N}_0)$, first note that

$$var\{\sum_{i\in U}T_0(X_i)\}=var\left\{\sum_{i\in U}\frac{e(X_i)}{p(X_i)}I_{i\in\mathcal{E}}\right\}=N\left\{\int_{\mathcal{X}}\frac{e^2}{p}f-\left(\int_{\mathcal{X}}ef\right)^2\right\}, \tag{4.32}$$

where in the integral, the dummy variable x is suppressed. And the integration over $\mathcal{X}$ is defined by understanding $f(x)$ as the Radon-Nikodym derivative with respect to the product measure of Lebesgue measure and counting measure. In the following derivations, the summations are taking over the set of population U unless otherwise is stated. Define

$$s_1(X_i)=\frac{e(X_i)I_{i\in\mathcal{E}}}{p(X_i)g(X_i)f(X_i)}, s_2(X_i)=\frac{e(X_i)I_{i\in\mathcal{E}}}{p^2(X_i)g(X_i)f(X_i)} \tag{4.33}$$

and let

$$Q_1^{i,j} = s_1(X_i)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i, X_j)I_{j\in\mathcal{P}}, Q_2^{i,j} = s_2(X_i)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i, X_j)I_{j\in\mathcal{P}}I_{j\in\mathcal{E}}. \tag{4.34}$$

Let $\lambda_a^{(b)} = \prod_{k=1}^{d_u} \lambda_{ak}^b$ for $a = 1, 2$ and $b = 1, 2, 3, 4$ and $\lambda_1 = \max_{1\le k\le d_u}(\lambda_{1k})$. If $k = l$,

$$\begin{aligned}
E(Q_1^{i,k}Q_1^{j,l}) &= \int_{\mathcal{X}} s_1(x_1)s_1(x_2)\mathcal{K}_{h_1,\vec{\lambda}_1}(x_1, x_3)\mathcal{K}_{h_1,\vec{\lambda}_1}(x_2, x_3)g(x_3)f(x_1)f(x_2)f(x_3)dx_1dx_2dx_3 \\
&= \lambda_1^{(2)} \sum_{y^u\in\mathcal{X}^u} \int_{\mathcal{X}^c} \{s_1(x_1^c, y^u)s_1(x_2^c, y^u)g(x_3^c, y^u)f(x_1^c, y^u)f(x_2^c, y^u)f(x_3^c, y^u)\times \\
&\quad h_1^{-2d_c} K\left(\frac{x_1^c - x_3^c}{h_1}\right) K\left(\frac{x_2^c - x_3^c}{h_1}\right)\Big\} dx_1^c dx_2^c dx_3^c + O\{(1-\lambda_1)\} \\
&= \lambda_1^{(2)} \int_{\mathcal{X}} \frac{e^2 f}{g} + O(h^2) + O\{(1-\lambda_1)\}.
\end{aligned}$$

Similarly, by considering the case when the indices i, j, k and l are pairwise equal and ignoring smaller order terms,

$$E(Q_a^{i,k}Q_b^{j,l}) = \lambda_1^{(2)}\mu_{ab}^{i,j,k,l} + o(1), \text{ for } a, b \in \{1, 2\}, \tag{4.35}$$

where

$$\mu_{11}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i = j \\ \int_{\mathcal{X}} e^2 f & \text{if } i = l \text{ or } j = k \\ \int_{\mathcal{X}} \frac{e^2 f}{g} & \text{if } k = l \end{cases}, \mu_{12}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i = j \text{ or } i = l \\ \int_{\mathcal{X}} e^2 f & \text{if } j = k \\ \int_{\mathcal{X}} \frac{e^2 f}{pg} & \text{if } k = l \end{cases},$$

$$\mu_{21}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i = j \text{ or } j = k \\ \int_{\mathcal{X}} e^2 f & \text{if } i = l \\ \int_{\mathcal{X}} \frac{e^2 f}{pg} & \text{if } k = l \end{cases}, \mu_{22}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i = j \text{ or } i = l \\ & \text{or } j = k \\ \int_{\mathcal{X}} \frac{e^2 f}{pg} & \text{if } k = l \end{cases}.$$

Then by substitute (4.35) into following,

$$\begin{aligned} var\{\sum_i T_1(X_i)\} &= \sum_i \sum_j cov\{T_1(X_i), T_1(X_j)\} \\ &= \sum_i \sum_j [cov\{T_{11}(X_i), T_{11}(X_j)\} + cov\{T_{11}(X_i), T_{12}(X_j)\} \\ &+ cov\{T_{12}(X_i), T_{11}(X_j)\} + cov\{T_{12}(X_i), T_{12}(X_j)\}] \\ &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_1^{i,k}Q_1^{j,l} - Q_1^{i,k}Q_2^{j,l} - Q_2^{i,k}Q_1^{j,l} + Q_2^{i,k}Q_2^{j,l})\right\} \\ &= N\lambda_1^{(2)}\int_{\mathcal{X}} \frac{e^2(1-p)f}{gp} + o(N). \end{aligned} \tag{4.36}$$

In the above derivation, all terms except the case associated with $k = l$ cancel each other in the leading order. Further,

$$\begin{aligned} cov\left\{\sum_i \frac{e(X_i)I_{i\in\mathcal{E}}}{p(X_i)}, \sum_j T_1(X_i)\right\} &= N^{-1}E\left[\sum_{i,j,k}\left\{\frac{e(X_i)I_{i\in\mathcal{E}}(Q_1^{jk} - Q_2^{jk})}{p(X_i)}\right\}\right] \\ &= -N\lambda_1^{(1)}\int_{\mathcal{X}} \frac{e^2(1-p)f}{p} + o(N). \end{aligned} \tag{4.37}$$

By following exactly the same steps in establishing (4.36),

$$var\{\sum_i T_2(X_i)\} = N\lambda_2^{(2)}\int_{\mathcal{X}} \frac{e(1-e)f}{p} + o(N). \tag{4.38}$$

Let

$$s_3(X_i) = \frac{I_{i\in\mathcal{E}}(X_i)}{p^2(X_i)f(X_i)}, s_4(X_i) = \frac{e(X_i)I_{\mathcal{E}}(X_i)}{p^2(X_i)f(X_i)}, \tag{4.39}$$

$$Q_3^{i,j} = s_3(X_i)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i, X_j)I_{j\in\mathcal{E}}I_{j\in\tilde{\mathcal{E}}} \text{ and } Q_4^{i,j} = s_4(X_i)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i, X_j)I_{j\in\mathcal{E}}. \tag{4.40}$$

