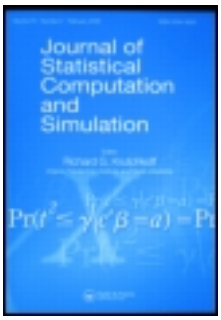

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Difference-based variance estimator for nonparametric regression in complex surveys

Yan Lu [a]

[a] Department of Mathematics and Statistics, University of New Mexico, MSC01 1115, 1, Albuquerque, NM, USA

Version of record first published: 30 Jul 2012.

To cite this article: Yan Lu (2012): Difference-based variance estimator for nonparametric regression in complex surveys, Journal of Statistical Computation and Simulation, DOI:10.1080/00949655.2012.708344

To link to this article: http://dx.doi.org/10.1080/00949655.2012.708344

Journal of Statistical Computation and Simulation
iFirst, 2012, 1–9

Difference-based variance estimator for nonparametric regression in complex surveys

Yan Lu*

Department of Mathematics and Statistics, MSC01 1115, 1 University of New Mexico, Albuquerque, NM, USA

(*Received 24 February 2012; final version received 27 June 2012*)

A difference-based variance estimator is proposed for nonparametric regression in complex surveys. By using a combined inference framework, the estimator is shown to be asymptotically normal and to converge to the true variance at a parametric rate. Simulation studies show that the proposed variance estimator works well for complex survey data and also reveals some finite sample properties of the estimator.

Keywords: asymptotic property; combined inference framework; complex surveys; difference-based variance estimator; nonparametric regression; simulation

1. Introduction

Consider a general nonparametric regression model

$$y_i = \mu(t_i) + \epsilon_i, \quad i = 1, 2, \ldots n, \tag{1}$$

where $\mu(\cdot)$ is an unknown function and ϵ_i are zero mean random errors with common variance σ^2. If $\mu(\cdot)$ is smooth and t ordinates are closely spaced, it is possible to remove the effect of the unknown function by differencing the data appropriately. So variance could be estimated without having to estimate the underlying regression curve.

Von Neumann [1] first proposed using successive differences to obtain variance estimator that could be used in situations where mean was not constant. A similar idea was used by Rice [2] for the estimation of the variance term in a bandwidth selection criterion for nonparametric kernel regression. Gasser *et al.* [3] employed Rice's [2] suggestion of a pseudo-residual estimator in the case of nonparametric regression and showed that variance could be estimated with parametric efficiency without having to estimate the underlying regression curve. Since then, the idea of pseudo-residuals attracted many interests from statisticians. Pseudo-residuals of similar form were used in Müller and Stadtmüller [4] for estimating heteroscedasticity in regression analysis. Hall *et al.* [5] suggested and computed asymptotically optimal difference sequence for estimating

*Email: luyan@math.unm.edu

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2012.708344
http://www.tandfonline.com

error variance in homoscedastic nonparametric regression. Buckley *et al.* [6] considered a wide class of estimators of the residual variance in nonparametric regression and derived the minimax mean-squared error estimator over a natural class of regression curve. Eubank *et al.* [7], and Klipple and Eubank [8] extended work from Gasser *et al.* [3] to partially linear models.

Nonparametric regression in complex surveys has gained much attention in recent years. A complex survey may include strata and clusters at the design stage. Naively ignoring survey weights may lead to biased inferences or undesired outcome in survey sampling practice. Classical nonparametric regression estimators and methods have been extended and investigated in survey area. Korn *et al.* [9] discussed a smoothed empirical cumulative distribution function to obtain estimates of the underlying percentiles. Modifications were also discussed for survey data. Korn and Graubard [10] suggested nonparametric smoothing for estimating conditional means and percentile curves. Bellhouse and Stafford [11,12] developed estimators for density estimation and regression functions. Breidt and Opsomer [13] proposed local polynomial regression estimators for estimating population totals and proved that their estimator was asymptotically design unbiased and consistent. Buskirk and Lohr [14] presented finite-sample and asymptotic properties under several approaches for inference of a modified density estimator introduced by Buskirk [15] and Bellhouse and Stafford [11]. Opsomer and Miller [16] studied the selection of the amount of smoothing for the nonparametric regression component of a model-assisted estimator using a cross-validation criterion. Breidt *et al.* [17] and Goga [18] proposed estimators of the population totals using splines. Harms and Duchesne [19] considered nonparametric regression analysis between two variables when data were sampled through a complex survey and derived asymptotic mean-squared error of kernel estimators using a combined inference framework.

