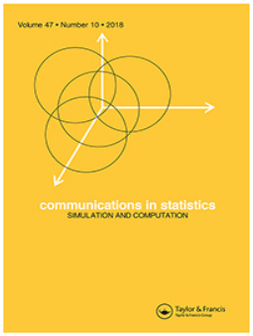

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Nonparametric regression estimators in dual frame surveys

Yan Lu, Ye Fu & Guoyi Zhang

To cite this article: Yan Lu, Ye Fu & Guoyi Zhang (2019): Nonparametric regression estimators in dual frame surveys, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1568474

To link to this article: https://doi.org/10.1080/03610918.2019.1568474

Published online: 05 Feb 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1568474

Nonparametric regression estimators in dual frame surveys

Yan Lu[a], Ye Fu[b], and Guoyi Zhang[c]

[a]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA;
[b]Department of Information Management and Information Systems, Fudan University, Shanghai, China;
[c]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA

ABSTRACT

Dual frame surveys, in which independent samples are selected from two frames to decrease survey costs or to improve coverage, can present challenges for regression estimation because of complex designs and unknown degree of overlap. In this research, we developed three nonparametric regression estimators in dual frame surveys and investigated their asymptotic properties. Simulation results show that all the proposed methods work well.

ARTICLE HISTORY
Received 3 August 2018
Accepted 25 December 2018

KEYWORDS
Cross validation; Dual frame surveys; Nonparametric regression estimator; Prediction error; Simulations

1. Introduction

Traditionally, large surveys use a single sampling frame from which the sample is selected. As the population and methods used to collect survey data change, single frame surveys may miss parts of the population. For example, random digit dialing is a popular sampling method. However, as mentioned in Keeter, Dimock, and Christian (2010), "The number of Americans who rely solely or mostly on a cell phone has been growing for several years, posing an increasing likelihood that public opinion polls conducted only by landline telephone will be biased". In order to obtain better coverage of the population of interest and cost less, a number of surveys employ dual frame design, in which independent samples are taken from two overlapping sampling frames. In a general case, each frame can contain units the other frame does not have as well as units in common as depicted in Figure 1. For example, frame A can be a landline frame and frame B can be a cell phone frame. The overlap domain ab includes elements with both landline and cellphone. A dual frame survey presents additional challenges to those from a single frame survey because there are now two samples, each with a possibly complex sampling design and may have an unknown degree of overlap.

Researchers have proposed methods for combining information from the two independent samples in a dual frame survey to estimate population quantities such as total, mean and gross flows. These include but not limited to Hartley (1962, 1974), Bankier (1986), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996) and Lu and Lohr (2010) etc. Lohr and Rao (2000) summarized estimators used for estimating population total in cross-sectional dual frame surveys.

CONTACT Yan Lu luyan@math.unm.edu

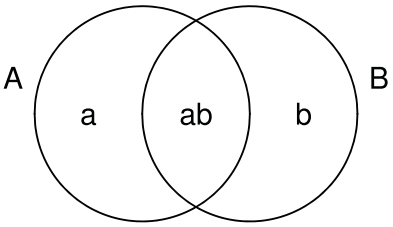

Figure 1. Frames A and B are both incomplete and overlapping.

For applications in economic and health surveys, the main interest is usually in analysis and comparison rather than in the estimation of population mean or totals. Relationships among the variables, prediction of new observations, or imputation of missing values are often of interest. Merkouris (2004) considered combing independent regression estimators from multiple surveys. Metcalf and Scott (2009) introduced a simple class of procedures for analyzing dual frame surveys, which can be extended to regression analysis. Lu (2014) proposed several approaches to estimate the linear regression coefficients in dual frame surveys. In practice, however, the underlying model may not be linear. To solve this problem, we propose nonparametric regression estimators in dual frame surveys.

Classical nonparametric regression estimators and methods have been extended and investigated in survey area, which includes Korn and Graubard (1998), Bellhouse and Stafford (1999, 2001), Breidt and Opsomer (2000), Buskirk (1998), Buskirk and Lohr (2005), Opsomer and Miller (2005), Breidt, Claeskens, and Opsomer (2005), Goga (2005), Zhang, Christensen, and Zheng (2015) etc. Harms and Duchesne (2010) introduced a completely data driven optimal bandwidth for local linear estimator in complex surveys, and derived the asymptotic mean squared error of the kernel estimators. In this research, we apply local linear estimator to dual frame surveys using optimal bandwidth suggested by Harms and Duchesne (2010).

This article is organized as follows. In Sec. 2, we review the frame work and pseudo maximum likelihood (PML) estimators in a dual frame survey, and the bandwidth selection method for local linear estimator in complex surveys (Harms and Duchesne 2010). In Sec. 3, we propose local linear estimators in dual frame surveys and examine their asymptotic properties. In Sec. 4, we present simulation studies. Finally, we summarize our research in Sec. 5.

