International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2011), 79, 3, 336–361 doi:10.1111/j.1751-5823.2011.00150.x

Properties of Census Dual System Population Size Estimators

Song Xi Chen[1,2] and Cheng Yong Tang[3]

[1]*Department of Statistics, Iowa State University, Iowa, USA*
[2]*Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, China*
E-mail: songchen@iastate.edu
[3]*Department of Statistics and Applied Probability, National University of Singapore, Singapore*
E-mail: statc@nus.edu.sg

Summary

We study parametric and non-parametric approaches for assessing the accuracy and coverage of a population census based on dual system surveys. The two parametric approaches being considered are post-stratification and logistic regression, which have been or will be implemented for the US Census dual system surveys. We show that the parametric model-based approaches are generally biased unless the model is correctly specified. We then study a local post-stratification approach based on a non-parametric kernel estimate of the Census enumeration functions. We illustrate that the non-parametric approach avoids the risk of model mis-specification and is consistent under relatively weak conditions. The performances of these estimators are evaluated numerically via simulation studies and an empirical analysis based on the 2000 US Census post-enumeration survey data.

Key words: Capture-recapture; discrete covariate; erroneous enumeration; kernel smoothing; model bias; population size estimation.

1 Introduction

The US decennial Census is a major information source providing counts of the entire population and those sub-populations defined by states, congressional districts, and demographic groups. A comprehensive account of the US Census is given in Anderson & Feinberg (2001). Since the Census is such a large-scale data collection process, two types of survey errors are inevitably present. One type is omission error which occurs when genuine Census persons are missed (omitted), causing a population under-count. The other type is enumeration error (EE) due to invalid records in the Census, such as fictitious or duplicate persons, and those enumerated in wrong locations. The EEs inflate the Census count. Both errors can affect the accuracy of the census population counts significantly (Hogan, 2003). Since the omission error has been steadily reduced in recent US Censuses by improved survey effort, design, and execution, the impacts of the EEs have become more apparent (US Census Bureau, 2004).

To assess Census coverage and accuracy, the US Census Bureau has conducted post-enumeration surveys since the 1980 Census (Anderson & Feinberg, 2002; US Census Bureau,

2004). These post-enumeration surveys have attempted to estimate census undercount and and overcount; see Hogan (1993, 2003). One part of the post-enumeration survey is largely the Census itself but restricted to randomly selected sample block clusters. The Census enumerations obtained over the selected block clusters constitute the E sample. A primary purpose of the E sample is to identify and measure the EEs via extensive record checking and follow-up. The second part of the post-enumeration survey is an independently conducted enumeration of the same block-clusters occupied by the E sample, which gives rise to the P sample. The E and P samples support a capture-recapture estimation methodology via a comprehensive matching process for the records between the two samples. For assessing the omission errors, the matched enumerations (recaptures) that appeared in both samples together with those enumerations (captures) only in the P sample are used to estimate the E-sample enumeration probability. See Hogan (1993, 2000a, 2000b), Haberman *et al.* (1998), Bell (1993), Darroch *et al.* (1993), Chao & Tsay (1998), Brown & Zhao (2008), and Chen *et al.* (2010) for more specific discussions and approaches for the Census dual system methodology. Anderson & Feinberg (1999) contains a comprehensive overview of some critical issues and controversies surrounding the Census. Wolter (1986) and Pollock (1991) provide the theory behind capture-recapture based population size estimation. Besides the United States, countries including Australia, New Zealand, Turkey, Switzerland and the United Kingdom also utilize similar dual system methods to evaluate their national censuses for the coverage and accuracy; see Dunstan *et al.* (2001), Ayhan & Ekni (2003), Abbott (2007), and Brown *et al.* (1999) among others.

As commonly encountered in surveying either a human or wildlife population, omissions and EEs do not occur homogeneously across the population. Certain sub-populations are more prone to errors than others. For the US census, Racial Origin (RO), Age (A), Sex (S), housing Tenure (T), and geographical Region (R) (hence ROASTR) are variables known to substantially contribute to the heterogeneity in those errors for the US Census (Hogan, 1993, 2003). Therefore, modelling the probabilities of these two errors as functions of the variables (for instance ROASTR) is a main task of the US Census accuracy and coverage assessment.

The main approach used to address heterogeneity in dual system estimation prior to the 2010 US Census had been post-stratification (PS, Hogan, 1993). Given a set of variables, for instance ROASTR, PS subdivides the support of the variables into non-overlapping post-strata. In each post-stratum, a population size estimate is obtained by applying the classical Petersen model (Petersen, 1896). The underlying assumption of the PS approach is that the omission and EEs occurred homogeneously within each post-stratum. However, Hogan (1993, 2003) showed that the PS failed to eliminate the homogeneity within individual post-stratum, especially with respect to continuous variables like age. To overcome the limitations of the PS, the Census Bureau has decided to implement a logistic regression approach in the 2010 US Census coverage assessment (Bell & Cohen, 2008). Logistic regression provides more flexible parametric modelling than the PS and allows extrapolation in areas with sparse observations. It was applied in analysing the 1990 US Census dual system surveys by Alho *et al.* (1993); see also Mule *et al.* (2007). However, since logistic regression, like PS, is model based, there is a risk of model mis-specification, which may produce a systematic bias in the population estimation.

In this paper we first study the properties of PS and logistic regression approaches in a unified framework incorporating the features of the US Census post-numeration surveys. In particular, we attempt to evaluate the impacts of both omission and EEs. Despite their presence, EEs have been less studied in most of the conventional capture-recapture surveys mainly either due to limited data information on the EEs or lack of awareness. Our work extends the conventional theory on population size estimation (Wolter, 1986; Pollock, 1991) that focused primarily on omission errors. We first show results illustrating that population estimators based on PS and

logistic regression are generally biased unless their respective model assumptions are correct. To alleviate the bias due to model mis-specification, we then carry out a study on a non-parametric local post-stratification (local PS) method recently proposed by Chen *et al.* (2010) in an empirical study using the 2000 US Census post-enumeration survey data. Instead of having fixed post-strata as in the PS, the local PS produces local post-strata via non-parametric kernel smoothing method without specifying a parametric model. A local post-stratum shrinks when the number of observations gets larger around it. This feature leads to the removal of the heterogeneity and consistent population size estimation under much weaker conditions than those for the PS and the logistic regression.

The rest of this paper is structured as follows. Section 2 overviews population size estimation based on dual system surveys. The properties of the population size estimates using PS, logistic regression, and local PS are presented in Sections 3, 4, and 5, respectively. Section 6 reports some simulation results. An empirical study on 2000 US Census post-enumeration survey 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, $\mathcal{E}$ and $\mathcal{P}$ to denote the sets of enumerations by the E and P samples, respectively, and X denote 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 i-th person in $\mathcal{U}$ with covariates X_i is $p(X_i) = P(i \in \mathcal{C}|X_i)$. If the enumeration function $p(x)$ was known and there were no EEs, the Horvitz–Thompson type estimator $\sum_{i\in\mathcal{C}} p^{-1}(X_i)$ would be a consistent and unbiased estimator for the size of $\mathcal{U}$. However, $p(x)$ is unknown in reality. Its estimation requires a capture-recapture approach supported by the E and P samples along with the assumption that the E and P samples are conditionally independent given each person to be enumerated.

The input to capture-recapture is obtained via a comprehensive matching operation that includes computer matching, field follow-up, and clerical checks. The purpose is to match each P-sample person to an E-sample person or otherwise, classifying each P-sample person as (i) a match to an E-sample person (recapture) and (ii) not a match. There are actually unresolved matches when match status cannot be established. These cases can be viewed as missing response variables, an issue we ignore in this paper to simplify our exposition without altering the main conclusions of the paper. We want to point out that in the US Census coverage measurement operation, while P-sample records are matched to those in the E sample, no attempt has been made to systematically match E-sample persons to the P-sample ones. This is because doing so would require additional field operation and resources. There are also difficulties due to the fact that the P sample is a survey so that those design aspects of the survey can make the matching a rather complicated operation. Hence, the capture-recapture used in the US Census dual system estimation is only one way, rather than two ways as is more common practiced for many conventional capture-recapture studies. As a consequence, we only consider estimation of the E-sample enumeration probability rather than the probability of being enumerated by either the E or the P sample. The latter is needed for the more efficient two way estimation; see Alho *et al.* (1993) and Chen & Lloyd (2002) for discussions. We also note that the entire US Census is being computer matched to identify duplicates. However, such information has not been used for the dual system estimation to our best knowledge.

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 covariates X_i matches (or does not match) to an

Table 1
Cross-classifications of (a) counts and (b) expected values of counts for the population $\mathcal{U}$, E sample $\mathcal{E}$, P sample $\mathcal{P}$, and erroneous enumeration $\mathcal{U}_e$.

