Model Averaging in Survey Estimation

Xiaoxi Li[†] and J. D. Opsomer[*]
Center for Survey Statistics and Methodology, and
Department of Statistics, Iowa State University, Ames, IA 50011

Abstract:
Model averaging is a widely used method as it accounts for uncertainties in model selection. However, its applications in survey estimation are much to be explored. We investigate a model-averaging (MA) regression estimator for the population total, for the case of a nonparametric regression model. Different ways to obtain this estimator are explored through simulation studies.

KEY WORDS: Local polynomial regression, cross-validation, regression estimation.

1. Regression estimator

Regression estimation is often used in survey estimation. It makes use of the auxiliary information about the population to provide efficient estimation. Suppose we have l study variables $\mathbf{Y}_j \in \Re^l$ and p auxiliary variables $\mathbf{X}_j \in \Re^p$. Both $\mathbf{Y}_j$ and $\mathbf{X}_j$ are row vectors. Let Y_j denote the jth element of one of the study variables. Consider the following linear regression model

$$Y_j = \mathbf{X}_j \boldsymbol{\beta} + \varepsilon_j,$$

where ε_j's are independent random variables with mean zero and variance σ_j^2. We observe the study variable Y_j for $j \in S$, and the auxiliary variables $\mathbf{X}_j$ for $j \in U$. The population U is of size N and the sample S is of size n. Suppose the quantity of interest is the population total $t_y = \sum_{j\in U} Y_j$. Särndal et al. (1992) describe a regression estimator for the population total of the form

$$\hat{t}_{reg} = \sum_{j\in S} \frac{Y_j - \hat{Y}_j}{\pi_j} + \sum_{j\in U} \hat{Y}_j,$$

where π_j is the inclusion probability for the jth element in sample S and $\hat{Y}_j = \mathbf{X}_j \hat{\boldsymbol{\beta}}_S$ is the predicted value for Y_j where $\hat{\boldsymbol{\beta}}_S$ is defined as

$$\hat{\boldsymbol{\beta}}_S = \left(\sum_{j\in S} \frac{\mathbf{X}_j^T \mathbf{X}_j}{\sigma_j^2 \pi_j} \right)^{-1} \sum_{j\in S} \frac{\mathbf{X}_j^T Y_j}{\sigma_j^2 \pi_j}.$$

This form shows that the regression estimator is a sum of population total of fitted values, $\sum_{j\in U} \hat{Y}_j$, and an adjustment term $\sum_{j\in S}(Y_j - \hat{Y}_j)/\pi_j$.

The efficiency of regression estimator is measured by its asymptotic variance, since its asymptotic bias is negligible. Using Taylor linearization, Särndal et al. (1992) showed that the approximate variance for $\hat{t}_{reg}$ is

$$\begin{aligned} AV(\hat{t}_{reg}) &= \sum_{j\in U}\sum_{i\in U} (\pi_{ji} - \pi_j \pi_i) \frac{(Y_j - \mathbf{X}_j \mathbf{B})}{\pi_j} \\ &\quad \cdot \frac{(Y_i - \mathbf{X}_i \mathbf{B})}{\pi_i}, \end{aligned}$$

where $\mathbf{B}$ is defined as

$$\mathbf{B} = \left(\sum_{j\in U} \frac{\mathbf{X}_j^T \mathbf{X}_j}{\sigma_j^2} \right)^{-1} \sum_{j\in U} \frac{\mathbf{X}_j^T Y_j}{\sigma_j^2}.$$

The variance estimator for $\hat{t}_{reg}$ is

$$\hat{V}(\hat{t}_{reg}) = \sum_{j\in S}\sum_{i\in S} \frac{\pi_{ji} - \pi_j \pi_i}{\pi_{ji}} \frac{(Y_j - \hat{Y}_j)}{\pi_j} \frac{(Y_i - \hat{Y}_i)}{\pi_i},$$

where π_{ji} is the probability of including both the jth and the ith element in sample S.