2. Background

In this section, we review the frame work and pseudo maximum likelihood (PML) estimators for population totals in a dual frame survey. We also review the completely data driven bandwidth selection method for local linear estimator in a single frame complex surveys suggested by Harms and Duchesne (2010).

2.1. Frame work and PML in a dual frame survey

As depicted in Figure 1, in a dual frame survey, frame A and frame B together cover the population of interest. Domain a includes the elements contained only in frame A.

Domain b includes the elements contained only in frame B. The overlap domain ab includes the elements contained in both frame A and frame B. The population sizes for the frames and domains are denoted by N_A, N_B, N_a, N_b, and N_{ab}, where $N_A = N_a + N_{ab}$, and $N_B = N_b + N_{ab}$. Two independent samples $\mathcal{S}_A$ and $\mathcal{S}_B$ are taken from frame A and frame B respectively according to specified probability sampling designs. The probability of unit i being included in $\mathcal{S}_A$ is $\pi_{iA} = p\{i \in \mathcal{S}_A\}$. The probability of unit i being included in $\mathcal{S}_B$ is $\pi_{iB} = p\{i \in \mathcal{S}_B\}$. The sample sizes for the frames and domains are n_A, n_B, n_a, n_b, n_{ab}^A and n_{ab}^B with $n_A = n_a + n_{ab}^A, n_B = n_b + n_{ab}^B$ and $n = n_A + n_B$, where n is the sample size from the union of frame A and frame B, n_{ab}^A and n_{ab}^B represent the sample sizes for the elements of domain ab that were originally taken from frames A and B respectively.

Skinner and Rao (1996) modified the maximum likelihood estimator from a simple random sample to obtain a Pseudo Maximum Likelihood (PML) estimator for complex designs and suggested the following estimator for population total:

$$\hat{Y}_{PML} = \frac{N_A - \hat{N}_{ab,PML}}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab,PML}}{\hat{N}_{ab}}\hat{Y}_{ab} + \frac{N_B - \hat{N}_{ab,PML}}{\hat{N}_b}\hat{Y}_b, \tag{1}$$

where $\hat{N}_a, \hat{Y}_a, \hat{N}_b$ and $\hat{Y}_b$ are standard basic estimators, $\hat{Y}_{ab} = \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B$, with $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ the estimators of Y_{ab} by using elements of domain ab that were originally taken from frames A and frame B respectively. $\hat{N}_{ab} = \theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B$, where $\hat{N}_{ab}^A$ and $\hat{N}_{ab}^B$ are the estiamtors of N_{ab} by using elements of domain ab that were originally taken from frames A and frame B respectively. The estimator $\hat{N}_{ab,PML}$ is a function of $\hat{N}_{ab}^A, \hat{N}_{ab}^B$ and θ, and is the smaller root of the quadratic equation

$$\left[\frac{\theta}{N_B} + \frac{(1-\theta)}{N_A}\right]x^2 - \left[1 + \theta\frac{\hat{N}_{ab}^A}{N_B} + (1-\theta)\frac{\hat{N}_{ab}^B}{N_A}\right]x + \left[\theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B\right] = 0, \tag{2}$$

where

$$\theta_P = \frac{\hat{N}_a N_B \nu\left(\hat{N}_{ab}^B\right)}{\hat{N}_a N_B \nu\left(\hat{N}_{ab}^B\right) + \hat{N}_b N_A \nu\left(\hat{N}_{ab}^A\right)} \tag{3}$$

is chosen to minimize the asymptotic variance of $\hat{N}_{ab,PML}(\theta)$.

2.2. Local linear estimator using completely data driven bandwidth selection methods in complex surveys

Consider the general nonparametric regression model

$$y_i = \mu(t_i) + \varepsilon_i, i = 1, 2, ..., n, \tag{4}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of independent, identically distributed random variables with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{E}(\varepsilon_i^2) = \sigma^2, \mu(\cdot)$ is an unknown smooth regression curve to be estimated. Without loss of generality, we take $t_i \in [0, 1], i = 1, 2, ..., n$ and for simplicity we assume that $0 < t_1 < \cdots < t_n < 1$.

Let U be the union of frame A and frame B, S be a survey sample, N be the population size, n_S be the sample size (Note that n_S is random with $E(n_S) = n$), and let π_k be

the first order inclusion probability with $\pi_k = p(\text{unit} \quad k \in S)$. Sample weight d_k is the reciprocal of inclusion probability π_k, i.e. $d_k = 1/\pi_k$ for $k \in S$. Let $\hat{N}$ be the estimate of population size N i.e. $\hat{N} = \sum_{k=1}^{n_S} d_k$ and let r be the sampling rate defined as $r = n_S/N$. If every unit in the finite population was available, a Nadaraya–Watson regression smoother assuming that $\varepsilon_k, k = 1, 2 \cdots N$ are iid would be

$$\hat{\mu}_{(t,h)} = \frac{N^{-1} \sum_U K_h(t - t_k) y_k}{N^{-1} \sum_U K_h(t - t_k)}, \tag{5}$$

where h is bandwidth, $K_h(\cdot)$ is the kernel function, and $K_h(t - t_k) = h^{-1} K\{(t - t_k)/h\}$.

