Properties of Census Dual System Population Size Estimators [1]

Song Xi Chen[2] and Cheng Yong Tang[3]

[2]Guanghua Scool of Management and Center for Statistical Science, Peking University and Department of Statistics, Iowa State University

Email: csx@gsm.pku.edu.edu

[3] Department of Statistics and Applied Probability, National University of Singapore Singapore 117546 Email:statc@nus.edu.sg

SUMMARY

We evaluate three population size estimators, including the post-stratification and logistic regression estimators which has been or will be implemented in the US Census dual system surveys. Conditions that ensure consistency of these two Census population size estimators are provided. We also study a local post-stratification estimator based on a nonparametric kernel estimates to the Census enumeration functions, which is shown to be consistent under weaker conditions. The performances of these estimators are evaluated numerically via simulation studies and an empirical analysis based on the 2000 Census Accuracy and Coverage Evaluation data.

Key words: Capture-recapture; Correlation Bias; Erroneous enumeration; Kernel smoothing; Model bias; Population size estimation.

1 INTRODUCTION

The decennial US Census is a major source of information providing counts of the whole population and sub-populations defined by states, congressional districts and variables from demographical and racial information. A comprehensive account on the Census taking in the US is given in Anderson and Feinberg (2002). As a large scale data collection, two types

[1]We thank the US Census Bureau for supporting our research and allowing access to the ACE research data. The project was supported by a National Science Foundation grant SES-0518904.

of survey errors are inevitably present in the Census. One type is the error of omission that occurs when genuine Census persons are missed (omitted) and causes a net population under-count. The another type is the enumeration errors (EEs) due to enumerations of invalid Census persons, for instance fictitious or duplicate persons and valid Census persons but enumerated in wrong location. The EEs tend to inflate the Census count. Both errors can reduce the accuracy of the census population counts significantly (Hogan, 2003).

To gain information on the extent of these errors, the Census Bureau has conducted dual system capture-recapture surveys, as early as from the 1950 Census (Anderson and Feinberg, 2002). The early dual system surveys focused on estimating the population undercounts caused by the error of omission. Later surveys since 1990 had an added agenda of detecting EEs; see Hogan (1993, 2003). The first of the dual systems surveys is largely the Census itself but restricted to randomly selected sample block clusters of the Census. The Census enumerations obtained over the selected block clusters form the E sample. A primary purpose of the E sample is to identify and measure the EEs via extensive records checking and follow-up. The second survey is an independent post-census enumeration on the same sample block-clusters occupied by the E sample, which gives rise to the P sample. The E and P samples form a capture-recapture design by carrying out comprehensive matching between E and P samples. The enumerations appeared in both samples are "recaptures", which together with enumerations appeared only in the P sample are used to estimate the E sample enumeration (capture) probability, which can be used to quantify the error of omission. See Hogan (1993, 2000a, 2000b), Haberman et al. (1998), Bell (1993), Darroch et al. (1993), Chao and Tsay (1998), Brown and Zhao (2008) and Chen et al. (2010) for more specific discussions and methods of estimation on the Census dual system surveys. We also refer to Wolter (1986) and Pollock (1991) on capture-recapture based population size estimation. Anderson and Feinberg (1999) contains a comprehensive account on some critical issues and controversies surrounding the Censuses. The US is not the only country that conducts the dual system surveys to gain information on the accuracy of the census counts. Australia, New Zealand, Turkey, Switzerland and the UK also carry out similar dual system surveys to evaluate their national censuses; see (Census

Customer Service, 2002; Dunstan et al., 2001; Ayhan and Ekni, 2003; Rhind, 2003).

Like many population surveys, human or wildlife, the errors of omission and enumerations do not occur homogeneously across the population. Certain sections of the population are more prone to the errors than others, making the errors to be heterogeneous. Racial Original (RO), Age (A), Sex (S), housing Tenure (T), and geographical Region (R) (hence ROASTR) are covariates which are known to contribute to the heterogeneity for the US population (Hogan, 1993, 2003). Modeling the probabilities of the two errors as functions of the covariates (for instance ROASTR) has been a main task of the Census dual system estimation.

The method employed in the dual system estimation prior to the 2010 Census had been the post-stratification (PS) (Hogan, 1993). Given a set of covariates, for instance ROASTR, the PS subdivides the support of the covariates into non-overlapping post-strata where a population size estimate for each post-strata is obtained using the classical Petersen's estimator. The rationale is that the PS makes the enumerations within each post-strata more homogeneous. However, Hogan (1993, 2003) showed that the PS was not able to achieve satisfactory homogeneity within each post-stratum, especially with respect to continuous covariates like age. To overcome the limitation of the PS, the Census Bureau has decided to implement a logistic regression approach in the 2010 Census dual system estimation (Bell and Cohen, 2009).

Logistic regression offers more flexibility than the PS and allows extrapolation in areas with sparse observations. It was applied in analyzing the 1990 dual system data by Alho et al. (1993) followed by a set of extensive research as in Mule et al. (2007). However, as a model based approach, a risk of the logistic regression approach is using a mis-specified model, which may produce a systematic bias in the population estimation.

We study the properties of PS and logistic regression in a unified framework incorporating the features of the US Census dual system surveys. The conventional theory on population size estimation (Wolter, 1986; Pollock, 1991) concerns primarily with the omission error only. In this paper, we try to evaluate the impacts of both omission errors and

the EEs. We show that PS and logistic regression are subject to different forms of bias due to model mis-specifications unless rather stringent conditions are satisfied.

To alleviate the bias due to model mis-specification, we carry out a study on a nonparametric local post-stratification (local PS) recently proposed by Chen et al. (2010) in an empirical study for the 2000 Census Accuracy and Coverage Evaluation (A.C.E.) data. Instead of having fixed post-strata as in the PS, the local PS effectively produces local post-strata via nonparametric kernel smoothing method without a specific parametric model. The local post-stratum shrinks when the number of observations in the defined target stratum and its neighborhood gets larger, allowing the removal of the heterogeneity in the dual system surveys. We show that the local PS leads to consistent population size estimation, under much weaker conditions than those for the PS and the logistic regression.

This paper is structured as follows. Section 2 overviews the dual population size estimation. The properties of the population size estimators using the PS, the logistic regression and the local PS are reported in Sections 3, 4 and 5 respectively. Section 6 reports some simulation results. An empirical study on the 2000 Census A.C.E. data is given in Section 7. All technical details are deferred to the Appendix.

2 DUAL SYSTEM ESTIMATION FOR THE US CENSUS

Let $\mathcal{C}$ be the set of census records, $\mathcal{U}$ be the set of genuine persons on the Census day, and X denoting a set of covariates contributing to the heterogeneity in the Census enumerations. Estimating the size of $\mathcal{U}$ and sizes of sub-populations are the main objectives of the Census. The Census enumeration probability for the ith person in $\mathcal{U}$ with covariate X_i is $p(X_i) = P(i \in \mathcal{C}|X_i)$. Clearly, $1 - p(X_i)$ is the probability of omission for the i-th person. If the numeration function $p(x)$ were known and there were no erroneous enumerations, the Horvitz-Thompson type estimator $\sum_{i\in\mathcal{C}} p^{-1}(X_i)$ would be an consistent and unbiased estimator for the size of $\mathcal{U}$. However, $p(x)$ is unknown in reality and needs to be estimated based on the E and P samples with a capture-recapture design. Let $\mathcal{E}$ and $\mathcal{P}$ denote the sets of enumerations by the E and P samples respectively.

The enumerations appeared in both samples are "recaptures", which together with enumerations appeared only in the P sample are used to estimate the E-sample enumeration (capture) function $p(x)$ that can be used to quantify the omission error. After a comprehensive matching operation consisting of computer matching and fields follow-ups, each P-sample person is classified as (i) a match to an E-sample person (recapture), (ii) not a match or (iii) unresolved. Unresolved matches can be viewed as missing response variables, which we will ignore in this paper to simplify our expedition without altering the main conclusions of the paper. The matching process gives rise to the P-sample data $\{(Y_i, X_i)\}_{i=1}^{n_p}$ where n_p is the P-sample size and $Y_i = 1$ (or 0) if the $i \in \mathcal{P}$ with covariate X_i matches (or does not match) to an E-sample record. Since $E(Y_i|X_i) = p(X_i)$, the enumeration function $p(\cdot)$ can be estimated by binary regression.

Different from that in conventional capture-recapture experiments, the primary purpose of the E sample is to identify and measure the EEs via extensive record-checking and follow-ups. For each $i \in \mathcal{E}$, let e_i be the EE indicator such that $e_i = 1$ (0) if it is correct (erroneous) enumeration. We will ignore the missing values in the E sample from unresolved cases as well. Previous research (Hogan, 2003) revealed analogous heterogeneity in EEs caused by a set of covariates Z. Here Z can have different covariates from X in modeling the enumeration function $p(\cdot)$. For instance, the non-response follow-up code is a unique Z-covariate that provides more information on those who did not respond to the census mail-out questionnaires (Belin et al., 1993; Cantwell and Childers, 2001). We denote the E-sample data by $\{(e_i, Z_i)\}_{i=1}^{n_e}$ where n_e is the E-sample size. The correct enumeration function $e(Z_i) = P(e_i = 1|Z_i)$ quantifies the heterogeneity in the EEs and can be estimated by performing binary regression on the E sample data.

Let N be the size of the true population $\mathcal{U}$, and $\tilde{N}$ be the size of the nominal population $\tilde{\mathcal{U}}$ consisting of the true population $\mathcal{U}$ and the erroneous population $\mathcal{U}_e$. Let $U = Z \cup X$ be the combined covariates in the E and the P samples. To characterize the statistical properties of the estimators, we assume that $\{U_i\}_{i=1}^{\tilde{N}}$ is a random sample from a superpopulation. This assumption is commonly used in studying survey samples from finite populations (Fuller, 2009). Let $f_U(u)$ be the probability density function of U, and $f_Z(z)$

and $f_X(x)$ be the marginal density functions of Z and X respectively. The true population size is then given by

$$N = \tilde{N} \int e(z) f_Z(z) dz. \tag{2.1}$$

Given consistent estimators $\hat{e}(z)$ and $\hat{p}(x)$ based on the E and P samples respectively, the population size can be estimated by

$$\hat{N} = \sum_{i \in \mathcal{E}} \frac{\pi_i \hat{e}(Z_i)}{\hat{p}(X_i)} \tag{2.2}$$

where π_i is the known E-sample survey weight for the sample block cluster where the ith person resides. Moreover, let $N_{\mathcal{S}}$ be the population size of a small area $\mathcal{S}$, for instance a state or congressional district. Then, $N_{\mathcal{S}}$ can be estimated by

$$\hat{N} = \sum_{i \in \mathcal{E} \cap \mathcal{S}} \frac{\pi_i \hat{e}(Z_i)}{\hat{p}(X_i)}. \tag{2.3}$$

A comprehensive study on the dual system estimation for small areas is given in Brown and Zhao (2008). If the impact of EEs were ignored by regarding $e(z) \equiv 1$, then (2.2) would actually estimate $\tilde{N}$ instead of N, and hence would cause an over-estimation of the true population size. Despite their presence, EEs have been largely ignored in most of the conventional capture-recapture experiments mainly because data information on the EEs are not available, either due to limited resource or lack of awareness.

In this paper, we study population size estimator (2.2) based on the three methods in estimating the two enumeration functions $e(z)$ and $p(x)$: the PS, the logistic regression and the local PS. Without loss of generality, we set $\pi_i \equiv 1$, which effectively makes $\mathcal{U}$ to be the population occupied by the E sample. Evaluation for the estimator with survey weighs π_i can be made using the standard approach in survey sampling (Fuller, 2009).