The covariance between T_0 and T_2 is given by

$$\begin{aligned} cov\left\{\sum_i \frac{e(X_i)I_{i\in\mathcal{E}}}{p(X_i)}, \sum_j T_2(X_i)\right\} &= N^{-1}E\left[\sum_{i,j,k}\left\{\frac{e(X_i)I_{i\in\mathcal{E}}(Q_3^{jk} - Q_4^{jk})}{p(X_i)}\right\}\right] \\ &= o(N). \end{aligned} \tag{4.41}$$

Similar to (4.35), the following can be established.

$$E(Q_a^{i,k}Q_b^{j,l}) = \lambda_1^{(1)}\lambda_2^{(1)}\mu_{ab}^{i,j,k,l} + o(1), \text{ for } a \in \{1,2\} \text{ and } b \in \{3,4\}, \tag{4.42}$$

where

$$\mu_{13}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i=j \text{ or } i=l \\ \int_{\mathcal{X}} e^2 f & \text{if } j=k \\ \int_{\mathcal{X}} \frac{ef\xi}{pg} & \text{if } k=l \end{cases}, \mu_{14}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i=j \text{ or } i=l \\ \int_{\mathcal{X}} e^2 f & \text{if } j=k \text{ or } k=l \end{cases},$$

$$\mu_{23}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i=j \text{ or } i=l \\ & \text{or } k=j \\ \int_{\mathcal{X}} \frac{ef\xi}{pg} & \text{if } k=l \end{cases}, \mu_{24}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2 f}{p} & \text{if } i=j \text{ or } i=l \\ & \text{or } j=k \text{ or } k=l \end{cases},$$

and $\xi(X_i) = E\{I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} | X_i = x\}$. Therefore,

$$\begin{aligned} cov\left\{\sum_i T_1(X_i), \sum_j T_2(X_j)\right\} &= N^{-2} E\left\{\sum_{i,j,k,l} (Q_1^{i,k} Q_3^{j,l} - Q_1^{i,k} Q_4^{j,l} - Q_2^{i,k} Q_3^{j,l} + Q_2^{i,k} Q_4^{j,l})\right\} \\ &= -N\lambda_1^{(1)}\lambda_2^{(1)} \int_{\mathcal{X}} \frac{e(1-p)\rho f}{p^2 g}, \end{aligned} \tag{4.43}$$

where

$$\rho(x) = \xi(x) - p(x)g(x)e(x) = E\{I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} | X_i = x\} - E\{I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} | X_i = x\} E\{I_{i\in\tilde{\mathcal{E}}} | X_i = x\}$$

is the covariance function between the enumeration status and correct enumeration status. As a result of (4.32), (4.36), (4.38), (4.37), (4.41) and (4.43), we establish that

$$\begin{aligned} var(\hat{N}_0) &= N\left\{\int_{\mathcal{X}} \frac{e^2 f}{p} - \left(\int_{\mathcal{X}} ef\right)^2 + \lambda_1^{(2)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{pg} - 2\lambda_1^{(1)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{p}\right. \\ &+ \left. \lambda_2^{(2)} \int_{\mathcal{X}} \frac{e(1-e)f}{p} - 2\lambda_1^{(1)}\lambda_2^{(1)} \int_{\mathcal{X}} \frac{e(1-p)\rho f}{p^2 g}\right\}. \end{aligned} \tag{4.44}$$

Proof of Theorem 4.2

The proof of Theorem 4.2 is quite similar to that of Theorem 4.1.

$$\hat{N}_1 = \sum_{i\in\mathcal{E}} \frac{\hat{e}_2(X_i)}{\hat{p}_2(X_i)} = \sum_{i\in U} \frac{\hat{e}_2(X_i) I_{i\in\mathcal{E}}}{\hat{p}_2(X_i)}. \tag{4.45}$$

Let $\tilde{w}_p(x) = \int w_p(x,z) f(x,z) dz$, $\tilde{w}_e(x) = \int w_e(x,z) f(x,z) dz$, $t_i(X_i) = s_i(X_i)/\tilde{w}_p(X_i)$ for $i = 1, 2$ and $t_i(X_i) = s_i(X_i)/\tilde{w}_e(X_i)$ for $i = 3, 4$, where $s_i(X_i)$ $i = 1, \ldots, 4$ are defined

in (4.33) and (4.39). And let

$$\begin{aligned}
\hat{\theta}_3(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_1,\vec{\lambda}_1}(x,X_j)I_{j\in\mathcal{P}}I_{j\in\mathcal{E}}\delta_j,\\
\hat{\theta}_4(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_1,\vec{\lambda}_1}(x,X_j)I_{j\in\mathcal{P}}\delta_j,\\
\hat{\gamma}_3(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_2,\vec{\lambda}_2}(x,X_j)I_{j\in\mathcal{E}}I_{j\in\tilde{\mathcal{E}}}\eta_j \text{ and}\\
\hat{\gamma}_4(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_2,\vec{\lambda}_2}(x,X_j)I_{j\in\mathcal{E}}\eta_j,
\end{aligned}$$

$T_{31}(X_i)=t_1(X_i)\hat{\theta}_4(X_i)$, $T_{32}(X_i)=-t_2(X_i)\hat{\theta}_3(X_i)$, $T_{41}(X_i)=t_3(X_i)\hat{\gamma}_3(X_i)$, $T_{42}(X_i)=t_4(X_i)\hat{\gamma}_4(X_i)$, $T_3(X_i)=T_{31}(X_i)+T_{32}(X_i)$ and $T_4(X_i)=T_{41}(X_i)+T_{42}(X_i)$. Similar to (4.31), the following expansion for (4.45) is established.

$$\hat{N}_1=\sum_{i\in U}\{T_0(X_i)+T_3(X_i)+T_4(X_i)\}\{1+o_p(1)\}. \tag{4.46}$$

Similar to (4.34) and (4.40), let

$$Q_5^{i,j}=t_1(X_i)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i,X_j)I_{j\in\mathcal{P}}\delta_j, Q_6^{i,j}=t_2(X_i)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i,X_j)I_{j\in\mathcal{P}}I_{j\in\mathcal{E}}\delta_j, \tag{4.47}$$

$$Q_7^{i,j}=t_3(X_i)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i,X_j)I_{j\in\mathcal{E}}\eta_j, Q_8^{i,j}=t_4(X_i)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i,X_j)I_{j\in\mathcal{E}}I_{j\in\tilde{\mathcal{E}}}\eta_j. \tag{4.48}$$