The above discussion deals with a parametric approach for regression estimation. There are also nonparametric approaches. Let us consider the following model

$$Y_j = m(\mathbf{X}_j) + \varepsilon_j,$$

where m is a continuous and bounded function and ε_j's are independent random variables with mean zero and variance σ_j^2. Let $\hat{m}_j$ denote the predicted model function for $m(\mathbf{x}_j)$ using nonparametric regression. Breidt and Opsomer (2000) proposed a model-assisted local polynomial regression estimator of the form

$$\hat{t}_y = \sum_{j\in S} \frac{Y_j - \hat{m}_j}{\pi_j} + \sum_{j\in U} \hat{m}_j. \qquad (1)$$

[†] lixiaoxi@iastate.edu.
[*] jopsomer@iastate.edu.

We will take $\mathbf{x}_j$ to be univariate from now on. In that case,

$$\begin{aligned}\hat{m}_j &= \mathbf{e}_1^T\left(\mathbf{X}_{Sj}^T\mathbf{W}_{Sj}\mathbf{X}_{Sj}+\frac{\nu}{N^{q+1}}\mathbf{I}\right)^{-1} \\ &\quad \cdot\mathbf{X}_{Sj}^T\mathbf{W}_{Sj}\mathbf{Y}_S \qquad (2) \\ &= \mathbf{w}_{Sj}^T\mathbf{Y}_S,\end{aligned}$$

where q is the degrees of local polynomial regression, $\mathbf{e}_1$ is the $(q+1)\times 1$ vector having 1 in the first entry and all other entries 0, and

$$\mathbf{X}_{Sj}=\begin{pmatrix}1 & (x_1-x_j) & \cdots & (x_1-x_j)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_n-x_j) & \cdots & (x_n-x_j)^q\end{pmatrix},$$

$$\mathbf{W}_{Sj}=\text{diag}\left\{K\left(\frac{x_i-x_j}{h}\right)\frac{1}{\pi_j h},\quad i\in S\right\},$$

where h is the bandwidth and $K\left(\frac{x_i-x_j}{h}\right)$ is the kernel function. On the right hand side of expression (2), the adjustment term $\frac{\nu}{N^{q+1}}\mathbf{I}$, where $\nu>0$, is used to ensure the estimator $\hat{m}_j$ is well defined for all $S\subset U$.

Breidt and Opsomer (2000) showed that the asymptotic MSE of the local polynomial regression estimator $\hat{t}_y$ is equivalent to the variance of the generalized difference estimator, which is

$$\text{Var}_p(t_y^*)=\sum_{j\in U}\sum_{i\in U}(\pi_{ji}-\pi_j\pi_i)\frac{Y_j-m_j}{\pi_j}\frac{Y_i-m_i}{\pi_i},$$

where

$$t_y^*=\sum_{j\in S}\frac{Y_j-m_j}{\pi_j}+\sum_{j\in U}m_j,$$

and m_j is the local polynomial regression estimator at point x_j, based on the entire finite population, given by

$$\begin{aligned}m_j &= \mathbf{e}_1^T\left(\mathbf{X}_{Uj}^T\mathbf{W}_{Uj}\mathbf{X}_{Uj}\right)^{-1}\mathbf{X}_{Uj}^T\mathbf{W}_{Uj}\mathbf{Y}_U \\ &= \mathbf{w}_{Uj}^T\mathbf{Y}_U. \qquad (3)\end{aligned}$$

In expression (3), $\mathbf{X}_{Uj}$ and $\mathbf{W}_{Uj}$ are defined as follows:

$$\mathbf{X}_{Uj}=\begin{pmatrix}1 & (x_1-x_j) & \cdots & (x_1-x_j)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_N-x_j) & \cdots & (x_N-x_j)^q\end{pmatrix},$$

$$\mathbf{W}_{Uj}=\text{diag}\left\{K\left(\frac{x_i-x_j}{h}\right)\frac{1}{h},\quad i\in U\right\}.$$

Breidt and Opsomer (2000) also showed that the MSE of $\hat{t}_y$ is consistently estimated by

$$\hat{\text{V}}(\hat{t}_y)=\sum_{j\in S}\sum_{i\in S}\frac{\pi_{ji}-\pi_j\pi_i}{\pi_{ji}}\frac{Y_j-\hat{m}_j}{\pi_j}\frac{Y_i-\hat{m}_i}{\pi_i}. \qquad (4)$$