3 POST-STRATIFICATION

As in the case of the Census 2000 A.C.E. revision II (US Census Bureau, 2004), Z_i and X_i could differ such that the post-strata created for $e(z)$ and $p(x)$ may not be the same. Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{K_1}$ denote post-strata obtained by a stratification scheme on X, and

$\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_{K_2}$ be post-strata by stratifying Z. Let $U_i = X_i \cup Z_i$ be the combined covariates for the i-th person and $\mathcal{U}_1, \ldots, \mathcal{U}_K$ be the post-strata by stratifying with respect to U_i, where K is larger than K_1 and K_2 as the post-strata on U are finer. For each $\mathcal{U}_k$, let $\mathcal{X}(k)$ and $\mathcal{Z}(k)$ be respectively the X- and Z-strata that intercepts with $\mathcal{U}_k$. Stratifying based on two sets of covariates is more general and involved technically; and it also introduces extra sources of bias as we will show shortly.

At a given stratum $\mathcal{U}_k$, let n_k and $n_1(k)$ be the numbers of E-sample persons in $\mathcal{U}_k$ and $\mathcal{Z}(k)$ respectively, $n_{1c}(k)$ be the number of correct enumerations in $\mathcal{Z}(k)$. Then the PS estimator for the correct enumeration on $\mathcal{U}_k$ is $\hat{e}_k = n_{1c}(k)/n_1(k)$. Similarly, let $n_2(k)$ be the size of $\mathcal{P}$ in $\mathcal{X}(k)$ and $n_{12}(k)$ be the numbers of matches in $\mathcal{X}(k)$. The PS estimator for the enumeration probability in $\mathcal{U}_k$ is $\hat{p}_k = n_{12}(k)/n_2(k)$. Subsequently, the PS estimator of the total population size N is

$$\hat{N}_{ps} = \sum_{k=1}^{K} n_k \frac{\hat{e}_k}{\hat{p}_k}. \tag{3.1}$$

To characterize the property of (3.1), we define $\psi(z) = E\{p(X)|Z = z\}$ be the projection of $p(X)$ onto the space of Z, and $g(x) = P(i \in \mathcal{P}|X_i = x)$ be the P-sample enumeration function which is the P-sample counterpart of $p(x)$. To measure the heterogeneity on $\mathcal{U}_k$, we define

$$\eta_{1k} = \int_{\mathcal{Z}(k)} e(z)\psi(z) f_Z(z)dz, \eta_{2k} = \int_{\mathcal{Z}(k)} \psi(z) f_Z(z)dz, \eta_{3k} = \int_{\mathcal{X}(k)} p(x)g(x) f_X(x)dx,$$
$$\eta_{4k} = \int_{\mathcal{X}(k)} g(x) f_X(x)dx \text{ and } \eta_{5k} = \int_{\mathcal{U}_k} p(x) f_U(u)du.$$

A measure of heterogeneity on $\mathcal{U}_k$ is $\alpha_k = \eta_{1k}(\eta_{4k}/\eta_{3k})(\eta_{5k}/\eta_{2k})$, which generalizes the one proposed in Chen and Lloyd (2000) by incorporating EEs.

Theorem 1 *Under Conditions C.1-C.3 in the Appendix,*

$$E(\hat{N}_{ps}) \;=\; \tilde{N}\sum_{k=1}^{K} \alpha_k + O(1) \quad \textit{and} \quad Var(\hat{N}_{ps}) = \tilde{N}V + O(1). \tag{3.2}$$

where V is a bounded quantity whose expression is given in (A.4).

The above theorem indicates that $Var(\hat{N}_{ps}/N) = V\tilde{N}/N^2 + o(N^{-1}) = O(N^{-1})$ which converges to zero as $N \to \infty$. However, the form of the mean rings alarm. Since N is given

by (2.1), $\hat{N}_{ps}$ is asymptotically unbiased to N if and only if $\sum_{k=1}^{K} \alpha_k = \int e(z) f_Z(z) dz$ or equivalently

$$\sum_{k=1}^{K} \eta_{1k}(\eta_{5k}/\eta_{2k})(\eta_{4k}/\eta_{3k}) = \int e(z) f_Z(z) dz. \tag{3.3}$$

As η_{1k} is the averaged correct enumeration function over $\mathcal{U}_k$, a bias will occur on the population size estimator on $\mathcal{U}_k$ if the product of η_{4k}/η_{3k} and η_{5k}/η_{2k} is not one. Consequently, the relative bias of the PS estimator is

$$\tilde{N}\{\sum_{k=1}^{K} \alpha_k - \int e(z) f_Z(z) dz\}/N,$$

which does not converge to zero even as $N \to \infty$. The issue of bias becomes more apparent in the following two cases.

Case 1: $Z = X$, namely both enumeration functions $p(\cdot)$ and $g(\cdot)$ have identical covariates, which is typical for conventional PS. In this case, the post-strata $\mathcal{X}(k) = \mathcal{Z}(k) = \mathcal{X}_k$ and $\psi(z) = p(x)$. Hence, (3.3) reduces to

$$\sum_{k=1}^{K} \frac{\int_{\mathcal{X}_k} e(x)p(x)f_X(x)dx \int_{\mathcal{X}_k} g(x)f_X(x)dx}{\int_{\mathcal{X}_k} p(x)g(x)f_X(x)dx} = \int e(x) f_X(x) dx. \tag{3.4}$$

If $p(x)$ is piece-wise constant over $\{\mathcal{X}_k\}_{k=1}^{K}$, then regardless the functional form of $e(z)$ and the P-sample enumeration function $g(x)$, (3.4) is valid and hence the PS estimator $\hat{N}_{ps}$ will be asymptotically unbiased.

Case 2: $Z \neq X$. In this case, the post-strata $\mathcal{X}(k)$, $\mathcal{Z}(k)$ and $\mathcal{U}_k$ are different. To ensure (3.3), both $p(x)$ and $e(z)$ have to be piece-wise constant over their respective post-strata, which is more demanding than that in Case 1.

In summary, to ensure unbiasedness and hence consistency, the PS estimator requires either $p(x)$ or both $p(x)$ and $e(z)$ to be piece-wise constant over the post-strata. These requirements are difficult to attain in the presence of any continue covariates, for instance the age and mail return rates. The PS used in the 2000 A.C.E. had five age strata which is unable to capture the heterogeneity induced by the age. Figures 1 and 2 provide clear evidence of the age effect by plotting $p(x)$ and $e(z)$ as a function of the age for selected values of other ROASTR covariates using the 2000 A.C.E. data.

4 LOGISTIC REGRESSION

Logistic regression is an addition to the 2010 Census dual system estimation (Bell and Cohen, 2009) by assuming parametric logistic models for $e(z)$ and $p(x)$. Let $t(Z) = \{t_1(Z), \cdots, t_m(Z)\}^T$ and $s(X) = \{s_1(X), \ldots, s_q(X)\}^T$ denote some known transformations of the covariates Z and X respectively. Then, $e(z)$ and $p(z)$ are logistic in terms of $t(z)$ and $s(x)$ respectively, namely

$$e(z;\theta_1) = \frac{\exp\{t^T(z)\theta_1\}}{1+\exp\{t^T(z)\theta_1\}} \text{ and } p(x;\theta_2) = \frac{\exp\{s^T(x)\theta_2\}}{1+\exp\{s^T(x)\theta_2\}}$$

where θ_1 and θ_2 are respectively m and q dimensional unknown parameters. Mule et al. (2007) and Chen et al. (2010) considered logistic regression models with 86 main effects and interactions from the ROASTR to analyze the 2000 Census A.C.E data. Their models also include interactions between the four discrete variables (racial origin, sex, tenure and region) in ROASTR with six parametric polynomial splines. The parametric splines model the age effect continuously instead of keeping it fixed over post-strata as in the PS.

The conditional binary log likelihoods to estimate the unknown parameters θ_1 and θ_2 are

$$\begin{aligned} l_{n1}(\theta_1) &= \sum_{i\in\mathcal{E}}[e_i \log\{e(Z_i;\theta_1)\} + (1-e_i)\log\{1-e(Z_i;\theta_1)\}] \text{ and} \\ l_{n2}(\theta_2) &= \sum_{i\in\mathcal{P}}[Y_i \log\{p(X_i;\theta_2)\} + (1-Y_i)\log\{1-p(X_i;\theta_2)\}]. \end{aligned} \tag{4.1}$$

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the maximum likelihood estimates based on (4.1) based on the E and P samples respectively. the logistic regression population size estimator is

$$\hat{N}_l = \sum_{i\in\mathcal{E}} \frac{e(Z_i;\hat{\theta}_1)}{p(X_i;\hat{\theta}_2)}. \tag{4.2}$$

Let θ_1^* and θ_2^* be the probability limits of $\hat{\theta}_1$ and $\hat{\theta}_2$ as $N \to \infty$ and $R(z;\theta_2^*) = E\left\{\frac{p(X)}{p(X;\theta_2^*)}\middle| Z = z\right\}$. The properties of $\hat{N}_l$ is summarized in the following theorem.

Theorem 2 *Under Conditions C.1, C.2 and C.4 given in the Appendix,*

$$\begin{aligned} E(\hat{N}_l) &= \tilde{N}\int_{\mathcal{Z}} e(z;\theta_1^*)R(z;\theta_2^*)f_Z(z)dz + O(1), \\ Var(\hat{N}_l) &= \tilde{N}V_l + O(1) \end{aligned}$$

where $V_l = T_0 + T_1 + T_2 + T_3$ whose expression is given in (A.5) in the Appendix.

Like the PS estimator $\hat{N}_{ps}$, the variance of $\hat{N}_l$ is expected and does not create any issues. An potential issue comes with respect to the mean of $\hat{N}_l$. The condition that ensures the logistic regression estimator $\hat{N}_l$ to be asymptotically unbiased is

$$\int e(z;\theta_1^*)R(z;\theta_2^*)f_Z(z)dz = \int e(z)f_Z(z)dz. \tag{4.3}$$

Since $R(z;\theta_2^*)$ is the ratio of the true enumeration $p(x)$ and $p(x,\theta_2)$, this condition basically requires both $e(z;\theta_1)$ and $p(x;\theta_2)$ are the correct specifications of $e(z)$ and $p(x)$ respectively. This is clearer when $X = Z$ as (4.3) becomes

$$\int e(x;\theta_1^*)\frac{p(x)}{p(x;\theta_2^*)}f_X(x)dx = \int e(x)f_X(x)dx. \tag{4.4}$$

Requiring that both $e(z;\theta_1)$ and $p(x;\theta_2)$ are correct specifications maybe viewed as restrictive as requiring the two enumeration functions $e(z)$ and $p(x)$ to be piece-wise constant by the PS. When $e(z) \neq e(z;\theta_1)$ and/or $p(x) \neq p(x;\theta_2)$, $\hat{N}_l$ is subject to a systematic relative bias

$$\tilde{N}\left\{\int_{\mathcal{Z}} e(z;\theta_1^*)R(z;\theta_2^*)f_Z(z)dz - \int_{\mathcal{Z}} e(z)f_Z(z)dz\right\}/N$$

which does not diminish to zero even as $N \to \infty$. How to find reasonable parametric models for $p(x)$ and $e(x)$ based on the empirical data is the main challenge for the logistic regression estimation.

5 LOCAL POST-STRATIFICATION

We evaluate in this section the local post-stratification approach formulated based on nonparametric regression estimations of $e(z)$ and $p(x)$. We will demonstrate that, unlike the PS and the logistic regression estimators, the local PS estimator is free of the systematic bias and is consistent under much weaker conditions.

