NON-PARAMETRIC REGRESSION FOR ESTIMATING TOTALS IN FINITE POPULATIONS

Alan H. Dorfman, Bureau of Labor Statistics
2 Massachusetts Avenue N.E., Washington, D.C. 20212

Keywords: Auxiliary Information; bandwidth; kernel estimator; survey sample data; ratio estimator; regression estimator

Non-parametric regression provides computationally intensive estimation of unknown finite population quantities. Such estimation is frequently more flexible and robust than inference tied to design-probabilities (in design-based inference) or to parametric regression models (in model-based inference). The non-parametric regression based estimator for finite population totals is introduced. Its utility is supported by determination of its asymptotic bias and variance and by a simulation study on wage data.

1. **The problem**. We consider a fresh approach to the estimation of finite population "parameters" based on a sample from the population. Given a population P of N units for each of which there is a variable Y of interest, with values available on a sample s of P, we often wish to estimate a function of the population Y's, for example, the population total $T=\sum_P Y_i$, the population mean T/N, or the population distribution function $F(y)=N^{-1}\sum_P I(Y_i \le y)$; the last is useful in estimating population quantiles. In what follows we will focus on the total T; parallel work on the distribution function may be found in (Chambers, Dorfman, and Wehrly, 1992) and (Dorfman and Hall, 1992). We assume that an auxiliary variable x related to Y is available for the entire population.

Section 2 briefly reviews the model-based and design-based approaches to inference on totals in survey sampling and describes (a sample from) a dataset on wages collected in Boston by the Bureau of Labor Statistics. Section 3 reviews non-parametric regression. Section 4 introduces a non-parametric regression based estimator of the total, gives some theoretical properties, and compares it to the standard design-based expansion estimator. Section 5 gives empirical results on several design-based and model-based estimators of the total, based on simulations on the Boston Wage Data. The non-parametric regression estimator outperforms the design-based estimators. Section 6 states conclusions and points to questions requiring research.

2. **Rival approaches**. There are two incompatible approaches for making inference from sample to population: the more traditional designed-based approach, in which the probability structure of the procedure by which the sample s is selected serves as the basis for inference, and the model-based or predictive approach, in which a regression model of Y on x is used to predict the non-sample Y's and, by consequence, their total (or other function of interest).

For an example of the difference, consider the sample in Figure 1 from a population, the "Boston Wage Population", consisting of N=400 Boston establishments. Y is the total wages paid to workers in a selected group of occupations; x is the total number of workers in each establishment including those in positions outside the select group. The data is taken from the Bureau of Labor Statistics' 1991 White Collar Pay Survey for Boston. The sample is a stratified random sample: for h=1,2,3 , n_h=20 points were taken from each of three strata of sizes N_h= 202, 114, and 84 respectively. Three classes of company size, viz. $0 < x < 250$, $250 \le x < 1000$, and $1000 \le x$, determined the strata.

Then a designed-based estimator of total is the stratified expansion estimator

$$\hat{T}_{\exp} = \sum_h \pi_h^{-1} \sum_{s_h} Y_{hi}$$

where, for h=1,2,3, π_h are the probabilities of including units Y_{hi} in the sample component s_h of the hth stratum. The presence of inclusion probabilities is characteristic of design-based estimators.

If one assumes that Y is linear in x, for example, $Y_i = \alpha + \beta x_i + \sigma_i e_i$, $i = 1,\ldots N$, with the e_i IID with mean 0, then an appropriate model-based estimator is

$$\hat{T}_{lin} = \sum_s Y_i + \sum_{P-s} \left(\hat{\alpha} + \hat{\beta} x_j\right)$$

where $\hat{\alpha}$, $\hat{\beta}$ are the appropriate weighted least squares estimators of α, β. The model-based estimator ignores the selection probabilities.

For elaboration of the issues, see Royall and Cumberland (1981) with discussion, Hansen, Madow, and Tepping (1983) with discussion, and, at this conference, Smith (1992), with discussion.

One advantage of the design-based approach is its *automaticity*: once the planning is done and the sample is selected, estimation of the population quantity is determined, at least in principle (there are lots of ragged edges, such as non-response). It is

helpful to be able to say to interested, possibly powerful, non-statisticians that the analyst her/himself had no input into the result such as arises in the choosing of a model.

The use of non-parametric regression for inference on finite populations, discussed below, is firmly within the model-based tradition. However, it has a much greater degree of automaticity than is generally associated with model-based inference based on standard parametric models. Since the parameters are essentially nuisance parameters anyway (not α, β but T is of interest), this approach is a natural one to consider.

3. **Non-parametric regression**. The idea of non-parametric regression goes back to Nadaraya (1964) and Watson (1964). A current reference is Hardle (1990). There are many species of non-parametric regression; we here consider the simple Nadaraya-Watson kernel estimator.