The local linear kernel estimator incorporating sample weights has the following form

$$\hat{\mu}(t,h) = \frac{\sum_S \left\{\hat{s}_2(t,h) - \hat{s}_1(t,h)(t_k - t)\right\} d_k K_h(t_k - t) y_k}{\hat{s}_2(t,h)\hat{s}_0(t,h) - \hat{s}_1^2(t,h)}, \tag{6}$$

where $\hat{s}_i(t,h) = \sum_S d_k (t_k - t)^i K_h(t_k - t), i = 0, 1,$ and 2.

Let $\tilde{\mu}(t,h)$ be the classical local linear estimator without considering sample weights, Harms and Duchesne (2010) showed that

$$\text{Bias}(\hat{\mu}(t,h)) = \text{Bias}(\tilde{\mu}(t,h)) + o(h^2), \tag{7}$$

and

$$\text{Var}(\hat{\mu}(t,h)) = (\triangle + r)\text{Var}(\tilde{\mu}(t,h)) + o\big((Nh)^{-1}\big), \quad \triangle = n_S/N^2 \sum_U (d_k - 1). \tag{8}$$

By minimizing the asymptotic MSE, Harms and Duchesne (2010) derived the optimal bandwidth for $\hat{\mu}$ as follows

$$\hat{h}^{\text{opt}}(t) = (\triangle + r)^{1/5} \tilde{h}^{\text{opt}}, \tag{9}$$

where $\tilde{h}^{\text{opt}}$ is the optimal bandwidth for $\tilde{\mu}(t,h)$, $(\triangle + r)^{1/5}$ is the correction factor, which is a function related to the sampling plan (refer to Harms and Duchesne (2010) for more details).

3. Local linear estimators in a dual frame survey

In Sec. 2, we have reviewed PML population total estimator in a dual frame survey, and local linear estimator in a single frame survey. In this section, we study local linear estimators in a dual frame survey by using the idea of PML estimator. We propose three methods for nonparametric local linear regression estimators in dual frame surveys and investigate their asymptotic properties. Our main idea is to convert the two independent samples to a pseudo single sample by using adjusted weight d_i^* and perform regular local linear regression estimation. A challenge here is to find estimates of θ for the overlap domain and the optimal bandwidth h for nonparametric regression. Method 1 simply uses the optimal θ_P in Eq. (3) and the paired optimal bandwidth $\hat{h}$ derived by Eq. (9). Method 2 uses cross-validation (CV) to select θ, with the paired optimal bandwidth $\hat{h}$ derived by Eq. (9). Method 3, on the other hand, uses CV to select optimal bandwidth $\hat{h}$ when fixing θ. All the methods are trying to look for optimal estimates of the parameter pair (θ, h).

3.1. Methods/estimators

Let $\mathbf{D}^*$ be the diagonal matrix of the modified sample weights d_i^*, with

$$d_i^* = \begin{cases} d_i, & \text{if} \quad i \in a, \\ \theta d_i, & \text{if} \quad i \in ab \text{ and } i \in \mathcal{S}_A, \\ (1-\theta) d_i, & \text{if} \quad i \in ab \text{ and } i \in \mathcal{S}_B, \\ d_i, & \text{if} \quad i \in b. \end{cases} \tag{10}$$

Method 1 uses the optimal variable θ_P in Eq. (3) to reweight the observations in the overlap domain to construct a pseudo single sample. Optimal bandwidth $\hat{h}$ is derived by Eq. (9). $\hat{\mu}(t,h)$ in Eq. (6) is derived using modified weight d_i^* in Eq. (10) and the optimal bandwidth $\hat{h}$.

In Method 2, we use cross-validation (CV) to derive a fully data-driven θ selection procedure. The weighted prediction sum of squares is as follows

$$\begin{aligned} CV(\theta) &= \sum_{i \in \mathcal{S}_A \cup \mathcal{S}_B} d_i^* \left(y_i - \hat{y}_{(i)}\right)^2 \\ &= \sum_{i=1}^{n_a} d_i \left(y_i - \hat{y}_{(i)}\right)^2 + \sum_{i=1}^{n_{ab}^A} \theta d_i \left(y_i - \hat{y}_{(i)}\right)^2 \\ &\quad + \sum_{j=1}^{n_{ab}^B} (1-\theta) d_j \left(y_j - \hat{y}_{(j)}\right)^2 + \sum_{j=1}^{n_b} d_j \left(y_j - \hat{y}_{(j)}\right)^2, \end{aligned}$$

where $\hat{y}_{(i)}$ is the estimate computed without using the ith observation. The ith observation is treated as an additional observation for prediction and $CV(\theta)$ measures the quality of predictions. In practice, we set up a grid between (0, 1) for possible θ value. For each θ value, we use Eq. (9) to find corresponding optimal $\hat{h}$. We then minimize the CV quantity to find optimal $\hat{\theta}$.

