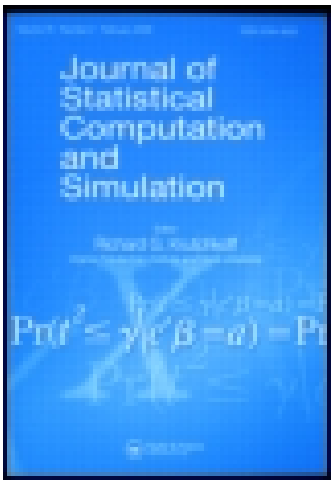

Journal of Statistical Computation and Simulation

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

Nonparametric regression estimators in complex surveys

Guoyi Zhang[a], Fletcher Christensen[b] & Wei Zheng[c]

[a] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-0001, USA

[b] Department of Statistics, University of California, Irvine, Irvine, CA 92697, USA

[c] Department of Mathematics, Indiana University-Purdue University Indianapolis, IN 46202, USA

Published online: 21 Nov 2013.

To cite this article: Guoyi Zhang, Fletcher Christensen & Wei Zheng (2013): Nonparametric regression estimators in complex surveys, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2013.860139

To link to this article: http://dx.doi.org/10.1080/00949655.2013.860139

Journal of Statistical Computation and Simulation, 2013
http://dx.doi.org/10.1080/00949655.2013.860139

Taylor & Francis
Taylor & Francis Group

Nonparametric regression estimators in complex surveys

Guoyi Zhang[a]*, Fletcher Christensen[b] and Wei Zheng[c]

[a]*Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-0001, USA;* [b]*Department of Statistics, University of California, Irvine, Irvine, CA 92697, USA;* [c]*Department of Mathematics, Indiana University-Purdue University Indianapolis, IN 46202, USA*

(*Received 1 August 2013; accepted 24 October 2013*)

In this article, we extend smoothing splines to model the regression mean structure when data are sampled through a complex survey. Smoothing splines are evaluated both with and without sample weights, and are compared with local linear estimator. Simulation studies find that nonparametric estimators perform better when sample weights are incorporated, rather than being treated as if iid. They also find that smoothing splines perform better than local linear estimator through completely data-driven bandwidth selection methods.

Keywords: complex surveys; local liner estimator; nonparametric regression; simulations; smoothing splines

1. Introduction

Consider the general nonparametric regression model

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

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 $t_i \in [0, 1], i = 1, 2, \ldots, n$ and for simplicity we assume that $0 < t_1 < \cdots < t_n < 1$.

A complex survey may include strata and clusters at the design stage, in which the iid assumption in Equation (1) is contradicted and standard nonparametric estimation methods cannot apply. For example, [1, p.171], suppose that we want to find out how many bicycles are owned by residents in a community of 10,000 households. We sample every household in each of 20 blocks/clusters selected at random from the 500 blocks in the community. Some blocks of the community may compose mainly of families (with more bicycles), whereas the residents of other blocks are mainly retirees (with fewer bicycles). Households selected in this example are not independent. In addition, ignoring the survey weights may lead to biased inferences or undesired outcome in the survey sampling practice. Classical nonparametric regression estimators and methods have been extended and investigated in survey area. Korn and Graubard [2] suggested nonparametric smoothing for estimating conditional means and percentile curves. Bellhouse and Stafford [3,4]

*Corresponding author. Email: gzhang123@gmail.com

developed estimators for density estimation and regression functions. Breidt and Opsomer [5] proposed local polynomial regression estimators for estimating population totals and proved that their estimator is asymptotically design unbiased and consistent. Buskirk and Lohr [6] presented finite-sample and asymptotic properties under several approaches for inference of a modified density estimator introduced by Buskirk [7] and Bellhouse and Stafford.[3] Opsomer and Miller [8] 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.[9] and Goga [10] proposed estimators of the population totals using smoothing splines. Harms and Duchesne [11] considered nonparametric regression and derived the asymptotic mean squared error (MSE) of the kernel estimators using a combined inference framework. Harms and Duchesne [11] first proposed a completely data-driven optimal bandwidth for use in local linear estimator in complex surveys.

