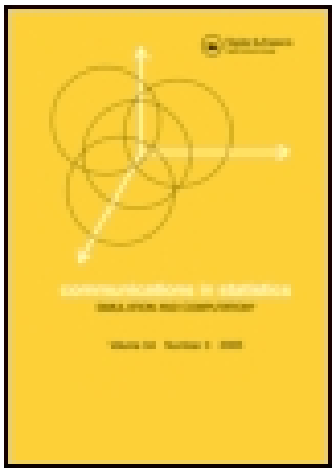

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Adjusted Confidence Bands for Complex Survey Data

Guoyi Zhang Assistant Professor[a], Maozhen Gong Graduate Student[b] & Yang Cheng Lead Scientist[c]

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

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

[c] U.S. Census Bureau, Suitland, MD 20746

Accepted author version posted online: 20 Jun 2014.

To cite this article: Guoyi Zhang Assistant Professor, Maozhen Gong Graduate Student & Yang Cheng Lead Scientist (2014): Adjusted Confidence Bands for Complex Survey Data, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2014.882946

To link to this article: http://dx.doi.org/10.1080/03610918.2014.882946

Adjusted Confidence Bands for Complex Survey Data

Guoyi Zhang[*], Maozhen Gong [†] and Yang Cheng [‡]

Abstract

Confidence bands in nonparametric regression have been studied for a long time with data assumed to be generated from independent and identically distributed (iid) random variables. The methods and theoretical results for iid data, however, do not directly apply to data from stratified multistage samples. In this paper, we extend the confidence bands introduced by Zhang and Lu (2008) for iid case to complex surveys based on an entirely data-driven procedure; the proposed confidence bands incorporate both the sampling weights and the kernel weights. Simulation studies show that the proposed method works well.

Key Words: Complex surveys, Confidence bands, Local linear estimator, Nonparametric regression, Simulations.

1 Introduction

A confidence band enables us to estimate the region in which the true function lies. It can also be used to determine the appropriateness of a fitted regression function. 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. Lohr (2010, page423) considered an unequal-probability sample (shorter men have smaller weights and taller men have larger weights) of 200 men taken from Macdonell (1901)'s data on length (cm)

[*]Guoyi Zhang, Assistant Professor, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001, gzhang123@gmail.com

[†]Maozhen Gong, Graduate Student, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, 87131-0001

[‡]Yang Cheng, Lead Scientist, U.S. Census Bureau, Suitland, MD 20746

of the left middle finger and height (inches) for 3000 criminals. To model the relationship between length of the left middle finger and height, the choice between a parametric regression line and a nonparametric curve is quite subjective. This research intends to provide a tool to determine the appropriateness of a fitted regression function. A lack of fit test is a possible application of confidence bands. For example, we wish to test the parametric linear null hypothesis of the form: $H_0 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ against a nonparametric alternative. If the regression function under the null hypothesis is not entirely contained in the confidence band, we can reject the null hypothesis.

Consider a general nonparametric regression model

$$y_i = m(x_i) + \epsilon_i, \ i = 1, 2, \cdots n, \tag{1}$$

where $m(\cdot)$ is an unknown function and ϵ_i are zero mean random errors with common variance σ^2. Without loss of generality, we assume that $x_i \in [0, 1], i = 1, 2, \ldots, n$. Given $\alpha \in (0, 1)$, to construct a confidence band, we need an estimator $\hat{m}(x)$ for $m(x)$ and a bound l_α, such that

$$P\{ |\hat{m}(x) - m(x)| \leq l_\alpha \quad \text{for all x}\} \geq 1 - \alpha. \tag{2}$$

A 100(1-α)% confidence band can be constructed as $\hat{m}(x) \pm l_\alpha$, in which the bias of $\hat{m}$ is considered to be negligible relative to its standard error. A bias-corrected confidence band in literature takes the form of $\hat{m}(x) - \widehat{\text{bias}} \pm l_\alpha$, which involves estimating the second derivative of the regression function. Zhang and Lu (2008) suggest a confidence band that takes the form of $\hat{m}(x) \pm c * l_\alpha$ and use a logistic regression model to estimate the coefficient c. The idea of incorporating logistic regression is to set $c_i (i = 1, 2, ...1000)$ as equally spaced points between some specific interval, say $(0, 1.5)$, and set h_i as a binary response variable with 1 indicating that the constructed confidence band from simulation contains the estimate of the regression function and 0 otherwise. Zhang and Lu (2008)'s method is efficient, simple to use and can be applied to fixed equal spaced design and randomly designed x_is.