In Method 3, given the optimal $\hat{\theta}$ from Method 2, we use the following CV criterion to minimize the prediction error to select optimal bandwidth.

$$\begin{aligned} CV(h) &= \sum_{i \in \mathcal{S}_A \cup \mathcal{S}_B} d_i^* \left(y_i - \hat{y}_{(i)h}\right)^2 \\ &= \sum_{i=1}^{n_a} d_i \left(y_i - \hat{y}_{(i)h}\right)^2 + \sum_{i=1}^{n_{ab}^A} \hat{\theta} d_i \left(y_i - \hat{y}_{(i)h}\right)^2 \\ &\quad + \sum_{j=1}^{n_{ab}^B} (1-\hat{\theta}) d_j \left(y_j - \hat{y}_{(j)h}\right)^2 + \sum_{j=1}^{n_b} d_j \left(y_j - \hat{y}_{(j)h}\right)^2, \end{aligned}$$

where $\hat{\theta}$ is the optimal estimate of θ from Method 2, $\hat{y}_{(i)h}$ is the estimate computed without using the ith observation given a fixed h.

In practice, we first use Eq. (9) to find optimal bandwidth $\hat{h}_{opt,method2}$ based on $\hat{\theta}$ selected from Method 2. Next we set up a grid around $\hat{h}_{opt,method2}$ to find the optimal $\hat{h}$ that minimize the CV quantity.

3.2. Asymptotic properties of the estimators

From all the three methods, the local linear estimators are in the following matrix form

$$\hat{\boldsymbol{\mu}} = \mathbf{e}_1^T \left(\mathbf{Z}(t)^T \mathbf{D}_S^* \mathbf{W}_S \mathbf{Z}(t) \right)^{-1} \left(\mathbf{Z}(t)^T \mathbf{D}_S^* \mathbf{W}_S \mathbf{y} \right), \tag{11}$$

where $\mathbf{e}_1$ is a vector with 1st element 1 and zero else, $\mathbf{D}_S^*$ is a diagonal matrix whose elements are the sampling weights, $\mathbf{W}_S$ is an $n \times n$ diagonal matrix whose elements are equal to $K_h(t-t_k)_{k\in S}$, and $\mathbf{Z}(t)$ is an $n \times 2$ matrix whose kth row is of the form $\mathbf{Z}(t)_k = (1, (t_k - t))_{k\in S}$.

We now examine the asymptotic properties of the estimators in Eq. (11). For this, we'v e made three assumptions for the population, sampling plan and bandwidth.

Assumption 1: For frame A, suppose that a finite population U_{1A} with size N_{1A} is a subset of a larger superpopulation U_{2A} with size N_{2A}, the superpopulation U_{2A} is a subset of a larger superpopulation U_{3A} with size N_{3A}, and so on, i.e., $N_{1A} < N_{2A} < \cdots < N_{iA} < \cdots$. N_{iA} goes to infinity as $i \to \infty$. Given a population U_{iA}, a sample of size n_{iA} is drawn according to a sampling plan. For frame B, suppose that a finite population U_{1B} with size N_{1B} is a subset of a larger superpopulation U_{2B} with size N_{2B}, the superpopulation U_{2B} is a subset of a larger superpopulation U_{3B} with size N_{3B}, and so on, i.e., $N_{1B} < N_{2B} < \cdots < N_{iB} < \cdots$. N_{iB} goes to infinity as $i \to \infty$. Given a population U_{iB}, a sample of size n_{iB} is drawn according to a sampling plan.

Assumption 2: For both frames, the sampling rate n_{iA}/N_{iA} and n_{iB}/N_{iB} converges with probability one (wp1) to a finite constant $r > 0$, as $i \to \infty$. The first order inclusion probabilities for sample from frame A are such that for any N_{iA}, $min_{k\in U_{iA}} \pi_{kA} \geq \lambda_A > 0$ wp1. The first order inclusion probabilities for sample from frame B are such that for any N_{iB}, $min_{k\in U_{iB}} \pi_{kB} \geq \lambda_B > 0$ wp1. The second order inclusion probabilities for sample from frame A satisfy $min_{k,l\in U_{iA}} \pi_{klA} \geq \lambda_A^* > 0$ and

$$\lim \sup_{i\to\infty} n_{iA} \max_{k,l\in U_{iA}: k\neq l} |\pi_{klA} - \pi_{kA}\pi_{lA}| < \infty, \text{wp1}.$$

The second order inclusion probabilities for sample from frame B satisfy $min_{k,l\in U_{iB}} \pi_{klB} \geq \lambda_B^* > 0$ and

$$\lim \sup_{i\to\infty} n_{iB} \max_{k,l\in U_{iB}: k\neq l} |\pi_{klB} - \pi_{kB}\pi_{lB}| < \infty, \text{wp1}.$$

In addition, assume that the number of psus $\tilde{n}_{iA}$ and $\tilde{n}_{iB}$ in the two samples both increase such that $\tilde{n}_{iA}/\tilde{n}_{iB} \to \gamma$ for some $0 < \gamma < 1$.