The method that supports the local PS is the nonparametric regression, which has been extensively studied (Härdle, 1990; Fan and Gijbels, 1996) for continuous covariate. The idea behind the nonparametric regression is the locally weighted least square by a kernel

function K and a smoothing bandwidth h that controls the amount of smoothness of the resulting nonparametric regression estimate. In the context of estimating $p(x)$, if all the covariates in X_i are continuous, one can choose a kernel $K(x)$ which is a radially symmetric probability density function in R^d and d is the dimension of X. The Nadaraya-Watson kernel estimator

$$\hat{p}(x) = \frac{\sum_{i\in\mathcal{P}} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i\in\mathcal{P}} K\left(\frac{x-X_i}{h}\right)} \tag{5.1}$$

is the locally weighted least square estimator that minimizes

$$\sum_{i\in\mathcal{P}} K\left(\frac{x-X_i}{h}\right)(Y_i - a)^2. \tag{5.2}$$

with respect to a. The applications of nonparametric regression methods in estimating the population size with continuous covariate were considered in Chen and Lloyd (2000, 2002) and Huggins and Hwang (2007).

As commonly encountered in surveys of social and economic studies, the covariate X_i and Z_i in the Census dual system surveys are mostly unordered discrete rather than being continuous. Indeed, four out the five covariates in ROASTR are discrete. Unordered discrete covariates can be also smoothed using the kernels proposed in Aitchison and Aitken (1976). Suppose the d-dimensional covariate $X_i = (X_i^c, X_i^u)$ where X_i^c is d_c-dimensional continuous and X_i^u is d_u-dimensional unordered discrete. Similarly, in estimating $e(z)$, we write $Z_i = (Z_i^c, Z_i^u)$ where Z_i^c is of q_c-dimensional continuous and Z_i^u is of q_u-dimensional unordered discrete.

Let X_{ij}^u denote the jth component of the unordered discrete $X_i^u = (X_{i1}^u, \cdots, X_{im_u}^u)$; and it takes c_j discrete values $\{0, 1, ..., c_j - 1\}$. Let λ_j be the smoothing bandwidth taking values in $[c_j^{-1}, 1]$. The kernel weight that smooths X_{ij}^u at a 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. Assigning $\lambda_j = c_j^{-1}$ leads to a uniform kernel 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 is the same as the standard frequency weight. The values between c_j^{-1} and 1 offer a range of choices for utilizing data information from the neighboring cells.

By multiplying the discrete kernel components, we have the productive kernel that "smooths" the entire categorical component X_i^u:

$$L(x^u, X_i^u; \vec{\lambda}) = \prod_{j=1}^{d_u} \{\lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \neq x_j^u)\}, \tag{5.3}$$

where $x^u = (x_1^u, \cdots, x_{m_u}^u)$ and $\vec{\lambda} = (\lambda_1, \cdots, \lambda_{d_u})$ is the bandwidth vector. The overall kernel weight drawn from $X_i = (X_i^c, X_i^u)$ at $x = (x^c, x^u)$ is

$$K\left(\frac{x^c - X_i^c}{h}\right) L(x^u, X_i^u; \vec{\lambda}). \tag{5.4}$$

When estimating $p(x)$, the kernel (5.4) effectively defines a local post-stratum around each $x = (x^c, x^u)$. Around each x^u (the central stratum), there exists a ring of post-strata which have only one different component from x^u. They are called the nearest neighbors of x^u. More generally, the kth nearest neighbors of x^u consists of those strata having k different components from x^u. The discrete kernels assigns the largest weight to the central stratum, and decreasing weights to other strata as their distances to x^u increase. This is similar in principle to continuous kernel weight allocation by $K\left(\frac{x^c - X_i^c}{h}\right)$ which allocates higher weights near x^c when $|(x^c - X_i^c)/h|$ is smaller.

Applying the kernel (5.4) instead of the continuous kernel K in (5.2), we have the kernel estimator of $p(x)$

$$\hat{p}(x) = \frac{\sum_{i \in \mathcal{P}} K_{h_1}(x^c - X_i^c) L(x^u, X_i^u; \vec{\lambda}_1) Y_i}{\sum_{i \in \mathcal{P}} K_{h_1}(x^c - X_i^c) L(x^u, X_i^u; \vec{\lambda}_1)} \tag{5.5}$$

where h_1 and $\vec{\lambda}_1 = (\lambda_{11}, \ldots, \lambda_{1d_u})$ are respectively the smoothing bandwidths. A similar operation on Z_i leads to

$$\hat{e}(z) = \frac{\sum_{i \in \mathcal{E}} K\left(\frac{z^c - Z_i^c}{h_2}\right) L(z^u, z_i^u; \vec{\lambda}_2) e_i}{\sum_{i \in \mathcal{E}} K\left(\frac{z^c - Z_i^c}{h_2}\right) L(z^u, Z_i^u; \vec{\lambda}_2)} \tag{5.6}$$

where $\vec{\lambda}_2 = (\lambda_{21}, \ldots, \lambda_{2q_u})$. Here without loss of generality we assume X and Z have the same number of continuous covariates. Two sets of bandwidths $(h_1, \vec{\lambda}_1)$ and $(h_2, \vec{\lambda}_2)$ are utilized in the kernel estimation of $p(x)$ and $e(z)$ respectively, reflecting that different levels

of smoothness may be applied to different functions. The smoothing parameters $(h_k, \vec{\lambda}_k)$ by minimizing respectively the cross-validation (CV) scores:

$$CV_p(h_1, \vec{\lambda_1}) = n^{-1} \sum_{i \in \mathcal{P}} \{Y_i - \hat{p}^{(-i)}_{h_1, \vec{\lambda}_1}(X_i)\}^2 \text{ and } \quad CV_e(h_2, \vec{\lambda}_2) = n^{-1} \sum_{i \in \mathcal{E}} \{e_i - \hat{e}^{(-i)}_{h_2, \vec{\lambda}_2}(X_i)\}^2,$$

where $\hat{p}^{(-i)}_{h_1, \vec{\lambda}_1}(x)$ and $\hat{e}^{(-i)}_{h_2, \vec{\lambda}_2}(x)$ are the estimators of $p(x)$ and $e(x)$ after excluding the i^{th} observation.

The most striking feature of the kernel estimators is that $p(x)$ and $e(z)$ are consistently estimated without relying on specific parametric assumptions as the kernel estimation allows data speak to tell what the underlying models are. The local PS population size estimator is

$$\hat{N}_{lp} = \sum_{i \in \mathcal{E}} \frac{\hat{e}(Z_i)}{\hat{p}(X_i)}. \tag{5.7}$$

This estimator was implemented in the empirical study reported in Chen et al. (2010) on the 2000 dual system A.C.E. data. We provide in this paper the properties of this estimator.

A key difference between the PS and the local PS is that the post-strata used in the PS are fixed whereas the local post strata are adaptive with their sizes shrinking as the amount of data information increases. The latter is achieved by letting the bandwidths $h_k \to 0$ and each $\lambda_{kj} \to 1$ when $N \to \infty$. The shrinking local post-strata leads to the removal of the bias caused by the heterogeneity as shown in the next theorem.

We need some notations first. Let $\mathcal{D}^1_{a^u} = \{y^u : \sum_{j=1}^{\dim(a^u)} I(y^u_j \neq a^u_j) = 1\}$ be a collection of the nearest neighboring cells to an a^u whose dimension is $\dim(a^u)$, which is d_u for x^u and q_u for z^u; and

$$\beta_\lambda(a^u, y^u) = \frac{1 - \sum_{j=1}^{\dim(a^u)} \lambda_j I(y^u_j \neq a^u_j)}{\sum_{j=1}^{\dim(a^u)} c_j I(y^u_j \neq a^u_j) - 1}$$

be the discrete kernel weight contributed from cell y^u to the cell a^u. We use ∇^k to denote the kth differential operator with respect to the continuous covariates. The following quantities are needed in describing the bias of $\hat{N}_{lp}$ from smoothing the continuous and the discrete

covariates in the estimation of $e(z)$:

$$b_c^e = \int \frac{\text{tr}\left[\nabla^2\{e(z)\psi(z)f_Z(z)\} - e(z)\nabla^2\{\psi(z)f_Z(z)\}\right]}{\psi(z)}dz \text{ and}$$
$$b_u^e(\vec{\lambda}_2) = \sum_{z^u\in\mathcal{Z}^u}\sum_{y^u\in\mathcal{D}^1_{z^u}} \beta_{\lambda_2}(z^u,y^u)\int_{\mathcal{Z}^c}\frac{\psi(z^c,y^u)f_Z(z^c,y^u)}{\psi(z^c,z^u)}\{e(z^c,y^u)-e(z^c,z^u)\}dz^c$$

where $\psi(z) = E\{p(x)|Z=z\}$. The corresponding terms from estimating $p(x)$ are

$$b_c^p = \int_{\mathcal{X}} \frac{\text{tr}\left[\nabla^2\{p(x)g(x)f_X(x)\} - p(x)\nabla^2\{g(x)f_X(x)\}\right]\phi(x)}{p(x)g(x)}dx \text{ and}$$
$$b_u^p(\vec{\lambda}_1) = \sum_{x^u\in\mathcal{X}^u}\sum_{y^u\in\mathcal{D}^1_{x^u}} \beta_{\lambda_1}(z^u,y^u)\int_{\mathcal{X}^c}\phi(x^c,x^u)\frac{g(x^c,y^u)f_X(x^c,y^u)}{p(x^c,x^u)g(x^c,x^u)}\{p(x^c,y^u)-p(x^c,x^u)\}dx^c$$

where $\phi(x) = E\{e(Z)|X=x\}$. Furthermore, let $\sigma_K^2 = \int t^2K(t)dt$, $R(K) = \int K^2(t)dt$, $h = \min(h_1,h_2)$, $1-\lambda_1 = \max_{1\le j\le d_u}(1-\lambda_{1j})$, $1-\lambda_2 = \max_{1\le j\le q_u}(1-\lambda_{2j})$ and $1-\lambda = \max(1-\lambda_1, 1-\lambda_2)$.

The properties of the local PS estimator $\hat{N}_{lp}$ is summarized in the following theorem.

Theorem 3 *Under the conditions C.1-C.3 and C.5 given in the Appendix,*

$$E(\hat{N}_{lp}) = N - \frac{1}{2}\sigma_K^2\tilde{N}(h_1^2b_c^p - h_2^2b_c^e) - \tilde{N}\{b_u^p(\vec{\lambda}_1) - b_u^e(\vec{\lambda}_2)\}$$
$$+ R(K)h^{-d_c}\int_{\mathcal{X}}\frac{\phi(x)\{1-p(x)\}}{p(x)g(x)}dx + o\{\tilde{N}(h^2+1-\lambda)+h^{-d_c}\}; \quad (5.8)$$
$$Var(\hat{N}_{lp}) = \tilde{N}\left\{\int_{\mathcal{X}}\frac{\phi(x)^2\{1-p(x)\}\{1-g(x)\}f_X(x)dx}{p(x)g(x)} + \int_{\mathcal{Z}}e^2(z)f_Z(z)dz\right.$$
$$\left. - \left(\int_{\mathcal{Z}}e(z)f_Z(z)dz\right)^2 + \int_{\mathcal{Z}}\frac{e(z)\{1-e(z)\}f_Z(z)}{\psi(z)}\right\} + O\{\tilde{N}(h^2+1-\lambda)\}. \quad (5.9)$$

As $\tilde{N} = N/\int e(z)f_z(z)dz = O(N)$, the variance of the local post-stratification estimator is $O(N)$ which is the same order as that of the parametric logistic regression estimator $\hat{N}_l$, despite the local PS is nonparametric. This is different from other situations where the kernel estimation has a slower rate of convergence than its parametric counterpart. The reason for the parametric and the local PS population size estimators having the same rate of convergence is due to the summation $\sum_{i\in\mathcal{E}}$ in (5.7).