	$\mathcal{P} \cap \mathcal{U}$	$\mathcal{P}^c \cap \mathcal{U}$	$\mathcal{P}^c \cap \mathcal{U}_e$	Total
		(a) Cross-classification of counts		
$\mathcal{E} \cap \mathcal{U}$	n_m	$n_{ce} - n_m$	0	n_{ce}
$\mathcal{E}^c \cap \mathcal{U}$	$n_p - n_m$	$N - n_p - (n_{ce} - n_m)$	0	$N - n_{ce}$
$\mathcal{E} \cap \mathcal{U}_e$	0	0	n_{ee}	n_{ee}
Total	n_p	$N - n_p$	n_{ee}	$N + n_{ee}$
		(b) Cross-classification of expected values of counts		
$\mathcal{E} \cap \mathcal{U}$	pn_p	$en_e - pn_p$	0	en_e
$\mathcal{E}^c \cap \mathcal{U}$	$(1-p)n_p$	$(1-p)p^{-1}(en_e - pn_p)$	0	$(1-p)p^{-1}en_e$
$\mathcal{E} \cap \mathcal{U}_e$	0	0	$(1-e)n_e$	$(1-e)n_e$
Total	n_p	$p^{-1}en_e - n_p$	$(1-e)n_e$	$\{p^{-1}e + (1-e)\}n_e$

E-sample record. The enumerations appearing in both samples are "recaptures", which together with enumerations appearing only in the P sample are used to estimate the Census enumeration (capture) probability $p(x)$ that can be used to quantify the omission error. Since $E(Y_i|X_i) = p(X_i)$, the enumeration function $p(\cdot)$ can be estimated by a binary regression.

Different from conventional capture-recapture experiments, the primary purpose of the E sample is to identify and measure the EEs via extensive record-checking and follow-up. For each $i \in \mathcal{E}$, let e_i be the EE indicator such that $e_i = 1$ (0) if it is a correct (erroneous) enumeration. We will ignore the missing values in the E sample from un-resolved cases as well. Previous research (Hogan, 2003) revealed analogous heterogeneity in EEs caused by a set of covariates Z. Here Z can contain different covariates from X. For instance, the non-response follow-up code is a unique Z-covariate that provides information on those who did not respond to the census mail-out questionnaires (Belin *et al.*, 1993; Cantwell & Childers, 2001) and has been shown to be influential in modelling EEs. 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, like the E-sample enumeration probability $p(x)$, can be estimated by performing binary regression, this time on the E-sample data. In this paper, we consider three estimators of $p(x)$ and $e(z)$ based on the PS, the logistic regression and the local PS.

Let N be the size of the true population $\mathcal{U}$, and $\tilde{N}$ be the size of the nominal population $\tilde{\mathcal{U}}$ that consists of the true population $\mathcal{U}$ and the erroneous part $\mathcal{U}_e$. 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 super-population. 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

$$N = \tilde{N} \int e(z) f_Z(z) dz. \tag{1}$$

To illustrate how the E and P samples are used together for the dual system estimation, we present in Table 1 the cross-classifications of counts and their expected values, respectively, regarding the enumeration, match, and EE status. To make the illustration simple, we ignore all the covariates. In Table 1 we use, respectively, n_e and n_p for the sizes of $\mathcal{E}$ and $\mathcal{P}$, n_{ce} and n_{ee} for the numbers of correct and erroneous enumerations in $\mathcal{E}$, and n_m for the number of matched persons between $\mathcal{P}$ and $\mathcal{E}$, while assuming the probability of having the EEs in the P-sample is zero. We define $e = E(n_{ce}|n_e)/n_e$ and $p = E(n_m|n_p)/n_p$ which are the simplified versions of the correct enumeration and enumeration probability functions $e(z)$ and $p(x)$, respectively. The

expected values of the counts given in Panel (b) of Table 1 are obtained under the assumption that

$$\frac{E(n_{ce} - n_m | n_e, n_p)}{E\{N - n_p - (n_{ce} - n_m) | n_e, n_p)\}} = \frac{p}{1-p}.$$

The fact that the expected value of N conditioning on n_e and n_p is $p^{-1}en_e$ motivates a general estimator

$$\hat{N} = \sum_{i \in \mathcal{E}} \frac{\pi_i \hat{e}(Z_i)}{\hat{p}(X_i)}, \tag{2}$$

where $\hat{e}(z)$ and $\hat{p}(x)$ are estimators of the enumeration functions and π_i is the known E-sample survey weight for the sample block cluster where the i-th person resides. Clearly, if the impact of EEs was ignored by regarding $e(z) \equiv 1$, (2) would actually estimate $\tilde{N}$ instead of N, and hence would cause an over-estimation of the true population size. If the omission and EEs occur homogeneously in the population, Table 1 implies that

$$\hat{N} = \frac{n_{ce} n_p}{n_e n_m} \sum_{i \in \mathcal{E}} \pi_i$$

is a consistent estimator of the true population size.

In this paper, we focus on the population size estimator (2) based on the three methods in estimating the two enumeration functions $e(z)$ and $p(x)$: PS, logistic regression, and local PS. Additionally, let N_S be the population size of a small area $\mathcal{S}$, for instance a state or congressional district. Then, N_S can be estimated by

$$\hat{N} = \sum_{i \in \mathcal{E} \cap \mathcal{S}} \frac{\pi_i \hat{e}(Z_i)}{\hat{p}(X_i)}. \tag{3}$$

A comprehensive study of dual system estimation for small areas is given in Brown & Zhao (2008). Without loss of generality, we set $\pi_i \equiv 1$ in our study, which effectively makes $\mathcal{U}$ to be the population occupied by the E sample. Evaluation for the estimator with survey weights π_i can be made using the standard approach in survey sampling (Fuller, 2009).

3 Post-Stratification

We first provide a general assessment of system estimation based on PS. To expedite the presentation, we consider a simplified case where $X = Z$, namely the covariates for the enumeration and correct enumeration functions are identical. An analysis of the general case where $X \neq Z$ is available in Chen & Tang (2011). Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K$ be the non-overlapping post-strata with respect to X. For each stratum $\mathcal{X}_k$, let $n_{e,k}$, $n_{ce,k}$, $n_{p,k}$, and $n_{m,k}$ be the sizes of E-sample enumerations, E-sample correct enumerations, P-sample enumerations, and matched enumerations, respectively. The PS estimators for the E-sample enumeration probability and correct enumeration probability for the stratum are, respectively,

$$\hat{p}_k = \frac{n_{m,k}}{n_{p,k}} \text{ and } \hat{e}_k = \frac{n_{ce,k}}{n_{e,k}}.$$

Then from (2), the dual system PS estimator of the population size is given by

$$\hat{N}_{ps} = \sum_{k=1}^{K} n_{e,k} \frac{\hat{e}_k}{\hat{p}_k} = \sum_{i=1}^{K} \frac{n_{ce,k} n_{p,k}}{n_{m,k}}. \tag{4}$$

Let $g(x) = P(i \in \mathcal{P}|X_i = x)$ be the P-sample enumeration probability function, and

$$\eta_{1k} = \int_{\mathcal{X}_k} e(x)p(x)f_X(x)dx \text{ and } \eta_{2k} = \frac{\int_{\mathcal{X}_k} p(x)g(x)f_X(x)dx}{\int_{\mathcal{X}_k} g(x)f_X(x)dx}.$$

Here η_{1k} and η_{2k} are affected by the aggregated heterogeneities in the enumeration functions $p(x)$ and $e(x)$ over stratum k. We also define an index

$$\alpha_k = \eta_{1k}/\eta_{2k},$$

which can be viewed as a version of the one proposed in Chen & Lloyd (2000). In particular, the α_k summarizes the contributions to the heterogeneity from the E-sample enumeration and correct enumeration functions from the post-stratum $\mathcal{X}_k$. If $p(x)$ is piece-wise constant over the post-strata, then $\alpha_k = \int_{\mathcal{X}_k} e(x)f_X(x)dx$. The following theorem quantifies the properties of $\hat{N}_{ps}$ under the broad assumption that P and E samples are conditionally independent given the covariates.

THEOREM 1. *Under Conditions C.1–C.3 in the Appendix,*

$$E(\hat{N}_{ps}) = \tilde{N}\sum_{k=1}^{K}\alpha_k + O(1) \text{ and } \text{var}(\hat{N}_{ps}) = \tilde{N}V + O(1), \tag{5}$$

where V is a bounded quantity whose expression is given in (A.1) in the Appendix.