In this article, we extend the variance estimator from Gasser *et al.* [3] by incorporating survey weights to nonparametric regression in complex surveys. A complex survey presents additional challenges to those from simple random sampling (SRS) mainly because of existence of weights. Questions arise such as: if we can ignore weights in some particular situation; how much will weights affect our estimator; are there effects from sampling rate on our estimator and so on. Furthermore, we want to derive asymptotic properties of the proposed estimator. A combined inference framework is used for this purpose. The combined approach has been used for a while, including Hartley and Sielken [20], Preffermann [21], Graubard and Korn [22], Buskirk and Lohr [14], and Harms and Duchesne [19]. Buskirk and Lohr [14] introduced a superpopulation structure for model-based inference that allows the population model to reflect the presence of clusters. In this paper, we adopt the combined approach from Buskirk and Lohr [14] that assumes a two-stage process: first, the elements in the finite population are supposed to be realizations of a N-dimensional (N is the population size) vector according to a joint probability distribution. Next, a sample is selected using a probability sampling design, from the elements in the finite population. Using the combined inference framework, the proposed estimator is shown to be asymptotically normal and to converge to the true variance at a parametric rate.

This paper is organized as follows. In Section 2, we give a brief review of the estimator in Gasser *et al.* [3]. In Section 3, we propose difference-based variance estimator for nonparametric regression in complex surveys and derive asymptotic properties of the estimator. In Section 4, we present simulation studies. Finally, we give our conclusions in Section 5.

2. Background

The standard nonparametric regression model (1) assumes that we observe responses $y_1, y_2, \ldots, y_n$ at points $0 < t_1 < \cdots < t_n < 1$. Suppose that $u(\cdot)$ is differentiable and t ordinates are closely spaced. Gasser *et al.* [3] suggested the following variance estimator without having to estimate

the underlying regression curve:

$$\hat{\sigma}_0^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}\tilde{\epsilon}_i^2, \tag{2}$$

where the $\tilde{\epsilon}_i$ are called pseudo-residuals defined by

$$\tilde{\epsilon}_i = d_{i0}y_i + d_{i1}y_{i+1} + d_{i2}y_{i+2}, \tag{3}$$

with

$$d_{i0} = \frac{-a_i}{\sqrt{1+a_i^2+b_i^2}}, \quad d_{i1} = \frac{1}{\sqrt{1+a_i^2+b_i^2}}, \quad d_{i2} = \frac{-b_i}{\sqrt{1+a_i^2+b_i^2}}$$

for

$$a_i = \frac{t_{i+2}-t_{i+1}}{t_{i+2}-t_i} \quad \text{and} \quad b_i = \frac{t_{i+1}-t_i}{t_{i+2}-t_i}.$$

The pseudo-residuals $\tilde{\epsilon}_i^2$ are from straight line fits involving triples of points in lieu of successive differences. They can be considered as weighted average of the observations that are asymptotically free of the response means. Gasser *et al.* [3] showed that when the ϵ_i are independent and identically distributed, $\sqrt{n}(\hat{\sigma}_0^2 - \sigma^2)$ has a limiting normal distribution.

3. Estimator and asymptotic properties

3.1. *Difference-based variance estimators for nonparametric regression in complex survey*