Note that $b_u^p(\vec{\lambda}_1) = O(1-\lambda_1)$ and $b_u^e(\vec{\lambda}_2) = O(1-\lambda_2)$, and $h\to 0$, $(1-\lambda_1)\to 0$ and $(1-\lambda_2)\to 0$ as $N\to\infty$. The leading order bias of $\hat{N}_{lp}$ as conveyed from (5.8) is at the order of $Nh^2 + N(1-\lambda) + h^{-d_c}$. And hence the relative bias $\{E(\hat{N}_{lp})-N\}/N = O\{h^2+$

$(1-\lambda)+(Nh)^{-d_c}\}$ which diminishes to 0 as $N\to\infty$, implying that $\hat{N}_{lp}$ is asymptotically unbiased. This together with the result on the variance (5.9) means that

$$E\left(\frac{\hat{N}_{lp}-N}{N}\right)^2\to 0 \quad \text{as} \quad N\to\infty.$$

Hence, $\hat{N}_{lp}$ is ratioly consistent to N. We have seen from Theorems 1 and 2 that the same ratio consistency of $\hat{N}_{ps}$ and $\hat{N}_l$ are attained only when Conditions (3.3) and (4.3) are met respectively. Theorem 3 shows that the consistency of the local PS estimator $\hat{N}_{lp}$ is achieved under very weak conditions which are no more than requesting the existence of certain derivatives with respect to the continuous covariates and without requiring the two enumeration functions $e(z)$ and $p(x)$ to be either piece-wise constant or a specific parametric form. Although the nonparametric estimation generally requires more data to allow the data themselves to tell us what the underlying model should be, this is not an issue here as there are good amount of data in the Census dual system surveys.

6 SIMULATION STUDIES

We report results from simulation studies which were designed to confirm the theoretical analyses reported in the previous sections. To reflect the Census covariates, we chose $X=(X_1,\ldots,X_5)$, where $X_1\in[0,70]$ mimicking the age, $X_2\in\{0,\ldots,6\}$ for the racial origins, $X_3\in\{0,\ldots,4\}$ for the region, and $X_4,X_5\in\{0,1\}$ are for sex and housing tenure respectively. Then the domain of X was $\mathcal{X}=[0,70]\times\{0,\ldots,6\}\times\{0,\ldots,4\}\times\{0,1\}^2$. The Z covariate was chosen such that $Z=(X^T,Z_6)^T$ with $Z_6\in\{0,1\}$. Here, Z_6 was for a covariate only observable in the E sample. Without loss of generality, we independently generated $\{X_i\}_{i=1}^{\tilde{N}}$ and $\{Z_i\}_{i=1}^{\tilde{N}}$ respectively from super-population densities f_X and f_Z. The super-population densities were formed by assuming independence and uniformly distribution among the components in X_i and Z_i. The four discrete covariates in X_i produced 112 cells, whereas the five in Z_i determined 224 cells.

We tried to generate heterogeneity in $p(x)$ and $e(z)$ functions to be responsive to that observed from the empirical estimates on the 2000 A.C.E data, for instance those displayed in Figures 1 and 2. Let $l(x)=16x_2+8x_3+4x_4+2x_5+1$ be a one-to-one mapping from

$\{0,\ldots,6\}^4 \times \{0,\ldots,4\} \times \{0,1\}^2$ to $\{1,\ldots,112\}$. We chose $p(x) = [1+\exp\{-b_p(x);\beta^{(p)}\}]^{-1}$ where

$$b_p(x;\beta^{(p)}) = \beta^{(p)}_{l(x)0} + \sum_{k=1}^{6} \beta^{(p)}_{l(x)k} B_k(x_1), \tag{6.1}$$

where $B_k(x)$, for $k = 1,\ldots,6$, are the basis functions of a cubic B-spline with knots at $\{10,20,30\}$. We first generated $\beta^{(p)}_{l(x)} = (\beta^{(p)}_{l(x)0},\ldots,\beta^{(p)}_{l(x)6})^T$ from $N(\mu_p,\Sigma)$ independently for $l(x) = 1,\ldots,112$ where

$$\mu_p = (3.0,.5,-.5,-3.5,3.0,-2.5,1.0)^T \text{ and } \Sigma = \text{diag}(.2,.001,.001,.2,.1,.1,.1).$$

These coefficients were kept fixed once generated. Similarly, let $m(z) = 32z_2+16z_3+8z_4+4z_5+2z_6+1$ be a 1-1 mapping from $\{0,\ldots,6\}^4 \times \{0,\ldots,4\} \times \{0,1\}^3$ to $\{1,\ldots,224\}$. We set $e(z) = [1+\exp\{-b_e(z;\beta^{(e)})\}]^{-1}$ where

$$b_e(z;\beta^{(e)}) = \beta^{(e)}_{m(z)0} + \sum_{k=1}^{6} \beta^{(e)}_{m(z)k} B_k(z_1). \tag{6.2}$$

To introduce heterogeneity from Z_6, we generated the coefficient vectors $\beta^{(e)}_1, \beta^{(e)}_3, \ldots, \beta^{(e)}_{223} \overset{iid}{\sim} N(\mu_{e1},\Sigma)$ and $\beta^{(e)}_2, \beta^{(e)}_4, \ldots, \beta^{(e)}_{224} \overset{iid}{\sim} N(\mu_{e2},\Sigma)$ with $\mu_{e1} = (3.0,.8,-.3,-3.5,3.0,-2,1)^T$ and $\mu_{e2} = (1.0,.8,-.3,-3.5,3.0,-2,1)^T$. The same Σ used for the $p(x)$ coefficients $\beta^{(p)}_{l(x)}$ was used here. The setting for μ_{e2} induced a much lower $e(z)$ than that by μ_{e1}. Again, the coefficients $\beta^{(e)}_{m(x)}$ were held fixed once generated. We created the P sample enumeration functions $g(x)$ the same way as $p(x)$.

We implemented the PS and logistic regression estimation as follows. The post-strata were constructed by subdividing the age range $[0,70]$ into 4 groups: $[0,10)$, $[10,25)$, $[35,50)$ and $[50,70]$. The four age strata together with the 112 or 224 cells with respect to the discrete covariates in the P or E sample produced 448 and 672 post-strata respectively. The PS population size estimates were obtained by (3.1). The logistic regressions estimation were performed under two models. One had the correct specification as in (6.1) and (6.2), and the other was mis-specified by replacing (6.1) and (6.2) with the cubic B-splines on x_1 and z_1 with knots at $\{10,20\}$. The smoothing bandwidths for the local PS were chosen by the CV method and the population size estimates were obtained by applying (5.7).

We considered two nominal population sizes $\tilde{N} = 500,000$ and $\tilde{N} = 1,000,000$ with 2000 replications respectively in the simulation. We applied the three estimators to estimate the true population size N and the 224 sub-population sizes determined by the Z-covariates. Figure 3 presents the average absolute relative bias and the relative root mean square errors (RMSE) for the sub-population size estimates.

Figure 3 shows that both the PS and logistic regression under the mis-specification (LR-Mis) endured much larger biases than the the local post-stratification (L-PS) and the logistic regression under the correct specification (LR-True). The biases with the PS and LR-Mis did not get smaller when $\tilde{N}$ was doubled to 1 million. These confirmed our theoretical findings of systematic bias with the PS and a mis-specified logistic regression model. The systematic bias was so significant that the relative root mean square errors (right panels) did not converge to zero, causing the two sets of estimates not consistent. At the same time, the bias and the root mean square errors of the local PS and the logistic regression estimates got smaller as $\tilde{N}$ was increased indicating the consistency of these two estimators. The logistic regression under the true specification enjoyed the smallest relative bias due to its using the true models, with the local PS the second best. The relative RMSE were largely the same among the four methods when $\tilde{N} = 500,000$. However, when $\tilde{N}$ was increased, both the local PS and the logistic regression using the correct models were noticeably better due to their reduced bias.

For the estimation of the overall population size N, the absolute average biases (the standard errors) for PS, LR-True, LR-Mis and L-PS were respectively 1.3(0.16), 0.5(0.19), 1.2(0.16) and 0.8(0.18) for $\tilde{N} = 500,000$, and 1.2(0.10), 0.4(0.11), 1.1(0.10) and 0.6(0.11) for $\tilde{N} = 1,000,000$. Despite doubling the population size, the biases associated with the PS and the logistic regression using a wrong model were still very large. This is largely consistent with the results for sub-population estimation reported in Figure 3.

7 ANALYSIS OF THE 2000 A.C.E. DATA

In this section, we applied the three estimation approach on the 2000 A.C.E revision II data. The covariates used in modeling the E sample enumeration $p(x)$ were ROASTR,

and those for the correct enumeration $e(Z)$ were ROASTR plus the match coding group (MCG). Since the MCG covariate is not available outside the E sample, we constructed the population estimates

$$\hat{N} = \sum_{i \in \mathcal{C}} \frac{\hat{e}(X_i)}{\hat{p}(X_i)}$$

based on the projected correct enumeration function $e(X_i) = E\{e(Z_i)|X_i\}$ that can also be estimated by a modified local PS method discussed in Chen et al. (2010).

The PS we used had 280 post-strata after combining strata for small racial domains and for people under 18; see Schindler (2008) for details. The logistic regression employed in the analysis was the one used in Mule et al. (2007) which has 86 main effects and interactions, and four spline pieces for the age effect. The smoothing bandwidths h and $\vec{\lambda}$ in the local post-stratification were chosen by the CV method as spelt out in Section 5.

Figures 1 and 2 display the estimates of $p(x)$ and $e(x)$ based on the PS, the logistic regression and the local PS with respect to age while having the discrete covariates fixed at (Hispanic, Male, Owner, West) and (Asian, Male, Renter, Northeast). The heterogeneity is clearly seen from the fitted curves, especially from the local PS estimates. By comparing the two panels in each figure, we see that the Asian Male renters in the Northeast was less likely to be enumerated and correctly enumerated than the Hispanic Male owner in the West. The heterogeneous age effect was quite apparent as shown by dips in both functions between age 17-25 which is known to be the most difficult part of the US population to be enumerated by the Census. It is observed that the local PS estimates had some agreement with those of the PS in a global sense. However, the local PS can pick up local heterogeneity within each stratum. At the same time, the estimates by the logistic regression were much influenced by the shapes of the age splines used. And the estimates from the PS and the local PS estimates deviated substantially for the renters.

For a sub-population defined by the discrete covariate y^u, its population size can be estimated by

$$\hat{N}(y^u) = \sum_{i \in \mathcal{E}(y^u)} \frac{\hat{e}(Z_i)}{\hat{p}(X_i)}$$

where $\mathcal{E}(y^u)$ denotes the set of enumerations of the sub-population in the E sample. Let

$|\mathcal{C}(y^u)|$ be the census count. A commonly used empirical measure on the Census is the percentage of undercount $u(y^u) = \{\hat{N}(y^u) - |\mathcal{C}(y^u)|\}/\hat{N}(y^u)$. The percentages of undercount (standard errors) for the two cells considered in Figures 1 and 2 were 1.22 (0.65) for the PS, 1.71 (0.65) for the logistic regression and 1.31 (0.63) for the local PS for Hispanic Male Owner in the West, and 3.05 (1.40) for the PS, 4.16 (2.50) for the logistic regression, and 3.66 (1.66) for the local PS for Asian Male renters in the Northeast. The standard errors were obtained by the Jackknife variance estimation, (Shao and Tu, 1995; Wolter, 2007). Table 1 provides population undercount estimates for sixteen selected states. While the three estimates were largely comparable for most of the states, we do see substantial difference among them in Hawaii, Florida and VA (full spelling here and in the table). While the local PS and the logistic regression estimates were close in NH and New Jersey, they were a little different from the PS estimates. While part of the difference may be attributed to random variation, some can be due to the built-in bias associated with mis-specification by the PS and logistic regression.