The above theorem indicates that $\text{var}(\hat{N}_{ps}/N) = V\tilde{N}/N^2 + o(N^{-1}) = O(N^{-1})$ which converges to zero as $N \to \infty$. However, the form of the mean rings alarm because $\hat{N}_{ps}$ is asymptotically unbiased for N if and only if

$$\sum_{k=1}^{K}\alpha_k = \sum_{k=1}^{K}\frac{\int_{\mathcal{X}_k} e(x)p(x)f_X(x)dx \int_{\mathcal{X}_k} g(x)f_X(x)dx}{\int_{\mathcal{X}_k} p(x)g(x)f_X(x)dx} = \int e(x)f_X(x)dx. \tag{6}$$

Otherwise, the relative bias of the PS estimator is

$$\tilde{N}\left\{\sum_{k=1}^{K}\alpha_k - \int e(x)f_X(x)dx\right\}/N,$$

which does not converge to zero as $N \to \infty$.

The requirement (6) is satisfied as long as $p(\cdot)$ is a piece-wise constant function over the post-strata, while the form of the correct enumeration function $e(x)$ is left general. This is because if $p(x)$ is constant over the post-strata, the ratios of the integrals appearing on the left of (6) can be summed up matching the global integral on the right, even if $e(x)$ is heterogeneous. This finding is consistent with one in Chen & Lloyd (2000) for mark-recapture experiments without EEs.

In the general case of $X \neq Z$, an analysis reported in Chen & Tang (2011) indicates that to ensure that the PS estimator is asymptotically unbiased, both $p(x)$ and $e(z)$ need to be constant over their respective post-strata. This higher demand than that for the case of identical covariates is due to the fact that the post-strata for X generally differ from those for Z when the covariates are not identical. As a result, to satisfy a similar condition to (6), we need to require $e(z)$ to be homogeneous within the stratum as well. This is certainly harder to attain in the presence of

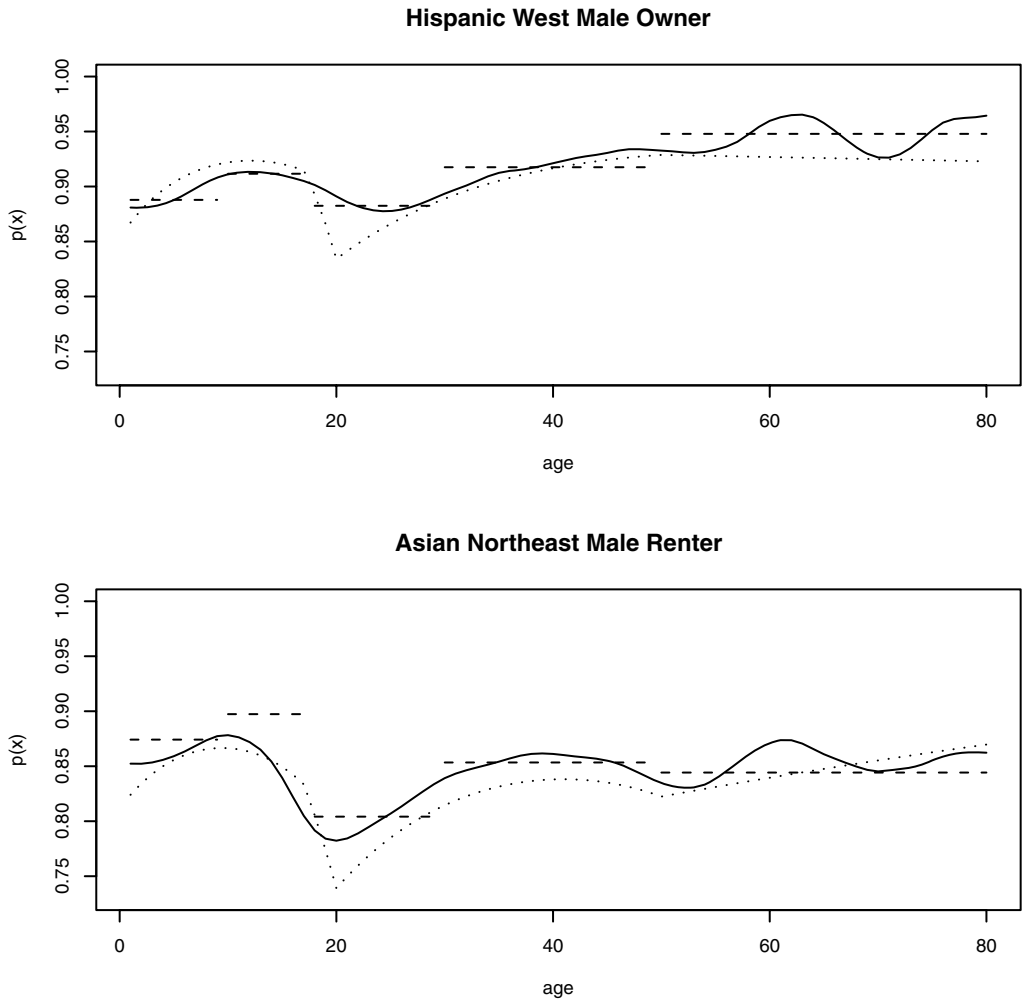

Figure 1. *Estimation of $p(x)$ by post-stratification(PS), logistic regression(LR), and the local post-stratification(L-PS). Dashed line: PS, Dotted line: LR and Solid Line: L-PS.*

continuous variables, for instance for the age and mail return rates. The PS used in the 2000 US Census post-enumeration data analysis had five age strata which is unable to capture the full heterogeneity induced by age. This age effect is clearly demonstrated in Figures 1 and 2 which plot $p(x)$ and $e(z)$.

4 Logistic Regression

Logistic regression is an alternative approach to the PS that is being implemented in the 2010 US Census dual system estimation (Bell & Cohen, 2008) via fitting parametric logistic regression models for the enumeration functions $e(z)$ and $p(x)$. Let $t(Z) = \{t_1(Z), \ldots, t_m(Z)\}^T$ and $s(X) = \{s_1(X), \ldots, s_q(X)\}^T$ be some transformations of the covariates Z and X, respectively.

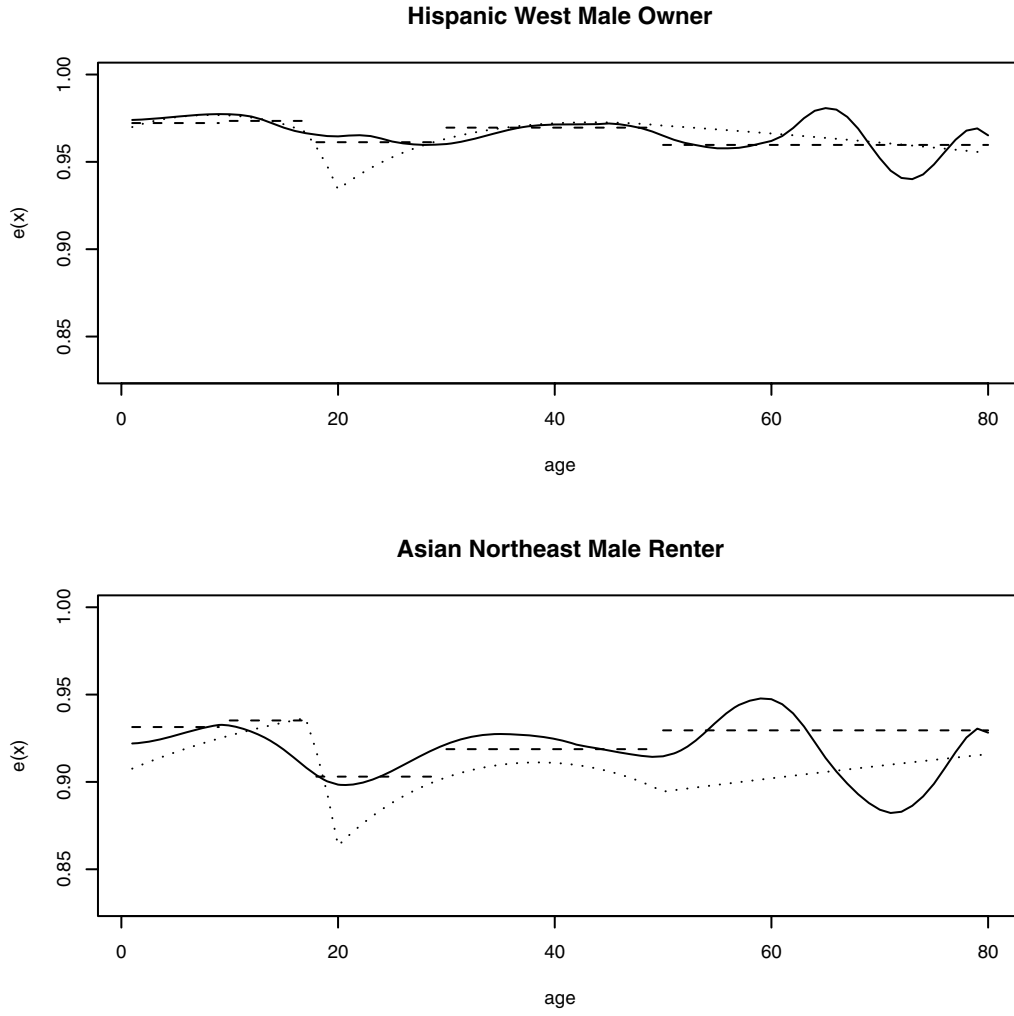

Figure 2. *Estimation of e(x) by post-stratification(PS), logistic regression(LR), and the local post-stratification(L-PS). Dashed line: PS, Dotted line: LR, and Solid Line: L-PS.*