This research extends the confidence bands suggested by Zhang and Lu (2008) for iid case to complex surveys. A complex survey may include strata and clusters at the design stage, in which

the general iid assumption in (1) is contradicted and standard nonparametric estimation methods do not apply. In addition, ignoring the survey weights may lead to biased inferences, or undesired outcome in the survey sampling practice. In the following, we review the nonparametric regression estimators and difference-based variance estimators in complex surveys.

Classical nonparametric regression estimators and methods have been extended and investigated in survey area. Korn and Graubard (1998) suggested nonparametric smoothing for estimating conditional means and percentile curves. Bellhouse and Stafford (1999, 2001) developed estimators for density estimation and regression functions. Breidt and Opsomer (2000) proposed local polynomial regression estimators for estimating population totals and proved that their estimator is asymptotically design unbiased and consistent. Buskirk and Lohr (2005) presented finite-sample and asymptotic properties under several approaches for inference of a modified density estimator introduced by Buskirk (1998) and Bellhouse and Stafford (1999). Harms and Duchesne (2010) derived the asymptotic mean squared error of the kernel estimators using a combined inference framework. They first proposed a completely data driven optimal bandwidth for use in local linear estimator for complex surveys.

If $m(\cdot)$ in (1) is smooth and the x 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. Gasser, Sroka, and Jennen-Steinmetz (1986) employed Rice (1984) suggestion of a pseudo-residual estimator in the case of nonparametric regression and showed that the 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 (1987) for estimating heteroscedasticity in regression analysis. Hall, Kay, and Titterington (1990) suggested and computed asymptotically optimal difference sequence for estimating error variance in homoscedastic nonparametric regression. Buckley, Eagleson, and Silverman (1988) 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, Kambour, Kim, Klipple, and Reese (1998), and Klipple and Eubank (2007) extended work from Gasser et al. (1986) to partially linear models. Lu (2014) extends the variance estimator from Gasser et al. (1986) by incorporating survey weights to nonparametric regression in complex surveys and derived the asymptotic properties.

This paper is organized as follows. In Section 2, we review the local linear estimator by Harms and Duchesne (2010) and difference based estimator by Lu (2014). In Section 3, we propose the adjusted confidence bands for nonparametric regression in complex surveys. In Section 4, we perform simulation studies. Section 5 gives the conclusion.

2 Background

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

The classical bandwidth of 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 that takes into account the sampling plan, Harms and Duchesne (2010) 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_i be the first order inclusion probability with $\pi_i = p(\text{unit} \quad i \in S)$. Sample weight d_i is the reciprocal of the inclusion probability π_i, i.e. $d_i = 1/\pi_i$ for $i \in S$. Let $\hat{N}$ be an estimate of population size N, i.e. $\hat{N} = \sum_{i=1}^{n_S} d_i$ 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

the following

$$\hat{m}(x,h) = \frac{\sum_S \{\hat{s}_2(x,h) - \hat{s}_1(x,h)(x_k - x)\} d_k y_k K(\frac{x_k - x}{h})/h}{\hat{s}_2(x,h)\hat{s}_0(x,h) - \hat{s}_1^2(x,h)}, \tag{3}$$

where $\hat{s}_i(x,h) = \sum_S d_k (x_k - x)^i K(\frac{x_k - x}{h})/h$, $i = 0, 1,$ and 2, and $K(\cdot)$ is the kernel function.