8 DISCUSSION

In this paper, we have assumed that the covariates X_i and Z_i, and the statue responses Y_i and e_i are all observed completely. In reality, these variables are subject to missing values. However, the conclusion of our analysis regarding the bias of the three population size estimators, will remain valid when the missing values are replaced by their imputations (Chen et al., 2010).

We have evaluated the properties of three dual system population size estimators. While all three estimators have comparable variance at the order of N, the properties of their biases are quite different. Our analysis reveals that there can be systematic biases for the PS and logistic regression estimators when the model assumptions for the two enumeration functions deviate from the underlying true model. The logistic regression is designed to improve the PS with an actively parametric modeling on the covariates effects. It represents a methodological improvement over the PS. Our analysis reveals that the effectiveness of the logistic regressions in the dual system estimation depend on using logistic models which

are close to the real underlying models. Hence, selecting a logistic model that is close to the true model is a crucial step in implementing the approach.

Our analysis shows that the local PS estimation is model robust as it produces consistent estimates without specific model assumptions for the two enumeration functions $p(x)$ and $e(z)$. In addition to producing model-robust population size estimates, the local PS estimates can be used as empirical checks on the reasonableness of the logistic regression estimates. This is what we can infer from the case study reported in Section 7, which showed that at the State level the employed logistic regression may not be too mis-specified for most states. However, at different population aggregates, for instance the two chosen in Figures 1 and 2, the estimates can be quite different. This points to some lack of fits for the logistic regression model. The current plan in the 2010 dual system CCM study is to use the PS to evaluate the logistic regression estimates. Given the analysis done in this paper, we advocate for using the local PS instead of the PS to evaluate the logistic regression estimation.

APPENDIX: TECHNICAL DETAILS

In the Appendix, we use $I_{i\in\mathcal{E}}$ as the indicator for enumeration such that $I_{i\in\mathcal{E}} = 1$ of the individual i is enumerated in the E sample. And similarly, $I_{i\in\mathcal{P}}$ is the indicator for enumeration in the P sample. Consequently, $I_{i\in\mathcal{E}}I_{i\in\mathcal{P}} = 1$ implies i is a recapture (match). We assume the following conditions in our analysis.

C.1 Let $U_i = X_i \cup Z_i$ be the combined covariates. We assume $\{U_i\}_{i=1}^{\tilde{N}}$ is a sequence of independent and identically distributed random variables from a super-population with density f_U.

C.2 (i) The sampling of individuals in $\mathcal{E}$ and in $\mathcal{P}$ are conditional independent given the combined covariate X, namely $I_{i\in\mathcal{E}}$ and $I_{i\in\mathcal{P}}$ are independent namely conditioning on X_i so that $E(I_{i\in\mathcal{E}}I_{i\in\mathcal{P}}|X_i) = E(I_{i\in\mathcal{E}}|X_i)E(I_{i\in\mathcal{P}}|X_i) = p(X_i)g(X_i)$; (ii) $E(Y_i|U_i) = p(X_i)$ and $E(e_i|U_i) = e(Z_i)$ where $U = Z_i \cup X_i$ is the combined covariates; (iii) that an individual in E sample being an EE and its enumeration by the Census is conditional

independent given U so that $E(e_i I_{i\in\mathcal{E}}|U_i) = E(e_i|U_i)E(I_{i\in\mathcal{E}}|U_i) = e(Z_i)p(X_i)$.

C.3 $p(x)$, $g(x)f_X(x)$ and $\psi(z)f_Z(z)$ are all bounded from below by some $C > 0$ for all x and z in their support.

C.4 For estimating θ_1 and θ_2 under the logistic regression models $e(x;\theta_1)$ and $p(x;\theta_2)$, let

$$\begin{aligned}\ell_{i1}(\theta_1) &= \frac{e^{(1)}(Z_i;\theta_1)\{e_i - e(X_i,\theta_1)\}I_{i\in\mathcal{E}}}{e(Z_i;\theta_1)\{1-e(Z_i;\theta_1)\}}\\ \ell_{i2}(\theta_2) &= \frac{p^{(1)}(X_i;\theta_2)\{Y_i - p(X_i,\theta_2)\}I_{i\in\mathcal{P}}}{p(X_i;\theta_2)\{1-p(X_i;\theta_2)\}}.\end{aligned}$$

There exist unique θ_1^* and θ_2^* such that $E\{\ell_{i1}(Z_i,\theta_1^*)\} = 0$ and $E\{\ell_{i2}(X_i,\theta_2^*)\} = 0$, $E\{\ell_{i1}(\theta_1^*)\ell_{i1}^T(\theta_1^*)\}$ and $E\{\ell_{i2}(\theta_2^*)\ell_{i2}^T(\theta_2^*)\}$ are positive definite. In addition, $E\{\ell_{i1}^{(1)}(\theta_1^*)\}$ and $E\{\ell_{i2}^{(1)}(\theta_2^*)\}$ are full rank, $\ell_{i1}(Z_i;\theta_1)$ and $\ell_i(X_i;\theta_2)$ are both twice continuously differentiable with respect to θ_1 and θ_2 in neighborhoods of θ_1^* and θ_2^*, respectively. In addition, we assume that for θ_2 in a neighborhood of θ_2^*, $p(x;\theta_2) > C_2$ for some $C_2 > 0$ for all x in its support.

C.5 Assume that $p(x)$ and $e(z)$ are twice continuously differentiable in their support. The continuous kernel $K(u)$ is symmetric probability density function that has finite second moment. Let $h = \min(h_1, h_2)$, $1-\lambda_1 = \max_{1\le j\le d_u}(1-\lambda_{1j})$, $1-\lambda_2 = \max_{1\le j\le q_u}(1-\lambda_{2j})$ and $1-\lambda = \max\{1-\lambda_1, 1-\lambda_2\}$. We assume as $N\to\infty$, $h\to 0$, $Nh^{d_c}/\log^2(N)\to\infty$ and $N(1-\lambda)/\log^2(N)\to\infty$.

C.1 defines the super-population. C.2 specifies the conditional independent between the selection of person in both samples and defines the enumeration and correct enumeration functions $p(x)$ and $e(z)$ in light of X and Z may differ. C.3. ensures that for each $k = 1,\ldots,K$, $\int_{\mathcal{U}_k} p(x)f_U(u)du > C_1$ and $\int_{\mathcal{U}_k} g(x)f_U(u)du > C_1$ for some $C_1 > 0$. C.4 contains some standard conditions for the asymptotic analysis on the maximum likelihood estimation. The conditions in C.5 are commonly assumed in the nonparametric regression.

Proof of Theorem 1

Let $\hat{\eta}_{1k} = \tilde{N}^{-1}\sum_{i\in\mathcal{U}} I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} I_{i\in\mathcal{Z}(k)}$, $\hat{\eta}_{2k} = \tilde{N}^{-1}\sum_{i\in\mathcal{U}} I_{i\in\mathcal{E}} I_{i\in\mathcal{Z}(k)}$ where $\mathcal{Z}(k)$ is the stratum for estimating $e(\cdot)$. Then $\hat{e}_k = \hat{\eta}_{1k}/\hat{\eta}_{2k}$. By the law of large numbers, $\hat{\eta}_{1k} \xrightarrow{p} \eta_{1k}$ where

$$\eta_{1k} = \int e(z)p(x)I_{i\in\mathcal{Z}(k)}f(u)du = \int_{\mathcal{Z}(k)} e(z)\psi(z)f_Z(z)dz.$$

Similarly $\hat{\eta}_{2k} \xrightarrow{p} \eta_{2k} = \int\limits_{\mathcal{Z}(k)} \psi(z)f_Z(z)dz$. Let $\hat{\eta}_{3k} = \tilde{N}^{-1}\sum_{i\in\mathcal{U}} I_{i\in\mathcal{E}} I_{i\in\mathcal{P}} I_{i\in\mathcal{X}(k)}$ and $\hat{\eta}_{4k} = \tilde{N}^{-1}\sum_{i\in\mathcal{U}} I_{i\in\mathcal{P}} I_{i\in\mathcal{X}(k)}$, then $\hat{p}_k = \hat{\eta}_{3k}/\hat{\eta}_{4k}$. Apply the law of large numbers, we show $\hat{\eta}_{3k} \xrightarrow{p} \eta_{3k} = \int_{\mathcal{X}(k)} p(x)g(x)f_X(x)dx$ and $\hat{\eta}_{4k} \xrightarrow{p} \eta_{4k} = \int_{\mathcal{X}(k)} g(x)f_X(x)dx$, and $\hat{p}_k \xrightarrow{p} \eta_{3k}/\eta_{4k}$. Furthermore, $\hat{\eta}_{5k} = N^{-1}\sum_{i\in\mathcal{U}} I_{i\in\mathcal{E}} I_{i\in\mathcal{U}_k} \xrightarrow{p} \eta_{5k} = \int_{\mathcal{U}_k} p(x)f(u)du$. Define $\hat{N}_k = n_k\hat{e}_k/\hat{p}_k = \tilde{N}\{\hat{\eta}_{5k}(\hat{\eta}_{1k}/\hat{\eta}_{2k})/(\hat{\eta}_{3k}/\hat{\eta}_{4k})\}$, then $\hat{N}_p = \sum_{k=1}^K \hat{N}_k$. Note that

$$\begin{aligned}\hat{N}_k = \tilde{N}\Bigg\{&\frac{\eta_{5k}\eta_{1k}\eta_{4k}}{\eta_{2k}\eta_{3k}} + \frac{\eta_{1k}\eta_{4k}}{\eta_{2k}\eta_{3k}}(\hat{\eta}_{5k}-\eta_{5k}) + \frac{\eta_{5k}\eta_{1k}}{\eta_{2k}\eta_{3k}}(\hat{\eta}_{4k}-\eta_{4k}) + \frac{\eta_{5k}\eta_{4k}}{\eta_{2k}\eta_{3k}}(\hat{\eta}_{1k}-\eta_{1k}) \\ &- \frac{\eta_{5k}\eta_{1k}\eta_{4k}}{\eta_{2k}\eta_{3k}^2}(\hat{\eta}_{3k}-\eta_{3k}) - \frac{\eta_{5k}\eta_{1k}\eta_{4k}}{\eta_{2k}^2\eta_{3k}}(\hat{\eta}_{2k}-\eta_{2k}) + r_k\Bigg\},\end{aligned} \tag{A.1}$$

where r_k is a negligible term so that Nr_k has finite second moment and hence $Var(r_k) = O(N^{-2})$. Therefore, (A.1) implies $E(\hat{N}_k) = N\alpha_k + O(1)$ and thus $E(\hat{N}_p) = \tilde{N}\sum_{k=1}^K \alpha_k + O(1)$. This derives the bias of $\hat{N}$. Define $\gamma_{jk} = \eta_{jk}(1-\eta_{jk})$, $j = 1,\ldots,5$. It is seen

$$var(\hat{\eta}_{jk}) = \tilde{N}^{-1}\gamma_{jk},\ j = 1,\ldots,5. \tag{A.2}$$

Let $\eta_{6k} = E\{I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} I_{i\in\mathcal{P}} I_{i\in\mathcal{U}_k}\}$, $\eta_{7k} = E(I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} I_{i\in\mathcal{U}_k})$, $\eta_{8k} = E(I_{i\in\mathcal{E}} I_{i\in\mathcal{P}} I_{i\in\mathcal{U}_k})$ and

$$\omega_{ij,k} = \tilde{N}cov(\hat{\eta}_{ik},\hat{\eta}_{jk}) \text{ for } i,j = 1,\ldots,5 \text{ and } i \neq j. \tag{A.3}$$