Assumption 3: Bandwidth $h_i = h_i(N_i, n_i)$, is such that $h_i \to 0$ and $N_i h_i \to \infty$, as $i \to \infty$, where $N_i = N_{iA} + N_{iB} - N_{iAB}$ and $n_i = n_{iA} + n_{iB}$.

Based on the assumptions, follow from Theorem 1 in Harms and Duchesne (2010), we have the following properties of the estimators.

Asymptotic Properties: the asymptotic bias and variance of the local linear estimators in Eq. (11) are as follows

$$\text{Bias}(\hat{\mu}(t,h)) = \frac{1}{2} h^2 \mu''(t) \mu_2(K) + o(h^2), \tag{12}$$

$$\mathrm{Var}(\hat{\mu}(t,h)) = \frac{1}{nh}(\triangle + r)\frac{R(k)\sigma^2}{f_t(t)} + o\big((Nh)^{-1}\big), \tag{13}$$

where $\triangle = n/N^2 \sum_U (d_k^* - 1), R(k) = \int K^2(z)dz, \mu_2(K) = \int_{-\infty}^{\infty} z^2 K(z)dz$, and $f_t(t)$ is the density function of t.

Note that Eq. (12) shows that the first term of bias is related to bandwidth h. The effect of sampling weight is of $o(h^2)$ order, and is negligible. While, Eq. (13) shows that sampling weights are closely related to variance of $\hat{\mu}(t,h)$. We can see that the larger the bandwidth h, the larger the bias and the smaller the variance. Therefore, the optimal bandwidth $\hat{h}$ is selected to balance the bias and variance.

4. Simulation studies

In this section, we perform a simulation study to investigate finite sample properties of the nonparametric regression estimators in dual frame surveys. We will compare among linear regression (Lu 2014) estimator and the three proposed nonparametric local linear estimators. Method 1 in Lu (2014) was used in our simulation study, which uses θ_P in Eq. (3) to reweight the samples for a pseudo single sample and perform linear regression estimation.

Simulation setup follows Harms and Duchesne (2010) and Zhang, Christensen, and Zheng (2015). At the super model stage, the following equation is used to generate the population

$$y_i = \mu_k(t_i) + \epsilon_i \;\; i = 1, ..., 1000 \text{ and } k = 1, 2, 3, 4, \tag{14}$$

where each population has $N = 1000$ values of t_i which are equally spaced in the interval $[0, 1]$ and random errors are normally distributed with mean 0 and constant variance σ^2. First, we generate population of $A \cup B$ by setting $t \in [0, 1]$. Frame A is defined by setting $t \in [0, 0.7]$ and frame B is defined by setting $t \in [0.3, 1]$. Note, when $t \in [0.3, 0.7]$, frame A and frame B overlapped.

Four functions are used to generate populations at the super model stage:

Härdle : $\mu_1(t) = \sin^3(2\pi t^3)$ Härdle (1991),

Bump : $\mu_2(t) = 1 + 2(t - 0.5) + \exp(-200(t - 0.5)^2)$ Breidt and Opsomer (2000),

Exponential : $\mu_3(t) = \exp(-8t)$ Breidt and Opsomer (2000),

Slow sine : $\mu_4(t) = 2 + \sin(2\pi t)$ Opsomer and Miller (2005). Note that all the four functions are not linear. As a result, Lu (2014) method are not appropriate in most settings since they assume an underlying linear funtion.

At the sampling design stage, we consider different sampling rates and poisson sampling scheme. (1) Sampling rate: 10% and 20%; (2) Sampling plan: Poisson sampling scheme (unequal probability design). The sample weights w_i of poisson sampling scheme have been chosen such that weights are proportional to the auxiliary variable $z_i = (y_i + 2)(t_i + 2)$ and $\sum_U 1/w_i = E(n_S) = N * r$;

For each setting, we did $L = 500$ simulations. Each time, we generate a population based on one of the four super models. Next, we use Poisson sampling to draw two samples from frame A and frame B respectively. We evaluate the estimators by bias, variance and MSE. Let $\hat{\mu}(t)$ be an estimator of $\mu(t)$. Assume $\hat{\mu}^{(i)}(t)$ represents the estimator of $\mu(t)$ from the ith sample, $i = 1, ..., L$. The Monte Carlo mean E_{MC}, the Monte

Table 1. Comparison of Bias, Variance and MSE under Poisson Sampling Scheme between Linear estimator (LE) and Local linear estimator by Method 1 (LLE1).