In practice, we may want to discover a relationship between diastolic blood pressure as a function of age and gender from a survey data. For applications where prediction is the objective, such as imputing missing values, regression estimation provides a useful tool. Smoothing splines are important statistical tools for nonparametric regression function estimation. From a computational perspective, smoothing splines are the most efficient method. Standard smoothing spline methods have been well studied. However, there is no literature on smoothing splines in survey data. Our research is inspired by Harms and Duchesne.[11] We introduce smoothing splines to survey data in estimating the regression function by incorporating sampling weights. This article is organized as follows. In Section 2, we review the completely data-driven bandwidth selection method for local linear estimator in complex surveys suggested by Harms and Duchesne.[11] In Section 3, we extend smoothing spline estimator to complex surveys. In Section 4, we present simulation studies. In Section 5, we give an example of application. Finally, we summarize our research in Section 6.

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

The classical bandwidth in nonparametric regression relies on an estimator of the optimal bandwidth for iid data and is of the plug-in type. By modifying the bandwidth by a correction factor, which takes into account the sampling plan, Harms and Duchesne [11] proposed a bandwidth selector of the local linear estimator for use in complex surveys.

Let 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 $\pi_k = p(\text{unit } k \in S)$. 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 $r = n_S/N$.

The local linear kernel estimator incorporating sample weights has a simple explicit formula as shown in the following:

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

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

Let $\tilde{\mu}(t,h)$ be the classical local linear estimator which ignores the sample weights, Harms and Duchesne [11] showed that

$$\text{Bias}(\hat{\mu}(t,h)) = \text{Bias}(\tilde{\mu}(t,h)), \tag{3}$$

and

$$\mathrm{Var}(\hat{\mu}(t,h)) = (\triangle + r)\mathrm{Var}(\tilde{\mu}(t,h)), \quad \triangle = \frac{n_S}{N^2 \sum_U (d_k - 1)}. \tag{4}$$

By using Equations (3) and (4), Harms and Duchesne [11] derived the optimal bandwidth for $\hat{\mu}$ by minimizing the asymptotic MSE as the following:

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

where $\tilde{h}^{\mathrm{opt}}$ is the optimal bandwidth for $\tilde{\mu}(t,h)$, $(\triangle + r)^{1/5}$ is called the correction factor. The correction factor is a function related to the sampling plan and can be interpreted as a multiplicative factor taking into account the information concerning the survey design. Details can be found from Harms and Duchesne.[11]

3. Smoothing splines in complex surveys

Consider model (1), suppose $\mu(\cdot)$ is an unknown, smooth regression curve in the second-order Sobolev space $W_2^2[0,1]$. That is, μ and μ' are absolutely continuous and μ'' is a Lebesgue integrable function. A natural cubic spline is a smooth piecewise cubic polynomial under certain boundary constraints. The space of natural splines corresponding to $t_1, \ldots, t_n$ is a linear space of dimension n. The natural cubic smoothing spline can be obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 (f''(t))^2 \, \mathrm{d}t \tag{6}$$

over $f \in W_2^2[0,1]$, where $\lambda > 0$ is the smoothing parameter which controls the tradeoff between smoothness and goodness-of-fit. The first term in Equation (6) is the residual sum of squares which is a standard measure of goodness-of-fit to the data. The second term in Equation (6) is a natural measure of curvature of the function. The smoothing spline estimator is as follows:

$$\hat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1), \ldots, \hat{\mu}_\lambda(t_n))^{\mathrm{T}} = \mathbf{S}_\lambda \mathbf{y}, \tag{7}$$

where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + n\lambda\Omega)^{-1}\mathbf{X}^{\mathrm{T}}$, $\mathbf{X} = \{x_j(t_i)\}_{i,j=1,\ldots,n}$, $\Omega = \{\int_0^1 x_i(t)x_j(t)\,\mathrm{d}t\}_{i,j=1,\ldots,n}$, $\mathbf{y} = (y_1, \ldots, y_n)$ and $x_1, x_2, \ldots, x_n$ is a basis for the set of natural cubic splines with knots at $t_1, \ldots, t_n$ (details can be found in Eubank [12]).

3.1. *Generalization of smoothing splines to survey data*

The original generalization of smoothing splines is motivated by the case that each observation has different variance. A generalized smoothing criterion is to minimize

$$n^{-1}\sum_{i=1}^{n} w_i(y_i - f(t_i))^2 + \lambda \int_0^1 (f''(t))^2 \, \mathrm{d}t, \tag{8}$$

where $w_i = [\mathrm{var}(y_i)]^{-1}$ for $i = 1, \ldots, n$. Hence, the smoothing spline estimator is

$$\hat{\boldsymbol{\mu}}_\lambda^{(2)} = (\hat{\mu}_\lambda^{(2)}(t_1), \ldots, \hat{\mu}_\lambda^{(2)}(t_n))^{\mathrm{T}} = \mathbf{S}_\lambda^{(2)}\mathbf{y}, \tag{9}$$

where $\mathbf{S}_\lambda^{(2)} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + n\lambda\Omega)^{-1}\mathbf{X}^{\mathrm{T}}$, and $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_n)$.