The goal of this study is to estimate the random error in nonparametric regression based on a sample S drawn through a complex sampling plan without estimating the unknown function $\mu(\cdot)$. Let $\pi_i = P(i \in S)$ be the first-order inclusion probability, so $1/\pi_i$ represents the sampling weight. A natural extension of the estimator from Gasser *et al.* [3] is as follows:

$$\hat{\sigma}_w^2 = \frac{\sum_{i=1}^{n-2}(1/\pi_i)\tilde{\epsilon}_i^2}{\sum_{i=1}^{n-2}(1/\pi_i)}, \tag{4}$$

or in a matrix form as

$$\hat{\sigma}_w^2 = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{D}\mathbf{y}}{\mathrm{tr}(\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{D})}, \tag{5}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{W}$ is a diagonal matrix with $w_{ii} = 1/\pi_i$, and tr is the trace function for a square matrix. The $(n-2) \times n$ matrix $\mathbf{D}$ has the ith row $[\mathbf{0}_{i-1}, d_{i0}, d_{i1}, d_{i2}, \mathbf{0}_{n-i-2}]$ with $\mathbf{0}_r$ representing a r-vector with all zero elements.

3.2. *Asymptotic properties of our estimator*

Let U be a finite population of size N and $(T_i, Y_i), i \in U$, correspond to the auxiliary and response variables associated with each unit i in the population U. The unknown regression function $\mu(\cdot)$ represents the relationship between the predictor and the response through model (1). At the model stage, assume a sequence of populations U_j with population size N_j, such that N_j goes to infinity as $j \to \infty$. The finite populations U_j are generated independently, such that (T, Y) has the joint density $f_{TY}(t, y)$. At the sampling stage, a sample of size $n_{s,j}$ is drawn from U_j according to a

sampling design. More details of combined inference framework can be found from Buskirk and Lohr [14].

THEOREM 1 *Given the following conditions*,

(1) *The sampling rate* $f_j = n_{s,j}/N_j$ *converges with probability one to a finite constant* f, *as* $j \to \infty$. *The first-order inclusion probabilities are positive and bounded with* $min_{k\in U_j}\pi_k \geq \lambda^* > 0$ *and* $max_{k\in U_j}\pi_k \leq \lambda^{**} < 1$.
(2) *The successive differences* $t_i - t_{i-1}$ *decay to zero at the rate* n^{-1} *for all* U_j.
(3) *The random error* $\{\epsilon_i\}$ *are independent and identically distributed with mean zero and variance* σ^2 *and fourth moment bounded.*
(4) *The unknown regression function* μ *is differentiable*, *we obtain*

$$\theta^{-1}(\hat{\sigma}_w^2 - \sigma^2) \xrightarrow{d} N(0, \sigma^4), \tag{6}$$

where

$$\theta = \left\{ \frac{2\text{tr}((\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D})^2)}{(\text{tr}(\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D}))^2} + \frac{(m_4 - 3)\sum_{i=1}^n s_i^2}{(\text{tr}(\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D}))^2} \right\}^{1/2},$$

s_i *is the ith diagonal element of matrix* $\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D}$ *and* $E(\epsilon^4) = m_4\sigma^4$.

Proof First, use mean value theorem to show that

$$\tilde{\epsilon}_i = d_{i0}\epsilon_i + d_{i1}\epsilon_{i+1} + d_{i2}\epsilon_{i+2} + O(n^{-1}). \tag{7}$$

Asymptotic normality then follows from a proof of the central limit theorem for m-dependent sequences by Orey [23] and convergence of weighted averages of independent random variables by Jamison *et al.* [24]. Other results are directly from Theorem 1 in Gasser *et al.* [3]. ■

PROPOSITION 1 *Suppose conditions* (1)–(4) *hold. The bias of* $\hat{\sigma}_w^2$ *is in the form of*

$$\text{Bias}(\hat{\sigma}_w^2) = \frac{\boldsymbol{\mu}^t\boldsymbol{D}^t\boldsymbol{W}\boldsymbol{D}\boldsymbol{\mu}}{\text{tr}(\boldsymbol{D}^t\boldsymbol{W}\boldsymbol{D})}, \tag{8}$$

and decays at rate n^{-2}, *where* $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_n))^{\text{T}}$.