2. Model averaging

In practical survey estimation problems, especially large-scale ones, usually multiple response variables are of interest and many auxiliary variables are available. In order to get a good regression estimator in terms of both efficiency and simplicity, a natural approach is to use model selection procedures. However, despite the nice theoretical properties, automated model selection procedures often result in rather unstable estimators in applications. A small variation of the data may produce a very different model. Therefore, if model selection were fully taken into account, regression estimators based on model selection might have unnecessarily large variance. In addition, when there are multiple study variables, it seems almost impossible to select one model that fits all the study variables well. We will show this in the simulation section of this chapter.

An alternative to model selection is model averaging. Intuitively, if two models are very close with respect to a selection criterion, proper weighting of the models can be better than choosing only one of them (an exaggerated $0-1$ decision). In this way, we can eliminate the uncertainty of model selection procedures. Various work has been done in the area of model mixing, such as Breiman (1996), LeBlanc and Tibshirani (1996) and Yang (2001). However, there are few applications in survey estimation.

In this article, we propose a model averaging estimator that can be properly applied to survey estimation problems. We focus on the local polynomial regression estimator $\hat{t}_y$ defined in (1) because nonparametric regression is flexible for a wide range of models and will not suffer from misspecifying the true model as much as parametric regression. The estimator (1) depends on the value of bandwidth h, so its MSE can be considered as a function of h. Selecting proper candidates for model averaging in this case is equivalent to selecting proper values of bandwidth h. See Opsomer and Miller (2005), for example, on optimal bandwidth selection. Note that estimator (1) can also be written in the form of a weighted sum of Y_j's, i.e. $\hat{t}_y=\sum_{j\in S}w_j^*Y_j$, so the weights w_j^* also depends on the value of bandwidth

h and each set of regression weights correspond to a different regression model procedure.

Suppose we have a finite collection of regression procedures to estimate the regression function m. Let the proposed procedures be $\delta_k,\ k = 1, \cdots, K$. Our goal is to provide a method that can properly mix the K regression procedures. The resulting model averaging estimator should be flexible and perform well for multiple study variables under a wide range of regression models. In other words, this model averaging method should be overall a good choice for all study variables. In practical survey problems, several sets of regression weights are often available, with each set being obtained from a certain regression procedure. Suppose we have K sets of regression weights, then for each study variable Y_j, there are K possible regression estimators for t_y, denoted by $\hat{t}_{yk},\ k = 1, \cdots, K$. This model averaging estimator should be appealing due to its flexibility. It should also significantly reduce the amount of work needed for estimation because it does not require a separate estimation procedure for each individual study variable. All that is needed is to use this method to average the K sets of regression weights and apply it to all study variables.

We consider regression procedure δ_k to be local polynomial regression with bandwidth h_k. The corresponding regression estimator $\hat{t}_{yk}$ is

$$\hat{t}_{yk} = \sum_{j\in S} \frac{Y_j - \hat{m}_{j,k}}{\pi_j} + \sum_{j\in U} \hat{m}_{j,k}, \tag{5}$$

where $\hat{m}_{j,k}$ is the regression predictor for model function m, using procedure δ_k, and

$$\begin{aligned} \hat{m}_{j,k} &= \mathbf{e}_1^T \left(\mathbf{X}_{Sj}^T \mathbf{W}_{Sj,k} \mathbf{X}_{Sj} + \frac{\nu}{N^{q+1}} \mathbf{I} \right)^{-1} \\ &\quad \cdot \mathbf{X}_{Sj}^T \mathbf{W}_{Sj,k} \mathbf{Y}_S \qquad (6) \\ &= \mathbf{w}_{Sj,k}^T \mathbf{Y}_S . \end{aligned}$$