Consider a more general case, where there are $n_i \geq 1$ responses at the design point t_i. We derive the following generalized theorem.

THEOREM 1 *Let $x_1, x_2, \ldots, x_n$, be a basis for the set of natural splines of order 2m with knots at $t_1, \ldots, t_n$ and define $\mathbf{X} = \{x_j(t_i)\}_{i,j=1,\ldots,n}$. Also let $w_i^* = [\mathrm{var}(y_i)/n_i]^{-1}$ and $n^* = \sum n_i$. The minimizing criterion is as follows:*

$$\frac{1}{n^*}\sum_{i=1}^{n} w_i^*(\bar{y}_i - f(t_i))^2 + \lambda \int_0^1 [f^m(t)]^2 \, \mathrm{d}t. \tag{10}$$

The unique minimizer of Equation (10) over $f \in W_2^m[0,1]$ (the mth order Sobolev space) is

$$\hat{\boldsymbol{\mu}}_\lambda^{(3)} = \sum_{j=1}^{n} b_{\lambda j} x_j, \tag{11}$$

where $\mathbf{b}_\lambda = (b_{\lambda 1}, \ldots, b_{\lambda n})^{\mathrm{T}}$ is the solution with respect to $\mathbf{c}$ of the equation system

$$(\mathbf{X}^{\mathrm{T}}\mathbf{W}^*\mathbf{X} + n^*\lambda\boldsymbol{\Omega})\mathbf{c} = \mathbf{X}^{\mathrm{T}}\bar{\mathbf{y}}, \tag{12}$$

with $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_n)^{\mathrm{T}}$. The smoothing spline estimators in matrix form are as follows:

$$\hat{\boldsymbol{\mu}}_\lambda^{(3)} = \mathbf{S}_\lambda^{(3)}\bar{\mathbf{y}}, \tag{13}$$

where $\mathbf{S}_\lambda{}^{(3)} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}^*\mathbf{X} + n^*\lambda\Omega)^{-1}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{W}^* = \mathrm{diag}(w_1^*, w_2^*, \ldots, w_n^*)$.

Proof follows from Wahba and Wendelberger [13] immediately. Theorem 1 provides a key to determining how to handle survey data. We describe it as follows,

COROLLARY *Suppose that sampling weight of unit i is d_i and $var(y_i) = \sigma^2$. $\hat{N}$ is the estimated total number of observations as defined in Section 2. To get a cubic smoothing spline for a survey data, the minimizing criterion (10) is equivalent to minimizing*

$$\frac{1}{\hat{N}}\sum_{i=1}^{n} d_i(y_i - f(t_i))^2 + \lambda \int_0^1 [f''(t)]^2 \, \mathrm{d}t. \tag{14}$$

In this case, we can consider that there are d_i responses at design point t_i. The unique minimizer of Equation (14) over $f \in W_2^2[0,1]$ has the same form as Equation (11) with coefficients $\mathbf{b}_\lambda = (b_{\lambda 1}, \ldots, b_{\lambda n})^{\mathrm{T}}$ the solution with respect to $\mathbf{c}$ of the equation system

$$(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X} + \hat{N}\lambda\boldsymbol{\Omega})\mathbf{c} = \mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{15}$$

The smoothing spline estimators in matrix form are as follows:

$$\hat{\boldsymbol{\mu}}_\lambda^{(4)} = \mathbf{S}_\lambda^{(4)}\mathbf{y}, \tag{16}$$

where $\mathbf{S}_\lambda{}^{(4)} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X} + \hat{N}\lambda\Omega)^{-1}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$.

3.2. *Selection of* λ

The smoothing parameter λ controls the tradeoff between smoothness and goodness-of-fit. We want to choose λ in such a way that it can balance the bias and the variance. The cross validation (CV) method and generalized cross validation (GCV) method are two frequently used technique for smoothing parameter selection. CV(λ) is generally biased for prediction risk.