Moreover, define $c_{1k} = \eta_{5k}\eta_{4k}/(\eta_{2k}\eta_{3k})$, $c_{2k} = \eta_{5k}\eta_{1k}\eta_{4k}/(\eta_{2k}^2\eta_{3k})$, $c_{3k} = \eta_{5k}\eta_{1k}\eta_{4k}/(\eta_{2k}\eta_{3k}^2)$, $c_{4k} = \eta_{5k}\eta_{1k}/(\eta_{2k}\eta_{3k})$ and $c_{5k} = \eta_{1k}\eta_{4k}/(\eta_{2k}\eta_{3k})$. By substituting (A.2) and (A.3) into the variance of (A.1), we establish

$$var(\hat{N}_l) = \sum_{i=1}^K\sum_{j=1}^K cov(\hat{N}_i,\hat{N}_j) = V\tilde{N} + O(1) \tag{A.4}$$

where $V = V_1 + V_2 + V_3 + V4$,

$$V_1 = \sum_{k=1}^{K_1} M_p(k)\{c_{3k}^2\gamma_{3k} + c_{4k}^2\gamma_{4k} - 2c_{3k}c_{4k}\omega_{34,k}\},$$

$$V_2 = \sum_{k=1}^{K_2} M_e(k)\{c_{1k}^2\gamma_{1k} + c_{2k}^2\gamma_{2k} - 2c_{1k}c_{2k}\omega_{12,k}\}, V_3 = \sum_{k=1}^{K} c_{5k}^2\gamma_{5k},$$

$$V_4 = -2\sum_{k=1}^{K}\{c_{1k}c_{3k}\omega_{13,k} - c_{1k}c_{4k}\omega_{14,k} - c_{1k}c_{5k}\omega_{15,k} - c_{2k}c_{3k}\omega_{23,k} + c_{2k}c_{4k}\omega_{24,k}$$

$$c_{2k}c_{5k}\omega_{25,k} + c_{3k}c_{5k}\omega_{35,k} - c_{4k}c_{5k}\omega_{45,k}\} \text{ and}$$

$M_p(k)$ or $M_e(k)$ is the total number of cells among $\mathcal{U}_1, \ldots, \mathcal{U}_K$ whose X or Z components belonging to $\mathcal{X}_k$ or $\mathcal{Z}_k$.

Proof of Theorem 2

We need to introduce some notations. Let $R_1(z;\theta_2^*) = E\left\{\frac{p(X)}{p^2(X;\theta_2^*)}\middle| Z = z\right\}$, denote $e^{(l)}(z;\theta_1)$ and $p^{(l)}(x;\theta_2)$ be the l-th derivatives of $e(z;\theta_1)$ and $p(x;\theta_2)$ with respect to θ_1 and θ_2. Similar to those in studying the PS approach, we define the projected enumeration function $\phi(x;\theta_1^*) = E\{e(Z;\theta_1^*)|X = x\}$, $\phi_2(x;\theta_1^*) = E\{e^2(Z;\theta_1^*)|X = x\}$ and

$$M_1 = \int_{\mathcal{Z}} e^{(1)}(z;\theta_1^*)R(z;\theta_2^*)f_Z(z)dz \text{ and } M_2 = \int_{\mathcal{X}} \frac{p^{(1)}(x;\theta_2^*)\phi(x;\theta_1^*)p(x)f_X(x)dx}{p^2(x;\theta_2^*)}.$$

The following quantities are defined to study the variance of (4.2):

$$D = \int \frac{e^{(1)}(z;\theta_1^*)\{e^{(1)}(z;\theta_1^*)\}^T[e(z)\{1 - 2e(z;\theta_1^*)\} + e^2(z;\theta_1^*)]\psi(z)f_Z(z)}{e^2(z;\theta_1^*)\{1 - e(z;\theta_1^*)\}^2}dz,$$

$$C = -\int \frac{e^{(2)}(z;\theta_1^*)\{e(z) - e(z;\theta_1^*)\}\psi(z)f_Z(z)}{e(z;\theta_1^*)\{1 - e(z;\theta_1^*\}}dz + D;$$

$$B = \int \frac{p^{(1)}(x;\theta_2^*)\{p^{(1)}(x;\theta_2^*\}^T[p(x)\{1 - 2p(x;\theta_2^*\} + p^2(x;\theta_2^*)]g(x)f_X(x)}{p^2(x;\theta_2^*)\{1 - p(x;\theta_2^*)\}^2}dx,$$

$$A = -\int \frac{p^{(2)}(x;\theta_2^*)\{p(x) - p(x;\theta_2^*)\}g(x)f_X(x)}{p(x;\theta_2^*)\{1 - p(x;\theta_2^*)\}}dx + B$$

$$T_0 = \int_{\mathcal{Z}} e^2(z;\theta_1^*)R_1(z;\theta_2^*)f_Z(z)dz - \left(\int_{\mathcal{Z}} e(z;\theta_1^*)R(z;\theta_2^*)f_Z(z)dz\right)^2,$$

$$T_1 = M_1^T C^{-1}DC^{-1}M_1, T_2 = M_2^T A^{-1}BA^{-1}M_2 \text{ and}$$

$$T_3 = -2\int_{\mathcal{X}} \phi_2(x;\theta_1^*)\{p^{(1)}(x;\theta_2^*)\}^T A^{-1}p^{(1)}(x;\theta_2^*)g(x)f_X(x)p^{-3}(x;\theta_2^*)dx. \tag{A.5}$$

Given Condition C.4, we can show that $\hat{\theta}_1$ and $\hat{\theta}_2$ converge in probability to θ_1^* and θ_2^* respectively as $N \to \infty$. We note that if the parametric models $e(\cdot, \theta_1)$ and $p(\cdot, \theta_2)$ are correctly specified, then θ_1^* and θ_2^* are the true parameters of the models. If the parametric models are mis-specified, θ_1^* and θ_2^* correspond to parameter values of certain parametric models that are closest to the mis-specified models under the Kullback-Leibler (KL) distance (White, 1982).

We shall only develop the expansion for $\hat{\theta}_2$ and note that the case for $\hat{\theta}_1$ follows in exactly the same way. By definition, the MLE $\hat{\theta}_2$ is the root of

$$0 = \tilde{N}^{-1} \sum_{i \in \mathcal{U}} \frac{p^{(1)}(X_i; \theta)\{I_{i \in \mathcal{E}} - p(X_i; \theta)\} I_{i \in \mathcal{P}}}{p(X_i; \theta)\{1 - p(X_i; \theta)\}} =: \tilde{N}^{-1} \sum_{i \in \mathcal{U}} \ell_i(\theta),$$

where $p^{(1)} = \partial p / \partial \theta$. We note that the limit of $\hat{\theta}_2$ denoted by θ_2^* satisfies

$$\int \frac{p^{(1)}(x; \theta_2^*)\{p(x) - p(x; \theta_2^*)\} g(x) f_X(x)}{p(x; \theta_2^*)\{1 - p(x; \theta^*)\}} = 0,$$

where the $p(x)$ and $g(x)$ are the enumeration functions of $\mathcal{E}$ and $\mathcal{P}$ samples, $f(x)$ is the density of the super-population. However, we note that $p(x)$ may not equal to $p(x; \theta^*)$ pointwise. This represents the fact that the parametric model may be mis-specified. We apply Taylor's expansion on the above equation in a neighborhood of θ^*,

$$0 = \tilde{N}^{-1} \sum_{i \in \mathcal{U}} \ell_i(\hat{\theta}_2) = \tilde{N}^{-1} \sum_{i \in \mathcal{U}}^{n} \ell_i(\theta_2^*) + \tilde{N}^{-1} \sum_{i \in \mathcal{U}} \ell_i^{(1)}(\theta_2^*)(\hat{\theta}_2 - \theta_2^*) + R_n(\theta_2) \tag{A.6}$$

where $R_n(\theta_2)$ is the remainder term whose kth component is given by

$$R_{nk} = \tilde{N}^{-1}(\hat{\theta}_2 - \theta_2^*)^T \{\partial^2 \ell_{ik}(\tilde{\theta}_2) / \partial \theta_2 \partial^T \theta_2\}(\hat{\theta}_2 - \theta_2^*)$$

where $||\tilde{\theta} - \theta^*|| \leq ||\hat{\theta} - \theta^*||$. Under the regularity conditions, $R_n = O_p(N^{-1})$ and $\tilde{N} R_n$ has bounded second moment. By law of large numbers, $\tilde{N}^{-1} \sum_{i \in \mathcal{U}} \ell_i^{(1)}(\theta_2^*) \xrightarrow{p} E\{\ell_i^{(1)}(\theta_2^*)\} =: A(\theta_2^*)$. Then, we have

$$\hat{\theta}_2 - \theta_2^* = A^{-1}(\theta_2)\{1 + o_p(1)\}\{\tilde{N}^{-1} \sum_{i \in \mathcal{U}} \ell_i(\theta_2^*) + R_n\}. \tag{A.7}$$

Let $B(\theta_2^*) = E\{\ell_i(\theta_2^*)\ell_i^T(\theta_2^*)\}$, then we have $var(\hat{\theta}) = \tilde{N}^{-1} A^{-1}(\theta_2^*) B(\theta_2^*) A^{-1}(\theta_2^*) + o(N^{-1})$. In particular, by letting $p_\theta(x) = p(x; \theta)$ and $p_\theta^{(2)} = \partial^2 p / \partial \theta \partial \theta^T$, we have

$$A(\theta) = -\int_X \frac{p_\theta^{(2)}(p - p_\theta) g f_X}{p_\theta(1 - p_\theta)} + B(\theta) \text{ and } B(\theta) = \int_X \frac{p_\theta^{(1)}\{p_\theta^{(1)}\}^T \{p(1 - 2p_\theta\} + p_\theta^2\} g f_X}{p_\theta^2 (1 - p_\theta)^2}$$

where the dummy variable in the integration is suppressed, i.e. $\int f(x)dx = \int f$. Next, we develop the following expansion for $\hat{N}_l$ given by (4.2). We have

$$\hat{N}_l = \sum_{i\in\mathcal{U}} I_{i\in\mathcal{E}} \left\{ \frac{e_i(Z_i;\theta_1^*)}{p_i(X_i;\theta_2^*)} + \frac{\{e_i^{(1)}(Z_i;\theta_1^*)\}^T(\hat{\theta}_1 - \theta_1^*)}{p_i(X_i;\theta_2^*)} \right.$$
$$\left. - \frac{e_i(Z_i;\theta_1^*)\{p_i^{(1)}(X_i;\theta_2^*)\}^T(\hat{\theta}_2 - \theta_2^*)}{p_i^2(X_i;\theta_2^*)} + O_p(N^{-1}) \right\}. \tag{A.8}$$

Then $E(\hat{N}_l)$ is established from (A.8). To derive the variance part of Theorem 2, we note that $I_{i\in\mathcal{E}}$ appears in (A.6) and thus a non-ignorable correlation between the first and third term is induced in (A.8). And we show that the remaining between terms correlations in (A.8) are negligible. Then by taking variance operation on (A.8), we established the variance part of Theorem 2.