Our proposed model averaging (MA) estimator is of the simple linear form

$$\hat{t}_y^{MA} = \sum_{k=1}^{K} \alpha_k \hat{t}_{yk}, \tag{7}$$

where $\hat{t}_{yk}$ is defined in (5), $\alpha_k \geq 0$ and $\sum_{k=1}^K \alpha_k = 1$. Here α_k is the "weight" that is assigned to procedure δ_k. We are interested in finding the appropriate α_k's for the model averaging estimator $\hat{t}_y^{MA}$. Note that

$$\begin{aligned} \hat{t}_y^{MA} &= \sum_{k=1}^{K} \alpha_k \hat{t}_{yk} \\ &= \sum_{k=1}^{K} \alpha_k \left(\sum_{j\in S} \frac{Y_j - \hat{m}_{j,k}}{\pi_j} + \sum_{j\in U} \hat{m}_{j,k} \right) \\ &= \sum_{j\in S} \sum_{k=1}^{K} \frac{\alpha_k Y_j - \alpha_k \hat{m}_{j,k}}{\pi_j} + \sum_{j\in U} \sum_{k=1}^{K} \alpha_k \hat{m}_{j,k} \\ &= \sum_{j\in S} \frac{Y_j - \sum_{k=1}^K \alpha_k \hat{m}_{j,k}}{\pi_j} + \sum_{j\in U} \sum_{k=1}^{K} \alpha_k \hat{m}_{j,k} \\ &= \sum_{j\in S} \frac{Y_j - \hat{m}_j^{MA}}{\pi_j} + \sum_{j\in U} \hat{m}_j^{MA}, \end{aligned}$$

where

$$\hat{m}_j^{MA} = \sum_{k=1}^{K} \alpha_k \hat{m}_{j,k}. \tag{8}$$

So choosing the proper α_k's for $\hat{t}_{yk}$ is equivalent to choosing proper α_k's for $\hat{m}_{j,k}$.

To proceed with model averaging, let us first consider selecting one best model, i.e. the optimal bandwidth h_{opt}. As stated in section 1, Breidt and Opsomer (2000) showed that the MSE of $\hat{t}_y$ is consistently estimated by $\hat{\mathrm{V}}(\hat{t}_y)$, where $\hat{\mathrm{V}}(\hat{t}_y)$ is defined in equation (4). It seems tempting to consider that h_{opt} can be estimated (asymptotically) by the bandwidth that minimizes $\hat{\mathrm{V}}(\hat{t}_y)$. However, this is not true. One can always choose arbitrarily small bandwidth h so that $\hat{m}_j$ is as close to Y_j as possible. Therefore, as a modification, Opsomer and Miller (2005) proposed a design-based cross-validation (CV) criterion:

$$\begin{aligned} \hat{\mathrm{V}}_{CV}(h_k) &= \sum_{j\in S} \sum_{i\in S} \frac{\pi_{ji} - \pi_j \pi_i}{\pi_{ji}} \frac{Y_j - \hat{m}_{j,k}^{(-)}}{\pi_j} \\ &\quad \cdot \frac{Y_i - \hat{m}_{i,k}^{(-)}}{\pi_i} . \qquad (9) \end{aligned}$$

where $\hat{m}_{j,k}^{(-)}$ is the "leave-one-out" estimator for m_j using procedure δ_k. To obtain this, we replace $\mathbf{w}_{Sj,k}$ in equation (6) by a modified vector $\mathbf{w}'_{Sj,k}$, whose elements are

$$w'_{Sji,k} = \begin{cases} \frac{w_{Sji,k}}{1 - w_{Sji,k}} & \text{if} \quad j \neq i \\ 0 & \text{if} \quad j = i, \end{cases}$$

where $w_{Sji,k}$ denotes the jth element of the vector $\mathbf{w}_{Sj,k}$, and set $\hat{m}_{j,k}^{(-)} = \sum_{j\in S} w'_{Sji,k} Y_j$.

For model averaging purposes, we will consider the following methods to combine models:

1. Take the average of C estimators that have the lowest C $\hat{V}_{CV}(h_k)$.

2. Choose the estimator with the lowest $\hat{V}_{CV}(h_k)$, then we also include estimators with $\hat{V}_{CV}(h_k)$ that are within, say, a $p\%$ window above the lowest one. Then we take average of these estimators.