Then, $e(z)$ and $p(z)$ are assumed to be logistic in terms of $t(z)$ and $s(x)$, respectively, namely

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

where θ_1 and θ_2 are, respectively, m- and q-dimensional unknown vector parameters. Mule *et al.* (2007) and Chen *et al.* (2010) considered logistic regression models with 86 main effects and interactions from the ROASTR to analyse the 2000 US Census post-enumeration data. Their models also include interactions between the four discrete variables (racial origin, sex, tenure, and region) with six parametric polynomial splines to account for the age effect. The parametric splines are designed to incorporate the age effect continuously instead of assuming it to be piece-wise constant over post-strata as in the PS approach.

The conditional binary log likelihoods to estimate the unknown parameters θ_1 and θ_2 can be written as

$$l_{n1}(\theta_1) = \sum_{i\in\mathcal{E}} [e_i \log\{e(Z_i;\theta_1)\} + (1-e_i)\log\{1-e(Z_i;\theta_1)\}] \text{ and} \tag{7}$$

$$l_{n2}(\theta_2) = \sum_{i\in\mathcal{P}} [Y_i \log\{p(X_i;\theta_2)\} + (1-Y_i)\log\{1-p(X_i;\theta_2)\}]. \tag{8}$$

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the maximum likelihood estimates based on (7) and (8) for the E and P samples, respectively. Then the population size estimator (2) in the context of the logistic regression models is

$$\hat{N}_l = \sum_{i\in\mathcal{E}} \frac{e(Z_i;\hat{\theta}_1)}{p(X_i;\hat{\theta}_2)}. \tag{9}$$

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

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

$$E(\hat{N}_l) = \tilde{N}\int_{\mathcal{Z}} e\big(z;\theta_1^*\big)R\big(z;\theta_2^*\big)f_Z(z)dz + O(1) \text{ and } \operatorname{var}(\hat{N}_l) = \tilde{N}V_l + O(1), \tag{10}$$

where $V_l = T_0 + T_1 + T_2 + T_3$ whose expression is given in (A.2) in the Appendix.

Similar to the PS estimator $\hat{N}_{ps}$, the variance of $\hat{N}_l$ is as expected. A potential issue arises with respect to the expected value of $\hat{N}_l$. It shows that in order for the the logistic regression estimator $\hat{N}_l$ to be asymptotically unbiased, it is required that

$$\int e\big(z;\theta_1^*\big)R\big(z;\theta_2^*\big)f_Z(z)dz = \int e(z)f_Z(z)dz. \tag{11}$$

Since $R(z;\theta_2^*)$ is the expectation of the ratio of the true enumeration function $p(x)$ to its parametric version $p(x,\theta_2^*)$, this condition basically requires both $e(z;\theta_1)$ and $p(x;\theta_2)$ to be the correct specifications of $e(z)$ and $p(x)$, respectively. This is clearer when we consider the identical covariate case of $X = Z$, where (11) becomes

$$\int e\big(x;\theta_1^*\big)\frac{p(x)}{p\big(x;\theta_2^*\big)}f_X(x)dx = \int e(x)f_X(x)dx. \tag{12}$$

Requiring both $e(z;\theta_1)$ and $p(x;\theta_2)$ to be correctly specified may be viewed as restrictive as requiring the two enumeration functions $e(z)$ and $p(x)$ to be constants over post-strata as in the PS approach. 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

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

which does not diminish to zero even when $N\to\infty$. Therefore, finding parametric models for $p(x)$ and $e(x)$ based on the empirical data so that Condition (11) is satisfied remains a challenge for the logistic regression approach.

5 Local Post-Stratification

We evaluate in this section the local post-stratification approach based on nonparametric regression estimations of $e(z)$ and $p(x)$ by smoothing over the covariates X_i and Z_i, respectively. We will demonstrate that, unlike the PS and the logistic regression estimators, the local PS estimator is free of the systematic bias due to model mis-specification and is consistent under much weaker conditions.

The method that supports the local PS is non-parametric regression, which has been extensively studied (Härdle, 1990; Fan & Gijbels, 1996) for continuous covariates. An approach taken for non-parametric regression is locally weighted least squares using a kernel function K and a smoothing bandwidth h that controls the amount of smoothness of the resulting non-parametric regression estimate. In the context of estimating $p(x)$, if all the variables in X_i are continuous, one can choose a kernel $K(x)$ which is a radially symmetric probability density function in R^d where d is the dimension of X. The Nadaraya–Watson kernel estimator

$$\hat{p}(x) = \frac{\sum_{i \in \mathcal{P}} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i \in \mathcal{P}} K\left(\frac{x - X_i}{h}\right)} \tag{13}$$

is the locally weighted least-squares estimator that minimizes

$$\sum_{i \in \mathcal{P}} K\left(\frac{x - X_i}{h}\right) (Y_i - a)^2 \tag{14}$$

with respect to a. The applications of non-parametric regression methods in estimating the population size with continuous covariates were considered in Chen & Lloyd (2000, 2002) and Huggins & Hwang (2007); see also Dorfman (2000) for estimation of finite population totals via the kernel estimator for the conditional mean with continuous covariates based on survey data.

As commonly encountered in surveys of social and economic studies, the covariates X and Z in the Census dual system surveys are mostly unordered discrete rather than continuous. Indeed, four out of the five variables in ROASTR are discrete. Unordered discrete covariates can be also smoothed using the kernels proposed in Aitchison & Aitken (1976). Suppose the d-dimensional covariates $X_i = (X_i^c, X_i^u)$, where X_i^c is d_c-dimensional continuous and X_i^u is d_u-dimensional unordered discrete. Analogously 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 j-th component of the unordered discrete $X_i^u = (X_{i1}^u, \ldots, X_{id_u}^u)$; and taking c_j discrete values $\{0, 1, \ldots, c_j - 1\}$. 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

$$\lambda_j I\left(X_{ij}^u = x_j^u\right) + \frac{1 - \lambda_j}{c_j - 1} I\left(X_{ij}^u \neq x_j^u\right), \tag{15}$$

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 information from the neighbouring cells. The choice of λ_j controls the level of smoothness in the j-th component, which depends on the

relative frequencies of the discrete components. To balance the bias and variance, a good choice of λ_j would take a relatively small value for a target x_j^u with high frequency, and a relatively larger value for a target x_j^u with low frequency. Furthermore, we note that the formulation of kernel (15) does not distinguish between all the neighbours with the same number of different components. For instance, if being Asian is closer to being white than to other races in enumeration behaviours, this kernel needs to be revised to reflect this.

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

$$L\left(x^u, X_i^u; \vec{\lambda}\right) = \prod_{j=1}^{d_u} \left\{ \lambda_j I\left(X_{ij}^u = x_j^u\right) + \frac{1-\lambda_j}{c_j - 1} I\left(X_{ij}^u \neq x_j^u\right) \right\}, \tag{16}$$

where $x^u = (x_1^u, \ldots, x_{m_u}^u)$ and $\vec{\lambda} = (\lambda_1, \ldots, \lambda_{d_u})$ 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

$$K\left(\frac{x^c - X_i^c}{h}\right) L\left(x^u, X_i^u; \vec{\lambda}\right). \tag{17}$$

When estimating $p(x)$, the kernel (17) effectively defines a local post-stratum around each $x = (x^c, x^u)$. For 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 neighbours of x^u. More generally, the k-th nearest neighbours 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 (17) instead of the continuous kernel K in (14), we have the kernel estimator of $p(x)$

$$\hat{p}(x) = \frac{\sum_{i \in \mathcal{P}} K\left(\frac{x^c - X_i^c}{h_1}\right) L\left(x^u, X_i^u; \vec{\lambda}_1\right) Y_i}{\sum_{i \in \mathcal{P}} K\left(\frac{x^c - X_i^c}{h_1}\right) L\left(x^u, X_i^u; \vec{\lambda}_1\right)}, \tag{18}$$

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

$$\hat{e}(z) = \frac{\sum_{i \in \mathcal{E}} K\left(\frac{z^c - Z_i^c}{h_2}\right) L\left(z^u, z_i^u; \vec{\lambda}_2\right) e_i}{\sum_{i \in \mathcal{E}} K\left(\frac{z^c - Z_i^c}{h_2}\right) L\left(z^u, Z_i^u; \vec{\lambda}_2\right)}, \tag{19}$$

where h_2 is the bandwidth for smoothing Z_i^c and $\vec{\lambda}_2 = (\lambda_{21}, \ldots, \lambda_{2q_u})$ for smoothing the discrete component. 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 when estimating different functions. The smoothing parameters $(h_k, \vec{\lambda}_k)$ can be estimated by minimizing,

respectively, the cross-validation (CV) scores:

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

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 greatest advantages 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 to speak for themselves regarding the proper models as shown in Chen & Tang (2009). The local PS population size estimator is

$$\hat{N}_{lp} = \sum_{i \in \mathcal{E}} \frac{\hat{e}(Z_i)}{\hat{p}(X_i)}. \tag{20}$$

This estimator was implemented in the empirical study reported in Chen *et al.* (2010) on the 2000 US Census post-enumeration data. See Dorfman (2000) for a related estimator for estimating a population total via the kernel estimator of a conditional mean function of a continuous covariate.