Proof of Theorem 3

We note that $\hat{N}$ can be written as $\hat{N} = \sum_{i\in U} \frac{\hat{e}(Z_i)}{\hat{p}(X_i)} I_{i\in\mathcal{E}}$ and by Taylor expansion,

$$\hat{N} = \sum_{i\in U} \frac{\hat{e}(Z_i)}{\hat{p}(X_i)} I_{i\in\mathcal{E}} = t_1 + t_2 - t_3 - t_4 + t_5\{1 + O_p(1)\}, \text{ where} \tag{A.9}$$

$$t_1 = \sum_{i\in U} \frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)}, t_2 = \sum_{i\in U} \frac{\{\hat{e}(Z_i) - e(Z_i)\}I_{i\in\mathcal{E}}}{p(X_i)}, t_3 = \sum_{i\in U} \frac{e(Z_i)\{\hat{p}(X_i) - p(X_i)\}I_{i\in\mathcal{E}}}{p^2(X_i)}$$
$$t_4 = \sum_{i\in U} \frac{I_{i\in\mathcal{E}}}{p^2(X_i)}\{\hat{e}(Z_i) - e(Z_i)\}\{\hat{p}(X_i) - p(X_i)\} \text{ and } t_5 = \sum_{i\in U} \frac{e(Z_i)I_{i\in\mathcal{E}}}{p^3(X_i)}\{\hat{p}(X_i) - p(X_i)\}^2.$$

Existing theory on nonparametric regression (Härdle, 1990) ensures that $\hat{p}(\cdot) \xrightarrow{p} p(\cdot)$ and $\hat{e}(\cdot) \xrightarrow{p} e(\cdot)$ uniformly over the supports of $e(\cdot)$ and $p(\cdot)$ under Condition C.5. Thus the expansion (A.9) is valid. Let $\mathcal{K}_{h,\vec{\lambda}}(x,y) = K_h(x^c - y^c)L(x^u, y^u, \vec{\lambda})$, we define

$$\hat{\eta}_1(z) = \tilde{N}^{-1} \sum_{j=1}^{N} \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j) I_{j\in\mathcal{E}} I_{j\in\tilde{\mathcal{E}}}, \ \hat{\eta}_2(z) = \tilde{N}^{-1} \sum_{j=1}^{N} \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j) I_{j\in\mathcal{E}},$$
$$\hat{\eta}_3(x) = \tilde{N}^{-1} \sum_{j=1}^{N)} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_j) I_{j\in\mathcal{P}} I_{j\in\mathcal{E}} \text{ and } \hat{\eta}_4(x) = \tilde{N}^{-1} \sum_{j=1}^{N} \mathcal{K}_{h_1,\vec{\lambda}_1}(x, X_j) I_{j\in\mathcal{P}}.$$

Therefore, we show that

$$E\{\hat{\eta}_1(z)\} = \int \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)p(X_i)e(Z_i)f(U_i)dU_i = \int \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)e(Z_i)\psi(Z_i)f_Z(Z_i)dZ_i$$
$$= e(z)\psi(z)f_Z(z) + \tfrac{1}{2}h_2^2\sigma_K^2 \mathrm{tr}[\nabla^2\{e(z)\psi(z)f_Z(z)\}]$$
$$+ \sum_{y^u \in D^1_{z^u}} \beta_{\lambda_2}(z^u, y^u)e_{y^u}(z^c)\psi_{y^u}(z^c)f_{Z,y^u}(z^c) + O(h_2^2) + O(1-\lambda_2). \quad \text{(A.10)}$$

We may derive $E\{\hat{\eta}_2(z)\}$, $E\{\hat{\eta}_3(x)\}$ and $E\{\hat{\eta}_4(x)\}$ all similarly. By letting $\eta_1(z) = e(z)\psi(z)f_Z(z)$, $\eta_2(z) = \psi(z)f_Z(z)$, $\eta_3(x) = p(x)g(x)f(x)$, $\eta_4(x) = g(x)f(x)$,

$$\hat{e}(z) = e(z) + \frac{\hat{\eta}_1(z) - \eta_1(z)}{\eta_2(z)} - \frac{\eta_1(z)\{\hat{\eta}_2(z) - \eta_2(z)\}}{\eta_2^2(z)}\{1 + o_p(1)\} \text{ and}$$
$$\hat{p}(x) = p(x) + \frac{\hat{\eta}_3(x) - \eta_3(x)}{\eta_4(x)} - \frac{\eta_3(x)\{\hat{\eta}_4(x) - \eta_4(x)\}}{\eta_4^2(x)}\{1 + o_p(1)\}. \quad \text{(A.11)}$$

We note that $E(t_4) = O(h^4) + O\{(1-\lambda)^2\}$ and

$$E(t_5) = R(K)h^{-d_c}\int_{\mathcal{X}} \frac{\phi(1-p)}{pg} + o(h^{-d_c}). \quad \text{(A.12)}$$

Hence, the bias part of Theorem 3 is concluded from (A.9) by summarizing (A.10), (A.11) and (A.12).

To establish the variance of $\hat{N}$, we need to derive $cov(t_i, t_j)$ for $i, j = 1, 2, 3$. We first show that

$$var(t_1) = \sum_{i=1}^{\tilde{N}} var\left\{\frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)}\right\} = \tilde{N}\left\{\int_{\mathcal{X}} \frac{\phi^2(z)}{p(x)}f_X(x)dx - \left(\int_{\mathcal{Z}} e(z)f_Z(z)dz\right)^2\right\}. \quad \text{(A.13)}$$

Define

$$\alpha_{1,ab} = \frac{I_{a\in\mathcal{E}}}{p(X_a)\eta_2(Z_a)}\mathcal{K}_{h_2,\vec{\lambda}_2}(Z_a, Z_b)I_{b\in\mathcal{E}}I_{b\in\tilde{\mathcal{E}}} \text{ and } \alpha_{2,ab} = \frac{e(Z_a)I_{a\in\mathcal{E}}}{p(X_a)\eta_2(Z_a)}\mathcal{K}_{h_2,\vec{\lambda}_2}(Z_a, Z_b)I_{b\in\mathcal{E}}.$$

By ignoring smaller order terms, we note from (A.9) and (A.11) that

$$var(t_2) = \tilde{N}^{-2}\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}}\sum_{k=1}^{\tilde{N}}\sum_{l=1}^{\tilde{N}} cov\{(\alpha_{1,ik} - \alpha_{2,ik}), (\alpha_{1,jl} - \alpha_{2,jl})\}.$$

By the definition of the kernel $\mathcal{K}(x, y)$ and the independence assumption, it is true that

$$var(t_2) = \left[\tilde{N}^{-2}\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}}\sum_{k=1}^{\tilde{N}} cov\{(\alpha_{1,ik} - \alpha_{2,ik}), (\alpha_{1,jk} - \alpha_{2,jk})\}\right]\{1 + O(1-\lambda_2)\}. \quad \text{(A.14)}$$

Furthermore, let $\lambda_a^{(b)} = \prod_{j=1}^{d_u} \lambda_{aj}^b$, we have

$$\begin{aligned} cov\{(\alpha_{1,ik} - \alpha_{2,ik}),(\alpha_{1,jk} - \alpha_{2,jk})\} &= \lambda_2^{(2)} \int_{\mathcal{Z}} \frac{e(z)\{1-e(z)\}f_Z(z)}{\psi(z)} + O(h^2) \\ &= \int_{\mathcal{Z}} \frac{e(z)\{1-e(z)\}f_Z(z)}{\psi(z)} + O(h_2^2) + O(1-\lambda_2) \end{aligned}$$

where $\lambda^{(2)} = 1 - 2\sum_{j=1}^{d_u}(1-\lambda_j) = 1 + O(1-\lambda_2)$. Therefore,

$$var(t_2) = \tilde{N} \int_{\mathcal{Z}} \frac{e(z)\{1-e(z)\}f_Z(z)}{\psi(z)} + O(\tilde{N}h_2^2) + O\{\tilde{N}(1-\lambda_2)\}. \tag{A.15}$$

Let

$$\alpha_{3,ab} = \frac{e(Z_a)I_{a\in\mathcal{E}}}{p^2(X_a)\eta_4(X_a)}\mathcal{K}(X_a, X_b)I_{b\in\mathcal{E}}I_{b\in\mathcal{P}} \text{ and } \alpha_{4,ab} = \frac{e(Z_a)I_{a\in\mathcal{E}}}{p(X_a)\eta_4(X_a)}\mathcal{K}(X_a, X_b)I_{b\in\mathcal{P}}.$$

Similar to (A.14),

$$\begin{aligned} var(t_3) &= \left[\tilde{N}^{-2}\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}}\sum_{k=1}^{\tilde{N}} cov\{(\alpha_{3,ik} - \alpha_{4,ik}), (\alpha_{3,jk} - \alpha_{4,jk})\}\right][1 + O\{A(\vec{\lambda}_1)\}] \\ &= \tilde{N}\int_{\mathcal{X}} \frac{\phi^2(x)\{1-p(x)\}f_X(x)dx}{p(x)g(x)} + O(\tilde{N}h_1^2) + O\{\tilde{N}(1-\lambda_1)\}. \end{aligned} \tag{A.16}$$

By Condition C.1,

$$cov(t_1, t_2) = O(\tilde{N}h_2^2) + O\{\tilde{N}(1-\lambda_2)\} \text{ and } cov(t_2, t_3) = O(\tilde{N}h_1^2) + O\{\tilde{N}(1-\lambda_2)\}. \tag{A.17}$$

Finally,

$$\begin{aligned} cov(t_1,t_3) &= \tilde{N}^{-1}\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}}\sum_{k=1}^{\tilde{N}} cov\left\{\frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)}, (\alpha_{3,jk} - \alpha_{3,jk})\right\} \\ &= \tilde{N}^{-1}\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{\tilde{N}} cov\left\{\frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)}, (\alpha_{3,ji} - \alpha_{3,ji})\right\}\{1 + O(1-\lambda_1)\} \\ &= \tilde{N}\int_{\mathcal{X}} \frac{\phi^2(x)\{1-p(x)\}f_X(x)dx}{p(x)g(x)} + O(\tilde{N}h_1^2) + O\{\tilde{N}(1-\lambda_1)\}. \end{aligned} \tag{A.18}$$

In summary of these results (A.13)-(A.18), we conclude the variance part of Theorem 3.

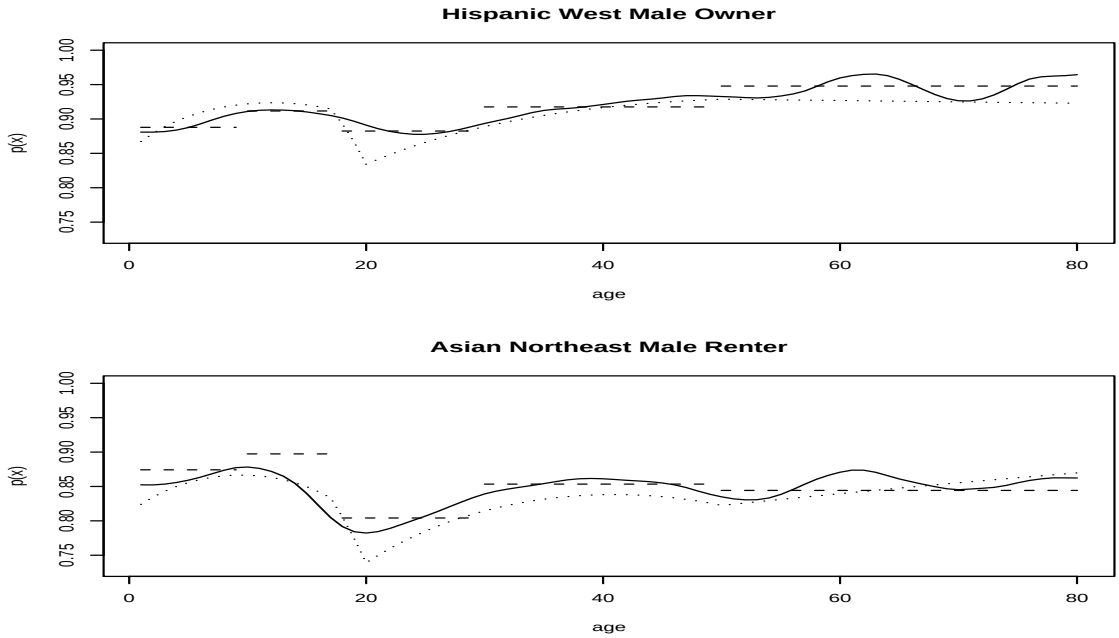

Figure 1: Estimated $p(x)$ by post-stratification(PS), logistic regression(LR) and the local post-stratification(L-PS). Dash line: PS, Dotted line: LR and Solid Line: L-PS.

