PARAMETRIC AND NON-PARAMETRIC PREDICTIVE MEAN REGRESSION ESTIMATES USING FUNCTION OF THE PROBABILITY OF SELECTION

Louis Rizzo, The University of Iowa
Department of Statistics and Actuarial Science, Iowa City, IA 52242

This paper is intended to be a development of the mixed framework for survey sampling. This framework is fundamentally a model-based, or superpopulation framework, but the effects of the complex survey sampling mechanism are studied explicitly. This viewpoint generates a variety of new possibilities for estimators that are easy to use and have superior properties to classical estimators under certain conditions. In this paper the range of possibilities are looked at, and examples are given from the National Educational Longitudinal Survey of 1988.

I. Theoretical Framework and Notation

In the mixed framework the basis for inference is a joint distribution $P(\alpha, X, Y)$ where Y is a response variable of interest, α is an indicator vector indicating which elements of a finite population are selected into a sample, and X will be all covariates related to Y including those used in defining the complex sampling process (for example a size variable in PPS sampling or a dummy variable indicating clusters). A possible reduction of this is $P(\alpha)\xi(Y)$, which would be the $P\xi$ unbiasedness framework (see Cassel, Särndal, and Wretman [1977]).

The underlying superpopulation model for Y is as follows:

$$y_i = f(\underset{\sim}{x}_i, \beta) + \epsilon_1 \tag{1.1}$$

where β is a parameter vector, and f is some parameterized vector function of $\underset{\sim}{x}_i$. The residual ϵ_i has mean 0 and is uncorrelated to $\underset{\sim}{x}_i$. Let W be a subset of the X covariates which are used in the construction of the survey sample design (called design covariates in the literature). Let π_i be the probability of selection of each unit i in the population $i = 1, \ldots, N$. Let $\underset{\sim}{\pi}$ be the vector of π_i. Let π_{ij}, π_{ijk}, etc. be the joint probabilities of selection (see Hajek [1981]). Define Π_∞ as the set of all nontrivial joint probabilities $\{\pi_i\} \cup \{\pi_{ij}\} \cup \cdots$ that define the sample design $P(s)$ fully. For example, in a fixed sample size systematic sampling process we can define the full sampling process either by assigning a probability to every sample p of size n or by defining joint probabilities of selection $\{\pi_i\}, \{\pi_{ij}\}, \cdots, \{\pi_{i_1,\ldots i_n}\}$. Π_∞ and $P(s)$ are equivalent. In this framework Π_∞ is assumed to be a deterministic function of W. Since W is a random variable Π_∞ will be also, though we will usually condition on the realized value of Π_∞, which is the actual sample design.

The problem to be studied is finding an estimator of the finite population mean $\bar{Y}$. The problem is viewed as a predictive problem: the response variable is known for the sampled population and need to be predicted for the unsampled population. For unbiased prediction of the unsampled $y_i's$ it would be sufficient to specify $\mathcal{E}(Y|\alpha = 0)$. This is usually difficult since there is only data from the $y_i's$ with $\alpha_i = 1$, i.e., the sample $y_i's$. However if $\mathcal{E}(Y|\Pi_\infty)$ can be specified then this can be used for unbiased prediction,

$$\mathcal{E}(Y|\Pi_\infty, \alpha) = \mathcal{E}(Y|\Pi_\infty) \tag{1.2}$$

i.e., the relationship between Y and Π_∞ in the sample will be the same as the relationship between Y and Π_∞ among the unsampled elements in the population, leading to unbiased prediction of $\bar{Y}$ based on the modeling of Y and Π_∞ from the sample. This is called ignorability of the sampling mechanism given Π_∞: for a demonstration of the validity of (1.2) see Rizzo [1990]. For good studies of the underlying ideas of the mixed framework see the comment of Rubin to Hanson, Madow, and Tepping [1983]. One can also cite Little and Rubin [1987], Chapter 12, and Little [1982], pp. 235 ff. A more extensive description is Sugden and Smith [1984] with related papers of Scott [1977] and Scott and Smith [1973].

In this paper as in Rizzo [1990] the following special assumptions are made:

$$\begin{aligned} &1) \quad \mathcal{E}(Y|\Pi_\infty) = \mathcal{E}(Y|\underset{\sim}{\pi}) \\ &2) \quad \mathcal{E}(y_i|\underset{\sim}{\pi}) = \mathcal{E}(y_i|\pi_i) \ . \end{aligned} \tag{1.3}$$

It is generally assumed that $\mathcal{E}(y_i|\pi_i)$ is some linear combination of functions of π_i for the parametric linear regression predictors below or a smooth function $f(\pi_i)$ for the nonparametric regression predictors. Under Assumptions 1 and 2 and (1.2), $\mathcal{E}(y_i|\pi_i, \alpha_i = 1)$ is equal to $\mathcal{E}(y_i|\pi_i, \alpha_i = 0)$, leading to unbiased predictors of $\bar{Y}$ based on specification of $\mathcal{E}(y_i|\pi_i)$ with the sampled elements (i.e., corresponding to $\alpha_i = 1$). For sufficient conditions giving the assumptions of (1.3) from (1.2) see Rizzo [1990].

