Nonparametric Regression and the Two Sample Problem

Alan H. Dorfman
Office of Survey Methods Research
U.S. Bureau of Labor Statistics
2 Massachusetts Ave., N.E.
Washington, D.C. 20212
dorfman.alan@bls.gov

Abstract
Nonparametric regression is the model-based sampler's method of choice when there is serious doubt about the suitability of a linear or other simple parametric model for the survey data at hand. It supersedes the need for use of design weights and standard design-based weights. Recognition of this is especially helpful in confronting problems in sampling situations where design weights are missing or questionable. One example is the case where we have data from two (or more) samples from a given population. We discuss this case.

Key Words: inclusion probabilities, post-stratification, model-based, model-assisted, bandwidth choice, twicing

1. Introduction: the two sample problem

The two sample problem is simply stated: Suppose two distinct surveys gather related information on a single population *U*. How do we best combine data from the two surveys to yield a single set of estimates of a population quantity ("population parameter") of interest?

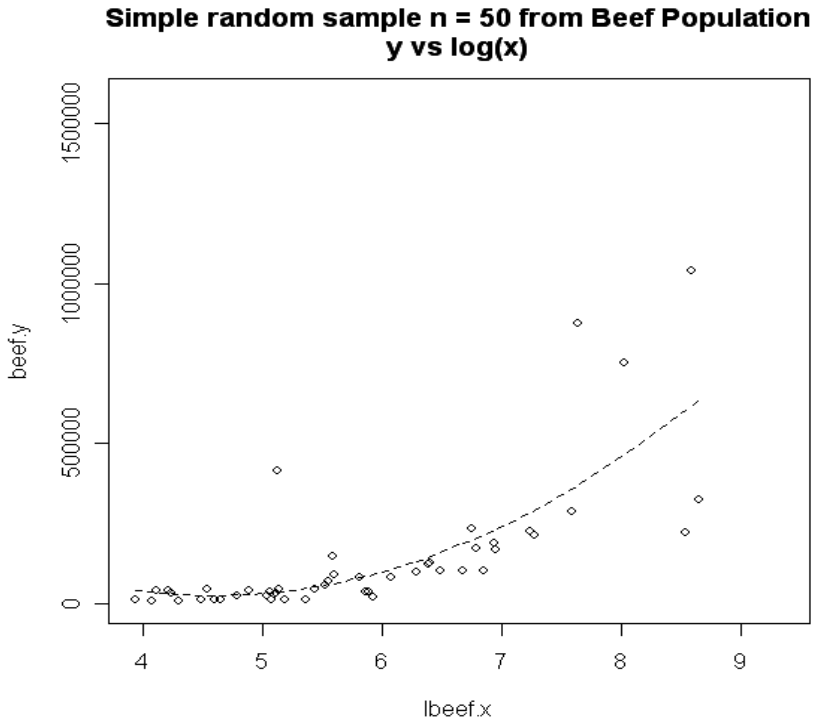
For example, at the Bureau of Labor Statistics (BLS), the survey of Occupational Employment Statistics (OES) and the national Compensation Survey (NCS) both collect data on occupational employment and wages in different ways and with different emphases. It has been an ongoing research question at BLS how the data from these two surveys might be unified into a single overall product.

In general, one possible way to address this situation is to get separate estimates from the two surveys and weight them together with weights the inverse of their estimated variances. See, for example, (Merkouris, 2004) and the references therein. Another possibility is to combine the two data sets into a single data set and modify the weights on individual sampled units in some appropriate fashion (Dorfman, 2008).

Consider the following example. The Beef Population (N = 410) is a collection of Australian Beef Farms (Cattle Ranches) that has been often studied beginning with (Chambers and Dunstan 1986). The auxiliary variable x is the Acreage of the farms, known prior to the survey for all farms in the population. The variable of interest y is the size of the herd sold in the particular year of interest. Values of y are available for those farms selected into the sample. Of interest is the total number of cattle sold across all farms, $T = \sum_{l \in U} y_l$ Now suppose a first sample ("Sample 1") is taken by simple random sampling (*srs*), with say sample size n_1 = 50, and the results are as in Figure 1. The line fitted to the curve is the result of a quadratic regression fit of y on log(x). The residuals

from the fit are given in Figure 2, giving some evidence that the variance increases with *x*. Some further investigation suggests the variance increases as x^d, with *d* somewhere between 1 and 2.

Figure 1



Now suppose there is a budget surplus and it is decided to take a second sample ("Sample 2") that capitalizes on the new knowledge of variance structure. Ideally this would be a probability proportional to size (*pps*) sample with size variable $x^{d/2}$, but as an approximation, a stratified simple random sample (*stratsrs*) is taken – ten strata, with cum($x^{3/4}$) equal in each, and 5 units taken at random in each. Thus there are now two samples, each of size 50, of different types, and composing a combined bigger sample $s = s_1 \cup s_2$. The number of distinct units in the combined sample would typically be less than 100, because of overlap.

How best make inference from the combined data? Several approaches seem worth considering.