Proof Rewriting $\hat{\sigma}_w^2$ as the following:

$$\hat{\sigma}_w^2 = \frac{\boldsymbol{\epsilon}^{\text{T}}\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D}\boldsymbol{\epsilon}}{\text{tr}(\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D})} + \frac{\boldsymbol{\epsilon}^{\text{T}}\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D}\boldsymbol{\mu}}{\text{tr}(\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D})} + \frac{\boldsymbol{\mu}^{\text{T}}\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D}\boldsymbol{\epsilon}}{\text{tr}(\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D})} + \frac{\boldsymbol{\mu}^{\text{T}}\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D}\boldsymbol{\mu}}{\text{tr}(\mathbf{D}^{\text{T}}\mathbf{W}\mathbf{D})}, \tag{9}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\text{T}}$. Notice that the expectation of the first term is σ^2 and the expectation of second and third terms is zero. The argument of decay rate can be derived immediately from Equation (7). ■

PROPOSITION 2 *Suppose that conditions* (1)–(4) *hold. The variance of* $\hat{\sigma}_w^2$ *is in the following form:*

$$\text{var}(\hat{\sigma}_w^2) = \frac{2\sigma^4\text{tr}((\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D})^2)}{(\text{tr}(\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D}))^2} + \frac{\sigma^4(m_4 - 3)\sum_{i=1}^n s_i^2}{(\text{tr}(\boldsymbol{D}^{\text{T}}\boldsymbol{W}\boldsymbol{D}))^2} + m_3 O(n^{-3}) + O(n^{-5}), \tag{10}$$

where $E(\epsilon^3) = m_3\sigma^3$.

Proof First, write the estimator of Gasser *et al.* [3] in Equation (2) in a matrix form

$$\hat{\sigma}_0^2 = \frac{\mathbf{y}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{y}}{\mathrm{tr}(\mathbf{D}^{\mathrm{T}}\mathbf{D})}. \tag{11}$$

The Proposition 2 is a special case of the result in Gasser *et al.* [3] by replacing $\mathbf{D}$ in Equation (11) by $D^* = \mathbf{W}^{1/2}\mathbf{D}$. ■

4. Simulation studies

In this section, a small simulation study has been conducted to investigate the finite sample properties of our proposed difference-based variance estimator. We also compare the proposed variance estimator with the estimator without weights information. The simulation set-up is similar as Harms and Duchesne [19]. The following equation is used to generate the population at the

Table 1. Simulation results under SRS.

Function		Sampling rate (%)	$\hat{\sigma}_w^2$	SE($\hat{\sigma}_w^2$)	$\sqrt{\widehat{V(\sigma_w^2)}}$
Härdle	$\sigma^2 = 1$	5	0.9789	0.2788	$0.2784_{(0.0793)}$
		10	1.0297	0.2030	$0.2056_{(0.0405)}$
		20	1.0189	0.1276	$0.1434_{(0.0179)}$
	$\sigma^2 = 4$	5	4.0568	1.1607	$1.1539_{(0.3302)}$
		10	4.2015	0.7849	$0.8390_{(0.1568)}$
		20	4.1563	0.5417	$0.5850_{(0.0762)}$
	$\sigma^2 = 16$	5	15.6324	4.2351	$4.4461_{(1.2052)}$
		10	16.0189	2.9176	$3.1990_{(0.5827)}$
		20	16.1420	2.0880	$2.2720_{(0.2939)}$
Bump	$\sigma^2 = 1$	5	0.9739	0.2803	$0.2769_{(0.0796)}$
		10	1.0134	0.1916	$0.2023_{(0.0382)}$
		20	1.1054	0.1432	$0.1555_{(0.0201)}$
	$\sigma^2 = 4$	5	4.2830	1.1919	$1.2180_{(0.3390)}$
		10	3.8470	0.7217	$0.7682_{(0.1442)}$
		20	3.7545	0.4960	$0.5284_{(0.0698)}$
	$\sigma^2 = 16$	5	16.0073	4.7051	$4.5520_{(1.3381)}$
		10	17.6056	3.4763	$3.5160_{(0.6943)}$
		20	17.1026	2.2759	$2.4072_{(0.3203)}$
Exponential	$\sigma^2 = 1$	5	0.9501	0.2461	$0.2702_{(0.0700)}$
		10	1.0339	0.2048	$0.2064_{(0.0409)}$
		20	0.9515	0.1275	$0.1339_{(0.0179)}$
	$\sigma^2 = 4$	5	4.0923	1.1402	$1.1637_{(0.3244)}$
		10	3.8319	0.7681	$0.7652_{(0.1534)}$
		20	3.8880	0.4772	$0.5472_{(0.0671)}$
	$\sigma^2 = 16$	5	16.3847	5.0943	$4.6597_{(1.4485)}$
		10	14.7465	2.9434	$2.9445_{(0.5878)}$
		20	16.8083	2.2148	$2.3657_{(0.3117)}$
Slow sine	$\sigma^2 = 1$	5	0.9904	0.2876	$0.2816_{(0.0817)}$
		10	1.0457	0.1968	$0.2088_{(0.0392)}$
		20	0.9986	0.1289	$0.1405_{(0.0181)}$
	$\sigma^2 = 4$	5	3.9214	1.1120	$1.1153_{(0.3164)}$
		10	3.6477	0.7011	$0.7284_{(0.1400)}$
		20	4.1882	0.5641	$0.5895_{(0.0794)}$
	$\sigma^2 = 16$	5	16.5579	5.1358	$4.7101_{(1.4613)}$
		10	15.0452	2.8274	$3.0044_{(0.5648)}$
		20	16.0492	2.0413	$2.2589_{(0.2874)}$