Therefore,

$$\begin{aligned}
var\left\{\sum_i T_3(X_i)\right\} &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_5^{i,k}Q_5^{j,l}-Q_5^{i,k}Q_6^{j,l}-Q_6^{i,k}Q_5^{j,l}+Q_6^{i,k}Q_6^{j,l})\right\}\\
&= N\lambda_1^{(2)}\int_{\mathcal{X}}\frac{e^2(1-p)f^2}{gp\tilde{w}_p}+o(N), \qquad (4.49)\\
var\left\{\sum_i T_4(X_i)\right\} &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_7^{i,k}Q_7^{j,l}-Q_7^{i,k}Q_8^{j,l}-Q_8^{i,k}Q_7^{j,l}+Q_8^{i,k}Q_8^{j,l})\right\}\\
&= N\lambda_1^{(2)}\int_{\mathcal{X}}\frac{e(1-e)f^2}{p\tilde{w}_e}+o(N), \qquad (4.50)\\
cov\left\{\sum_i T_3(X_i),\sum_j T_4(X_j)\right\} &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_5^{i,k}Q_7^{j,l}-Q_5^{i,k}Q_8^{j,l}-Q_6^{i,k}Q_7^{j,l}+Q_6^{i,k}Q_8^{j,l})\right\}\\
&= N\lambda_1^{(2)}\lambda_1^{(1)}\int_{\mathcal{X}}\frac{\tilde{w}_b}{\tilde{w}_p\tilde{w}_e}\frac{e(1-p)\rho}{p^2g}+o(N), \qquad (4.51)
\end{aligned}$$

where $\tilde{w}_b = \int w_b(x,z)f(x,z)dz$ and $w_b(x,z) = E(\delta_i\eta_i|X=x,Z=z)$. And

$$\begin{aligned} cov\left\{\sum_i T_0(X_i), \sum_j T_3(X_j)\right\} &= N^{-1}E\left\{\sum_{i,j,k}\frac{e(X_i)I_{i\in\mathcal{E}}(Q_5^{j,k}-Q_6^{j,k})}{p(X_i)}\right\} \\ &= -N\lambda_1^{(1)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{p}+o(N), \qquad (4.52) \\ cov\left\{\sum_i T_0(X_i), \sum_j T_4(X_j)\right\} &= N^{-1}E\left\{\sum_{i,j,k}\frac{e(X_i)I_{i\in\mathcal{E}}(Q_7^{j,k}-Q_8^{j,k})}{p(X_i)}\right\} \\ &= o(N). \qquad (4.53) \end{aligned}$$

Then by substitute (4.32), (4.49), (4.50), (4.51) (4.52) and(4.53) into the variance operation of (4.46), we have

$$\begin{aligned} var(\hat{N}_1) &= N\left\{\int_{\mathcal{X}}\frac{e^2f}{p}-\left(\int_{\mathcal{X}}ef\right)^2+\lambda_1^{(2)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{pg\tilde{w}_p}-2\lambda_1^{(1)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{p}\right. \\ &+ \left.\lambda_2^{(2)}\int_{\mathcal{X}}\frac{e(1-e)f}{p\tilde{w}_e}-2\lambda_1^{(1)}\lambda_2^{(1)}\int_{\mathcal{X}}\frac{\tilde{w}_b}{\tilde{w}_p\tilde{w}_e}\frac{e(1-p)\rho f^2}{p^2g}\right\}+o(N). \qquad (4.54) \end{aligned}$$

Proof of Theorem 4.3

We will first give the proof of the variance part.

$$\hat{N}_2 = \sum_{i\in E}\frac{\hat{e}_2(X_i)}{\hat{p}_2(X_i)} = \sum_{i\in U}\frac{\hat{e}_2(X_i)I_{i\in\mathcal{E}}}{\hat{p}_2(X_i)}. \qquad (4.55)$$

Let

$$\begin{aligned} \hat{\theta}_5(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_1,\vec{\lambda}_1}(x,X_j)I_{j\in\mathcal{P}}\{I_{j\in\mathcal{E}}\delta_j+\hat{p}_1(X_j)(1-\delta_j)\}, \\ \hat{\gamma}_5(x) &= N^{-1}\sum_{j\in U}\mathcal{K}_{h_2,\vec{\lambda}_2}(x,X_j)I_{j\in\mathcal{E}}\{I_{j\in\tilde{\mathcal{E}}}\eta_j+\hat{e}_1(X_j)(1-\eta_j)\} \text{ and} \end{aligned}$$

$T_{52}(X_i) = -s_2(X_i)\hat{\theta}_5(X_i)$, $T_{61}(X_i) = s_3(X_i)\hat{\gamma}_5(X_i)$, $T_5(X_i) = T_{11}(X_i)+T_{52}(X_i)$ and $T_6(X_i) = T_{61}(X_i)+T_{22}(X_i)$, we have the following expansion for (4.55)

$$\hat{N}_2 = \sum_{i\in U}\left\{T_0(X_i)+T_5(X_i)+T_6(X_i)\right\}\left\{1+o_p(1)\right\}. \qquad (4.56)$$

To develop $var(\hat{N}_2)$, the expansion of $\hat{p}_1(x)$ and $\hat{e}_1(x)$ will be used inside $\hat{\theta}_5(X_i)$ and $\hat{\gamma}_5(X_i)$. Using the same notations as those in (4.46), we have

$$\begin{aligned}\hat{p}_1(x) &= p(x) + \frac{\hat{\theta}_3(x)}{g(x)\tilde{w}_p(x)} - \frac{\hat{\theta}_4(x)}{g(x)\tilde{w}_p(x)}\{1+o_p(1)\} \text{ and} \\ \hat{e}_1(x) &= e(x) + \frac{\hat{\gamma}_3(x)}{p(x)\tilde{w}_e(x)} - \frac{\hat{\gamma}_4(x)}{p(x)\tilde{w}_e(x)}\{1+o_p(1)\}.\end{aligned}$$

Let $u_1(x) = 1/\{g(x)\tilde{w}_p(x)\}$ and $u_2(x) = 1/\{p(x)\tilde{w}_e(x)\}$, define

$$\begin{aligned}Q_9^{i,j} &= s_2(X_i)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i,X_j)I_{j\in\mathcal{P}}p(X_j)(1-\delta_j), \\ Q_{10}^{i,j} &= s_3(X_i)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i,X_j)I_{j\in\mathcal{E}}e(X_j)(1-\eta_j), \\ V_1^{i,j,k} &= s_2(X_i)u_1(X_j)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i,X_j)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_j,X_k)I_{j\in\mathcal{P}}I_{k\in\mathcal{P}}I_{k\in\mathcal{E}}\delta_k(1-\delta_j), \\ V_2^{i,j,k} &= s_2(X_i)u_1(X_j)p(X_j)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_i,X_j)\mathcal{K}_{h_1,\vec{\lambda}_1}(X_j,X_k)I_{j\in\mathcal{P}}I_{k\in\mathcal{P}}\delta_k(1-\delta_j), \\ V_3^{i,j,k} &= s_3(X_i)u_2(X_j)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i,X_j)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_j,X_k)I_{j\in\mathcal{E}}I_{k\in\mathcal{E}}I_{k\in\tilde{\mathcal{E}}}\eta_k(1-\eta_j), \\ V_4^{i,j,k} &= s_3(X_i)u_2(X_j)e(X_j)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_i,X_j)\mathcal{K}_{h_2,\vec{\lambda}_2}(X_j,X_k)I_{j\in\mathcal{E}}I_{k\in\mathcal{E}}\eta_k(1-\eta_j).\end{aligned}$$