1.1 Design-based approaches

1. *Mixture Approach.* Get separate estimates for each sample, and weight them together by the inverse of their estimated variances:

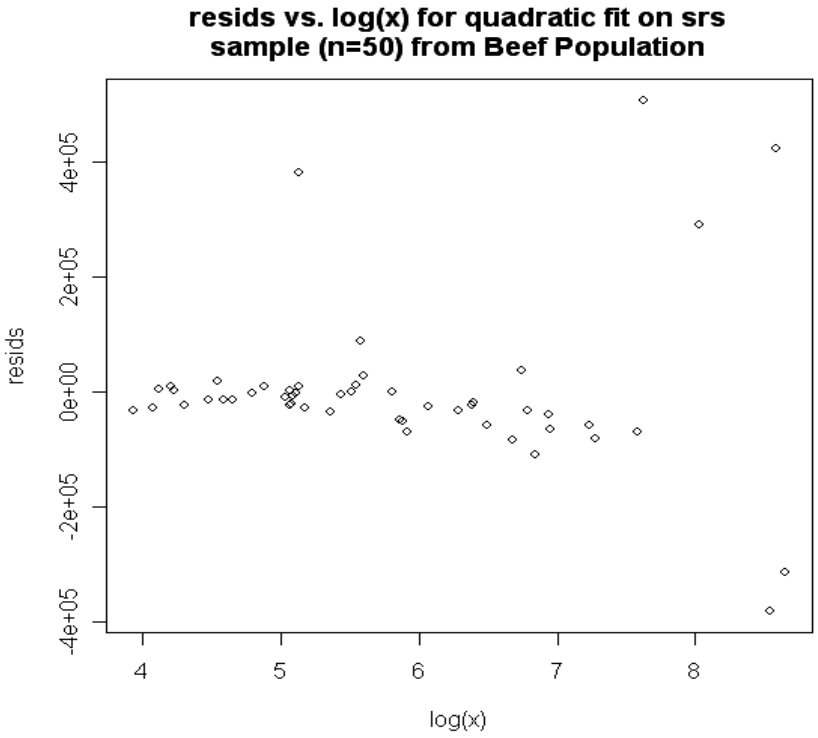
$$\hat{T}_{mix} = \hat{v}_1^{-1}\hat{T}_1 + \hat{v}_2^{-1}\hat{T}_2$$

Here the individual $\hat{T}_k$ would be estimators suitable for the respective samples, for example, their corresponding Horvitz-Thompson estimators.

2. *Using overall inclusion probabilities.* If we wish to base an estimate on the combined data *s*, and to proceed according to design-based methodology, then a first step is to get overall inclusion probabilities $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$. These can be incorporated into an Horvitz-Thompson estimator, a Hajak estimator, or a model-assisted estimator. For

example, the Horvitz-Thompson estimator would be $\hat{T}_{\pi^*} = \sum_{i \in s1 \cup s2} w_i^* y_i$,with $w_i^* = \pi_i^{*-1}$. We should note that units occurring in both samples ("overlaps units") appear only once in the expression. This assumes we can identify duplicates – if we cannot, or it is too expensive or time consuming to do so, we will be stuck with the mixture method.

Figure 2



1.1.1 A Disconcerting Example

Using the overall inclusion probabilities in combination with the Horvitz-Thompson estimator can lead to counter-intuitive results.

Suppose a population of size *N* is sampled twice (independently) by *srs.* Then for $i \in s1 \cup s2$, we have $\pi_i^* = n_1/N + n_2/N - n_1 n_2/N^2$. Suppose both samples take half of the population: $n_1 = n_2 = N/2$. The minimal size the combined sample could be is $N/2$ and the maximal is *N*; the expected sample size is (3/4)*N*. Then $\pi_i^* = 1/2 + 1/2 - 1/4 = 3/4$ and the HT estimator is $\hat{T}_{\pi^*} = \frac{4}{3}\sum_{i \in s1 \cup s2} y_i$. This is a bit strange: if the actual combined sample size is less than (3/4)*N*, we can anticipate the estimate will tend to be low. And contrariwise, if larger, $\hat{T}_{\pi^*}$ is likely to be too large. In the extreme case where the combined *n* = *N* (not likely in practice, but possible), we would *know* the estimate is 33% too high.

In general, design-based weights play a subtle dual role. Explicitly, they get used because they lead to attractive sampling properties: across all possible samples under the design, the average of estimates is the target ("unbiasedness") or is close to it ("near-unbiasedness"), etc. Implicitly, they seem appropriate for the particular sample at hand

insofar as they give a fair way for the units in the sample selected to represent to us what the population is like. For example in stratified sampling, if the strata are so chosen that *y* values within a stratum are likely to be near each other, then weighting up the sampled units by the ratio of the number in the stratum to the number sampled, makes good sense. Thus there are two aspects to sampling weights: (a) the across-potential-samples--their *sampling properties* and (b) the for-this-sample-within-this-population--their *representativeness*. In the above simple example, the representative aspect is lost. Question: Should we still want to use inclusion probabilities when their representative aspect –the way they tie *this* sample to the population sampled—goes by the wayside?