super model stage

$$y_i = f_k(t_i) + \epsilon_i \; i = 1, \ldots, 1000 \quad \text{and} \quad k = 1, 2, 3, 4, \tag{12}$$

where each population has $N = 1000$ values of t_i which is equally spaced in the interval $[0, 1]$ and random errors are from the normal distribution with mean 0 and constant variance σ^2. At the sampling design stage, different sampling rates and different sampling designs are considered.

The simulation study was performed with factors: (1) $\sigma^2 : 1, 4$ and 16; (2) sampling rate: 5%, 10% and 20%; (3) sampling plan: SRS and Poisson sampling scheme (unequal probability design). The sampling weights w_i of the Poisson sampling scheme have been chosen such that weights are proportional to the auxiliary variable $z_i = (y_i + 8)(t_i + 8)$ and $\sum_U 1/w_i = E(n_s) = N * f$ and (4) four functions are used to generate populations at the supermodel stage:

Härdle : $f_1(t) = \sin^3(2\pi t^3)$.
Bump : $f_2(t) = 1 + 2(t - 0.5) + \exp(-200(t - 0.5)^2)$.
Exponential : $f_3(t) = \exp(-8t)$.

Table 2. Simulation results under poisson sampling scheme.

Function		Sampling rate	$\hat{\sigma}_w^2$	SE($\hat{\sigma}_w^2$)	$\sqrt{\widehat{V(\sigma_w^2)}}$
Härdle	$\sigma^2 = 1$	5	0.9891	0.2831	$0.2856_{(0.0885)}$
		10	1.0529	0.2125	$0.2122_{(0.0441)}$
		20	0.9736	0.1271	$0.1381_{(0.0185)}$
	$\sigma^2 = 4$	5	4.0413	1.1773	$1.1826_{(0.3600)}$
		10	4.0958	0.9048	$0.8410_{(0.1950)}$
		20	3.7816	0.5288	$0.5446_{(0.0793)}$
	$\sigma^2 = 16$	5	15.8982	5.9930	$5.1079_{(2.6414)}$
		10	15.7047	4.5662	$3.5715_{(1.6306)}$
		20	14.9831	3.1844	$2.3622_{(1.0464)}$
Bump	$\sigma^2 = 1$	5	1.1074	0.3284	$0.3209_{(0.1001)}$
		10	0.9483	0.1943	$0.1915_{(0.0407)}$
		20	1.0203	0.1351	$0.1452_{(0.0198)}$
	$\sigma^2 = 4$	5	3.7615	1.1303	$1.0996_{(0.3481)}$
		10	4.0574	0.8216	$0.8299_{(0.1749)}$
		20	3.7355	0.5397	$0.5356_{(0.0797)}$
	$\sigma^2 = 16$	5	15.8620	5.4705	$4.9819_{(2.2765)}$
		10	14.6961	3.9217	$3.2782_{(1.3461)}$
		20	15.9398	3.2720	$2.5443_{(1.1814)}$
Exponential	$\sigma^2 = 1$	5	1.0243	0.2894	$0.2948_{(0.0862)}$
		10	1.0568	0.2134	$0.2129_{(0.0447)}$
		20	1.0581	0.1357	$0.1496_{(0.0198)}$
	$\sigma^2 = 4$	5	4.2840	1.2624	$1.2477_{(0.3850)}$
		10	4.2637	0.9680	$0.8706_{(0.2134)}$
		20	4.2728	0.6482	$0.6151_{(0.0969)}$
	$\sigma^2 = 16$	5	14.9327	5.7896	$4.8398_{(2.7160)}$
		10	17.0803	5.2429	$3.9398_{(2.2106)}$
		20	14.9830	3.0251	$2.4012_{(1.2957)}$
Slow sine	$\sigma^2 = 1$	5	1.0255	0.2956	$0.2958_{(0.0893)}$
		10	1.0322	0.2008	$0.2080_{(0.0417)}$
		20	0.9449	0.1233	$0.1337_{(0.0177)}$
	$\sigma^2 = 4$	5	3.7384	1.0632	$1.0853_{(0.3224)}$
		10	4.0252	0.8535	$0.8168_{(0.1816)}$
		20	4.0098	0.6574	$0.5721_{(0.0989)}$
	$\sigma^2 = 16$	5	15.7133	5.6594	$4.8222_{(2.1884)}$
		10	16.4389	5.0611	$3.6946_{(1.8175)}$
		20	15.5542	2.8251	$2.3698_{(0.8217)}$