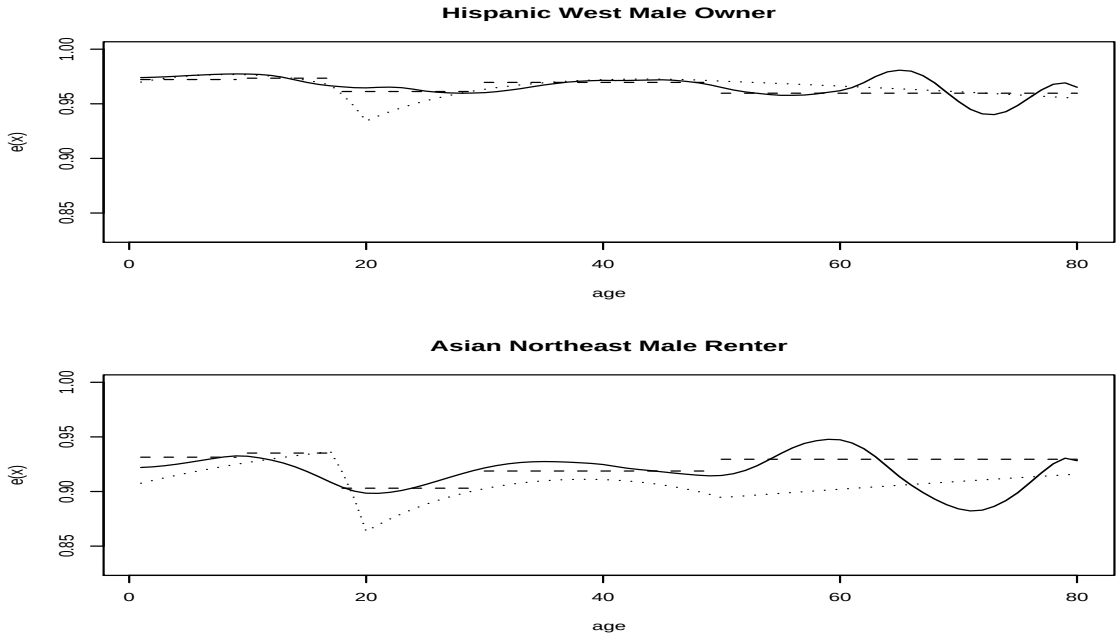

Figure 2: Estimated $e(x)$ by post-stratification(PS), logistic regression(LR) and the local post-stratification(L-PS). Dash line: PS, Dotted line: LR and Solid Line: L-PS.

STATE	L-PS	PS	LR
Northeast Region			
CT	1.07(0.26)	1.05(0.26)	1.1(0.27)
NH	-0.14(0.29)	-0.16(0.29)	-0.1(0.3)
NJ	0.86(0.26)	0.83(0.26)	0.94(0.28)
PA	0.71(0.26)	0.7(0.26)	0.74(0.27)
Midwest Region			
MI	0.59(0.18)	0.6(0.18)	0.58(0.18)
MN	0.29(0.17)	0.31(0.17)	0.28(0.17)
MO	0.3(0.17)	0.3(0.17)	0.28(0.18)
OH	0.73(0.17)	0.73(0.17)	0.72(0.18)
South Region			
FL	1.7(0.24)	1.72(0.24)	1.63(0.25)
MS	1.51(0.24)	1.48(0.24)	1.45(0.24)
OK	2.18(0.27)	2.18(0.26)	2.22(0.28)
VA	2.16(0.23)	2.09(0.22)	2.03(0.24)
West Region			
HI	1.13(0.86)	0.92(0.85)	0.87(0.84)
OR	1.21(0.32)	1.18(0.31)	1.2(0.32))
UT	1.28(0.32)	1.21(0.32)	1.32(0.33)
WA	1.2(0.31)	1.16(0.31)	1.19(0.31)

Table 1: State level research estimates of undercount percentage and their standard errors (in parentheses) for the local post-stratification (L-PS), the post-stratification (PS) and the logistic regression (LR). The abbreviation...

References

Aitchison, J. and Aitken, C. (1976), "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

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.

Anderson, M. and Feinberg, S. E. (1999), "To sample or not to sample? The 2000 Census Controversy," *Journal of Interdisciplinary History*, 30, 1–36.

— (2002), *Who Counts? The Politics of Census-Taking in Contemporary America*, Heldref Publications.

Ayhan, Ö. H. and Ekni, S. (2003), "Coverage error in population censuses: the case of Turkey," *Survey Methodology*, 29, 155–165.

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.

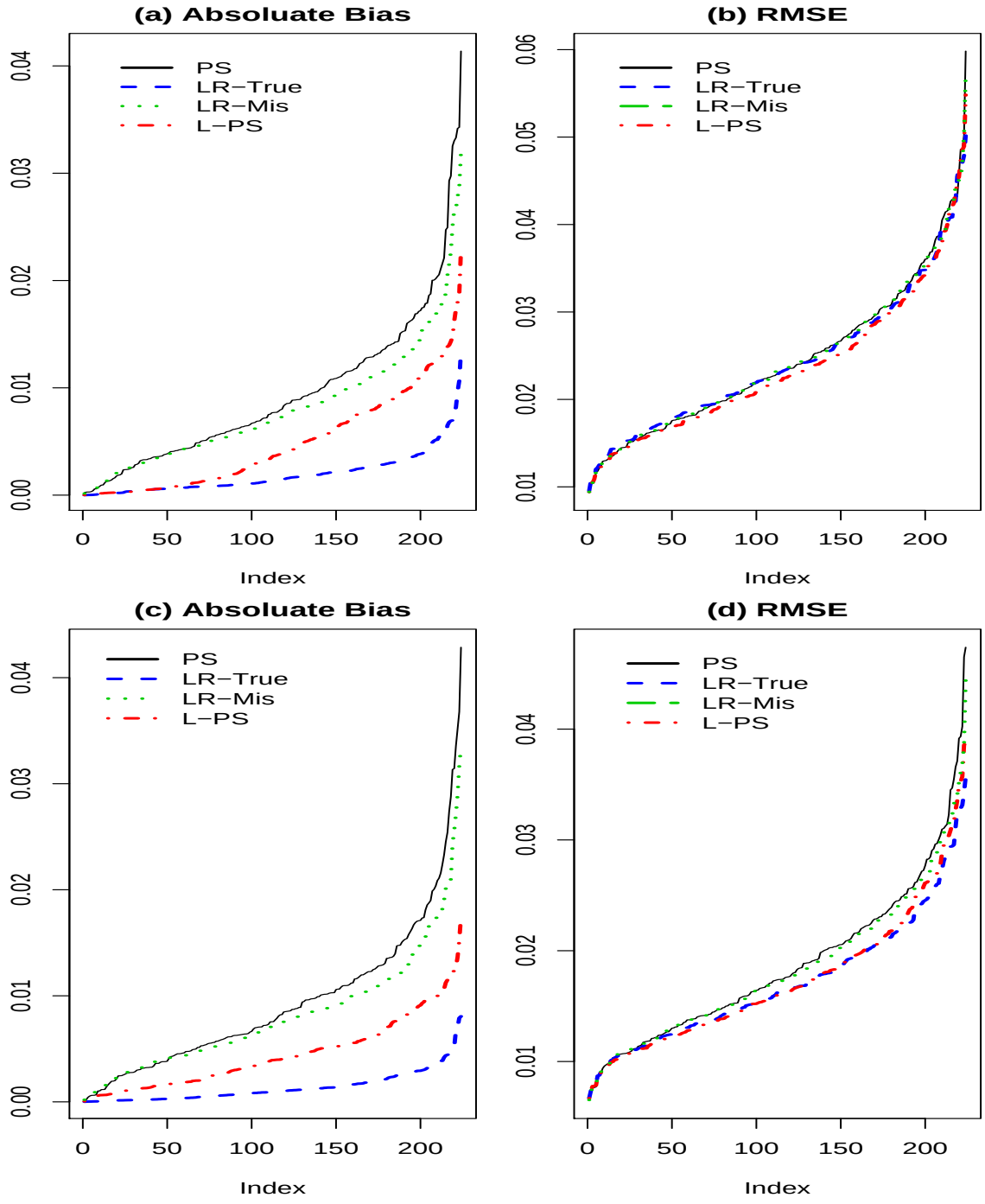

Figure 3: The relative absolute biases and the root mean square errors of four population size estimators: the post-stratification (PS), logistic regression with mis-specification (LR-Mis) and with the true specification (LR-true), and the local PS (L-PS) for the 224 cells; $\tilde{N} = 500,000$ in Panels (a) and (b) and $\tilde{N} = 1,000,000$ in Panels (c) and (d).

Bell, R. and Cohen, M. L. (eds.) (2009), *Coverage Measurement in the 2010 Census*, National Academies Press.

Bell, W. R. (1993), "Using information from Demographic analysis in post-enumeration

survey estimation," *Journal of the American Statistical Association*, 88, 1106–1118.

Brown, L. and Zhao, Z. (2008), "Alternative formulas for synthetic dual system estimation in 2000 census," *IMS Collections. Probability and Statistics: Essays in Honor of David A. Freedman*, 2, 90–113.

Cantwell, P. and Childers, D. (2001), "Accuracy and coverage evaluation survey: a change to the imputation cells to address unresolved resident and enumeration status," *DSSD Census 2000 Procedures and Operations Memorandum Series*, #Q-44.

Census Customer Service (2002), *Census coverage survey: evaluation report*, Office for National Statistics, UK.

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 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. (2010), "Local post-stratification and diagnostics in dual system accuracy and coverage evaluation for the U.S. Census," *Journal of the American Statistical Association*, 105, 105–119.

Darroch, J., Fienberg, S. E., Glonek, G. F., and Junker, B. W. (1993), "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability," *Journal of the American Statistical Association*, 88, 1137–1148.

Dunstan, K., Heyen, G., and Paice, J. (2001), "Measuring census undercount in Australia and New Zealand," *Demography working paper No. 99/4, Australian Bureau of Statistics.*

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

Fuller, W. A. (2009), *Sampling Statistics*, Wiley, New York.

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.*

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Hogan, H. (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*.

— (2003), "The accuracy and coverage evaluation: theory and design," *Survey Methodology*, 29, 129–138.

Huggins, R. and Hwang, W. H. (2007), "Non-parametric estimation of population size from capturerecapture data when the capture probability depends on a covariate," *Journal of Royal Statistical Society, Series C*, 56, 429–443.

Mule, T., Schellhamer, T., Malec, D., and Maples, J. (2007), "Using continuous variables as modeling covariates for net coverage estimation," *US Census Bureau DSSD 2010 Census Coverage Measurement Memorandum Series 2010-E-09-R1*.

Pollock, K. H. (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.

Rhind, D. (2003), *The 2001 Census in Westminister.*, Statistics Commission, UK.

Schindler, E. (2008), "Post-stratification by age for small intervals," *Census 2000 Procedures and Operations Memorandum Series Q-94*.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer.

US Census Bureau (2004), *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*, US Census Bureau.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

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.