1.2. Model-based

3. *Post-stratification.* Ignore sampling probabilities and use "post-stratified" weights, to form an estimator of the form $\hat{T}_{ps} = \sum_h N_h n_h^{-1} \sum_{i \in h \cap (s1 \cup s2)} y_i$, where $\{h\}$ are suitably chosen strata, N_h = number of population units in *h*, and n_h is the number of unique units in the combined sample that fall in stratum *h*. The original sampling does need not be stratified for this idea to be germane. Dorfman (2008) considers "iterated" post-stratification, where several post-stratified estimates, using systematically varied stratum boundaries, are averaged to give an overall estimate. This is implicitly a species of estimate based on non-parametric regression.

2. Nonparametric regression estimation

4. *Nonparametric regression estimation.* More generally, we can use weights that spring from a model that supposes only that the expected value of *y*, conditional on *x*, is continuous in *x*. Thus, suppose, for $i \in U$, that y_i depends on x_i through the model $y_i = m(x_i) + \varepsilon_i$, where $m(x_i)$ is an unspecified function assumed to be continuous, and the errors ε_i are independent with mean 0 and variance $v(x_i)$. We can refer to this set-up as *the weak model*. In this scheme, we can ignore which of the original samples, the y_i are available from and can ignore the inclusion probabilities.

The field of nonparametric regression offers a variety of ways to estimate $m(x) = E(y \mid x)$ for values of *x* within reasonable range of the sample values of x_i, in particular for non-sample $x = x_j$. Then the very simple idea is to get nonparametric estimates $\hat{m}(x_j)$ for all x_j in *U* and estimate *T* by $\hat{T} = \sum_{j \in U} \hat{m}(x_j)$.

This is reasonable, since we expect $\sum_{j \in U} m(x_j) \approx \sum_{j \in U} y_j$. A slightly better idea, in keeping with standard model-based practice, is to capitalize on the fact that the sample *y*'s are known, and take $\hat{T}_{np} = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \hat{m}(x_j)$.

Many ways are current for doing nonparametric regression estimation. The basic idea of all of them is that the auxiliary variable *x* provides some measure of *nearness* of points, so that we take as an estimate a weighted sum $\hat{m}(x_j) = \sum_{i \in s} w_{ij} y_i$, where the relative sizes of the w_{ij} depend on the distance of (sample) x_i to x_j. Probably the simplest version of this is (Nadayara-Watson) kernel estimation, where for *b* a chosen scaling factor (the

"bandwidth"), and K a symmetric density function, for example, the standard normal density function, we take $w_{ij}^{(b)} = \dfrac{K\left(\dfrac{x_i - x_j}{b}\right)}{\sum_{i' \in s} K\left(\dfrac{x_{i'} - x_j}{b}\right)}$.

Another possibility, with slightly better properties, is *local polynomial regression*.For each x_j, a weighted regression predictor is calculated with the weights on ith sample point dependent on distance of x_i from x_j, again scaled by a bandwidth b (Fan 1992; Ruppert and Wand 1994). Let $X(x) = \begin{bmatrix} 1\, x_1 - x \ldots (x_1 - x)^q \\ \vdots \quad \vdots \quad \vdots \\ 1\, x_n - x \ldots (x_n - x)^q \end{bmatrix}$ and

$W_b(x) = \begin{bmatrix} K\left(\dfrac{x_1 - x}{b}\right) 0 \ldots 0 \\ \vdots \quad \vdots \quad \vdots \\ 0\, 0 \ldots K\left(\dfrac{x_n - x}{b}\right) \end{bmatrix}$. The local polynomial regression estimate at x_j is then given by $\hat{m}_b(x_j) = [1\, 0 \ldots 0]\left(X(x_j)^T W_b(x_j) X(x_j)\right)^{-1} X(x_j)^T W_b(x_j) y_s$

$\equiv \sum_{i \in s} w_{ij}^b y_i$

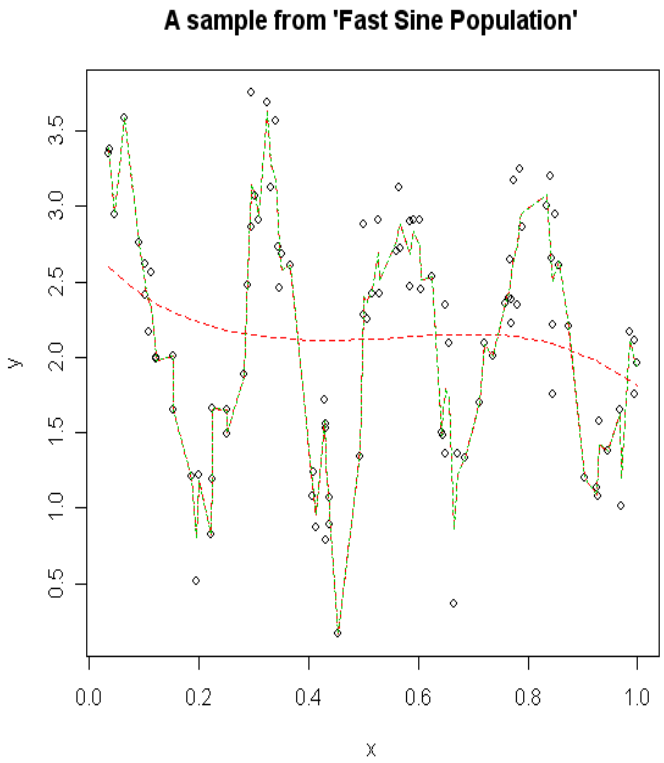
The (standard) use of nonparametric regressio*n*, outside of the context of survey sampling, is "smoothing"--getting a picture of the curvilinear relation of *y* on *x*, through a cloud of points. The choice of the bandwidth *b* makes a big difference in the degree of smoothness of the curve, and there exist many competing ways to do the selection. One that has a good deal of appeal is One Sided Cross Validation OSCV (Hart &Yi, 1998). The preferred bandwidth is a well chosen multiple of the bandwidth $\tilde{b}$ which minimizes the one-sided prediction error