Slow sine : $f_4(t) = 2 + \sin(2\pi t)$.

The first function has been considered in Härdle [25], the second and third functions are from Breidt and Opsomer [13] and the fourth one is taken from Opsomer and Miller [16].

Simulation does $L = 1000$ times for each setting. Each time, we generate a population based on one of the four supermodels, then draw a sample using either SRS or Poisson sampling. In Tables 1 and 2, $\hat{\sigma}_w^2$ is the average value of the proposed variance estimator from the 1000 replications; $\mathrm{SE}(\hat{\sigma}_w^2)$ is the sample standard error and is considered as the true standard deviation of $\hat{\sigma}_w^2$; $\sqrt{\widehat{V(\sigma_w^2)}}$ is the average of standard error estimate of $\hat{\sigma}_w^2$ using Equation (10) from the 1000 replication; Numbers in parenthesis are the sample standard error. Tables 1 and 2 report the performance of the proposed estimator under SRS and the Poisson sampling scheme for different settings.

Tables 1 and 2 show that our estimator performs very well under SRS and unequal probability sampling scheme. We see from the tables that $\hat{\sigma}_w^2$ is close to the true variance σ^2 and $\sqrt{\widehat{V(\sigma_w^2)}}$ is close to the true standard error $\mathrm{SE}(\hat{\sigma}_w^2)$ under all the settings. Note that under SRS, $\hat{\sigma}_0^2$ gives exactly the same results as $\hat{\sigma}_w^2$ since sample is self-weighting, so Equation (5) reduces to Equation (2).

Table 3. Comparison of EMSE and average of abbias between $\hat{\sigma}_w^2$ and $\hat{\sigma}_0^2$.