Consider the model

$$Y = m(x) + \sigma(x)e \qquad (1)$$

with $m(\cdot)$ a smooth function and the e_i independent with mean 0 and constant variance, and suppose we wish to estimate $m(\cdot)$. One possibility is to average the nearby values of Y_i, where "nearby" is measured in terms of the distances $|x_i - x|$. Let $K(u)$ be a symmetric density function, for example the standard normal. For a chosen scaling factor ("bandwidth") b, define $K_b(u) = b^{-1}K(u/b)$, and let the weights $w_i(x) = K_b(x_i - x) \Big/ \sum_{i=1}^{n} K_b(x_i - x)$. The larger b is, the flatter and broader the density function, and the more equal the weights. Then the Nadaraya-Watson estimator of $m(x)$ is

$$\hat{m}(x) = \sum_i w_i(x) Y_i. \qquad (2)$$

Under reasonable conditions on $m(x)$ and the design points x, $\hat{m}(x)$ will be consistent for $m(x)$, as $b \to 0$, $nb \to \infty$.

Figure 2 shows estimates of $m(x)$ for the sample of the wages data for three choices of bandwidth with $K(u)$ the standard normal density. A log transformation has been applied to the auxiliary to even out the spread. Note that the wider the bandwidth, the smoother the estimated function. From the figure, it appears that the smallest bandwidth is perhaps accomodating the data too much.

4. **Non-parametric regression based estimator of the total**. We can let $x = x_j$ for any point in the non-sample and so estimate $m(x_j)$. Then the following estimator of the total suggests itself:

$$\hat{T}_{np} = \sum_s Y_i + \sum_{P-s} \hat{m}(x_j)$$

As with model-based estimators generally, this estimator ignores sampling probabilities. (It also ignores stratum boundaries.) Except for the selection of bandwidth, and possible transformation of the auxiliary, it is an automatic estimator.

It is interesting to compare this estimator to the expansion estimator $\hat{T}_{\exp}$. $\hat{T}_{np}$ accumulates the non-sample values of $m(x_j)$ in lieu of the Y_j; $\hat{T}_{\exp}$ in effect does the same thing, replacing Y_j by the average of sample Y_i's in the corresponding stratum (compare Figure 3). The expansion estimator tacitly assumes a jump function, with jumps precisely at those points we happened to use for sample selection. This is a tighter model than merely assuming $m(x)$ smooth, so that in a sense $\hat{T}_{\exp}$ is more of a model-based estimator than $\hat{T}_{np}$!

It can be shown that, from the viewpoint of the model (1), the bias of $\hat{T}_{\exp}$ is of the same order as its variance (compare Cumberland and Royall(1988)). For the non-parametric regression estimator, we have the following proposition:

Proposition. ◊ Let $K(u)$ be a symmetric density function with $\int uK(u)du = 0$ and $k_2 \equiv \int u^2 K(u)du > 0$; let $\hat{m}(x)$ be defined as at (2) above; assume $m(x)$ has a continuous second derivative, and let $\beta(x) = d_s(x)m''(x) + 2d_s'(x)m'(x)$; assume n and N increase together such that $n/N \to \pi$, with $0 < \pi < 1$; assume sample and non-sample values of x are in the interval $[c,d]$ and are generated by design densities d_s and d_{P-s} respectively, both bounded away from zero on $[c,d]$, where d_s and d_{P-s} are defined by

$$n^{-1}\sum_i I(x_i \le x) \to \int_{-\infty}^{x} d_s(u)du \qquad \text{and}$$

$$(N-n)^{-1}\sum_j I(x_j \le x) \to \int_{-\infty}^{x} d_{P-s}(u)du \text{ respectively}$$

and are assumed to have continuous first derivatives; then

$$E\left(\hat{T}_{np}-T\right)=b^2(N-n)(k_2/2)\int \beta(x)d_s(x)^{-1}d_{P-s}(x)dx$$
$$+o\left(nb^2+b^{-1}\right) \text{ and}$$
$$\text{var}\left(\hat{T}_{np}-T\right)=(N-n)^2n^{-1}\int\sigma^2(x)d_s(x)^{-1}\left[d_{P-s}(x)\right]^2dx$$
$$+(N-n)\int\sigma^2(x)d_{P-s}(x)dx+o(n)\ \Diamond$$

The proof is omitted. We note the following consequences:

(i) The relative bias is $E\left(\hat{T}_{np}-T\right)/E(T)=O\left(b^2\right)+o\left(b^2+[nb]^{-1}\right)$; this goes to zero so long as the standard conditions $b\to 0$, $nb\to\infty$ are met. (ii) If $b=Cn^{\varepsilon}$ for $-1/2\le\varepsilon<-1/4$, then the ratio $E\left(\hat{T}_{np}-T\right)/\text{var}^{1/2}\left(\hat{T}_{np}-T\right)$ is asymptotically zero; this suggests that a fairly wide choice of bandwidth might be satisfactory in practice, yielding a better estimate than $\hat{T}_{\text{exp}}$. However, the proposition yields no practical prescription for choosing the bandwidth in a particular instance. (iii) These results on the bias hold whether or not the sample and non-sample design densities are the same; this suggests that *balance* (Cumberland and Royall 1988) plays a minor role with this estimator; on the other hand, if the sample x's are not spread throughout the non-sample x's, we can expect the kernel estimation process to run into difficulties.