$prederr(\tilde{b}) = \sum_{i=m}^{n} \left(\tilde{m}(x_i) - y_i\right)^2$, where $\tilde{m}(x_i)$ uses only data with $x < x_i$.

Figure 3 below shows some sample data from the Fast Sine Population (Breidt and Opsomer 2000), with curves fit from two bandwidths. The green curve is a local linear smoother using bandwidth $b = 0.005$, which was apparently too small, judging from the many gratuitous changes of direction of the curve; the red curve used $b = 0.2$, which was too large: clearly changes of direction in the data are being missed.

Figure 4 shows the same data, with the black curve determined by the OSCV selected bandwidth black, $b = .0233$; the true underlying curve which generated the data is shown is shown in red. The OSCV based curve seems like a pretty good match.

Figure 3



In the sampling context, if interest is in estimating T, the general closeness of $\hat{m}(x)$ to $m(x)$ across the range of x is not in itself of interest. We want, rather, that $\hat{T} = \sum_{i\in s} y_i + \sum_{j\in U\setminus s} \hat{m}(x_j)$ be close to T, or, equivalently, that $\hat{T}_r = \sum_{j\in U\setminus s} \hat{m}(x_j)$ be close to $T_r = \sum_{j\in U\setminus s} y_j$.

Now $\hat{T}_r = \sum_{j\in U\setminus s} \hat{m}_b(x_j) = \sum_{j\in U\setminus s}\sum_{i\in s} w_{ij}^b y_i = \sum_{i\in s} W_i^b y_i$, where $W_i^b = \sum_{j\in U\setminus s} w_{ij}^b$.

Therefore, we want b to minimize $E\left(\hat{T}_r - T_r\right)^2 = B^2 + v$, where the Bias B and Variance v of $\hat{T}$ are respectively

$$B = E\left(\hat{T} - T \mid \mathbf{x}_U\right) = E\sum_{j\in U\setminus s}\left\{\hat{m}(x_j) - m(x_j)\right\}$$

$$= \sum_{j\in U\setminus s} E\left\{\hat{m}(x_j) - m(x_j)\right\} \quad (1)$$

$$= \sum_{i\in s} W_i^b m(x_i) - \sum_{j\in U\setminus s} m(x_j) \quad (2)$$

$$v = \mathrm{var}\left(\hat{T} - T \mid \mathbf{x}_U\right) = \sum_{i\in s}\left(W_i^b\right)^2 v(x_i) + \sum_{j\in U\setminus s} v(x_j), \text{ where}$$

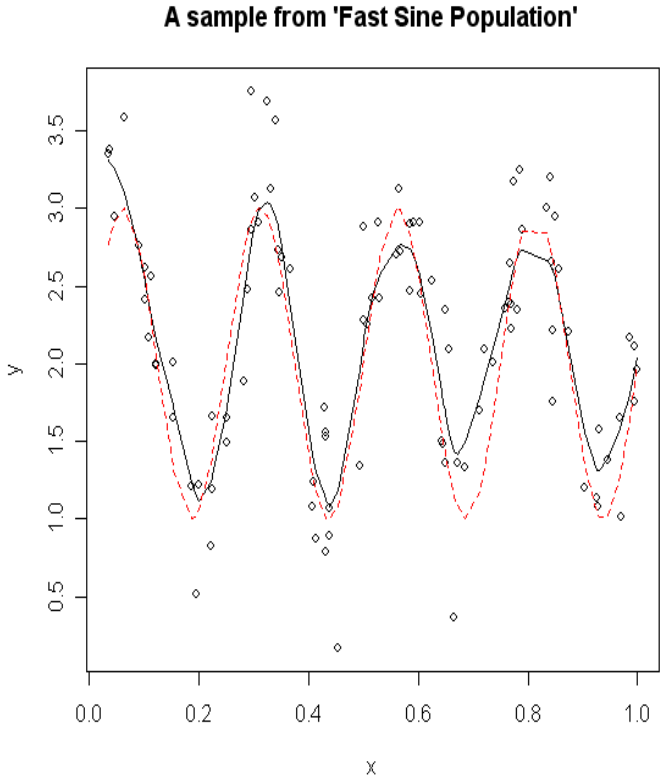
$$v(x_i) \equiv \mathrm{var}(\varepsilon_i \mid x_i).$$

We want b to minimize $B^2 + v_1$, where v_1 = first term of v, since 2nd term of v does not depend on b.