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 in the local PS are adaptive with their sizes shrinking as the amount of data increases. The latter is achieved by allowing the bandwidths $h_k \to 0$ and $\lambda_{kj} \to 1$ when $N \to \infty$, where $k = 1$ or 2 identifies smoothing in the P or E sample, j identifies the corresponding component in the discrete covariates. 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 notation 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 neighbouring 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

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

be the discrete kernel weight contributed from cell y^u to the cell a^u. We use ∇^k to denote the k-th 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 $e(z)$:

$$b^e_c = \int \frac{\operatorname{tr}\left[\nabla^2\{e(z)\psi(z)f_Z(z)\} - e(z)\nabla^2\{\psi(z)f_Z(z)\}\right]}{\psi(z)} dz \text{ and}$$

$$b^e_u\big(\vec{\lambda}_2\big) = \sum_{z^u \in \mathcal{Z}^u} \sum_{y^u \in \mathcal{D}^1_{z^u}} \beta_{\lambda_2}(z^u, y^u) \int_{\mathcal{Z}^c} \frac{\psi(z^c, y^u) f_Z(z^c, y^u)}{\psi(z^c, z^u)} \{e(z^c, y^u) - e(z^c, z^u)\} dz^c,$$

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

$$b_c^p = \int_{\mathcal{X}} \frac{\mathrm{tr}\left[\nabla^2\{p(x)g(x)f_X(x)\} - p(x)\nabla^2\{g(x)f_X(x)\}\right]\phi(x)}{p(x)g(x)}dx \text{ and}$$

$$b_u^p(\vec{\lambda}_1) = \sum_{x^u \in \mathcal{X}^u} \sum_{y^u \in \mathcal{D}^1_{x^u}} \beta_{\lambda_1}(z^u, y^u) \int_{\mathcal{X}^c} \phi(x^c, x^u) \frac{g(x^c, y^u) f_X(x^c, y^u)}{p(x^c, x^u) g(x^c, x^u)} \{p(x^c, y^u) - p(x^c, x^u)\} dx^c,$$

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

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

THEOREM 3. *Under the conditions C.1–C.3 and C.5 given in the Appendix,*

$$E(\hat{N}_{lp}) = N - \tfrac{1}{2}\sigma_K^2 \tilde{N}\left(h_1^2 b_c^p - h_2^2 b_c^e\right) - \tilde{N}\left\{b_u^p(\vec{\lambda}_1) - b_u^e(\vec{\lambda}_2)\right\} + R(K)h^{-d_c} \int_{\mathcal{X}} \frac{\phi(x)\{1 - p(x)\}}{p(x)g(x)}dx + o\left\{\tilde{N}(h^2 + 1 - \lambda) + h^{-d_c}\right\}; \quad (21)$$

$$\mathrm{var}(\hat{N}_{lp}) = \tilde{N}\left\{\int_{\mathcal{X}} \frac{\phi(x)^2\{1 - p(x)\}\{1 - g(x)\}f_X(x)dx}{p(x)g(x)} + \int_{\mathcal{Z}} e^2(z) f_Z(z)dz - \left(\int_{\mathcal{Z}} e(z) f_Z(z)dz\right)^2 + \int_{\mathcal{Z}} \frac{e(z)\{1 - e(z)\} f_Z(z)}{\psi(z)}\right\} + O\left\{\tilde{N}(h^2 + 1 - \lambda)\right\}. \quad (22)$$

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 $\hat{N}_l$, even though the local PS is non-parametric. This is largely due to the fact that we are estimating the population size instead of estimating, say, conditional mean functions like $p(x)$ or $e(z)$. It is known that non-parametric regression estimators of $p(x)$ or $e(z)$ have a larger variance compared to their parametric counterparts (Härdle, 1990; Chen & Tang, 2009). Specifically, despite the fact that the kernel estimators for $p(x)$ and $e(z)$ are involved in the population size estimator, the summation $\sum_{i \in \mathcal{E}}$ in (20) transfers the impact of the kernel smoothing to the second order so that the leading order variance is still $O(N)$; see Dorfman (2000) for a related result for smoothing a continuous covariate for estimating a population total.

We 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 (21) 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 (22) 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 (6) and (11) are met respectively. Theorem 3 shows that the consistency of the local PS estimator $\hat{N}_{lp}$ is achieved under very weak conditions requiring only 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 constants over post-strata or a specific parametric form. Although the non-parametric 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 for most census coverage measurement applications.

6 Simulation Studies

We report results from simulation studies which were designed to confirm the theoretical asymptotic analyses reported in the previous sections. To reflect the Census reality, 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, $X_3 \in \{0, \ldots, 3\}$ for the region, and $X_4, X_5 \in \{0, 1\}$ are for gender and housing tenure, respectively. Then the domain of X was $\mathcal{X} = [0, 70] \times \{0, \ldots, 6\} \times \{0, \ldots, 3\} \times \{0, 1\} \times \{0, 1\}$. 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 variable 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 a uniform distribution among the components in X_i and Z_i. The four discrete variables 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 for the 2000 US Census post-enumeration survey data, for instance those displayed in Figures 1 and 2. Let $l(x) = 16x_2 + 4x_3 + 2x_4 + x_5 + 1$ be a one-to-one mapping from $\{0, \ldots, 6\} \times \{0, \ldots, 3\} \times \{0, 1\} \times \{0, 1\}$ to $\{1, \ldots, 112\}$. We chose $p(x) = [1 + \exp\{-b_p(x); \beta^{(p)}\}]^{-1}$, where

$$b_p\big(x; \beta^{(p)}\big) = \beta^{(p)}_{l(x)0} + \sum_{k=1}^{6} \beta^{(p)}_{l(x)k} B_k(x_1), \tag{23}$$

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

$$b_e\big(z; \beta^{(e)}\big) = \beta^{(e)}_{m(z)0} + \sum_{k=1}^{6} \beta^{(e)}_{m(z)k} B_k(z_1). \tag{24}$$

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 or 896 post-strata, respectively. The PS population size