Define $\tilde{Q}_5^{i,j} = Q_5\tilde{w}_p(X_j)$, $\tilde{Q}_6^{i,j} = Q_6\tilde{w}_p(X_j)$, $\tilde{Q}_7^{i,j} = Q_7\tilde{w}_e(X_j)$ and $\tilde{Q}_8^{i,j} = Q_8\tilde{w}_p(X_j)$. Let

$$\mu_{15}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2\tilde{w}_p}{p} & \text{if } i=j \text{ or } i=l \\ \int_{\mathcal{X}} e^2\tilde{w}_p & \text{if } k=j \\ \int_{\mathcal{X}} \frac{e^2\tilde{w}_p}{g} & \text{if } k=l \end{cases}, \mu_{19}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{p} & \text{if } i=j \text{ or } i=l \\ \int_{\mathcal{X}} e^2(f-\tilde{w}_p) & \text{if } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \end{cases},$$

$$\mu_{55}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2\tilde{w}_p^2}{pf} & \text{if } i=j \text{ or } i=l \\ & \text{or } k=j \\ \int_{\mathcal{X}} \frac{e^2\tilde{w}_p}{pg} & \text{if } k=l \end{cases}, \mu_{59}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{pf} & \text{if } i=j \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{f} & \text{if } k=j \\ 0 & \text{if } k=l \end{cases},$$

$$\mu_{99}^{i,j,k,l} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=l \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \end{cases}, \text{ then we have}$$

$$E(Q_a^{i,k}Q_b^{j,l}) = \lambda_1^{(2)}\mu_{ab}^{i,j,k,l} + o(1), a, b \in \{1,5,9\}. \tag{4.57}$$

Let

$$\tau_{11}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{p} & \text{if } i=j \text{ or } i=m \\ \int_{\mathcal{X}} e^2(f-\tilde{w}_p) & \text{if } i=l \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \text{ or } k=m \end{cases},$$

$$\tau_{12}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{p} & \text{if } i=j \\ \int_{\mathcal{X}} e^2(f-\tilde{w}_p) & \text{if } i=l \text{ or } i=m \\ & \text{or } j=k \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \text{ or } k=m \end{cases},$$

$$\tau_{51}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{pf} & \text{if } i=j \text{ or } i=m \\ & \text{or } k=j \\ \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{f} & \text{if } i=l \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=m \\ 0 & \text{if } k=l \end{cases},$$

$$\tau_{52}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{pf} & \text{if } i=j \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2\tilde{w}_p(f-\tilde{w}_p)}{f} & \text{if } i=m \text{ or } k=l \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{pg} & \text{if } k=m \\ 0 & \text{if } k=l \end{cases},$$

$$\tau_{91}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \text{ or } i=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=l \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \\ 0 & \text{if } k=m \end{cases},$$

$$\tau_{92}^{i,j,k,l,m} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=l \text{ or } i=m \\ & \text{or } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=l \\ 0 & \text{if } k=m \end{cases} \text{, then}$$

$$E(Q_a^{i,k}V_b^{j,l,m}) = \lambda_1^{(3)}\tau_{ab}^{i,j,k,l,m} + o(1), a \in \{1,5,9\}, b \in \{1,2\}. \tag{4.58}$$

Similarly, let

$$\nu_{11}^{i,j,k,l,m,n} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \text{ or } i=m \text{ or } j=l \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=m \text{ or } k=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{gf} & \text{if } k=n \text{ or } l=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pg\tilde{w}_p} & \text{if } l=n \end{cases} ,$$

$$\nu_{12}^{i,j,k,l,m,n} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \text{ or } j=l \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=m \text{ or } k=j \text{ or } i=n \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{gf} & \text{if } k=n \text{ or } l=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{g\tilde{w}_p} & \text{if } l=n \end{cases} \text{and}$$

$$\nu_{22}^{i,j,k,l,m,n} = \begin{cases} \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{pf} & \text{if } i=j \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{f} & \text{if } i=m \text{ or } k=j \text{ or } j=l \text{ or } i=n \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)}{g} & \text{if } k=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{gf} & \text{if } k=n \text{ or } l=m \\ \int_{\mathcal{X}} \frac{e^2(f-\tilde{w}_p)^2}{g\tilde{w}_p} & \text{if } l=n \end{cases} \text{, then}$$

$$E(V_a^{i,k,l}V_b^{j,m,n}) = \lambda_1^{(4)}\nu_{ab}^{i,j,k,l,m,n} + o(1), a, b \in \{1,2\}. \tag{4.59}$$

By using results in (4.57), (4.58) and (4.59),

$$\begin{aligned} var\left\{\sum_i T_5(X_i)\right\} &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_1^{i,k}Q_1^{j,l} - Q_1^{i,k}\tilde{Q}_5^{j,l} - Q_1^{i,k}Q_9^{j,l} - \tilde{Q}_5^{i,k}Q_1^{j,l} + \tilde{Q}_5^{i,k}\tilde{Q}_5^{j,l}\right. \\ &+ \left.\tilde{Q}_5^{i,k}Q_9^{j,l} - Q_9^{i,k}Q_1^{j,l} + Q_9^{i,k}\tilde{Q}_5^{j,l} + Q_9^{i,k}Q_9^{j,l})\right\} \\ &+ 2N^{-3}E\left\{\sum_{i,j,k,l,m}(Q_1^{i,k}V_1^{j,l,m} - Q_1^{i,k}V_2^{j,l,m} + \tilde{Q}_5^{i,k}V_1^{j,l,m} - \tilde{Q}_5^{i,k}V_2^{j,l,m}\right. \\ &+ \left.Q_9^{i,k}V_1^{j,l,m} - Q_9^{i,k}V_2^{j,l,m})\right\} + N^{-4}E\left\{\sum_{i,j,k,l,m,n}(V_1^{i,k,l}V_1^{j,m,n}\right\} \\ &- V_1^{i,k,l}V_2^{j,m,n} - V_2^{i,k,l}V_1^{j,m,n} + V_2^{i,k,l}V_2^{j,m,n})\Big\} \\ &= N\int_{\mathcal{X}}\frac{e^2(1-p)}{pg}\left\{\lambda_1^{(2)}\tilde{w}_p + 2\lambda_1^{(3)}(f-\tilde{w}_p) + \lambda_1^{(4)}\frac{(f-\tilde{w}_p)^2}{\tilde{w}_p}\right\} + o(N). \end{aligned} \tag{4.60}$$