GCV was first proposed by Craven and Wahba [14] for use in the context of nonparametric regression. In the 1980s there were numerous theoretical and practical studies which demonstrated that GCV had a variety of statistical applications.[15] GCV is nearly an unbiased estimator of prediction risk. The GCV criterion is defined as

$$\mathrm{GCV}(\lambda) = \frac{(1/n)\sum_{i=1}^{n}(y_i - \mu_\lambda(t_i))^2}{((1/n)\mathrm{tr}(I - S_\lambda))^2}, \tag{17}$$

where tr(·) denotes the trace of the matrix and λ is chosen to minimize Equation (17). Zhang [16] showed that the GCV criterion is more likely to derive the estimate of λ with smaller variance for smoothing splines.

4. Simulation studies

In this section, a small simulation study has been conducted to investigate finite sample properties of the nonparametric regression estimators in complex surveys. We first consider the comparisons between the nonparametric estimators (local linear estimator and smoothing spline estimator) by incorporating sample weights or not. Another comparison is between local linear estimator and smoothing spline estimator for complex surveys. For simplicity, we did the three simulation studies separately. The two estimators are compared under completely data-driven methods: Harms and Duchesne [11]'s methods for local linear estimator and GCV for smoothing spline estimator. For these purposes, simulated data using different functions, different error variances and different sampling rates are considered. The simulation set-up is similar as in Harms and Duchesne.[11]

The following equation is used to generate the population at the supermodel stage:

$$y_i = \mu_k(t_i) + \epsilon_i \quad i = 1, \ldots, 1000 \quad \text{and} \quad k = 1, 2, 3, 4, \tag{18}$$

where each population has $N = 1000$ values of t_i which is equally spaced in the interval $[0, 1]$ and random errors are from a 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) σ^2 : 0.4 and 1; (2) sampling rate: 10% and 20%; (3) sampling plan: simple random sampling (SRS) and Poisson sampling scheme (unequal-probability design). In the SRS sampling plan, all elements of the population were assigned equal inclusion probabilities and sampled with replacement until $n_S = 200$ units were obtained. 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 $\sum_U 1/w_i = E(n_S) = N * r$; (4) Four functions are used to generate populations at the supermodel stage:

Härdle: $\mu_1(t) = \sin^3(2\pi t^3)$ Härdle,[17]
Bump : $\mu_2(t) = 1 + 2(t - 0.5) + \exp(-200(t - 0.5)^2)$ Breidt and Opsomer,[5]
Exponential : $\mu_3(t) = \exp(-8t)$ Breidt and Opsomer,[5]
Slowsine : $\mu_4(t) = 2 + \sin(2\pi t)$ Opsomer and Miller.[8]

Simulation done $L = 500$ 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.

Our primary goal is to evaluate the estimators with respect to 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, $i = 1, \ldots, L$. The Monte Carlo mean E_{MC}, the Monte Carlo bias B_{MC}, Monte Carlo variance V_{MC}, and the Monte Carlo MSE are given by the following formulas:

$$E_{\mathrm{MC}}\{\hat{\mu}(t)\} = L^{-1}\sum_{i=1}^{L}\hat{\mu}^{(i)}(t), \tag{19}$$

$$B_{\mathrm{MC}}\{\hat{\mu}(t)\} = E_{\mathrm{MC}}\{\hat{\mu}(t)\} - \mu(t), \tag{20}$$

$$V_{\mathrm{MC}}\{\hat{\mu}(t)\} = L^{-1}\sum_{m=1}^{L}[\hat{\mu}^{(i)}(t) - E_{\mathrm{MC}}\{\hat{\mu}(t)\}]^2, \tag{21}$$

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

$$\mathrm{MSE}_{\mathrm{MC}}\{\hat{\mu}(t)\} = L^{-1}\sum_{i=1}^{L}\{\hat{\mu}^{(i)}(t) - \mu(t)\}^2. \tag{22}$$

For each sample, the estimators have been evaluated at 200 equally spaced values, which is at each point $t_j = j/201, j = 1, \ldots, 200$. For each point t_j the Monte Carlo bias, variance and MSE have been calculated using the formulas (20)–(22), respectively. We summarize our findings by averaging over the bias, variance and MSE of the 200 values. Under SRS, inclusion probabilities and thus sample weights are equal for all units in the population. The estimators by incorporating sample weights or not are the same. Therefore, we only report the results under Poisson sampling.