2.1 Options for Choosing Bandwidth

1) Can use a standard method from the smoothing literature, for example OSCV.

Figure 4



2) *Plug-in estimate* of $B^2 + v_1$: Can plug in a "pilot estimate" $\hat{m}_{\tilde{b}}(x_j)$ for $m(x_j)$'s in B and, for first term of v, estimate $v(x_i) = E\left(\left[y_i - m(x_i)\right]^2\right)$ by $\hat{v}(x_i) = \left(y_i - \hat{m}_{\tilde{b}}^{-}(x_i)\right)^2$ where $\hat{m}_b^{-}(x_i) = \sum_{i' \in s, i' \neq i} w_{i'i}^b y_{i'} \Big/ \left(1 - w_{ii}^b\right)$ is an approximation to estimate of $m(x_j)$, which avoids using y_i, as in (Opsomer and Miller 2005)—henceforth O&M. For the pilot estimate we can use an estimate based on, for example, OSCV.
3) *Design-based cross-validation* (O&M), to be described below.

Note that (2) explicitly takes into account the non-sample x's, but (1) and (3) do not.

2.2 Asymptotics

For small b and large nb, under some fairly unrestrictive and simple conditions,

$$B = (1/2) c_K b^2 \sum_{j \in U \setminus s} m''(x_j) + \text{lower order terms}$$
$$= O\left(N_r b^2\right)$$
$$v = \frac{N_r^2}{n^2} \sum_{i \in s} \frac{f_r(x_i)^2}{f_s(x_i)^2} \sigma^2(x_i) \left(1 + O_p\left(b^2 + \frac{1}{\sqrt{nb}}\right)\right)$$
$$= O\left(N_r^2 / n\right)$$

Proof will be given elsewhere. Dorfman (1992, 1994) gives asymptotics for kernel-based estimators.

2.2.1 Some Implications of the asymptotics

The dominant component of the variance is independent of bandwidth choice and is of order $O\left(N_r^2/n\right)$, the same as for more conventional survey sample estimators of total.

To minimize the variance, the sample design should be such that sample density is proportional to standard deviation, i.e. the sample should be roughly a probability proportional to size sample, with size measure the standard deviation, again in keeping with typical survey sample theory.

The secondary terms of the variance will be of equal order provided the bandwidth b is of order $n^{-1/5}$, which is the same order that minimizes mean square error in ordinary smoothing. However, this is *not* the optimal order in the present case.

We see this by looking to the bias B. B^2 will be of lower order than the variance term only if b is of lower order than $n^{-1/4}$, which is narrower, of course, than $n^{-1/5}$. This implies that the secondary terms of the variance *cannot* be of equal order. For smaller bandwidth, the first term b^2 of the secondary terms will be smaller, and the second term larger. Mean square error will be minimized when the second term and the bias squared term are equal, which occurs when $b = O(n^{-1/3})$. Thus we would expect the best bandwidth for the estimation of totals to be narrower than that for standard smoothing.

2.3 Model-assisted non-parametric regression

One motive for turning to nonparametric regression based estimators for the two sample problem is the opportunity to avoid worrying about the complexities of inclusion probabilities in this context. Indeed, given sufficient population data, nonparametric regression seems a suitable replacement to standard design-based estimation.

But perhaps *model-assisted* nonparametric regression estimation, that is, estimation which makes use of the weak model but relies on inclusion probabilities nonetheless would be better, as suggested in simulation studies by (Breidt & Opsomer 2000) — henceforth B&O. Maybe the inclusion probabilities are still indeed necessary, even if the model is weak?

B&O note that if the actual means $m\left(x_j\right)$ were known for $j \in U$, then a design unbiased estimator of T would be the model-assisted estimator $\hat{T}^* = \sum_s \frac{y_i - m\left(x_i\right)}{\pi_i} + \sum_U m\left(x_j\right)$.

The weighted up residual adjustment in the first term allows for the fact that there is a difference between the true targets, the y_i, and the mean values $m\left(x_j\right)$, which can be expected merely to be *near* the y_i.

Since the $m\left(x_j\right)$ are not known, the basic idea is to use *np* regression to estimate $m\left(x_j\right)$ and plug into the above expression:

$$\hat{T}_{np,tw} = \sum_s \frac{y_i - \hat{m}\left(x_i\right)}{\pi_i} + \sum_U \hat{m}\left(x_j\right) \ . \qquad (3)$$

Their estimator modifies the simple nonparametric regression estimator in two ways:

1) By "twicing", the addition of the weighted sum of residuals, with the weights given by the inverse inclusion probabilities. To the extent that the inclusion probabilities are *representative*, this makes sense, and one could equally well use, for example, post-stratified weights. In general, twicing seems like not a bad idea, although one need not be locked into using the inverse inclusion probabilities (cf. Chambers, et al, 1993).