In exactly the same fashion, we establish that

$$\begin{aligned} var\left\{\sum_i T_6(X_i)\right\} &= N^{-2}E\left\{\sum_{i,j,k,l}(Q_3^{i,k}Q_3^{j,l} - Q_3^{i,k}\tilde{Q}_7^{j,l} - Q_3^{i,k}Q_{10}^{j,l} - \tilde{Q}_7^{i,k}Q_3^{j,l} + \tilde{Q}_7^{i,k}\tilde{Q}_7^{j,l}\right. \\ &+ \left.\tilde{Q}_7^{i,k}Q_{10}^{j,l} - Q_{10}^{i,k}Q_3^{j,l} + Q_{10}^{i,k}\tilde{Q}_7^{j,l} + Q_{10}^{i,k}Q_{10}^{j,l})\right\} \\ &+ 2N^{-3}E\left\{\sum_{i,j,k,l,m}(Q_3^{i,k}V_3^{j,l,m} - Q_3^{i,k}V_4^{j,l,m} + \tilde{Q}_7^{i,k}V_3^{j,l,m} - \tilde{Q}_7^{i,k}V_4^{j,l,m}\right. \\ &+ \left.Q_{10}^{i,k}V_3^{j,l,m} - Q_{10}^{i,k}V_4^{j,l,m})\right\} + N^{-4}E\left\{\sum_{i,j,k,l,m,n}(V_3^{i,k,l}V_3^{j,m,n}\right\} \\ &- V_3^{i,k,l}V_4^{j,m,n} - V_4^{i,k,l}V_3^{j,m,n} + V_4^{i,k,l}V_4^{j,m,n})\Big\} \\ &= N\int_{\mathcal{X}}\frac{e(1-e)}{p}\left\{\lambda_2^{(2)}\tilde{w}_e + 2\lambda_2^{(3)}(f-\tilde{w}_e) + \lambda_2^{(4)}\frac{(f-\tilde{w}_e)^2}{\tilde{w}_e}\right\} + o(N). \end{aligned} \tag{4.61}$$

And

$$cov\{\sum_i T_5(X_i), \sum_j T_6(X_j)\} = N^{-2}E\left\{\sum_{i,j,k,l}(Q_1^{i,k}Q_3^{j,l} - Q_1^{i,k}\tilde{Q}_7^{j,l} - Q_1^{i,k}Q_{10}^{j,l} - \tilde{Q}_5^{i,k}Q_3^{j,l}\right.$$
$$\left.+\tilde{Q}_5^{i,k}\tilde{Q}_7^{j,l} + \tilde{Q}_5^{i,k}Q_{10}^{j,l} - Q_9^{i,k}Q_3^{j,l} + Q_9^{i,k}\tilde{Q}_7^{j,l} + Q_9^{i,k}Q_{10}^{j,l})\right\} + N^{-3}E\left\{\sum_{i,j,k,l,m}(Q_1^{i,k}V_3^{j,l,m}\right\}$$
$$-Q_1^{i,k}V_4^{j,l,m} + \tilde{Q}_5^{i,k}V_3^{j,l,m} - \tilde{Q}_5^{i,k}V_4^{j,l,m} + Q_9^{i,k}V_3^{j,l,m} - Q_9^{i,k}V_4^{j,l,m}$$
$$\left.+Q_3^{i,k}V_1^{j,l,m} - Q_3^{i,k}V_2^{j,l,m} + \tilde{Q}_7^{i,k}V_1^{j,l,m} - \tilde{Q}_7^{i,k}V_2^{j,l,m} + Q_{10}^{i,k}V_1^{j,l,m} - Q_{10}^{i,k}V_2^{j,l,m})\right\}$$
$$+N^{-4}E\left\{\sum_{i,j,k,l,m,n}(V_1^{i,k,l}V_3^{j,m,n} - V_1^{i,k,l}V_4^{j,m,n} - V_2^{i,k,l}V_3^{j,m,n} + V_2^{i,k,l}V_4^{j,m,n})\right\}$$
$$= -N\int_{\mathcal{X}}\frac{e(1-p)\rho}{p^2 g}\left\{\lambda_1^{(1)}\lambda_2^{(1)}\tilde{w}_b + \lambda_1^{(2)}\lambda_2^{(1)}\frac{(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p} + \lambda_1^{(1)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)\tilde{w}_b}{\tilde{w}_e}\right.$$
$$\left.+\lambda_1^{(2)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p\tilde{w}_e}\right\} + o(N), \tag{4.62}$$

$$cov\{\sum_i T_0(X_i), \sum_j T_5(X_j)\} = N^{-1}E\left\{\sum_{i,j,k}\frac{e(X_i)I_{\mathcal{E}}(X_i)(Q_1^{j,k} + \tilde{Q}_5^{j,k} + Q_9^{j,k})}{p(X_i)}\right\}$$
$$+N^{-3}E\left\{\sum_{i,j,k,l}\frac{e(X_i)I_{\mathcal{E}}(X_i)(V_1^{j,k,l} - V_2^{j,k,l})}{p(X_i)}\right\}$$
$$= -N\lambda_1^{(2)}\int_{\mathcal{X}}\frac{e^2(1-p)f}{p} + o(N), \tag{4.63}$$

$$cov\{\sum_i T_0(X_i), \sum_j T_6(X_j)\} = N^{-1}E\left\{\sum_{i,j,k}\frac{e(X_i)I_{\mathcal{E}}(X_i)(Q_3^{j,k} + \tilde{Q}_7^{j,k} + Q_{10}^{j,k})}{p(X_i)}\right\}$$
$$+N^{-3}E\left\{\sum_{i,j,k,l}\frac{e(X_i)I_{\mathcal{E}}(X_i)(V_3^{j,k,l} - V_4^{j,k,l})}{p(X_i)}\right\}$$
$$= o(N). \tag{4.64}$$