Tables 1 and 2 give the simulation results of local linear estimator and smoothing spline estimator with sample weights and without sample weights, respectively. Tables 1 and 2 show that both local linear estimator and smoothing splines with sample weights are effective in reducing bias. The overall measurement of effectiveness MSE from both estimators is smaller than MSE of the estimators without incorporating sample weights. Tables 1 and 2 also reveal interesting findings. In general, the estimators incorporating weights perform better than the estimators without

Table 1. Simulation results under Poisson sampling scheme using local linear estimator, 'wo' means without weight information, 'w' means sample weights are incorporated into estimator.

			Bias		Variance		MSE	
Function		Sampling rate (%)	wo	w	wo	w	wo	w
Härdle	$\sigma = 0.4$	10	0.0818	0.0372	0.2344	0.2326	0.2464	0.2388
		20	0.0991	0.0347	0.0461	0.0470	0.0630	0.0544
	$\sigma = 1$	10	0.4424	0.1571	0.3195	0.3508	0.5367	0.3977
		20	0.4698	0.1438	0.0470	0.0889	0.2893	0.1306
Bump	$\sigma = 0.4$	10	0.0383	0.0068	0.2939	0.2851	0.2996	0.2894
		20	0.0509	0.0059	0.0454	0.0465	0.0498	0.0481
	$\sigma = 1$	10	0.3674	0.1358	0.4716	0.4817	0.6171	0.5105
		20	0.3783	0.1150	0.3291	0.3644	0.4903	0.3906
Exponential	$\sigma = 0.4$	10	0.0530	−0.0032	0.0230	0.0245	0.0269	0.0255
		20	0.0880	0.0189	0.0059	0.0070	0.0155	0.0095
	$\sigma = 1$	10	0.4619	0.1223	0.1082	0.1562	0.3229	0.1729
		20	0.4797	0.1258	0.0279	0.0580	0.2609	0.0770
Slow sine	$\sigma = 0.4$	10	0.0457	0.0146	0.0714	0.0705	0.0777	0.0746
		20	0.0556	0.0206	0.0099	0.0102	0.0137	0.0111
	$\sigma = 1$	10	0.2437	0.0374	0.2687	0.2945	0.3456	0.3064
		20	0.2155	−0.0275	0.0420	0.0554	0.0927	0.0621

Table 2. Simulation results under Poisson sampling scheme using smoothing splines, 'wo' means without weight information, 'w' means sample weights are incorporated into estimator.

			Bias		Variance		MSE	
Function		Sampling rate (%)	wo	w	wo	w	wo	w
Härdle	$\sigma = 0.4$	10	0.1065	0.0348	0.0200	0.0269	0.0404	0.0345
		20	0.0709	−0.0046	0.0089	0.0145	0.0197	0.0177
	$\sigma = 1$	10	0.4767	0.1551	0.4897	0.5858	0.7533	0.6339
		20	0.4960	0.1697	0.0267	0.0986	0.2960	0.1397
Bump	$\sigma = 0.4$	10	0.0068	0.0511	0.0272	0.0184	0.0300	0.0244
		20	−0.0054	0.0458	0.0120	0.0095	0.0143	0.0139
	$\sigma = 1$	10	0.3221	0.0773	0.0590	0.1443	0.1912	0.1646
		20	0.3621	0.0865	0.0308	0.0831	0.1841	0.1026
Exponential	$\sigma = 0.4$	10	0.0780	0.0065	0.0086	0.0107	0.0171	0.0128
		20	0.1076	0.0270	0.0040	0.0061	0.0170	0.0081
	$\sigma = 1$	10	0.4278	0.1194	0.0291	0.1198	0.2149	0.1368
		20	0.3926	0.0465	0.0138	0.0694	0.1760	0.0774
Slow sine	$\sigma = 0.4$	10	0.0373	−0.0026	0.0099	0.0115	0.0137	0.0135
		20	0.0377	−0.0024	0.0051	0.0060	0.0071	0.0066
	$\sigma = 1$	10	0.2789	0.0565	0.0544	0.0843	0.1541	0.1014
		20	0.2368	−0.0130	0.0216	0.0448	0.1207	0.0707

Table 3. Comparison between local linear estimator and smoothing splines with sample weights under Poisson sampling scheme, 'LLE' means local linear estimator with sample weights incorporated and 'SS' means smoothing splines with sample weights incorporated.