			$Bias^2$		Variance		MSE	
Function		Sampling rate	LE	LLE1	LE	LLE1	LE	LLE1
Härdle	$\sigma = .4$	10%	0.1855	0.0057	0.0063	0.136	0.1919	0.1417
		20%	0.1866	0.0050	0.0031	0.0147	0.1898	0.0198
	$\sigma = 1$	10%	0.1969	0.0638	0.0410	0.2455	0.2379	0.3094
		20%	0.1951	0.0274	0.0268	0.0679	0.2220	0.0953
Bump	$\sigma = .4$	10%	0.0729	0.0028	0.0029	0.0693	0.0759	0.0722
		20%	0.0732	0.0032	0.0012	0.0192	0.0744	0.0225
	$\sigma = 1$	10%	0.0764	0.0133	0.0219	0.2230	0.0984	0.2363
		20%	0.0732	0.0150	0.0109	0.0605	0.0841	0.0756
Exponential	$\sigma = .4$	10%	0.0198	0.0009	0.0018	0.0086	0.0217	0.0096
		20%	0.0198	0.0010	0.0008	0.0036	0.0207	0.0046
	$\sigma = 1$	10%	0.0215	0.0120	0.0194	0.0741	0.0409	0.0861
		20%	0.0298	0.0216	0.0164	0.0435	0.0463	0.0651
Slow sine	$\sigma = .4$	10%	0.1962	0.0017	0.0043	0.0169	0.2006	0.0187
		20%	0.1966	0.0015	0.0018	0.0048	0.1984	0.0064
	$\sigma = 1$	10%	0.1970	0.0077	0.0162	0.0729	0.2133	0.0807
		20%	0.2030	0.0115	0.0091	0.0287	0.2122	0.0403

Carlo bias B_{MC}, Monte Carlo variance V_{MC}, and the Monte Carlo MSE are given by the following formulas

$$E_{MC}\{\hat{\mu}(t)\} = L^{-1}\sum_{i=1}^{L}\hat{\mu}^{(i)}(t), \tag{15}$$

$$B_{MC}\{\hat{\mu}(t)\} = E_{MC}\{\hat{\mu}(t)\} - \mu(t), \tag{16}$$

$$V_{MC}\{\hat{\mu}(t)\} = L^{-1}\sum_{m=1}^{L}\left[\hat{\mu}^{(i)}(t) - E_{MC}\{\hat{\mu}(t)\}\right]^2, \tag{17}$$

and the main criterion for determining efficiency: Monte Carlo MSE is defined by

$$MSE_{MC}\{\hat{\mu}(t)\} = L^{-1}\sum_{i=1}^{L}\left\{\hat{\mu}^{(i)}(t) - \mu(t)\right\}^2. \tag{18}$$

For each point $t_j, j = 1, \ldots, 200$, we calculate Monte Carlo bias, variance and MSE using formulas Eqs. (16), (17) and (18) respectively. The averages of 200 bias, variance and MSE are reported.

In the following, we use LE to denote the linear estimator from Lu (2014) and use LLE1 to LLE3 to denote the proposed local linear estimators by Method 1 to Method 3 respectively. Tables 1–3 give the simulation results of bias, variance and MSE from the comparison between LE and LLE1, comparison between LLE1 and LLE2, and comparison between LLE1 and LLE3 respectively.

All the four super models Härdle, Bump, Exponential and Slow sine that we used to generate data are not linear. As a result, LE method that use regular linear regression fitting will produce larger bias than the nonparametric fitting using LLE1-LLE3. This systematic trend can be seem from Tables 1–3.

On the other hand, for the cases with large error variance $\sigma = 1$ and small sampling rate 10%, the trade off of small bias by using nonparametric methods LLE1-3 is that they produce larger variances of the estimators than LE. For example, in the case of

Table 2. Comparison of Bias, Variance and MSE under Poisson Sampling Scheme between local linear estimator by Method 1 (LLE1) and local linear estimator by Method 2 (LLE2).