3. LeBlanc and Tibshirani (1996) suggested using

$$\alpha_k = \frac{\hat{\sigma}_k^{-n}}{\sum_{j=1}^K \hat{\sigma}_j^{-n}}$$

for a normal model, where $\hat{\sigma}_k^2$ is the resubstitution estimate of prediction error for model k. We consider a slightly different model averaging coefficient, α_k^*, where

$$\alpha_k^* = \frac{\hat{\sigma}_k^{-1}}{\sum_{j=1}^K \hat{\sigma}_j^{-1}},$$

and $\hat{\sigma}_k^2$ is defined as

$$\hat{\sigma}_k^2 = \frac{1}{n}\sum_{j\in S}(Y_j - \hat{m}_{j,k}^{(-)})^2.$$

4. With constraint $\alpha_k \geq 0$ and $\sum_{k=1}^K \alpha_k = 1$, choose α_k to minimize the following criteria:

$$\hat{\sigma}^2 = \sum_{j\in S}\frac{1}{\pi_j}(Y_j - \hat{m}_j^{MA})^2, \qquad (10)$$

$$\hat{\sigma}_{CV}^2 = \sum_{j\in S}\frac{1}{\pi_j}(Y_j - \hat{m}_j^{MA(-)})^2, \qquad (11)$$

$$\hat{\mathrm{V}}(t_y^{MA}) = \sum_{j\in S}\sum_{i\in S}\frac{\pi_{ji}-\pi_j\pi_i}{\pi_{ji}}\frac{Y_j - \hat{m}_j^{MA}}{\pi_j} \cdot\frac{Y_i - \hat{m}_i^{MA}}{\pi_i}, \qquad (12)$$

$$\hat{\mathrm{V}}_{CV}(t_y^{MA}) = \sum_{j\in S}\sum_{i\in S}\frac{\pi_{ji}-\pi_j\pi_i}{\pi_{ji}}\frac{Y_j - \hat{m}_j^{MA(-)}}{\pi_j} \cdot\frac{Y_i - \hat{m}_i^{MA(-)}}{\pi_i}. \qquad (13)$$

The estimators $\hat{m}_j^{MA}$ in the methods co nsidered in 4 use the MA version defined in (8). Equation (12) uses a similar idea of minimizing $\hat{\mathrm{V}}(\hat{t}_y)$ in (4). Equation (13) borrows the idea of the CV criterion in (9), except that the minimum is now computed over the α_k instead of over the bandwidth directly. Equation (10) and (11) are similar to (12) and (13), respectively, except that they do not fully incorporate the sampling design. Among equation (10) to (13), (13) will probably provide the best estimator for population total t, but it is the most computational intensive one. So we also investigate (10) to (12). Equation (10) requires the least amount of computation, so if it works decently well, we may choose it over other methods. However, as we have discussed before, (10) can be minimized by choosing arbitrarily small bandwidth values. So it will probably not produce a good estimator. Same reasoning applies to (12). Equation (11) is an improved version of (10). But neither (10) or (11) fully incorporate the design.

In the following sections, we will illustrate the properties of different model averaging estimators through a large-scale simulation study.

3. Simulation setup

To evaluate the properties of Model Averaging (MA) estimators, we generate a single finite population and draw samples repeatedly from it. Specifically, we generate $N = 2000$ values of model variable X from the uniform distribution on $[0, 1]$, and 2000 values of error ε from $N(0, 1)$. This set of errors are used for all populations, up to multiplication by σ. We examine eight populations of Y:

$$Y_{jl} = m_l(x_j) + \varepsilon_j, \qquad 1 \leq i \leq 2000, \qquad 1 \leq l \leq 8$$

where $m_l(x_j)$ are defined on the third column of Table 1. We vary the value of σ to achieve high and low *coefficient of determination*, denoted by R^2. Specifically, we let $R^2 = 0.75$ and 0.25, where R^2 is defined as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{j\in U}\varepsilon_j^2}{\sum_{j\in U}(Y_j - \bar{Y}_U)^2}.$$

We also consider two sampling designs. One is simple random sampling without replacement (SRS) and the other is random stratified sampling (STSRS), that is, we draw an SRS sample within each stratum. We choose two sample sizes, $n = 500$ and $n = 100$. It is easy to see that for SRS design, each element has equal selection probability, which is n/N. For STSRS design, we assign a different selection probability to each stratum. Specifically, we create two equally sized strata in each finite population by the values of model variable x. Then from the first stratum, we draw $n/4$ points and from the second stratum, we draw $3n/4$ points. So the ratio of selection probabilities are 1:3 for these two strata.