By combining results of (4.32), (4.60), (4.61), (4.62), (4.63) and (4.64), we have

$$\begin{aligned} var(\hat{N}_2) &= N\left[\int_{\mathcal{X}} \frac{e^2 f}{p} - \left(\int_{\mathcal{X}} ef\right)^2 - 2\lambda_1^{(2)} \int_{\mathcal{X}} \frac{e^2(1-p)f}{p}\right. \\ &+ \int_{\mathcal{X}} \frac{e^2(1-p)}{pg}\left\{\lambda_1^{(2)}\tilde{w}_p + 2\lambda_1^{(3)}(f-\tilde{w}_p) + \lambda_1^{(4)}\frac{(f-\tilde{w}_p)^2}{\tilde{w}_p}\right\} \\ &+ \int_{\mathcal{X}} \frac{e(1-e)}{p}\left\{\lambda_2^{(2)}\tilde{w}_e + 2\lambda_2^{(3)}(f-\tilde{w}_e) + \lambda_2^{(4)}\frac{(f-\tilde{w}_e)^2}{\tilde{w}_e}\right\} \\ &- 2\int_{\mathcal{X}} \frac{e(1-p)\rho}{p^2 g}\left\{\lambda_1^{(1)}\lambda_2^{(1)}\tilde{w}_b + \lambda_1^{(2)}\lambda_2^{(1)}\frac{(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p} + \lambda_1^{(1)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)\tilde{w}_b}{\tilde{w}_e}\right. \\ &+ \left.\left.\lambda_1^{(2)}\lambda_2^{(2)}\frac{(f-\tilde{w}_e)(f-\tilde{w}_p)\tilde{w}_b}{\tilde{w}_p\tilde{w}_e}\right\} + o(N). \end{aligned} \tag{4.65}$$

Bias of Nonparametric Population Size Estimation

In deriving the bias of $\hat{N}_0$, $\hat{N}_1$ and $\hat{N}_2$, we use a different expansion as follows. Note that $\hat{p}_a(x)$ and $\hat{e}_a(x)$, $a = 0, 1, 2$, are consistent estimators of $p(x)$ and $e(x)$,

$$\begin{aligned} \hat{N}_a &= \sum_{i\in U} I_{i\in\mathcal{E}} \frac{\hat{e}_a(X_i)}{\hat{p}_a(X_i)} = \sum_{i\in U} I_{i\in\mathcal{E}}\left[\frac{e(X_i)}{p(X_i)} + \frac{\hat{e}_a(X_i) - e(X_i)}{p(X_i)} - \right. \\ &\quad \frac{e(X_i)\left\{\hat{p}_a(X_i) - p(X_i)\right\}}{p^2(X_i)} - \frac{\left\{\hat{e}_a(X_i) - e(X_i)\right\}\left\{\hat{p}_a(X_i) - p(X_i)\right\}}{p^2(X_i)} + \\ &\quad \left.\frac{e(X_i)}{p^3(X_i)}\left\{\hat{p}_a(X_i) - p(X_i)\right\}^2\left\{1 + o_p(1)\right\}\right] \\ &:= A_{1a} + A_{2a} + A_{3a} + A_{4a} + A_{5a}\left\{1 + o_p(1)\right\}, \end{aligned} \tag{4.66}$$

where$A_{ja} = \sum_{i\in U}\{I_{\mathcal{E}}(X_i)A_{ija}\}$, $j \in \{1,\ldots,5\}$, $a \in \{0,1,2\}$, $A_{i1a} = \frac{e(X_i)}{p(X_i)}$, $A_{i2a} = \frac{\hat{e}_a(X_i)-e(X_i)}{p(X_i)}$, $A_{i3a} = -\frac{e(X_i)\{\hat{p}_a(X_i)-p(X_i)\}}{p^2(X_i)}$, $A_{i4a} = \frac{\{\hat{e}_a(X_i)-e(X_i)\}\{\hat{p}_a(X_i)-p(X_i)\}}{p^2(X_i)}$ and $A_{i5a} = \frac{e(X_i)\{\hat{p}_a(X_i)-p(X_i)\}^2}{p^3(X_i)}$. Firstly, we note that for $a = 0, 1, 2$,

$$E(A_{1a}) = N\int_{\mathcal{X}} e(x)f(x)dx := \tilde{N}, \tag{4.67}$$

where $\tilde{N}$ is defined to be the population size. Then,

$$E\left\{\hat{p}_a(x) - p(x)\right\} = b_{a,c}(x;h_1) + b_{a,u}(x,\vec{\lambda}_1) + o\left\{h_1^2 + A(\vec{\lambda}_1)\right\}, \tag{4.68}$$

where $A(\lambda) = \min_{1\le j\le p_u}(1-\lambda_j)$ from the proof in Chapter 3. Similarly, we have

$$E\left\{\hat{e}_a(x) - e(x)\right\} = b_{a,c}(x;h_2) + b_{a,u}(x;\vec{\lambda}_2) + o\left\{h_2^2 + A(\vec{\lambda}_2)\right\}. \tag{4.69}$$

Further, let

$$\begin{aligned} r_0(x) &= \xi_{1,K}\lambda_1^{(1)}\lambda_2^{(1)}\frac{\{1-p(x)\}\rho(x)}{gfp(x)},\ r_1(x) = \xi_{1,K}\lambda_1^{(1)}\lambda_2^{(1)}\frac{\tilde{w}_b\{1-p\}\rho(x)}{g\tilde{w}_p p\tilde{w}_e(x)}, \\ r_2(x) &= \frac{\rho(x)}{f^2(x)}\sum_{l=1}^{4}\left[\xi_{l,K}\left\{r_{2,l}(x)\right\}\right], \end{aligned}$$

where $r_{2,1}(x) = \lambda_1^{(1)}\lambda_2^{(1)}\tilde{w}_b(x)$, $r_{2,2}(x) = \lambda_1^{(1)}\lambda_2^{(2)}\frac{\tilde{w}_b(1-\tilde{w}_e)(x)}{\tilde{w}_e(x)}$, $r_{2,3}(x) = \lambda_1^{(2)}\lambda_2^{(1)}\frac{\tilde{w}_b(1-\tilde{w}_p)(x)}{\tilde{w}_p(x)}$ and $r_{2,4}(x) = \lambda_1^{(2)}\lambda_2^{(2)})\frac{\tilde{w}_b(1-\tilde{w}_p)(1-\tilde{w}_e)(x)}{\tilde{w}_p\tilde{w}_e(x)}$, $\xi_{1,K} = \int K(t)K(\frac{h_1}{h_2}t)dt$, $\xi_{2,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(\frac{h_1}{h_2}t_1 - t_2)dt_1dt_2$, $\xi_{3,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1-t_2)dt_1dt_2$, $\xi_{3,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1-t_2)dt_1dt_2$ and $\xi_{4,K} = \int K(t_1)K(\frac{h_1}{h_2}t_2)K(t_1-t_2)K(\frac{h_1}{h_2}t_1-t_3)dt_1dt_2dt_3$. Then we have

$$\begin{aligned} cov\{\hat{p}_a(x), \hat{e}_a(x)\} &= \frac{cov\{\hat{\alpha}_{1a}(x), \hat{\alpha}_{2a}(x)\}}{\beta_{1a}(x)\beta_{2a(x)}} - \frac{cov\{\hat{\alpha}_{1a}(x), \hat{\beta}_{2a}(x)\}}{\beta_{1a}(x)\beta_{2a}^2(x)} \\ &- \frac{cov\{\hat{\beta}_{1a}(x), \hat{\alpha}_{2a}(x)\}}{\beta_{1a}^2(x)\beta_{2a}(x)} + \frac{Cov\{\hat{\beta}_{1a}(x), \hat{\beta}_{2a}(x)\}}{\beta_{1a}^2(x)\beta_{2a}^2(x)}\{1+o(1)\} \\ &= \frac{r_a(x)}{Nh_2^{d_c}} + o(N^{-1}h_2^{-d_c}). \end{aligned} \tag{4.70}$$