Let $\tilde{m}(x,h)$ be the classical local linear estimator ignoring sample weights. Harms and Duchesne (2010) showed that

$$\text{Bias}\left[\hat{m}(x,h)\right] = \text{Bias}\left[\tilde{m}(x,h)\right], \tag{4}$$

and

$$\text{Var}\left[\hat{m}(x,h)\right] = (\triangle + r)\text{Var}\left[\tilde{m}(x,h)\right], \quad \triangle = n_S/N^2 \sum_U (d_k - 1), \tag{5}$$

where subscript U denotes summing over the population elements.

By using (4) and (5), Harms and Duchesne (2010) derived the optimal bandwidth for $\hat{m}$ by minimizing the asymptotic MSE as the following

$$\hat{h}^{\text{opt}}(t) = (\triangle + r)^{1/5} \tilde{h}^{\text{opt}}, \tag{6}$$

where $\tilde{h}^{\text{opt}}$ is the optimal bandwidth for $\tilde{m}(x,h)$, $(\triangle + r)^{1/5}$ is called the correction factor. The correction factor is a function that can be interpreted as a multiplicative factor taking into account the information concerning the survey design. Details can be found from Harms and Duchesne (2010).

2.2 Difference Based Variance Estimator for Nonparametric Regression in Complex Survey

The goal of difference based estimator is to estimate the random error in nonparametric regression based on a sample S drawn according to a complex sampling plan without estimating the unknown regression function $m(\cdot)$. Recall that $\pi_i = P(i \in S)$ is the first order inclusion probability, and $1/\pi_i$

represents the sampling weight. Lu (2014) extended the estimator from Gasser et al. (1986) to complex surveys as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n-2} \frac{1}{\pi_i} \tilde{\epsilon}_i^2}{\sum_{i=1}^{n-2} \frac{1}{\pi_i}}, \tag{7}$$

or in a matrix form as

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T \mathbf{D}^T \mathbf{W} \mathbf{D} \mathbf{y}}{\text{tr}(\mathbf{D}^T \mathbf{W} \mathbf{D})}, \tag{8}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$, $\mathbf{W}$ is a diagonal matrix with ith diagonal element $1/\pi_i$, tr is the trace function for a square matrix, $\tilde{\epsilon}_i$ are called pseudo-residuals defined by

$$\tilde{\epsilon}_i = d_{i0} y_i + d_{i1} y_{i+1} + d_{i2} y_{i+2}, \tag{9}$$

with

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

for

$$a_i = \frac{x_{i+2} - x_{i+1}}{x_{i+2} - x_i} \text{ and } b_i = \frac{x_{i+1} - x_i}{x_{i+2} - x_i},$$

and 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. Given some certain conditions, Lu (2014) showed that

$$\theta^{-1}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} \text{N}(0, \sigma^4), \tag{10}$$

where $\theta = \{\frac{2\text{tr}((\mathbf{D}^T \mathbf{W} \mathbf{D})^2)}{(\text{tr}(\mathbf{D}^T \mathbf{W} \mathbf{D}))^2} + \frac{(m_4 - 3) \sum_{i=1}^{n} s_i^2}{(\text{tr}(\mathbf{D}^T \mathbf{W} \mathbf{D}))^2}\}^{1/2}$, s_i is the ith diagonal element of matrix $\mathbf{D}^T \mathbf{W} \mathbf{D}$ and $E(\epsilon^4) = m_4 \sigma^4$. Lu (2014) also derived the formulas for bias and variance of the estimator $\hat{\sigma}^2$.

3 Adjusted Confidence Bands in Complex Surveys

In classical nonparametric regression, the confidence bands proposed by Bickel and Rosenblatt (1973, 1975), Eubank and Speckman (1993) and Xia (1998) are commonly used. Eubank and

Speckman (1993) and Xia (1998) estimated the bias and recentered the confidence bands. However, one disadvantage of the above confidence bands is the slow convergence rate. In practice, we have finite samples, such as $n = 200$ (this is commonly considered as a large sample size in regression problems). Since the convergence rate is very slow, this sample size is not large enough to approximate the population quantities by sample estimators. Besides recentering the bands by correction of bias, another approach is to expand the bands to account for the bias. The confidence bands therefore take the form of $\hat{m}(x) \pm c * l_\alpha(x)$ (Zhang & Lu, 2008).