For model averaging estimation purpose, we examine five regression procedures. Specifically, we consider local polynomial regression with five different bandwidth values. We choose bandwidth from 0.01 to 0.5, equally spaced on the natural logarithm scale. These values are $h_1 = 0.01$, $h_2 = 0.027$, $h_3 = 0.071$, $h_4 = 0.188$, and $h_5 = 0.5$.

The finite population quantities of interest are $t_{yl} = \sum_{j \in U} Y_{jl}$ for each l. For each simulation, $B = 10000$ samples are drawn from each population. For each sample, we obtain five different estimators for t, denoted by $\{\hat{t}_{yk}\}_{k=1}^{5}$, where $\hat{t}_{yk}$ is defined in (5). Then we consider different methods to compute model averaging estimator $\hat{t}_y^{MA}$ described in the previous section. The details are listed in Table 2. Note that the last five rows in Table 2 are simply $\{\hat{t}_{yk}\}_{k=1}^{5}$. We list them here mainly for two reasons. One is to understand the behavior of each regression procedure, and the other is to compare them with other estimators to see if there are advantages to use model averaging. Loosely speaking, we will call all 13 estimators listed in Table 2 model averaging estimators. The last five rows can be regarded as model averaging of one regression procedure.

In summary, there are eight mean functions, two coefficients of determination ($R^2 = 0.75$ and 0.25), two sampling designs (SRS and STSRS), two sample sizes ($n = 500$ and 100), and 13 estimators for each population total. We report here the results for the cases with size $n = 500$ and $R^2 = 0.75$. Further results are in Li (2006).

Relative Bias (RB) and Mean Squared Error (MSE) are computed for each estimator. Let $\{\hat{t}_{yr}^{MA}\}_{r=1}^{13}$ denote the thirteen estimators listed in Table 2, then

$$\begin{aligned} \mathrm{RB}_r &= \frac{\mathrm{E}(\hat{t}_{yr}^{MA}) - t_y}{t_y}, \\ \text{and} \quad \mathrm{MSE}_r &= \mathrm{E}(\hat{t}_{yr}^{MA} - t_y)^2. \end{aligned}$$

4. Simulation results

Relative bias tables (not shown here) suggest that all 13 estimators have very small biases. We choose to show the relationship among the MESs of 13 estimators.

Table 3 and Table 4 report the MSEs of 12 estimators relative to the MSE of method CV minus one for eight population totals where SRS and STSRS samples are drawn, respectively, to calculate $\hat{t}_{yr}^{MA}$. Specifically, the values in Table 3 and Table 4 are calculated as

$$\frac{\text{MSE(12 estimators)}}{\text{MSE(CV)}} - 1$$

and reported as percentages. Note that this quantity shows how much higher (positive values) or lower (negative values) one method is relative to method CV in terms of MSE. For example, if a value is 50, it means that the correspond method's MSE is 50% higher than that of method CV.

If we examine these tables more closely, we can observe the following facts:

1. In all cases, method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ have smaller MSEs than method MIN$\{\hat{\sigma}^2\}$ and MIN$\{\hat{\mathrm{V}}\}$. So it is better to use leave-one-out CV estimators for m_j in terms of the variability of model averaging estimation.

2. MIN$\{\hat{\sigma}^2_{CV}\}$ is slightly better than, or at least as good as MIN$\{\hat{\mathrm{V}}_{CV}\}$ in terms of MSE.

3. Method CV, CV3, CVp20, MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ return similar MSEs. However, method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ give very similar MSEs, and are either the smallest among CV, CV3, CVp20, MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ or close to the smallest one. We like the fact that method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ are consistently good. Other methods can be inconsistent. They can either be the best for a certain population, or the worst for another. For instance, in Table 3, among method CV, CV3, CVp20, MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$, CV is the best for population "Normal CDF", "Exponential", "Slow sine" and "Fast sine", but the worst for population "Bump".