			$Bias^2$		Variance		MSE	
Function		Sampling rate	LLE1	LLE2	LLE1	LLE2	LLE1	LLE2
Härdle	$\sigma = .4$	10%	0.0047	0.0047	0.0742	0.0743	0.0790	0.0791
		20%	0.0049	0.0049	0.013	0.0131	0.0180	0.0181
	$\sigma = 1$	10%	0.0465	0.0467	0.3254	0.3252	0.3720	0.3720
		20%	0.0353	0.0360	0.0690	0.0689	0.1044	0.1050
Bump	$\sigma = .4$	10%	0.0031	0.0031	0.0831	0.0831	0.0862	0.0863
		20%	0.0016	0.0016	0.0118	0.0119	0.0134	0.0135
	$\sigma = 1$	10%	0.0155	0.0156	0.2170	0.2170	0.2326	0.2326
		20%	0.0133	0.0137	0.0655	0.0654	0.0789	0.0792
Exponential	$\sigma = .4$	10%	0.0010	0.0010	0.0088	0.0089	0.0098	0.0099
		20%	0.0012	0.0012	0.0041	0.0041	0.0054	0.0054
	$\sigma = 1$	10%	0.0164	0.0171	0.1289	0.1288	0.1454	0.1460
		20%	0.0102	0.0107	0.0363	0.0358	0.0465	0.0466
Slow sine	$\sigma = .4$	10%	0.0015	0.0016	0.0154	0.0155	0.0170	0.0171
		20%	0.0018	0.0018	0.0094	0.0094	0.0112	0.0113
	$\sigma = 1$	10%	0.0108	0.0111	0.0884	0.0887	0.0993	0.0998
		20%	0.0153	0.0150	0.0329	0.0328	0.0482	0.0479

Table 3. Comparison of Bias, Variance and MSE under Poisson Sampling Scheme between local linear estimator by Method 1 (LLE1) and local linear estimator by Method 3 (LLE3).

			$Bias^2$		Variance		MSE	
Function		Sampling rate	LLE1	LLE3	LLE1	LLE3	LLE1	LLE3
Härdle	$\sigma = .4$	10%	0.0070	0.0071	0.1283	0.1284	0.1353	0.1355
		20%	0.0047	0.0048	0.0098	0.0101	0.0146	0.0149
	$\sigma = 1$	10%	0.0633	0.0643	0.1229	0.1230	0.1863	0.1873
		20%	0.0563	0.0576	0.0932	0.0973	0.1496	0.1549
Bump	$\sigma = .4$	10%	0.0021	0.0023	0.0619	0.0621	0.0641	0.0644
		20%	0.0032	0.0033	0.0121	0.0128	0.0153	0.0161
	$\sigma = 1$	10%	0.0297	0.0299	0.1284	0.1292	0.1582	0.1591
		20%	0.0176	0.0181	0.0507	0.0520	0.0683	0.0702
Exponential	$\sigma = .4$	10%	0.0023	0.0023	0.0107	0.0110	0.0130	0.0134
		20%	0.0027	0.0027	0.0044	0.0045	0.0072	0.0072
	$\sigma = 1$	10%	0.0080	0.0081	0.1502	0.1511	0.1583	0.1592
		20%	0.0104	0.0105	0.0328	0.0334	0.0432	0.0439
Slow sine	$\sigma = .4$	10%	0.0043	0.0045	0.0233	0.0233	0.0276	0.0278
		20%	0.0022	0.0023	0.0057	0.0058	0.0079	0.0082
	$\sigma = 1$	10%	0.0107	0.0108	0.0796	0.0800	0.0904	0.0908
		20%	0.0077	0.0078	0.0342	0.0349	0.0420	0.0428

Bump function with $\sigma = 1$ and sampling rate of 10%, variance of LLE1 is 0.2230 compared to variance of LE of 0.0219. This results a larger MSE of 0.2363 by LLE1 compared with MSE of 0.0984 by LE. This phenomena can be seen from all the cases with $\sigma = 1$ and sampling rate of 10% except for the slow sine function. Slow sine function is the one with the greatest variability among the four functions, which makes the bias of LE dominating and variance of LLE1 more close to that of LE. Therefore, MSE of LLE1 is smaller than MSE of LE. With a larger sample size (sampling rate increase to 20%), we observe that the variances by LLE1 is close to the variances of LE and MSE by LLE1 is smaller than MSE of LE. Local linear regression method, at this point, reflects its advantage by fitting linear regression in local neighborhood instead of the whole data set, which allows great flexibility in the possible form of the regression curve, when the super model is not linear.

In general, nonparametric local linear estimators from the three methods perform better than linear estimators since our super model functions are nonlinear. For example, in Table 1, with Härdle function, $\sigma = 0.4$ and sampling rate of 20%, MSE from LL1 is only 0.0198 compared to 0.1898 from Lu's method.

Table 2 reports the comparison between performance of LLE1 and LLE2 (with data driven θ selection procedure instead of PML estimator $\hat{\theta}_P$ used by LLE1) under Poisson sampling scheme for different settings. We notice the squared bias, variance and MSE of the estimators are almost identical from LLE1 and LLE2. Same findings are found from Table 3 when comparing LLE1 with LLE3 (full data driven bandwidth h selection procedure). These finding show that using data driven selection procedure for θ or for h didn't help improve the efficiency of the estimators compared to simply using θ_P and $\hat{h}_P$ derived by Eq. (9). This means that θ_P may be close to the true value and formula Eq. (9) is efficient. We would suggest using LL1 (local linear estimator by method 1) for regression estimation in dual frame surveys.