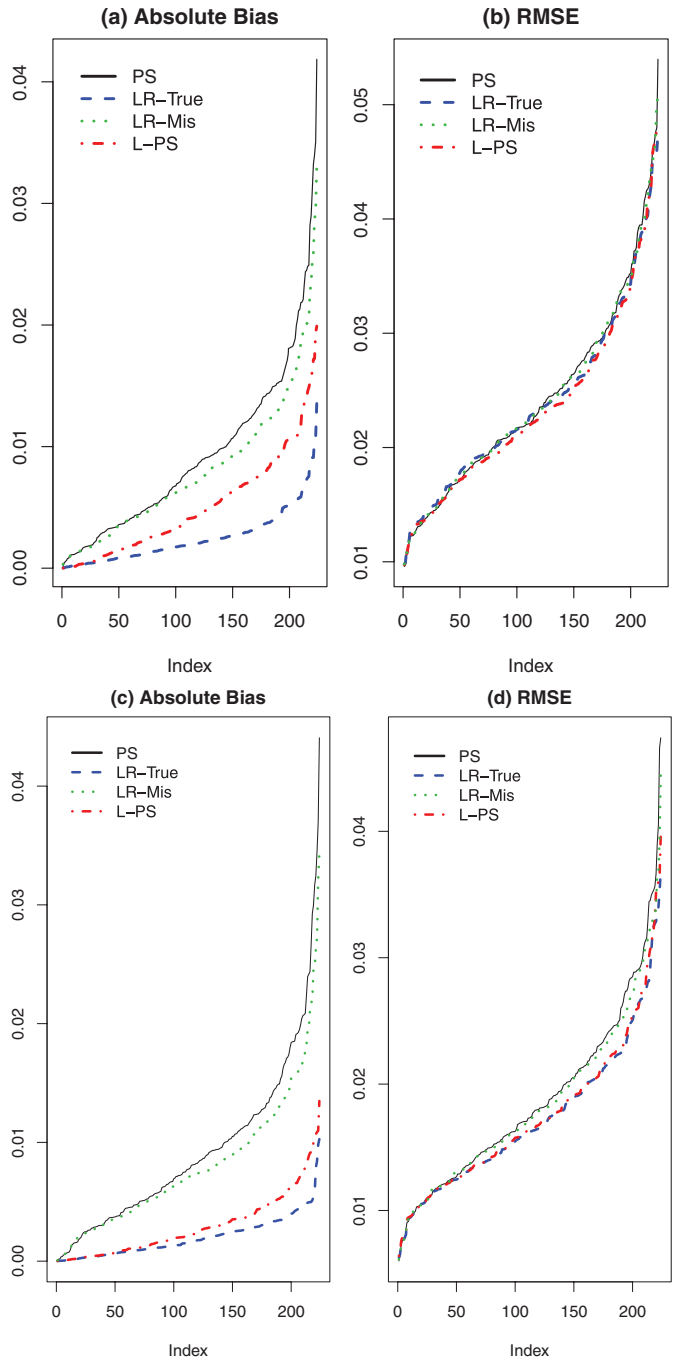

Figure 3. *The relative absolute biases and the root mean square errors of four population size estimators: post-stratification (PS), logistic regression with mis-specification (LR-Mis), and with the true specification (LR-true), and local PS (L-PS) for the* 224 *cells;* $\tilde{N} = 500\,000$ *in Panels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

estimates were obtained by (4). The logistic regressions estimation were performed under two models. One had the correct specification as in (23) and (24), and the other was mis-specified by replacing (23) and (24) 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 (20).

We considered two nominal population sizes $\tilde{N} = 5\,00\,000$ and $\tilde{N} = 1\,000\,000$ with 2 000 replications, respectively, in the simulation. We applied the three types of 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 the logistic regression under the mis-specification (LR-Mis) endured much larger biases than the local post-stratification (L-PS) and the logistic regression under the correct specification (LR-True). The biases with the PS and the LR-Mis did not become smaller when $\tilde{N}$ was doubled to 1 million. These confirmed our theoretical findings of a systematic bias with the PS and using a mis-specified logistic regression model. The systematic bias was so significant that the relative RMSE (right panels) did not decrease with increased sample size. At the same time, the bias and the RMSE 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 RMSE were very much 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. We note that the logistic regression employed was a relatively complex model with several hundred variables from main factors and B-splines of age, which may have resulted in a higher level of variability.

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 consistent with the results for sub-population estimations reported in Figure 3.

7 Analysis of US Census Data

In this section, we applied the three approaches to the 2000 US Census post-enumeration survey research data. The covariates used in modelling 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

$$\hat{N} = \sum_{i \in \mathcal{C}} \frac{\hat{e}(X_i)}{\hat{p}(X_i)} \tag{25}$$

based on the projected correct enumeration function $e(X_i) = E\{e(Z_i)|X_i\}$ that can also be obtained by a modified local PS method discussed in Chen *et al.* (2010). The reason by taking the conditional expectation is to integrate out the extra covariate MCG in Z as MCG is not available for Census records outside the E sample. If we just use the E sample based estimator (2), such an operation is not needed. We employed the estimator (25) in the case study as it is less variable than (2) since (25) utilizes all Census records $\mathcal{C}$.

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 described 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 were 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 and 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 covariates 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 undercounts (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 & Tu, 1995; Wolter, 2007). Table 2 provides population undercount estimates for 16 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 Virginia. While the local PS and the logistic regression estimates were close in New Hampshire and New Jersey, they were a little different from the PS estimates. While part of the difference may be attributed to random variation, some was due to the built-in bias associated with the potential mis-specification by the PS and the logistic regression.

8 Discussion

In this paper, we have assumed that the covariates X_i and Z_i, and the status 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 the logistic regression estimators when the model assumptions for the two enumeration functions deviate

Table 2
State level census research estimates of undercount percentage and their standard errors (in parentheses) for local post-stratification (L-PS), post-stratification (PS) and logistic regression (LR).

STATE	L-PS	PS	LR
	Northeast Region		
Connecticut	1.07 (0.26)	1.05 (0.26)	1.1 (0.27)
New Hampshire	−0.14 (0.29)	−0.16 (0.29)	−0.1 (0.3)
New Jersey	0.86 (0.26)	0.83 (0.26)	0.94 (0.28)
Pennsylvania	0.71 (0.26)	0.7 (0.26)	0.74 (0.27)
	Midwest Region		
Missouri	0.59 (0.18)	0.6 (0.18)	0.58 (0.18)
Minnesota	0.29 (0.17)	0.31 (0.17)	0.28 (0.17)
Montana	0.3 (0.17)	0.3 (0.17)	0.28 (0.18)
Ohio	0.73 (0.17)	0.73 (0.17)	0.72 (0.18)
	South Region		
Florida	1.7 (0.24)	1.72 (0.24)	1.63 (0.25)
Mississippi	1.51 (0.24)	1.48 (0.24)	1.45 (0.24)
Oklahoma	2.18 (0.27)	2.18 (0.26)	2.22 (0.28)
Virginia	2.16 (0.23)	2.09 (0.22)	2.03 (0.24)
	West Region		
Hawaii	1.13 (0.86)	0.92 (0.85)	0.87 (0.84)
Oregon	1.21 (0.32)	1.18 (0.31)	1.2 (0.32)
Utah	1.28 (0.32)	1.21 (0.32)	1.32 (0.33)
Washington	1.2 (0.31)	1.16 (0.31)	1.19 (0.31)

from the underlying true model. The logistic regression is designed to improve the PS using parametric modelling for the covariates effects. It represents a methodological improvement over the PS. Our analysis suggests that the effectiveness of logistic regression in dual system estimation depends 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 when 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 for the State level sizes, the employed logistic regression may not severely deviate from the truth 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 US Census coverage measurement study is to use the PS to evaluate the logistic regression estimates. Given the analysis done in this paper, we advocate using the local PS instead of the PS to evaluate the logistic regression estimation.

References

Abbott, O. (2007). 2011 UK Census coverage assessment and adjustment strategy. *Popul. Trends*, **127**, 7–14.

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

Alho, J.M., Mury, M.H., Wurdeman, K. & Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *J. Amer. Statist. Assoc.*, **88**, 1130–1136.

Anderson, M. & Feinberg, S.E. (1999). To sample or not to sample? The 2000 Census Controversy. *J. Interdiscipl. Hist.*, **30**, 1–36.

Anderson, M. & Feinberg, S.E. (2001). *Who Counts? The Politics of Census-Taking in Contemporary America*. New York: Russell Sage Foundation.

Ayhan,Ö.H. & Ekni, S. (2003). Coverage error in population censues: the case of Turkey. *Survey Methodol.*, **29**, 155–165.

Belin, T.R., Diffendal, G.J., Mack, S., Rubin, D.B., Schafer, J.L. & Zaslavsky, A. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussions). *J. Amer. Statist. Assoc.*, **88**, 1149–1166.

Bell, R. & Cohen, M.L. (eds.) (2008). *Coverage Measurement in the 2010 Census*. Washington, D.C.: National Academies Press.

Bell, W.R. (1993). Using information from Demographic analysis in postenumeration survey estimation. *J. Amer. Statist. Assoc.*, **88**, 1106–1118.

Brown, J., Diamond, I., Chambers, R., Buckner, L. & Teague, A. (1999). A methodological strategy for a one-number Census in the UK. *J. R. Statist. Soc., Series A*, **162**, 247–267.

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

Cantwell, P. & 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. Washington, D.C..: US Census Bureau.

Chao, A. & Tsay, P.K. (1998). A sample coverage approach to multiple-system estimation with application to census undercounts. *J. Amer. Statist. Assoc.*, **93**, 283–293.

Chen, S.X. & Lloyd, C.J. (2000). A non-parametric approach to the analysis of two stage mark-recapture experiments. *Biometrika*, **87**, 633–649.

Chen, S.X. & Lloyd, C.J. (2002). Estimation of population size based on biased samples using nonparametric binary regression. *Statistica Sinica*, **12**, 505–518.

Chen, S.X. & Tang, C.Y. (2009). Nonparametric regression with discrete covariates and missing values. *Technical Report*.

Chen, S.X. & Tang, C.Y. (2011). Properties of census dual system population size estimators. *Technical report*.

Chen, S.X. Tang, C.Y. & Mule, V.T. (2010). Local post-stratification and diagnostics in dual system accuracy and coverage evaluation for the U.S. Census. *J. Amer. Statist. Assoc.*, **105**, 105–119.