Function	σ^2	Sampling rate (%)	EMSE of $\hat{\sigma}_w^2$	EMSE of $\hat{\sigma}_0^2$	abbias of $\hat{\sigma}_w^2$	abbias of $\hat{\sigma}_0^2$
Härdle	1	5	0.1375	0.0848	0.2837	0.2466
		10	0.0657	0.0581	0.2003	0.2110
		20	0.0739	0.0700	0.1708	0.2447
	4	5	2.6172	3.0730	1.3444	1.6460
		10	2.1150	3.0264	1.2740	1.6768
		20	1.2102	2.4085	0.9642	1.5204
	16	5	48.5939	51.0600	6.2579	6.6970
		10	41.8003	59.0586	5.9868	7.5646
		20	48.7224	78.9899	6.7042	8.8537
Bump	1	5	0.1086	0.0606	0.2466	0.1989
		10	0.0714	0.0314	0.1865	0.1446
		20	0.0386	0.0401	0.1472	0.1746
	4	5	2.8081	2.4513	1.3659	1.3736
		10	1.5770	1.4987	1.0205	1.0812
		20	0.7807	1.3336	0.7110	1.0990
	16	5	45.5221	51.1168	5.9461	6.6883
		10	31.9839	48.7306	5.0639	6.7707
		20	32.4934	55.8585	5.3200	7.3999
Exponential	1	5	0.1483	0.0734	0.2770	0.2241
		10	0.0597	0.0519	0.1887	0.1938
		20	0.0367	0.0348	0.1393	0.1624
	4	5	2.4500	2.2074	1.2837	1.3293
		10	1.6035	2.1831	1.1161	1.4203
		20	0.7945	1.5538	0.7580	1.2038
	16	5	55.6085	68.7597	6.8147	7.9763
		10	44.5243	61.5709	6.0611	7.6845
		20	50.3562	82.5501	6.8338	9.0536
Slowsine	1	5	0.0733	0.0601	0.2092	0.1974
		10	0.0572	0.0353	0.1813	0.1523
		20	0.0239	0.0181	0.1206	0.1078
	4	5	1.9549	1.8209	1.1337	1.1789
		10	1.4184	1.3023	0.8781	1.0186
		20	0.4627	0.8422	0.5505	0.8360
	16	5	45.5060	52.2923	5.8344	6.7550
		10	21.8972	32.0708	4.0306	5.3480
		20	20.0894	31.6304	4.0610	5.4932

Empirical mean-squared error (EMSE) and average of absolute bias (abbias) are used to compare our developed variance estimator $\hat{\sigma}_w^2$ with $\hat{\sigma}_0^2$ (without weight information), and to discover some finite sample properties of our estimator. Given an estimator $\hat{Y}$, the average of the abbias is defined as

$$\text{abbias} = \frac{1}{R}\sum_{r=1}^{R}|\hat{Y}_r - Y|, \tag{13}$$

and EMSE is defined as

$$\text{EMSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_r - Y)^2, \tag{14}$$

where $\hat{Y}_r$ is the value of $\hat{Y}$ for the rth simulation run. In our simulation study, we used $R = 1000$. Table 3 gives interesting results of variance estimator for unequal probability sampling. Our proposed estimator performs similar as $\hat{\sigma}_0^2$ when $\sigma^2 = 1$; performs slightly better than $\hat{\sigma}_0^2$ when $\sigma^2 = 4$; and performs better than $\hat{\sigma}_0^2$ when $\sigma^2 = 16$. Variance reflects the spread out of the data. Under relatively large variance, weights become important in estimating the parameters. We also observe a pattern that the estimator performs better as sampling rate increases. This is because the finite population correction is not negligible when sampling rate increases. When the sampling rate reaches 20% and variance reaches 16 (such a condition is usually met in practical survey data), for function Härdle, EMSE of the proposed estimator is 48.7224 compared with 78.9899 (from unweighted estimator), average of abbias is 6.7042 compared with 8.8537; for the function Bump, EMSE of the proposed estimator is 32.4934 compared with 55.8585, average of abbias is 5.32 compared with 7.3999, etc.

5. Conclusion

In this article, we extended the difference-based variance estimator for nonparametric regression in [3] to complex surveys. By using a combined inference framework, the estimator is shown to be asymptotically normal and to converge to the true variance at a parametric rate. Simulation studies give interesting finite sample results. The proposed variance estimator performs well under all the settings. When the variance is relatively small, there is almost no difference between the proposed method and the estimator without weights information. When the variance is large, our estimator outperforms the one without incorporating weights information. The efficiency of our estimator is also related to sampling rate. Our estimator performs better as sampling rate increases.

Acknowledgements

The author thanks the associate editor and referees for their insightful and helpful comments.