2) Standard design-based theory will also lead to altering the internal kernel weights. The weights $w_{\pi ij}$ which get applied to the *i*th sample point in estimating $\hat{m}\left(x_j\right)$ in local linear regression depend on kernel weights $W_{\pi ib}\left(x_j\right)=\frac{1}{\pi_i b}K\left(\frac{x_i-x_j}{b}\right)$ which depend on the inclusion probabilities. This means that the less likely to be selected sample points will tend to have proportionately greater weight than they would under standard nonparametric regression. In practice, this does not seem to make much difference. But it is interesting to note a small discomfiting fact, what might be called the "paradox of the leaping neighbor": situations can arise where say $x_i < x_k < x_j$, but $W_{\pi ib}\left(x_j\right) > W_{\pi kb}\left(x_j\right)$.

Bandwidth selection is also, of course, a concern for model-assisted nonparametric regression estimation of totals. The estimator (3) will be virtually unbiased, so the aim is to choose a bandwidth which minimizes the design variance. (Opsomer and Miller 2005) [O&M] find *b* to minimize

$$\hat{v} \;=\; \hat{\mathrm{var}}\left(\hat{T}\right) \approx \sum\nolimits_{i,j\in s} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{y_i-\hat{m}_b^-\left(x_i\right)}{\pi_i} \frac{y_j-\hat{m}_b^-\left(x_j\right)}{\pi_j}, \quad (4)$$

where $\hat{m}_b^-\left(x_j\right)=\sum_{i\in s} w'_{\pi ij} y_i$, with

$$w'_{\pi ij}=\begin{cases}\dfrac{w_{\pi ij}}{1-w_{\pi ij}}, & j\neq i\\ 0, & j=i\end{cases}$$

The idea is to avoid using y_i as an ingredient in $\hat{m}_b^-\left(x_i\right)$; otherwise, the residuals could be made to shrink by taking *b* small. We may call the procedure which minimizes (4) *pi-weighted cross-validation*. Indeed, in the case of simple random sampling, with the π_i's all equal, it is readily seen that this procedure is equivalent to standard cross-validation.

3. Simulation Study on Beef Population

The simulation study of B&O showed the model-assisted nonparametric estimators outperforming by a fair margin competing non-model-assisted estimators of Dorfman (1992) and Chambers et al. (1993). This study considered the estimators at two somewhat arbitrarily selected bandwidths. Indeed, since the sample design considered was simple random sampling, where the π's wash out in the local linear weights, what B&O were actually comparing were a twiced local linear estimator against a simple kernel estimator and a linear-model-based estimator with a kernel adjustment of residuals. We do not expect the same bandwidth to be optimal for this array of estimators, and the logical possibility exists that at their optimal bandwidths, the relative merits of the estimators might shift. In any case, since the π's played no role, the study was not, strictly speaking, a comparison between design-based and model-based methodology.

We here report on a set of simulation studies, with a mix of survey designs, and employing various approaches to selecting the best bandwidths. We employ the Beef Population U of Australian farms, of size $N = 410$, with auxiliary variable x = Acreage, known for all the farms, and variable of interest y = Herd size, known only on the sampled farms. Of interest is the population total of cattle $T = \sum_{l \in U} y_l$.

We suppose, for each run of the experiment that two samples are taken:
Sample 1: *srs* ($n_1 = 50$)
Sample 2: *stratsrs*--approximation to *pps*($x^{3/4}$) – ten strata, with cum($x^{3/4}$) equal in each, and 5 units taken at random in each, so that $n_2 = 50$, also.

The total experiment consists of 500 runs of pairs of samples.
The median number of distinct units in the combined samples was 94, and ranged between 88 and 99.

The following table gives the estimators considered.

Estimator	*Formula*	*Comment*
Nonparametric regression	$\hat{T}_{np} = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \hat{m}\left(x_j\right)$	
Nonparametric regression, twiced	$\hat{T}_{np,tw} = \sum_s \frac{y_i - \hat{m}\left(x_i\right)}{\pi_i} + \sum_U \hat{m}\left(x_j\right)$	Two versions: (a) post-stratified (b) $\pi_i = \pi_i^*$ (as in HT)
Poststratified	$\hat{T}_{ps} = \sum_h N_h n_h^{-1} \sum_{i \in h \cap (s1 \cup s2)} y_i$	Using strata from s_2
Horvitz-Thompson	$\hat{T}_{\pi^*} = \sum_{i \in s1 \cup s2} w_i^* y_i$,	$w_i^* = \pi_i^{*-1}$ $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$
Hajek	$\hat{T}_{\pi^*,cal} = N \sum_{i \in s1 \cup s2} w_i^* y_i \Big/ \sum_{i \in s1 \cup s2} w_i^*$	same w_i^* as HT
mixture of separate est'rs	$\hat{T}_{mix} = \hat{v}_1^{-1} \hat{T}_{np,tw;1} + \hat{v}_2^{-1} \hat{T}_{np,tw;2}$	$\hat{v}_k =$ $\hat{\text{var}}_\pi\left(\hat{T}_{np,tw;k}\right)$ k = 1, 2