As results from Chapter 3,

$$var\{\hat{p}_a(x)\} = \frac{v_a(x)}{Nh_1^{d_c}} + o(N^{-1}h_1^{-d_c}), \tag{4.71}$$

where $v_0(x) = \lambda_1^{(2)}K^{(2)}(0)\frac{p(1-p)(x)}{gf(x)}$, $v_1(x) = \lambda_1^{(2)}K^{(2)}(0)\frac{p(1-p)(x)}{g\tilde{w}_p(x)}$ and

$$\begin{aligned} v_2(x) &= \frac{p(1-p)(x)}{gf^2(x)}\left\{\lambda_1^{(2)}K^{(2)}(0)\tilde{w}_p(x) + \lambda_1^{(3)}2K^{(3)}(0)\{f-\tilde{w}_p(x)\}\right. \\ &+ \left.\lambda_1^{(4)}K^{(4)}(0)\frac{\{f-\tilde{w}_p(x)\}^2}{\tilde{w}_p(x)}\right\}. \end{aligned}$$

Hence, (4.66), (4.68), (4.69), (4.70) and (4.71) imply that

$$\begin{aligned} E(\hat{N}_a) &= \tilde{N} + N\int_{\mathcal{X}}\sum_{j=1}^{2}S_j(x)\left\{b_{a,c}(x;h_j) + b_{a,u}(x;\vec{\lambda}_j)\right\}f(x)dx - \frac{1}{h_2^{d_c}}\int_{\mathcal{X}}\frac{r_a(x)}{p(x)}f(x)dx \\ &+ \frac{1}{h_1^{d_c}}\int_{\mathcal{X}}\frac{e(x)v_a(x)}{p^2(x)}f(x)dx + o\left[N\left\{h^2 + A(\vec{\lambda})\right\} + \frac{1}{h^{d_c}}\right], \end{aligned} \tag{4.72}$$

where $S_1(x) = -e(x)/p(x)$, $S_2(x) = 1$, $h = \max(h_1, h_2)$ and $A(\vec{\lambda}) = \max\{A(\vec{\lambda}_1), A(\vec{\lambda}_2)\}$.

BIBLIOGRAPHY

Aït-Sahalia, Y. (1996), "Testing continuous-time models of the spot interest rate," *Review of Financial Studies*, 9, 385–426.

— (1999), "Transition densities for interest rate and other nonlinear diffusions," *Journal of Finance*, 54, 1361–1395.

— (2002), "Maximum likelihood estimation of discretely sample diffusion: A close form approximation approach," *Econometrica*, 70, 223–262.

Aït-Sahalia, Y. and Kimmel, R. (2002), "Estimating affine multifactor term structure models using closed-Form likelihood expansions," *Working Paper.*

Aït-Sahalia, Y. and Mykland, P. (2003), "The Effects of Random and Discrete Sampling When Estimating Continuous-Time Diffusions," *Econometrica*, 71, 483–549.

— (2004), "Estimating diffusions with discretely and possibly Randomly spaced data: a general theory," *Annals of Statistics*, 32, 2186–2222.

Aitchison, J. and Aitken, C. (1976), "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Alho, J. M. (1990), "Logistic regression in capture-recapture models," *Biometrics*, 46, 623–635.

Alho, J. M., Mury, M. H., Wurdeman, K., and Kim, J. (1993), "Estimating heterogeneity in the probabilities of enumeration for dual-system estimation," *Journal of the American Statistical Association*, 88, 1130–1136.

Ball, C. and Torous, W. (1996), "Unit roots and the estimation of interest rate dynamics," *Journal of Empirical Finance*, 3, 215–238.

Bandi, F. M. and Phillips, P. C. B. (2003), "Fully nonparametric estimation of Scalar diffusion models," *Econometrica*, 71, 241–283.

Barndorff-Nielsen, O. E. and Shephard, N. (2001), "Non-Gaussian OrnsteinUhlenbeck-based models and some of their uses in financial economics," *Journal of the Royal Statistical Society: Series B*, 63, 167.

— (2002), "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," *Journal of the Royal Statistical Society: Series B*, 64, 253.

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. (1993), "Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussions)," *Journal of the American Statistical Association*, 88, 1149–1166.

Bell, R. and Cohen, M. L. (eds.) (2007), *Research and plans for coverage measurement in the 2010 Census: interim assessment.*, National Academies Press.

Bell, W. R. (1999), "Accuracy and coverage evaluation survey: ratio adjusting logistic regression DSEs (target model) using 1990 census counts," *DSSD Census 2000 Procedures And Operations Memorandum Series # Q-11.*

Bergstrom, A. R. (1984), "Continuous time stochastic models and issues of aggregation over time," *Handbook of Econometrics (Z. Griliches and M.D. Intriligator Eds) Vol 2.*

Bibby, B. and Sørensen, M. (1995), "Martingale estimating functions for discretely observed diffusion processes," *Bernoulli*, 1, 17–39.

Black, F. and Scholes, M. (1973), "The pricing of options and corporate liabilities," *Journal of Political Economy*, 81, 637–654.

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, Springer.

Brown, R. (1828), "A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *Philosophical Magazine*, 4, 161–173.

Cai, Z. and Hong, Y. (2003), "Nonparametric methods in continuous-time finance: a selective review," *Recent Advances and Trends in Nonparametric Statistics (M. G. Akritas and D. N. Politis, eds.)*, 283–302.

Chao, A. and Tsay, P. K. (1998), "A sample coverage approach to multiple-system estimation with application to census undercounts," *Journal of the American Statistical Association*, 93, 283–293.

Chen, S. X. and Gao, J. (2007), "An adaptive empirical likelihood test for time series models," *Journal of Econometrics*, 141, 950–972.

Chen, S. X. and Lloyd, C. J. (2000), "A non-parametric approach to the analysis of two stage mark-recapture experiments," *Biometrika*, 87, 633–649.

— (2002), "Estimation of population size based on biased samples using nonparametric binary regression," *Statistica Sinica*, 12, 505–518.