			Bias		Variance		MSE	
Function		Sampling rate (%)	SS	LLE	SS	LLE	SS	LLE
Härdle	$\sigma = 0.4$	10	0.0198	0.0315	0.0261	0.2178	0.0318	0.2229
		20	0.0123	0.0201	0.0145	0.0270	0.0192	0.0313
	$\sigma = 1$	10	0.2194	0.2453	0.1521	0.3993	0.2292	0.4865
		20	0.1386	0.1453	0.0975	0.09821	0.1379	0.1393
Bump	$\sigma = 0.4$	10	0.0209	0.0208	0.02765	0.2571	0.0281	0.2575
		20	0.0196	0.0254	0.0127	0.0758	0.0166	0.0798
	$\sigma = 1$	10	0.0427	0.0700	0.1716	1.0363	0.1872	1.0607
		20	0.0366	0.0464	0.1016	0.1430	0.1196	0.1610
Exponential	$\sigma = 0.4$	10	−0.0030	0.0077	0.0123	0.0294	0.0143	0.0319
		20	−0.0025	0.0029	0.0053	0.0074	0.0069	0.0085
	$\sigma = 1$	10	0.1742	0.1828	0.1479	0.1648	0.1811	0.2010
		20	0.0425	0.1266	0.0624	0.0602	0.0642	0.0762
Slow sine	$\sigma = 0.4$	10	0.0006	0.0079	0.0110	0.0464	0.0123	0.0482
		20	−0.0160	−0.0124	0.0048	0.0099	0.0067	0.0116
	$\sigma = 1$	10	0.0961	0.1225	0.0683	0.2046	0.0916	0.2283
		20	0.0104	0.0195	0.0483	0.0594	0.0551	0.0703

incorporating weights, especially under the situation when the variance σ^2 is large. For example, in Table 1, with Härdle function, $\sigma = 0.4$, and sampling rate 20%, MSE without incorporating weights is 0.0630 compared with the one with weights 0.0544; while under $\sigma = 1$ with sampling rate 20%, MSEs are 0.2893 compared with 0.1306. The difference in MSE incorporating wights or not increases when variance increases. The increase in variance reflects the spread out of data. Under relatively large variance, weights become more important in estimating the parameters. Table 3 reports the comparison between performance of local linear estimator and smoothing splines under the Poisson sampling scheme for different settings. We notice from Table 3, for function Härdle, $\sigma = 0.4$ and sampling rate 10%, MSE of smoothing spline estimator is much

smaller than that of local linear estimator, which is 0.0318 compared with 0.2229; for sampling rate 20%, MSEs are closer, which is 0.0192 and 0.0313, respectively. This pattern has been found through all the four functions under different settings. Smoothing splines work better than local linear estimator, particularly when sample size is small or sampling rate is small. This is because Harms and Duchesne [11] derived the optimal bandwidth for $\hat{\mu}$ by minimizing the asymptotic MSE. Therefore, when sample size is small, Harms and Duchesne [11]'s method does not work very well. The proposed smoothing splines use GCV to select the smoothing parameter λ, which is suitable even for small sample sizes. In summary, Table 3 shows that by incorporating sample weights, smoothing spline estimator works better than the local linear estimator in reducing bias, variance and MSE, especially for small sample size, therefore suggested for nonparametric regression estimation in complex surveys.

5. Application

In this section, we use an example to illustrate the use of our proposed smoothing splines in complex surveys. At the end of the nineteenth century, it was widely thought that criminal tendencies might be expressed in physical characteristics that were distinguishable from the physical characteristics of noncriminal classes. Macdonell [18]'s data on criminals gave the length (cm) of the left middle finger and height (inches) for 3000 criminals. Suppose an unequal-probability sample [1, p.423] of 200 men is taken from the 3000 criminals and that the selection probabilities are higher for the shorter men (shorter men have smaller weights). Figure 1 shows a scatterplot of the data from this unequal-probability sample, along with the weighted least-square regression line and smoothing spline curve by Equation (16). Since taller men have smaller inclusion probabilities or larger weights, the slope of the regression line incorporating weights is drawn upward. On the other hand, the smoothing spline describes the regression relationship by a smooth curve. The choice between a parametric regression line and a nonparametric smoothing curve is quite subjective. The smoothing splines give estimates of μ that allow great flexibility in the possible form of the regression curve and, in particular, make no assumptions about a parametric form.