Thus there were two versions of the twiced estimator (in case of combined sample)
(a) *Poststratified* ("np-twice^P")

$$\hat{T}_{np,tw} = \sum_s \frac{y_i - \hat{m}\left(x_i\right)}{\pi_{h(i)}} + \sum_U \hat{m}\left(x_j\right)$$

$\pi_{h(i)} = n_h / N_h$, $i \in h$, strata taken to be strata that defined s_2

(b) *Strict Model-Assisted* ("np-twice*")

$\hat{T}^*_{np,tw} = \sum_s \frac{y_i - \hat{m}\left(x_i\right)}{\pi_i^*} + \sum_U \hat{m}\left(x_j\right)$, with $\pi_i^* = \pi_{1i} + \pi_{2i} - \pi_{1i}\pi_{2i}$.

For the bandwidths selection formula (4), we also need the overall joint probability of inclusion of units *i* and *j*, which works out to be

$$\pi_{ij}^* = \pi_{1ij} + \pi_{2ij} + \pi_{1ij}\pi_{2ij} + \pi_{1i}\pi_{2j} + \pi_{2i}\pi_{1j} - \pi_{1ij}\left(\pi_{2i} + \pi_{2j}\right) - \pi_{2ij}\left(\pi_{1i} + \pi_{1j}\right).$$

For bandwidth selection, we investigated several methods:

(1) Hart-Yi's one sided cross validation, *oscv*
(2) plug-in to equation (2) using *oscv* as pilot
(2*) plug-in using a "near-*oscv*" pilot estimator, where the pilot bandwidth was the smallest *b* that gave one sided prediction error within 5% of the minimum prediction error, under their algorithm. This was in accord with a suggestion of (Ruppert 1997) that smaller pilot bandwidths are preferable for minimizing the bias of the pilot estimator.
(3) The design-plug-in of (Opsomer-Miller) , equation (4) above.
(4) plug-in using expression for bias and variance of *twiced* estimator, which is the analogue to (2) for the twiced estimator.

For an estimator $\hat{T}$ we considered three measures of relative success across the 500 runs:
(I) bias measured as ratio of mean value (across runs) to target

$$bias = \sum_{run=1,500}\left(\hat{T}_{(r)} - T\right) / T$$

(II) root mean square error divided by target

$$rmse = \sqrt{\sum_{run=1,500}\left(\hat{T}_{(r)} - T\right)^2} / T$$

(III) Frequencies of being a lesser or equal distance to target than the simple nonparametric regression based estimator using plug-in based bandwidth

$$freq = \#\left\{\left|\hat{T}_{(r)} - T\right| \le \left|\hat{T}_{np(r)} - T\right|\right\} / \#Runs$$

3.1 Results of Beef Simulation

We offer results for
(a) sample 1, gotten through simple random sampling, not a very good design given the heteroscedasticity of *y*-errors
(b) sample 2, a stratified version of *pps*($x^{3/4}$) sampling – a pretty good design
(c) the combination of sample 1 and sample 2

Results are tabulated in the tables below.

Simulation Results for s_1 = *srs* (498 runs)*

estimator	bandwidth	mean/*T*	100rmse/*T*	100ratio.nearer
np	oscv	1.05	18.33	49.8
	plug-in	0.98	*13.74*	------
	plug-in-nr	0.98	14.31	------
	design-cv	1.04	36.07	48.4
	plug-in-tw	0.99	14.11	63.5
np-twice	oscv	1.00	16.56	50.2
	plug-in	0.97	***13.48***	44.2
	plug-in-nr	0.97	14.06	44.4
	design-cv	1.00	35.14	50.8
	plug-in-tw	0.97	*13.50*	46.0
poststrat		0.87	19.35	30.7
Hajek		1.00	20.66	30.3
HT		1.00	20.66	30.3
mix		------	--------	-------

*two samples were omitted where, because of uneven spread of *x*-values, the *np* estimator could not be calculated (and other estimators gave distorted results)

Simulation Results for s_2 = stratified (~*pps*($x^{3/4}$))

estimator	bandwidth	mean/*T*	100rmse/*T*	100ratio.nearer
np	oscv	0.96	12.01	53.0
	plug-in	0.96	11.35	------
	plug-in-nr	0.99	*9.88*	67.4
	design-cv	1.03	10.88	54.8
	plug-in-tw	0.98	13.58	65.4
np-twice	oscv	0.99	10.80	57.4
	plug-in	0.99	*10.06*	61.2
	plg-in-nr-tw	0.99	***9.72***	62.4
	design-cv	1.00	10.38	59.4
	plug-in-tw	0.99	10.24	60.8
poststrat		1.00	*9.99*	60.6
Hajek		1.00	*9.99*	60.6
HT		1.00	*9.99*	60.6
mix		------	------	--------