5. **Empirical Results**. The non-parametric regression based estimator was compared to several design-based estimators of the total, namely the expansion estimator, and the combined and separate ratio and regression estimators (see Cochran (1977)), and also the linear-model based estimator with different assumed variance structures, in a series of 100 stratified random samples from the Boston Wage Population, with strata selected as described in section 2. The auxiliary variable was log-transformed for the non-parametric estimator, and, for comparison, for some of the design-based estimators. Three bandwidths were used which were judged to give reasonable results, based on visual inspection of fits on a single sample (Figure 2 above).

Table 1 gives summary results in the form of the average relative error $\sum_{r=1}^{100}T^{-1}\left(\hat{T}_r-T\right)/100$ and the average squared error $\sum_{r=1}^{100}\left(\hat{T}_r-T\right)^2/100$, where $\hat{T}_r$ is one of the estimators of T computed for sample r.

We note the non-parametric regression based estimator is more efficient (i.e. has smaller average squared error) than the best of the design-based estimators, at the two larger bandwidths. It has about the same efficiency as the expansion estimator at the smaller bandwidth. The non-parametric estimator at its best is about 20% more efficient than the best of the design-based estimators.

Greatest efficiency was achieved by the model based estimator relying on a linear model, with variance assumed proportional to x^2, but there is a drop in efficiency with other variance structures, well below the non-parametric estimator at larger bandwidth.

6. **Conclusions and suggestions for further work.** The above results suggest that, even in the current undeveloped state of the art, the non-parametric regression based estimator of a finite population total is a potent rival to familiar design-based estimators. It has the quality of automaticity we associate with design based estimators, but can better reflect the actual structure of the data, yielding greater efficiency. It can be costly in computer power, and will probably not do as well as a parametric-model based estimator, when the modelling process is done carefully.

Further work on the non-parametric regression based estimator is desirable. Is there a good way to select bandwidth automatically? How should its variance be estimated, and how satisfactory are the consequent confidence intervals? In the Boston Wage Data, the sample showed clear signs of heteroscedasticity, which we ignored in constructing the non-parametric regression based estimator; can its efficiency be improved by incorporating a reasonable assumption of variance structure into the non-parametric regression methodology?

References

Chambers, R. L., Dorfman, A. H. &Wehrly, T. E. (1992), Bias robust estimation in finite populations using nonparametric calibration, *J. Am Statist. Assoc.*,to appear

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley

Cumberland, W. G. and Royall, R. M. (1988), Does Simple Random Sampling Provide Adequate balance?, *J. Royal Statistical Society, Ser. B* 50, 118-124

Dorfman, A. H. and Hall, P. (1992) Estimators of the finite population distribution function using non-parametric regression.

Hardle, W. (1990), *Appled Nonparametric Regression Analysis*, Cambridge: Cambridge University Press

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), An evaluation of model-dependent and probability sampling inferences in sample surveys, *J. Am Statist. Assoc.* 78, 776-93

Nadaraya, E. A. (1964), On estimating regression, *Theory of Prob. and Applic.* 9, 141-142

Royall, R. M., and Cumberland, W. G. (1981), An empirical study of the ratio estimator and estimators of its variance, *J. Am Statist. Assoc.* 76, 66-77

Smith, T. M. F. (1992) Sample Surveys 1975-1990: An age of reconciliation?

Watson, G. S. (1964), Smooth regression analysis, *Sankhya*, Ser. A, 359-372

Table 1. Summary Statistics for Estimators of Total in Boston Wage Population

Estimator	Average Relative Bias	Root Average Squared Error/10^6	RASE($\hat{T}$)/RASE($\hat{T}_{exp}$)
expansion	0.035	6.34	1.00
combined ratio	0.040	6.22	0.98
separate ratio	0.042	6.32	1.00
combined regression	0.070	7.56	1.19
combined regression(log)	0.033	6.16	0.97
separate regression	0.069	7.71	1.22
separate regression (log)	0.032	6.33	1.00
linear model $(\sigma^2(x_i) \propto x_i^0)$	0.102	6.72	1.06
linear model $(\sigma^2(x_i) \propto x_i)$	0.067	6.94	1.10
linear model $(\sigma^2(x_i) \propto x_i^2)$	-0.063	4.56	0.72
non-parametric reg'n (b=0.25)	0.040	6.50	1.02
non-parametric reg'n (b=0.50)	0.013	5.67	0.89
non-parametric reg'n (b=0.75)	0.001	5.40	0.85