5. Conclusions

It is becoming more difficult, for a single sampling frame to include the entire population of interest and to be inexpensive to sample. Dual frame surveys, therefore, are becoming more popular. Such surveys require new methods for analyzing the regression aspects of the data.

In this research, we propose three nonparametric regression estimators for use in a dual frame survey. Simulation results show that all the three proposed methods work well and perform similarly to each other. In general, all the three methods perform better than the linear estimator proposed by Lu (2014) since the underlying functions are non-linear. Method 1, using PML estimator θ_P and optimal bandwidth by formula Eq. (9) suggested by Harms and Duchesne (2010) is therefore preferred and recommended for use in a dual frame survey regression estimation, since it is simple to use and is performing similarly as the other two methods.

Our research is done in the context of survey sampling, but they also apply to other settings in which data could be combined from two independent sources and could be extended to more than two surveys.

Acknowledgments

The authors thank the referees for their careful reading of the manuscript and would like to express sincere appreciation for their insightful and helpful comments.

References

Bankier, M. D. 1986. Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association* 81 (396):1074–9. doi:10.1080/01621459.1986.10478376.

Bellhouse, D. R., and J. E. Stafford. 1999. Density estimation from complex surveys. *Statistica Sinica* 9:407–24.

Bellhouse, D. R., and J. E. Stafford. 2001. Local polynomial regression in complex surveys. *Survey Methodology* 27 (2):197–203.

Breidt, F. J., G. Claeskens, and J. D. Opsomer. 2005. Model-assisted estimation for complex surveys using penalised splines. *Biometrika* 92 (4):831–46. doi:10.1093/biomet/92.4.831.

Breidt, F. J., and J. D. Opsomer. 2000. Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics* 28 (4):1026–53. doi:10.1214/aos/1015956706.

Buskirk, T. D. 1998. Nonparametric density estimation using complex survey data. Proceedings of the survey research methods section. American statistical association, 799–801.

Buskirk, T. D., and S. L. Lohr. 2005. Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference* 128 (1):160–90.

Fuller, W. A., and L. F. Burmeister. 1972. Estimators for samples selected from two overlapping frames. ASA proceedings of the social statistics section, 245–9. American Statistical Association.

Goga, C. 2005. Variance reduction in surveys with auxiliary information: A nonparametric approach involving regression splines. *Canadian Journal of Statistics* 33 (2):163–80. doi:10.1002/cjs.5550330202.

Härdle, W. 1991. *Smoothing techniques with implementation in S*. New York: Springer.

Harms, T., and P. Duchesne. 2010. On kernel nonparametric regression designed for complex survey data. *Metrika* 72 (1):111–38. doi:10.1007/s00184-009-0244-5.

Hartley, H. O. 1962. Multiple frame surveys. In ASA Proceedings of the Social Statistics Section, American Statistical Association, 203–6.

Hartley, H. O. 1974. Multiple frame methodology and selected applications. *Sankhyā, Series C* 36(3):99–118.

Keeter, S., M. Dimock, and L. Christian. 2010. The growing gap between landline and dual frame election polls. Pewresearch.org, November 22.

Korn, E. L., and B. I. Graubard. 1998. Scatterplots with survey data. *The American Statistician* 52: 58–69. doi:10.2307/2685570.

Lohr, S. L., and J. N. K. Rao. 2000. Inference from dual frame surveys. *Journal of the American Statistical Association* 95 (449):271–80. doi:10.1080/01621459.2000.10473920.

Lu, Y. 2014. Regression coefficient estimation in dual frame surveys. *Communications in Statistics - Simulation and Computation* 43 (7):1675–84. doi:10.1080/03610918.2012.752835.

Lu, Y., and S. L. Lohr. 2010. Gross flow estimation in dual frame surveys. *Survey Methodology* 36:13–22.

Merkouris, T. 2004. Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association* 99 (468):1131–9. doi:10.1198/016214504000000601.

Metcalf, P., and A. Scott. 2009. Using multiple frames in health surveys. *Statistics in Medicine* 28 (10):1512–23.

Opsomer, J. D., and C. P. Miller. 2005. Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Journal of Nonparametric Statistics* 17 (5):593–611. doi:10.1080/10485250500054642.

Skinner, C. J. 1991. On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association* 86 (415):779–84. doi:10.1080/01621459.1991.10475109.

Skinner, C. J., and J. N. K. Rao. 1996. Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association* 91 (433):349–56. doi:10.1080/01621459.1996.10476695.

Zhang, G., F. Christensen, and W. Zheng. 2015. Nonparametric regression estimators in complex surveys. *Journal of Statistical Computation and Simulation* 85 (5):1026–34. doi:10.1080/00949655.2013.860139.