*Observation*s on *srs* sample s_1
1. the plug-in is good, with twicing giving improvement
2. plug-in using "near *oscv*" not as good
3. the *oscv* and especially design-*cv*, which is just standard cross-validation in this case, have serious problems with respect to *rmse* (but note the *ratio-nearer* numbers)
4. the poststratified estimator was not so good, but the only stratification scheme used was the one already in place for s_2. It's possible that a more adaptable post-stratification would have done better (cf. Dorfman 2008)

Observations on the stratsrs sample s_2
1. twicing clearly desirable
2. using near-oscv as pilot improves things
3. simple plug-in not so good, but otherwise plug-in best
4. it is puzzling that, for the twiced estimator, the straight plug-in and plug-in using "near" *oscv* did better than a plug-in formula based on the twiced formula itself – but the differences are not great
5. design-*cv* and plug-in-twice on a par
6. straight use of *oscv* in second rank
7. post-stratified, which in this case is just the stratified, among best; it can be improved but not, it seems, by much

Our major interest lies in the Combined Sample, given below.

Observations on the *combined* sample $s = s_1 \cup s_2$
1. the model-assisted estimator is best by a small margin. That is does better than similar versions with post-stratified weights is surprising, given the degree to which the simple post-stratified estimator outperforms the Hajek or HT.
2. for twicing, all methods of bandwidth selection about same
3. plug-in-near, without twicing, is competitive with twiced estimators
4. simple poststratification estimator is among best
5. Hajek and Horvitz-Thompson not so good
6. the mixture estimator altogether deficient – it pays to be able to treat the combined sample as one sample

Simulation Results for $s = s_1 \cup s_2$

estimator	bandwidth	mean/*T*	100rmse/*T*	100ratio.nearer
Np	oscv	1.06	9.63	36.2
	plug-in	1.00	7.26	--------
	plug-in-nr	0.99	*7.04*	--------
	design-cv	1.07	9.86	33.4
	plug-in-tw	1.02	8.27	48.6
Np-twice[P]	oscv	1.00	*6.97*	56.0
	plug-in	0.99	*7.05*	53.4
	plg-in-nr-tw	0.99	*6.92*	55.6
	design-cv	1.00	*7.05*	56.2
	plug-in-tw	1.00	*6.92*	57.0
Np-twice[*]	design-cv	1.00	***6.86***	59.8
poststrat		1.00	*6.92*	53.6
Hajek		1.00	7.67	44.0
HT		1.00	8.94	38.0
mix		0.97	18.53	41.6

4. Discussion

1. One question we must face is: Are design weights and design-based approach necessary, as appeared to be the case in the B&O simulation study? The answer as evidenced here is No. Although the model-assisted twicing estimator using design-*cv* was best in the combined sample by a small margin, it is clear that nonparametric regression based estimators do quite well without these weights.
2. Twicing with post-stratified weights, using plug-in choice of bandwidth, was always competitive.
3. The plug-in approach is a viable means of selecting bandwidth. The improvements that arose by modifying the pilot estimator suggest that there is room for improvement
4. poststratification, which can be regarded as a simple version of np regression, did quite well in the combined sample. It is worth revisiting the iterated poststratified estimator of (Dorfman 2008) – it is basically a kind of variable bandwidth estimator
5. conventional cross validation and design-cross validation gave variable results – did surprisingly well on the combined sample, but were second rate in the individual *srs* and *stratified* samples
6. Straight Horvitz-Thompson, Hajek, and especially the mixture method should be avoided if sufficient information is available to apply one of the other estimators
7. Overall, *np* regression is a suitable way to handle the two sample problem.

Acknowledgements

The views expressed in this paper are the author's and do not reflect Bureau of Labor Statistics policy

References

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

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993), Bias robust estimation in finite populations using nonparametric calibration, *Journal of the American Statistical Association*, 88, 268-277

Chambers, R.L. and Dunstan, R. (1986), Estimating distribution functions from survey data, *Biometrika*, 73, 597-604

Dorfman, A.H. (1992), Non-parametric regression for estimating totals in finite populations, *Proceedings of the Joint Statistical Meetings, Section of Survey Research Methods*, 622-625

Dorfman, A.H. (1994), Open questions in the appliocation of smoothing methods to finite population inference, *Interface: Computing Science and Statistics, Volume 26*, 201-205

Dorfman, A.H. (2008), The two sample problem, *Proceedings of the Joint Statistical Meetings, Section of Survey Research Methods*

Fan, J. (1992), Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, 87, 998-1004

Hart, J.D. and Yi, S. (1998) One-sided cross-validation, *Journal of the American Statistical Association*, 93, 620-631

Merkouris, T. (2004), Combining independent regression estimators from multiple surveys, *Journal of the American Statistical Association*, 99, 1131-1139

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

Ruppert, D. (1997), Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *Journal of the American Statistical Association*, 92, 1049-1062

Ruppert, D. and Wand, M.P. (1994), Multivariate locally weighted least squares regression, *The Annals of Statistics*, 22, 1346-1370