Darroch, J., Fienberg, S.E., Glonek, G.F. & Junker, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Amer. Statist. Assoc.*, **88**, 1137–1148.

Dorfman, A.H. (2000). Non-parametric regression for estimating totals in finite populations. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 622–625. Alexandria: American Statistical Association.

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

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

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

Haberman, S., Jiang, W. & 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. *J. Amer. Statist. Assoc.*, **88**, 1047–1060.

Hogan, H. (2000a). Accuracy and coverage evaluation 2000: decomposition of dual system estimate components. *DSSD Census 2000 Procedures and Operation Memorandum Series B-8*.

Hogan, H. (2000b). Accuracy and coverage evaluation 2000: dual system estimate results. *DSSD Census 2000 Procedures and Operation Memorandum Series B-9*.

Hogan, H. (2003). The accuracy and coverage evaluation: theory and design. *Survey Methodol.*, **29**, 129–138.

Huggins, R. & Hwang, W.H. (2007). Non-parametric estimation of population size from capture–recapture data when the capture probability depends on a covariate. *J. R. Statist. Soc., Series C*, **56**, 429–443.

Mule, T., Schellhamer, T., Malec, D. & 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*.

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.

Pollock, K.H. (1991). Modeling capture-recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *J. Amer. Statist. Assoc.*, **86**, 225–238.
Schindler, E. (2008). Post-stratification by age for small intervals. *Census 2000 Procedures and Operations Memorandum Series Q-94*.
Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*. New York: 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. *J. Amer. Statist. Assoc.*, **81**, 338–346.
Wolter, K.M. (2007). *Introduction to Variance Estimation*. New York: Springer.

Résumé

Nous étudions des méthodes paramétriques et non paramétriques d'évaluation de la précision et de la couverture d'un recensement fondé sur un double système d'enquêtes. Deux approches paramétriques, la post-stratification et la régression logistique, sont considérées; ces approches ont été ou seront mises en pratique dans le cadre du double système d'enquêtes utilisé pour le recensement des Etats-Unis. Nous montrons que ces méthodes sont généralement biaisées lorsque le modèle sur lequel elles se fondent est incorrectement spécifié. Nous étudions ensuite une approche de post-stratification locale fondée sur un estimateur à noyau des fonctions d'énumération du recensement. Nous illustrons le fait que cette approche non paramétrique évite les risques liés aux erreurs de spécification des modèles utilisés, et converge sous des conditions assez générales. Les performances des estimateurs qui en résultent sont évaluées de façon numérique à partir de simulations et d'une analyse empirique fondée sur les résultats de l'enquête post-censitaire du recensement des Etats-Unis de 2000.

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 conditionally independent given the combined covariates X, namely $I_{i\in\mathcal{E}}$ and $I_{i\in\mathcal{P}}$ are independent 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) the joint occurrence of an individual in the E sample being an EE and being enumerated by the Census are conditionally 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

$$\ell_{i1}(\theta_1) = \frac{e^{(1)}(Z_i;\theta_1)\{e_i - e(X_i,\theta_1)\}I_{i\in\mathcal{E}}}{e(Z_i;\theta_1)\{1 - e(Z_i;\theta_1)\}},$$

$$\ell_{i2}(\theta_2) = \frac{p^{(1)}(X_i;\theta_2)\{Y_i - p(X_i,\theta_2)\}I_{i\in\mathcal{P}}}{p(X_i;\theta_2)\{1 - p(X_i;\theta_2)\}}.$$

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 neighbourhoods of θ_1^* and θ_2^*, respectively. In addition, we assume that for θ_2 in a neighbourhood 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 a 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 independence between the selection of person in both samples and defines the enumeration and correct enumeration functions $p(x)$ and $e(z)$ in light of that 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 maximum likelihood estimation. The conditions in C.5 are commonly assumed in non-parametric regression.

Proof of Theorem 1

We first note that $\hat\eta_{1k} = \tilde N^{-1} n_{ce,k} = \tilde N^{-1}\sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{E}} I_{i\in\tilde{\mathcal{E}}} I_{i\in\mathcal{X}_k} \xrightarrow{p} \eta_{1k}$,

$$\frac{n_{m,k}}{n_{p,k}} = \frac{\tilde N^{-1}\sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{E}} I_{i\in\mathcal{P}} I_{i\in\mathcal{X}_k}}{\tilde N^{-1}\sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{P}} I_{i\in\mathcal{X}_k}} \xrightarrow{p} \eta_{2k}.$$

We then develop the following expansion:

$$\tilde N^{-1} n_{ce,k} n_{p,k}/n_{m,k} = \frac{\eta_{1k}}{\eta_{2k}} + \eta_{2k}^{-1}(\hat\eta_{1k}-\eta_{1k}) + \frac{\eta_{1k}}{\omega_{1k}}(\hat\omega_{2k}-\omega_{2k}) - \frac{\eta_{1k}}{\eta_{2k}}(\hat\omega_{1k}-\omega_{1k}) + o_p(\tilde N^{-1/2}),$$

where $\omega_{1k} = \int_{\mathcal{X}_k} p(x)g(x)f_X(x)dx$, $\omega_{2k} = \int_{\mathcal{X}_k} g(x)f_X(x)dx$, $\hat\omega_{1k} = \tilde N^{-1}\sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{P}} I_{i\in\mathcal{E}} I_{i\in\mathcal{X}_k}$, and $\hat\omega_{2k} = \tilde N^{-1}\sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{P}} I_{i\in\mathcal{X}_k}$. Then the asymptotic bias and variance in Theorem 1 follow by defining V in Theorem 1 as

$$V = \sum_{k=1}^{K} V_k, \tag{A.1}$$

and $V_k = \eta_{2k}^{-1}\mathrm{var}(\hat\eta_{1k}) + \eta_{1k}^2\omega_{1k}^{-2}\mathrm{var}(\hat\omega_{2k}) + \eta_{1k}^2\eta_{2k}^{-2}\mathrm{var}(\hat\omega_{1k}) + 2\eta_{1k}\eta_{2k}^{-1}\omega_{1k}^{-1}\mathrm{cov}(\hat\eta_{1k},\hat\omega_{2k}) - 2\eta_{1k}\eta_{2k}^{-2}\mathrm{cov}(\hat\eta_{1k},\hat\omega_{1k}) - 2\eta_{1k}^2\eta_{2k}^{-1}\omega_{1k}^{-1}\mathrm{cov}(\hat\omega_{1k},\hat\omega_{2k})$. Further it is straightforward to show

$$\mathrm{var}(\hat\eta_{1k}) = \int_{\mathcal{X}_k} pef\left(1-\int_{\mathcal{X}_k} pe\ f\right), \mathrm{var}(\hat\omega_{1k}) = \int_{\mathcal{X}_k} pgf\left(1-\int_{\mathcal{X}_k} pgf\right),$$

$$\mathrm{var}(\hat\omega_{2k}) = \int_{\mathcal{X}_k} gf\left(1-\int_{\mathcal{X}_k} gf\right), \mathrm{cov}(\hat\eta_{1k},\hat\omega_{1k}) = \int_{\mathcal{X}_k} pgef - \left(\int_{\mathcal{X}_k} pef\right)\left(\int_{\mathcal{X}_k} pgf\right),$$

$$\mathrm{cov}(\hat\eta_{1k},\hat\omega_{2k}) = \int_{\mathcal{X}_k} pgef - \left(\int_{\mathcal{X}_k} pef\right)\left(\int_{\mathcal{X}_k} gf\right), \text{ and}$$

$$\mathrm{cov}(\hat\omega_{1k},\hat\omega_{2k}) = \int_{\mathcal{X}_k} pgf - \left(\int_{\mathcal{X}_k} pgf\right)\left(\int_{\mathcal{X}_k} gf\right),$$

where the dummy variable x is suppressed in the integrations.