In this section, we extend the confidence bands suggested by Zhang and Lu (2008) for iid case to complex surveys. The extended confidence bands are in similar forms as those with iid data but with some modifications. We describe them as follows

$$\hat{m}(x) \pm c * l_\alpha(x),$$

with

$$l_\alpha(x) = \frac{\hat{\sigma}(x)V}{(\sum_{i\in S} w_i(x))^{1/4}} \left\{ \sqrt{-2\log(\hat{h})} + \frac{1}{\sqrt{-2\log(\hat{h})}}(A - X_\alpha) \right\}, \tag{11}$$

$$A = \log\left\{ \frac{1}{2\pi} \left(\int K'(u)^2 du / \int K(u)^2 du \right)^{1/2} \right\},$$

$$X_\alpha = \log\left\{ \frac{-\log(1-\alpha)}{2} \right\},$$

$$w_i(x) = K\left(\frac{x_i - x}{h}\right)\left(s_2 - \frac{x_i - x}{h} s_1\right),$$

$$s_l = \sum_{i\in S} d_i K\left(\frac{x_i - x}{h}\right)\left(\frac{x_i - x}{h}\right)^l \quad l = 1, 2,$$

$$\hat{m}(x) = \sum_{i\in S} w_i(x) y_i / \sum_{i\in S} w_i(x),$$

$$V = \sqrt{\int K(u)^2 du},$$

where $K(\cdot)$ is a kernel function, $\hat{h}$ is the selected bandwidth, $\hat{\sigma}(x)$ is the consistent estimator of $\sigma(x)$

using (7), $w_i(x)$ is a weight function that combines the kernel weights and sampling weights, and

$$c = \inf_{b \in R} \{P(|\hat{m}(x) - m(x)| \leq b * l_\alpha \quad \text{for all x }) \geq 1 - \alpha\}. \tag{12}$$

The adjustment constant c has the following property:

Proposition 1. *Assume that the sampling rate $r_j = n_{S,j}/N_j$ converges with probability one to a finite constant γ, as $j \to \infty$. The adjustment value c is a monotone decreasing function of sample size n_S and the limit of the adjustment value is 1.*

Proof. From (11), the bound l_α has two parts. The first part, $\hat{\sigma}(x)\hat{V}/(\sum_{i \in S} w_i(x))^{1/4}$, is the estimated standard deviation of $\hat{m}$. The second part, $\left\{ \sqrt{-2\log(\hat{h})} + \frac{1}{\sqrt{-2\log(\hat{h})}}(A - X_\alpha) \right\}$, is a decreasing function of the bandwidth. As the sample size increases, the selected bandwidth decreases. The bias and the first part of the bound decay at the same rate: namely, $n_S^{-\frac{2}{5}}$. The second part of the bound increases as the sample size increases. Hence, the bias decays faster than the bound. We only need a smaller c to compensate for the bias when we use a bigger sample size. In other words, the adjustment value c is a monotone decreasing function of the sample size n_S. It is also clear from (11) that $l_{\alpha_1}/l_{\alpha_2} \to 1$ as $n_S \to \infty$ for any fixed $0 < \alpha_1, \alpha_2 < 1$. The effect of α on the width of the band l_α is of second order. Consequently, using the definition of (12), the limit of the adjustment value is 1. □

The remaining problem is to find an appropriate estimator of c. First we did pilot simulation to look for some properties of the adjustment value c. The lower bound of c is obviously 1. We did simulation for the fixed equal space design using function $m(x) = 2 + sin(2 * \pi * x_i)$ and tried different possible upper bound values. The possible upper bounds of c are summarized in Table (1). We see from Table (1) that the coverage is close and greater than the nominal level, which suggests the proposed c values are the possible upper bounds. We also notice that c decreases as the sample size increases.