4. The method Relative Fit behaves well except for population "Fast sine". For example, in Table 3 the MSE of method Relative Fit is 75.37% higher than the MSE of method CV.

5. From the last five rows in Table 3 and Table 4, we can see that the MSEs of the five regression procedures vary greatly for each population. Specifically, a regression procedure can be very good for a certain population, but very bad for another. For example, in Table 3, FIXt(0.188) has the smallest MSE for population "Normal CDF" (0.38% lower than CV), but for population "Fast sine," it is almost the worst (222.08% higher than CV), where CV has the smallest MSE among all 13 estimators.

5. Simulation conclusions

Based on the previous discussion and the results that are in Li (2006), we draw the following conclusions.

1. As far as biases are concerned, all 13 estimators perform very well. It is hard to choose from them if we only consider their biases.

2. In terms of MSEs, if sample size is large ($n = 500$), model selection through CV is important because it performs as well as the best available model averaging methods. If sample size is smaller ($n = 100$), we consider method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ to be the overall good choices, as they behave well in all cases. Method CV, CV3 and CVp20 are good, but they are not as consistent as method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ for different cases and different study variables.

3. We can draw the same conclusions for both equal selection probability design, SRS, and unequal selection probability design, STSRS. The only difference we can see is that relative biases and MSEs for SRS are smaller than the corresponding ones for STSRS.

4. If model averaging is carried out in a proper way, it can eliminate uncertainty of model selection procedures and produce more reliable estimators. If one is to draw a sample to estimate the population total but do not have prior knowledge about the population model, we recommend the method MIN$\{\hat{\sigma}^2_{CV}\}$ and MIN$\{\hat{\mathrm{V}}_{CV}\}$ to be the safer bets than other methods because they give good estimators under all cases.

References

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

Breiman, L. (1996). Stacked regressions. *Machine Learning 24*, 49 – 64.

LeBlanc, M. and R. Tibshirani (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association 91*, 1641–1650.

Li, X. (2006). *Applications of nonparametric regression in survey statistics*. Ph. D. thesis. Iowa State University.

Opsomer, J. and C. Miller (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Nonparametric Statistics 17*, 593 – 611.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1992). *Model assisted survey sampling*. Springer-Verlag Inc.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association 96*(454), 574–588.

Name	Abbreviation	Expression
(1) Linear	LINE	$2x$
(2) Quadratic	QUAD	$1 + 2(x - 0.5)^2$
(3) Bump	BUMP	$2x + \exp(-200(x - 0.5)^2)$
(4) Jump	JUMP	$\begin{cases} 2x & \text{if} \quad x \leq 0.65 \\ 0.65 & \text{if} \quad x > 0.65 \end{cases}$
(5) Normal CDF	NCDF	$\Phi^{-1}(1.5 - 2x)$
(6) Exponential	EXPO	$\exp(-8x)$
(7) Slow sine	SLOW	$2 + \sin(2\pi x)$
(8) Fast sine	FAST	$2 + \sin(8\pi x)$

Table 1: List of population functions.

Method	Description
CV	Choose the estimator that has the lowest $\hat{V}_{CV}(h_k)$.
CV3	Take the average of 3 estimators that have the lowest 3 $\hat{V}_{CV}(h_k)$.
CVp20	Choose the estimator that has the lowest $\hat{V}_{CV}(h_k)$. Also include estimators having $\hat{V}_{CV}(h_k)$ that are $\leq 20\%$ bigger than the lowest one.
Relative Fit	Use α_k's as described in method 3. i.e. $\alpha_k = \frac{\hat{\sigma}_k^{-1}}{\sum_{j=1}^{K} \hat{\sigma}_j^{-1}}$
MIN$\{\hat{\sigma}^2\}$	With constraint $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$, choose α_k to minimize (10).
MIN$\{\hat{\sigma}^2_{CV}\}$	With constraint $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$, choose α_k to minimize (11).
MIN$\{\hat{\mathrm{V}}\}$	With constraint $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$, choose α_k to minimize (12).
MIN$\{\hat{\mathrm{V}}_{CV}\}$	With constraint $\alpha_k \geq 0$ and $\sum_k \alpha_k = 1$, choose α_k to minimize (13).
FIXt(0.010)	Choose the estimator that uses bandwidth $h_1 = 0.010$.
FIXt(0.027)	Choose the estimator that uses bandwidth $h_2 = 0.027$.
FIXt(0.071)	Choose the estimator that uses bandwidth $h_3 = 0.071$.
FIXt(0.188)	Choose the estimator that uses bandwidth $h_4 = 0.188$.
FIXt(0.500)	Choose the estimator that uses bandwidth $h_5 = 0.500$.