Proof of Theorem 2

We need to introduce some notation. 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

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

The following quantities are defined to study the variance of (9):

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

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

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

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

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

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

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

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

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

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

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

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 be equal to $p(x;\theta^*)$ point-wise. This is because the parametric model may be mis-specified. We apply Taylor's expansion for

the above equation in a neighbourhood of θ^*,

$$0 = \tilde{N}^{-1}\sum_{i\in\tilde{\mathcal{U}}}\ell_i(\hat{\theta}_2) = \tilde{N}^{-1}\sum_{i\in\tilde{\mathcal{U}}}^{n}\ell_i\big(\theta_2^*\big) + \tilde{N}^{-1}\sum_{i\in\tilde{\mathcal{U}}}\ell_i^{(1)}\big(\theta_2^*\big)\big(\hat{\theta}_2-\theta_2^*\big) + R_n(\theta_2), \quad \text{(A.3)}$$

where $R_n(\theta_2)$ is the remainder term whose k-th component is given by

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

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

$$\hat{\theta}_2-\theta_2^* = A^{-1}(\theta_2)\big\{1+o_p(1)\big\}\left\{\tilde{N}^{-1}\sum_{i\in\tilde{\mathcal{U}}}\ell_i\big(\theta_2^*\big)+R_n\right\}. \quad \text{(A.4)}$$

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

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

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 (9). We have

$$\hat{N}_l = \sum_{i\in\tilde{\mathcal{U}}} I_{i\in\mathcal{E}}\left\{\frac{e_i\big(Z_i;\theta_1^*\big)}{p_i\big(X_i;\theta_2^*\big)} + \frac{\big\{e_i^{(1)}\big(Z_i;\theta_1^*\big)\big\}^T\big(\hat{\theta}_1-\theta_1^*\big)}{p_i\big(X_i;\theta_2^*\big)} - \frac{e_i\big(Z_i;\theta_1^*\big)\big\{p_i^{(1)}\big(X_i;\theta_2^*\big)\big\}^T\big(\hat{\theta}_2-\theta_2^*\big)}{p_i^2\big(X_i;\theta_2^*\big)} + O_p(N^{-1})\right\}. \quad \text{(A.5)}$$

Then $E(\hat{N}_l)$ is established from (A.5). To derive the variance part of Theorem 2, we note that $I_{i\in\mathcal{E}}$ appears in (A.3) and thus a non-ignorable correlation between the first and third term is induced in (A.5). And we show that the remaining between terms correlations in (A.5) are negligible. Then by taking the variance operation over (A.5), we established the variance part of Theorem 2.

Proof of Theorem 3

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

$$\hat{N} = \sum_{i\in\tilde{\mathcal{U}}}\frac{\hat{e}(Z_i)}{\hat{p}(X_i)}I_{i\in\mathcal{E}} = t_1+t_2-t_3-t_4+t_5\{1+O_p(1)\}, \text{ where} \quad \text{(A.6)}$$

$$t_1 = \sum_{i\in\tilde{\mathcal{U}}}\frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)},\ t_2 = \sum_{i\in\tilde{\mathcal{U}}}\frac{\{\hat{e}(Z_i)-e(Z_i)\}I_{i\in\mathcal{E}}}{p(X_i)},\ t_3 = \sum_{i\in\tilde{\mathcal{U}}}\frac{e(Z_i)\{\hat{p}(X_i)-p(X_i)\}I_{i\in\mathcal{E}}}{p^2(X_i)},$$

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

Existing theory on non-parametric 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.6) 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

$$\hat{\eta}_1(z) = \tilde{N}^{-1}\sum_{j=1}^{\tilde{N}} \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)I_{j\in\mathcal{E}}I_{j\in\tilde{\mathcal{E}}},\ \hat{\eta}_2(z) = \tilde{N}^{-1}\sum_{j=1}^{\tilde{N}} \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)I_{j\in\mathcal{E}},$$

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

Therefore, we show that

$$\begin{aligned} E\{\hat{\eta}_1(z)\} &= \int \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)p(X_i)e(Z_i)f(U_i)dU_i = \int \mathcal{K}_{h_2,\vec{\lambda}_2}(z, Z_j)e(Z_i)\psi(Z_i)f_Z(Z_i)dZ_i \\ &= e(z)\psi(z)f_Z(z) + \tfrac{1}{2}h_2^2\sigma_K^2\mathrm{tr}[\nabla^2\{e(z)\psi(z)f_Z(z)\}] \\ &\quad + \sum_{y^u\in D^1_{z^u}} \beta_{\lambda_2}(z^u, y^u)e_{y^u}(z^c)\psi_{y^u}(z^c)f_{Z,y^u}(z^c) + O(h_2^2) + O(1-\lambda_2). \end{aligned} \tag{A.7}$$

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

$$\begin{aligned} \hat{e}(z) &= e(z) + \frac{\hat{\eta}_1(z) - \eta_1(z)}{\eta_2(z)} - \frac{\eta_1(z)\{\hat{\eta}_2(z) - \eta_2(z)\}}{\eta_2^2(z)}\{1 + o_p(1)\} \text{ and} \\ \hat{p}(x) &= p(x) + \frac{\hat{\eta}_3(x) - \eta_3(x)}{\eta_4(x)} - \frac{\eta_3(x)\{\hat{\eta}_4(x) - \eta_4(x)\}}{\eta_4^2(x)}\{1 + o_p(1)\}. \end{aligned} \tag{A.8}$$

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

$$E(t_5) = R(K)h^{-d_c}\int_{\mathcal{X}} \frac{\phi(1-p)}{pg} + o(h^{-d_c}). \tag{A.9}$$

Hence, the bias part of Theorem 3 is concluded from (A.6) by summarizing (A.7), (A.8), and (A.9).

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

$$\mathrm{var}(t_1) = \sum_{i=1}^{\tilde{N}} \mathrm{var}\left\{\frac{e(Z_i)I_{i\in\mathcal{E}}}{p(X_i)}\right\} = \tilde{N}\left\{\int_{\mathcal{X}} \frac{\phi^2(z)}{p(x)}f_X(x)dx - \left(\int_{\mathcal{Z}} e(z)f_Z(z)dz\right)^2\right\}. \tag{A.10}$$

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.6) and (A.8) that

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

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

$$\operatorname{var}(t_2) = \left[\tilde{N}^{-2} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \sum_{k=1}^{\tilde{N}} \operatorname{cov}\{(\alpha_{1,ik} - \alpha_{2,ik}), (\alpha_{1,jk} - \alpha_{2,jk})\} \right] \{1 + O(1 - \lambda_2)\}. \quad \text{(A.11)}$$

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

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

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

$$\operatorname{var}(t_2) = \tilde{N} \int_{\mathcal{Z}} \frac{e(z)\{1 - e(z)\} f_Z(z)}{\psi(z)} + O\big(\tilde{N} h_2^2\big) + O\{\tilde{N}(1 - \lambda_2)\}. \quad \text{(A.12)}$$

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.11),

$$\begin{aligned} \operatorname{var}(t_3) &= \left[\tilde{N}^{-2} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \sum_{k=1}^{\tilde{N}} \operatorname{cov}\{(\alpha_{3,ik} - \alpha_{4,ik}), (\alpha_{3,jk} - \alpha_{4,jk})\} \right] [1 + O\{A(\vec{\lambda}_1)\}] \\ &= \tilde{N} \int_{\mathcal{X}} \frac{\phi^2(x)\{1 - p(x)\} f_X(x) dx}{p(x) g(x)} + O\big(\tilde{N} h_1^2\big) + O\{\tilde{N}(1 - \lambda_1)\}. \end{aligned} \quad \text{(A.13)}$$

By Condition C.1,

$$\operatorname{cov}(t_1, t_2) = O(\tilde{N} h_2^2) + O\{\tilde{N}(1 - \lambda_2)\} \text{ and } \operatorname{cov}(t_2, t_3) = O(\tilde{N} h_1^2) + O\{\tilde{N}(1 - \lambda_2)\}. \quad \text{(A.14)}$$

Finally,

$$\begin{aligned} \operatorname{cov}(t_1, t_3) &= \tilde{N}^{-1} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \sum_{k=1}^{\tilde{N}} \operatorname{cov}\left\{ \frac{e(Z_i) I_{i \in \mathcal{E}}}{p(X_i)}, (\alpha_{3,jk} - \alpha_{3,jk}) \right\} \\ &= \tilde{N}^{-1} \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \operatorname{cov}\left\{ \frac{e(Z_i) I_{i \in \mathcal{E}}}{p(X_i)}, (\alpha_{3,ji} - \alpha_{3,ji}) \right\} \{1 + O(1 - \lambda_1)\} \\ &= \tilde{N} \int_{\mathcal{X}} \frac{\phi^2(x)\{1 - p(x)\} f_X(x) dx}{p(x) g(x)} + O(\tilde{N} h_1^2) + O\{\tilde{N}(1 - \lambda_1)\}. \end{aligned} \quad \text{(A.15)}$$

In summary of these results (A.10)–(A.15), we conclude the variance part of Theorem 3.

Acknowledgements

We thank the Chief Editor, the Editor in charge, the AE, and a referee for their insightful comments and suggestions which have improved the presentation of the paper. We would like to thank the US Census Bureau for allowing access to the 2000 post-enumeration survey research data; and we are most grateful to Tom Mule, Bill Bell, Pat Cantwell, Phil Gbur, Richard Griffin, Doug Olson, Don Malec, Mary Mulry, and Tommy Wright for insightful comments and discussions. Of course, the views expressed here are entirely ours. Our research was supported by a National Science Foundation grant SES-0518904 and a US Census Bureau contract.

[*Received July 2010, accepted June 2011*]