Given true regression function $m(x)$, sample size n and σ, using the definition of (12), we can find c. In practice, the true regression function $m(x)$ is generally not available. We define c^* be the random variable corresponding to the estimator of the regression function $\hat{m}(x)$, sample size n and $\hat{\sigma}$ from the given data. It is expected that c^* is close to c. First, we generate random samples using the estimator of the regression function $\hat{m}(x)$, sample size n_S and $\hat{\sigma}$. Next, set a grid for c from 1 to an upper bound and perform simulations at each value. The estimator of c^* is the value for which the simulation coverage is closest and greater than or equal to the nominal coverage. Unfortunately, this process is time consuming and not efficient at all. Hence, Zhang and Lu (2008) developed a process to estimate the value of c by incorporating logistic regression model. We describe this procedure, called procedure 1 as follows.

Given a data set, we estimate $m(x)$ by (3) and σ by (7), say $\hat{m}$, $\hat{\sigma}$. We then generate a number of samples, say 1000, from $\hat{m}$, $\hat{\sigma}$ and the fixed xs. Between the lower bound 1 and upper bound of c, we set c_i $(i = 1, 2, ...1000)$ as equally spaced points. The 1000 different c_is are evaluated at the 1000 different samples respectively, i.e. c_i evaluated at sample i. If the confidence band $\hat{m}(x) \pm c_i l_\alpha$ contains the estimator of the regression function, set $h_i = 1$ and otherwise zero. This process will generate a sequence of binary responses $h_i(i = 1, 2, ...1000)$. We use the logistic regression model to fit the data $\{(h_i, c_i)\}$ for $i = 1, 2, \ldots, 1000$. The estimate of c is the value c_i corresponding to the probability 100(1-α)%.

4 Simulation Studies

In this section, a small simulation study has been conducted to investigate the performance of the proposed confidence bands. The simulation set up follows from Harms and Duchesne (2010) and Zhang and Lu (2008). The following equation is used to generate the population at the super model stage

$$y_i = 2 + sin(2 * \pi * x_i) + \epsilon_i, \ i = 1, ..., 1000, \tag{13}$$

where population size $N = 1000$. Random errors are from a normal distribution with mean 0 and constant variance σ^2. At the sampling design stage, Poisson sampling scheme (unequal probability design) is considered. The sample weight d_i of poisson sampling scheme have been chosen such that weights are proportional to the auxiliary variable $z_i = (y_i + 2)(x_i + 2)$ and $\sum_U 1/d_i = E(n_S) = n$. The simulation study was performed with factors: (1) standard deviation σ : .05 and .1; (2) nominal levels α: 0.1 and 0.05; (3) sampling sizes: $n = 100$ and $n = 200$; The confidence bands are investigated under different settings with different sample sizes, variances σ^2 and α levels.

The preferred simulation would generate 1000 samples by Poisson sampling. For each generated sample, we use procedure 1 to estimate the constant c, say $\hat{c}$, to construct the confidence band, which leads to a success or failure result. However, it takes 2 hours to estimate c using procedure 1. Consequently, we need 10000 hours. Practically speaking, it is impossible to carry out this simulation. Notice that when the sample size is bigger than 100, $\hat{m}(x)$ and $\hat{\sigma}$ are very close to the true function and σ. The c value for estimated settings using $\hat{m}(x)$ and $\hat{\sigma}$ (We defined it c^* previously) should be very close to the c value for the true settings. Hence, we did following simulations to investigate this method.

We use the Epanechnikov kernel for the local linear smoother $\hat{m}(x)$. Cross-validation method is used to select the bandwidth and Lu (2014)'s method is used to estimate the σ. The grid of 1000 evenly spaced points in the interval [0,1] is used. We generated 1000 samples by poisson sampling scheme under each experimental setting and estimated c by procedure 1. The estimates of c, called $\hat{c}$ were given in Table 2. Next, we did 5000 simulations for each experimental setting using the estimated c value. The experimental coverage results were shown in Table 3. Notice that from Table 3, the empirical coverage are very close to nominal level. Hence, we believe logistic regression model is appropriate to be used here. The simulation we did shows that in practice $\hat{c}^*$ from the logistic regression model is very close to c^*, which suggest $\hat{c}^*$ could be a good estimator of c when c^* is very close to c.