Chen, S. X., Tang, C. Y., and Mule, V. T. (2008), "Local post-stratification and diagnostics in dual system accuracy and coverage evaluation for US Census," *Working Paper*.

Cox, J., Ingersoll, J., and Ross, S. (1985), "A theory of the term structure of interest rates," *Econometrica*, 53, 385–407.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife." *Annals of Statistics*, 7, 1–26.

Einstein, A. (1956), *Investigations on the Theory of Brownian Movement*, Dover.

Fan, J. (2005), "A selective overview of nonparametric methods in financial econometrics (with discussion)," *Statistical Science*, 20, 317–357.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

Fan, J. and Zhang, C. (2003), "A re-examination of Stanton's diffusion estimations with applications to financial model validation," *Journal of American Statistical Association*, 98, 118–134.

Fricks, J. (2004), "Biomolecular motors and diffusion ratchets," *Ph.D. Dissertation, University of North Carolina at Chapel Hill*.

Fuller, W. A. (1996), *Introduction to Statistical Time Series*, Wiley, New York, 2nd ed.

Genon-catalot, V., Jeantheau, T., and Larédo, C. (2000), "Stochastic volatility models as hidden markov models and statistical applications," *Bernoulli*, 6, 1051–1079.

Gouriérous, C. (1997), *ARCH Models and Financial Applications*, Springer.

Haberman, S., Jiang, W., and Spencer, B. (1998), "Activity 7: develop methodology for evaluating model-based estimates of the population size for States. Final Reports," *Technical report, US Census Bureau.*

Hall, P. (1981), "On nonparametric multivariate binary discrimination," *Biometrika*, 68, 287–294.

— (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, 99, 1015–1026.

Hansen, L. P. and Scheinkman, J. A. (1995), "Back to the future: generating moment implications for continuous-time markov processes," *Econometrica*, 63, 767–804.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Härdle, W. and Mammen, E. (1993), "Comparing nonparametric versus parametric regression fits," *Annals of Statistics*, 21, 1926–1947.

Hogan, H. (1992), "The 1990 post-enumeration surve: an overview," *The American Statistician*, 46, 261–269.

— (1993), "The 1990 post-enumeration survey: operations and results," *Journal of the American Statistical Association*, 88, 1047–1060.

— (2000a), "Accuracy and coverage evaluation 2000: decomposition of dual system estimate components," *DSSD Census 2000 Procedures and Operation Memorandum Series B-8.*

— (2000b), "Accuracy and coverage evaluation 2000: dual system estimate results," *DSSD Census 2000 Procedures and Operation Memorandum Series B-9*.

Huggins, R. M. (1989), "On the statistical analysis of capture experiments," *Biometrika*, 76, 133–140.

Jiang, G. J. and Knight, J. L. (1997), "A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model," *Econometric Theory*, 13, 615–645.

Karlin, S. and Taylor, H. M. (1975), *A First Course in Stochastic Processes*, Academic Press, 2nd ed.

Kendall, M. G. (1954), "Note on bias in the estimation of autocorrelation," *Biometrika*, 41, 403–404.

Kloeden, P. and Platen, E. (2000), *Numerical Solution of Stochastic Differential Equations*, Springer.

Lahiri, S. (2003), *Resampling Methods for Dependent data*, Springer-Verlag, Inc.

Little, R. and Rubin, D. (2002), *Statistical Analysis With Missing Data*, Wiley, 2nd ed.

Lo, A. (1988), "Maximum likelihood estimation of generalized Ito processes with discretely sampled data," *Econometric Theory*, 4, 231–247.

Marriott, F. H. C. and Pope, J. A. (1954), "Bias in the estimation of autocorrelations," *Biometrika*, 41, 390–402.

Merton, R. C. (1971), "Optimum consumption and portfolio rules in a continuous-time model," *Journal of Economic Theory*, 3, 373–413.

Nowman, K. (1997), "Gaussian estimation of single-factor continuous time models of the term structure of interest rates," *Journal of Finance*, 52, 1695–1706.

Petersen, C. (1896), "The yearly immigration of young plaice into the Limfjord from the German sea," *Report of the Danish Biological Station*, 6, 1–48.

Phillips, P. C. and Yu, J. (2005), "Jackknifing bond option prices," *Review of Financial Studies*, 18, 707–742.

Pollock, K. H. (1976), "Building models of capture-recapture experiments," *The Statistician*, 25, 253–260.

— (1991), "Modeling capture-recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future;," *Journal of the American Statistical Association*, 86, 225–238.

Racine, J. S. and Li, Q. (2004), "Nonparametric estimation of regression functions with both catagorical and continuous data," *Journal of Econometrics*, 119, 99–130.

Rice, J. (1984), "Boundary modification for kernel regression," *Communications in Statistics*, 13, 893–900.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Sargan, J. (1976), "Econometric estimators and the edgeworth approximation," *Econometrica*, 44, 421–448.

Seber, G. (2002), *Estimation of Animal Abundance*, The Blackburn Press., 2nd ed.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley.

Shao, J. and Sitter, R. (1996), "Bootstrap for imputed survey data," *Journal of the American Statistical Association*, 91, 1278–1288.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer.

Shao, J. and Wu, C. F. J. (1989), "A general theory for Jackknife varaince estimation," *The Annals of Statistics*, 17, 1176–1197.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Simonnoff, J. S. (1995), "Smoothing categorical data," *Journal of Statistical Planning and Inference*, 47, 41–69.

Stanton, R. (1997), "A nonparametric model of term structure dynamics and the market price of interest rate risk," *The Journal of Finance*, 52, 1973–2002.

Stroock, D. and Varadhan, S. (1979), *Multidimensional Diffusion Processes*, Springer.

Sundaresan, S. M. (2000), "Continuous-time methods in finance: a review and an assessment," *Journal of Finance*, 55, 1569–1622.

US Census Bureau (2004), *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*, US Census Bureau.

Vasicek, O. (1977), "An equilibrium characterization of the term structure," *Journal of Financial Economics*, 5, 177–186.

Wolter, K. (1986), "Some coverage error models for census data," *Journal of the American Statistical Association*, 81, 338–346.

Wolter, K. M. (2007), *Introduction to Variance Estiamtion*, Springer.

Wu, C. F. J. (1986), "Jackknife, bootstrap and other resampling methods in regression analysis (with discussion)," *Annals of Statistics*, 14, 359–372.

Yokoyama, R. (1980), "Moment bounds for stationary mixing sequences," *Probability Theory and Related Fields*, 52, 45–57.

Yu, J. and Phillips, P. C. (2001), "A gaussian approach for estimating continuouse time models of short term interest rates," *The Econometrics Journal*, 4, 211–225.