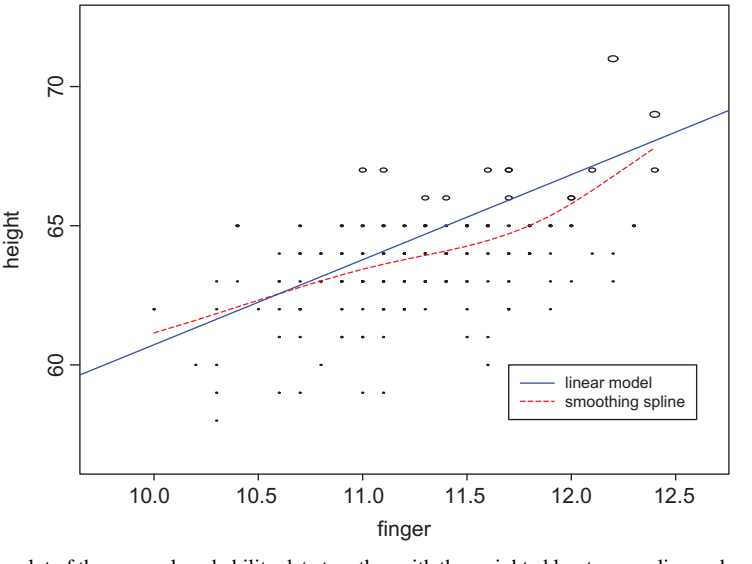

Figure 1. Scatterplot of the unequal-probability data together with the weighted least square line and smoothing spline curve.

While parametric regression modelling is the most prevalent approach to regression analysis when appropriate. In our example, both regression line and smoothing curve look good.

6. Conclusions

In this article, we extend smoothing splines to model regression structure in complex surveys. Both local linear estimator and smoothing splines are studied under the unequal-probability sampling scheme. Simulation studies show that the nonparametric estimators perform better when sample weights are incorporated. Simulation studies also show that by incorporating sample weights, smoothing splines perform significantly better than local linear estimator under completely data-driven methods, and therefore is suggested for use in complex surveys.

References

[1] Lohr S. Sampling: design and analysis. 2nd ed. Belmone, CA: Cengage Learning; 2009.
[2] Korn EL, Graubard BI. Scatterplots with survey data. Amer Statist. 1998;52:58–69.
[3] Bellhouse DR, Stafford JE. Density estimation from complex surveys. Statist Sinica. 1999;9:407–424.
[4] Bellhouse DR, Stafford JE. Local polynomial regression in complex surveys. Surv Methodol. 2001;27(2):197–203.
[5] Breidt FJ, Opsomer JD. Local polynomial regression estimators in survey sampling. Ann Statist. 2000;28(4): 1026–1053.
[6] Buskirk TD, Lohr SL. Asymptotic properties of kernel density estimation with complex survey data. J Statist Plann Inference. 2005;128(1):160–190.
[7] Buskirk TD. Nonparametric density estimation using complex survey data. Proceedings of the Survey Research Methods Section, American Statistical Association; 1998. p. 799–801.
[8] Opsomer JD, Miller CP. Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. J Nonparametr Stat. 2005;17:593–611.
[9] Breidt FJ, Claeskens G, Opsomer JD. Model-assisted estimation for complex surveys using penalised splines. Biometrika. 2005;92(4):831–846.
[10] Goga C. Variance reduction in surveys with auxiliary information: a nonparametric approach involving regression splines. Canad J Statist. 2005;33(2):163–180.
[11] Harms T, Duchesne P. On kernel nonparametric regression designed for complex survey data. Metrika. 2010;72(1):111–138.
[12] Eubank RL. Nonparametric regression and spline smoothing. New York, NY: Marcel Dekker Inc.; 1999. ISBN 0-8247-9337-4.
[13] Wahba G, Wendelberger J. Some new mathematical methods for variational objective analysis using splines and cross validation. Monthly Weather Rev. 1980;108:1122–1143.
[14] Craven P, Wahba G. Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of cross-validation. Numer Math. 1979;3:377–403.
[15] Wahba G. Spline models for observational data. Philadelphia, PA: SIAM; 1990.
[16] Zhang G. Smoothing splines using compactly supported, positive definite, radial basis functions. ProQuest. Ann Arbor: UMI Dissertation Publishing; 2011.
[17] Härdle W. Smoothing techniques with implementation in S. New York: Springer; 1991.
[18] Macdonell WR. On criminal anthropometry and the identification of criminals. Biometrika. 1901;1:177–227.