5 Concluding Remarks

In this paper, we proposed a new method to construct simultaneous confidence bands in nonparametric regression with complex survey data. The local linear smoother (Harms & Duchesne, 2010) is used to estimate the regression curve. Difference based variance estimator (Lu, 2014) is used to estimate the variance. Instead of subtracting the estimator of the bias, we expand the confidence bands by multiplying by an adjustment value c to account for the bias of the smoother. The confidence bands take the form of $\hat{m}(x) \pm c * l_\alpha(x)$, and incorporate both the sampling weights and the kernel weights. Logistic regression is used to estimate the bias adjustment constant c. The resulting procedure for constructing confidence bands is entirely data-driven. Simulations show that the proposed method works very well.

Acknowledgements

The authors thank the referees for their very careful reading of the manuscript, and their helpful comments and constructive suggestions to improve the manuscript.

References

Bellhouse, D. R., & Stafford, J. E. (1999). Density estimation from complex surveys. *Statistica Sinica*, *9*, 407–424.

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

Bickel, P. J., & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates (Corr: V3 p1370). *The Annals of Statistics*, *1*, 1071–1095.

Bickel, P. J., & Rosenblatt, M. (1975). Corrections to "on some global measures of the deviations of density function estimates". *The Annals of Statistics*, *3*, 1370–1370.

Breidt, F. J., & Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics*, *28*(4), 1026–1053.

Buckley, M. J., Eagleson, G. K., & Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, *75*, 189–199.

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., & Lohr, S. L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, *128*(1), 160–190.

Eubank, R. L., Kambour, E. L., Kim, J. T., Klipple, K., & Reese, C. S. (1998). Estimation in partially linear models. *Computational Statistics & Data Analysis*, *29*, 27–34.

Eubank, R. L., & Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, *88*, 1287–1301.

Gasser, T., Sroka, L., & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, *73*, 625–633.

Hall, P., Kay, J. W., & Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, *77*, 521–528.

Harms, T., & Duchesne, P. (2010). On kernel nonparametric regression designed for complex survey data. *Metrika*, *72*(1), 111–138.

Klipple, K., & Eubank, R. L. (2007). Difference based variance estimators for partially linear models. *Festschrift in Honor of Mir Masoom Ali on the occsion of his retirement, May 18-19*, 313-323.

Korn, E. L., & Graubard, B. I. (1998). Scatterplots with survey data. *The American Statistician*, *52*, 58–69.

Lohr, S. (2010). *Sampling: Design and analysis 2nd edition*. Cengage Learning.

Lu, Y. (2014). Difference based variance estimator for nonparametric regression in complex survey. *Journal of Statistical Computation and Simulation*, *84*, 335-343.

Macdonell, W. R. (1901). On criminal anthropometry and the identification of criminals. *Biometrika, 1*, 177–227.

Müller, H.-G., & Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics, 15*, 610–625.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics, 12*, 1215–1230.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society, Series B, Methodological, 60*, 797–811.

Zhang, G., & Lu, Y. (2008). Adjusted confidence bands in nonparametric regression. *Communications in Statistics-Simulation and Computation, 37*, 106-113.

Table 1: Potential Upper Bound for c, function used $m(x) = 2 + sin(2 * \pi * x_i)$

σ	Nominal coverage	c value	Actual coverage for n_S = 100	c value	Actual coverage for n_S = 200
0.05	0.90	c = 3	0.911	2	0.953
0.05	0.95	c = 4.5	0.952	2	0.972
0.1	0.90	c = 3	0.960	2	0.971
0.1	0.95	c = 3	0.971	2	0.977

Table 2: The Estimates of c by logistic regression, function $m(x) = 2 + sin(2 * \pi * x_i)$ is used

		$\hat{c}$ for the following n:	
σ	Nominal coverage	100	200
0.05	0.90	2.845619	1.675935
0.05	0.95	4.06597	1.785406
0.1	0.90	2.188657	1.603235
0.1	0.95	2.943163	1.666908

Table 3: Simulation study of coverage under different settings, function $m(x) = 2 + sin(2 * \pi * x_i)$ is used

		Actual coverage for the following n:	
σ	Nominal coverage	100	200
0.05	0.90	0.896	0.890
0.05	0.95	0.946	0.945
0.1	0.90	0.906	0.900
0.1	0.95	0.963	0.942