Table 2: List of methods and corresponding descriptions.

Averaging	Population Functions							
Method	LINE	QUAD	BUMP	JUMP	NCDF	EXPO	SLOW	FAST
CV3	-0.17	-0.28	0.70	-0.84	0.20	0.16	0.43	1.97
CVp20	0.76	0.64	1.00	-0.18	1.15	0.94	1.13	2.18
Relative Fit	0.90	2.24	4.11	4.70	1.34	2.39	4.90	75.37
MIN$\{\hat{\sigma}^2\}$	13.74	12.41	10.64	6.68	16.34	12.28	12.74	9.17
MIN$\{\hat{\sigma}^2_{CV}\}$	-0.09	-0.33	-0.42	-0.55	0.23	0.08	0.18	0.12
MIN$\{\hat{\mathrm{V}}\}$	13.74	12.41	10.64	6.68	16.34	12.28	12.74	9.17
MIN$\{\hat{\mathrm{V}}_{CV}\}$	-0.10	-0.31	-0.42	-0.55	0.30	0.13	0.18	0.12
FIXt(0.010)	13.74	12.41	10.64	6.68	16.34	12.28	12.74	9.17
FIXt(0.027)	3.75	2.55	0.94	-0.27	3.73	2.44	2.84	0.01
FIXt(0.071)	0.75	-0.34	-0.45	3.00	0.60	-0.17	-0.08	16.26
FIXt(0.188)	-0.04	-0.17	16.95	14.40	-0.38	2.50	3.60	222.08
FIXt(0.500)	-0.44	31.35	37.07	55.10	2.10	30.59	58.52	230.60

Table 3: Mean Squared Error (MSE) of 12 model averaging methods relative to method CV minus one (in percent) for eight populations ($R^2 = 0.75$), Simple Random Sampling (SRS) with sample size $n = 500$.

Averaging	Population Functions							
Method	LINE	QUAD	BUMP	JUMP	NCDF	EXPO	SLOW	FAST
CV3	-0.25	0.01	1.31	-1.78	0.29	-0.46	-0.56	1.66
CVp20	0.63	0.48	1.69	-1.08	1.20	0.09	-0.12	2.25
Relative Fit	2.23	2.68	4.16	-0.36	1.43	2.87	3.51	72.04
MIN$\{\hat{\sigma}^2\}$	23.38	21.56	18.61	15.36	12.45	19.85	20.70	13.78
MIN$\{\hat{\sigma}^2_{CV}\}$	-0.02	-0.17	-0.34	0.17	0.15	-0.33	-0.50	0.01
MIN$\{\hat{\mathrm{V}}\}$	23.38	21.56	18.61	15.36	12.45	19.85	20.70	13.78
MIN$\{\hat{\mathrm{V}}_{CV}\}$	0.09	-0.21	-0.11	-0.97	0.15	-0.10	-0.45	0.13
FIXt(0.010)	23.38	21.56	18.61	15.36	12.45	19.85	20.70	13.78
FIXt(0.027)	7.90	6.34	3.75	2.10	3.33	4.91	5.52	-0.15
FIXt(0.071)	1.31	-0.15	-0.94	-1.39	0.46	-0.99	-0.84	14.87
FIXt(0.188)	-0.31	-0.75	14.16	0.88	-0.20	2.20	-0.49	223.28
FIXt(0.500)	-0.47	30.51	36.01	20.05	4.78	48.30	47.02	223.37

Table 4: Mean Squared Error (MSE) of 12 model averaging methods relative to method CV minus one (in percent) for eight populations ($R^2 = 0.75$), Random Stratified Sampling (STSRS) with sample size $n = 500$.