References

[1] J. Von Neumann, *Distribution of the ratio of the mean squared successive difference to the variance*, Ann. Math. Statist. 12 (1941), pp. 367–395.
[2] J. Rice, *Bandwidth choice for nonparametric regression*, Ann. Statist. 12 (1984), pp. 1215–1230.
[3] T. Gasser, L. Sroka, and C. Jennen-Steinmetz, *Residual variance and residual pattern in nonlinear regression*, Biometrika 73 (1986), pp. 625–633.
[4] H.G. Müller and U. Stadtmüller, *Estimation of heteroscedasticity in regression analysis*, Ann. Statist. 15 (1987), pp. 610–625.
[5] P. Hall, J.W. Kay, and D.M. Titterington, *Asymptotically optimal difference-based estimation of variance in nonparametric regression*, Biometrika 77 (1990), pp. 521–528.

[6] M.J. Buckley, G.K. Eagleson, and B.W. Silverman, *The estimation of residual variance in nonparametric regression*, Biometrika 75 (1988), pp. 189–199.
[7] R.L. Eubank, E.L. Kambour, J.T. Kim, K. Klipple, and C.S. Reese, *Estimation in partially linear models*, Comput. Stat. Data Anal. 29 (1998), pp. 27–34.
[8] K. Klipple and R.L. Eubank, *Difference based variance estimators for partially linear models*, Festschrift in Honor of Distinguished Professor Mir Masoom Ali On the Occasion of his Retirement, Muncie, IN, USA. 18–19 May, 2007, pp. 313–323.
[9] E.L. Korn, D. Midthune, and B.I. Graubard, *Estimating interpolated percentiles from grouped data with large samples*, J. Official Stat. 13 (1997), pp. 385–399.
[10] E.L. Korn and B.I. Graubard, *Scatterplots with survey data*, Am. Stat. 52 (1998), pp. 58–69.
[11] D.R. Bellhouse and J.E. Stafford, *Density estimation from complex surveys*, Statist. Sinica 9 (1999), pp. 407–424.
[12] D.R. Bellhouse and J.E. Stafford, *Local polynomial regression in complex surveys*, Survey Methodol. 27(2) (2001), pp. 197–203.
[13] F.J. Breidt and J.D. Opsomer, *Local polynomial regresssion estimators in survey sampling*, Ann. Stat. 28(4) (2000), pp. 1026–1053.
[14] T.D. Buskirk and S.L. Lohr, *Asymptotic properties of kernel density estimation with complex survey data*, J. Statist. Plann. Inference 128(1) (2005), pp. 160–190.
[15] T.D. Buskirk, *Nonparametric Density Estimation Using Complex Survey Data*, Proceedings of the Survey Research Methods Section, American Statistical Association, Washington, DC, 1998, pp. 799–801.
[16] J.D. Opsomer and C.P. Miller, *Selecting the amount of smoothing in nonparametric regression estimation for complex surveys*, J. Nonparametric Statist. 17 (2005), pp. 593–611.
[17] F.J. Breidt, G. Claeskens, and J.D. Opsomer, *Model-assisted estimation for complex surveys using penalised splines*, Biometrika 92(4) (2005), pp. 831–846.
[18] C. Goga, *Variance reduction in surveys with auxiliary information: A nonparametric approach involving regression splines*, Can. J. Statist./La Revue Canadienne de Statistique 33(2) (2005), pp. 163–180.
[19] T. Harms and P. Duchesne, *On kernel nonparametric regression designed for complex survey data*, Metrika 72(1) (2010), pp. 111–138.
[20] H.O. Hartley and R.L. Sielken Jr, *A 'super-population viewpoint' for finite population sampling*, Biometrics 31 (1975), pp. 411–422.
[21] D. Preffermann, *The role of sampling weights when modeling survey data*, Int. Statist. Rev. 61 (1993), pp. 317–337.
[22] B.I. Graubard and E.L. Korn, *Inference for superpopulation parameters using sample surveys*, Statist. Sci. 17(1) (2002), pp. 73–96.
[23] S. Orey, *A central limit theorem for m-dependent random variables*, Duke Math. J. 52 (1958), pp. 543–546.
[24] B. Jamison, S. Orey, and W. Pruitt, *Convergence of weighted averages of independent random variables*, Z. Wahr. Verb. Gebiete 4 (1965), pp. 40–44.
[25] W. Härdle, *Smoothing Techniques with Implementation in S*, Springer, New York, 1991.