II. Alternative Estimators of $\bar{Y}$ Using Assumption (1.3)

The classical estimator of the finite population mean is the Horvitz-Thompson estimator when one has a complex survey with differing first-stage probabilities of selection. The version of this estimator studied here is a conditional version that has superior properties conditionally to the classical version. This estimator is

$$\begin{aligned} \hat{\bar{Y}}_{HT} &= \frac{\frac{1}{N}\sum_{i\in s} y_i/\pi_i}{\frac{1}{N}\sum_{i\in s} 1/\pi_i} \\ &= \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i \alpha_\pi} \end{aligned} \tag{2.1}$$

where $\alpha_\pi = \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_i}$.

Suppose that $\mathcal{E}(y_i|\pi_i)$ in Assumption (1.3) can be written as a linear combination of functions of π_i, i.e., $\mathcal{E}(y_i|\pi_i) = \Pi_i\gamma$ with $\Pi_i = [f_1(\pi_i)\cdots f_k(\pi_i)]$ for some functions $f_1, \cdots, f_k$. (Usually f_1 is just 1 corresponding to an intercept.) If we can specify this correctly based on the sample data the following predictive regression estimator can be an alternative:

$$\hat{\bar{Y}}_P = \frac{1}{N}\left[\sum_{i\notin s}\Pi_i\hat{\gamma} + \sum_{i\in s} y_i\right] \tag{2.2}$$

(see Royall [1976] for the predictive estimator based on a full model $\underset{\sim}{x}_i'\beta$). Essentially for each unsampled element in the population the unknown y_i is predicted using the conditional

expectation based on the $\Pi_i\gamma$ specification. $\hat{\gamma}$ is the usual least squares estimator of γ based on the sample.

$$\hat{\gamma} = (\Pi_s'\Pi_s)^{-1}\Pi_s' y_s \tag{2.3}$$

where y_s is the vector of sampled $y_i's$ and Π_s is the sample design matrix with rows Π_i , $i \in s$. For a use of the probability of inclusion as a response variable in the non-response situation see Little [1986] and David et al., [1983].

It will be shown below that under assumption (1.3) and the validity of the specified conditional expectation $\hat{\bar{Y}}_p$ is unbiased conditionally and unconditionally.

A general predictive estimator will also be studied based on weaker assumptions:

$$\hat{\bar{Y}}_{NP} = \frac{1}{N}\left[\sum_{i\notin s} y_i + \sum_{i\in s} y_i\right] . \tag{2.4}$$

The predictors in this case are based on a general $\mathcal{E}(y_i|\pi_i)$ specification: no parametric model is fit. The possibility studied in this paper is $\hat{\bar{Y}}_{NN}$. $\hat{\bar{Y}}_{NN}$ is a nearest neighbor nonparametric estimator defined to be as close to the classical $\hat{\bar{Y}}_{HT}$ as possible while having superior conditional properties. It is described in section 5 below.

III. Moment Properties of $\hat{\bar{Y}}_P$ and $\hat{\bar{Y}}_{HT,C}$ in the Mixed Framework

In this section the moment properties of $\hat{\bar{Y}}_P$ and $\hat{\bar{Y}}_{HT,C}$ are studied under the assumptions of (1.3) and a parametric form for $\mathcal{E}(y_i|\pi_i)$. We can write y_i as follows:

$$\begin{aligned} y_i &= \mathcal{E}(y_i|\pi_i) + e_i \\ &= \Pi_i\gamma + e_i \end{aligned} \tag{3.1}$$

for some row vector $\Pi_i = [f(\pi_i)\cdots f_k(\pi_i)]$ of functions of the probability of selection. Under (1.2) and (1.3) we have

$$\mathcal{E}(e_i) = \mathcal{E}(e_i|\alpha_i = 1) = \mathcal{E}(e_i|\alpha_i = 0) = 0 \tag{3.2}$$

This is the 'gain' of (1.2) and (1.3): under the full superpopulation model. e_i contains systematic elements, but with the simple $\mathcal{E}(y_i|\pi_i)$ specified, the expectations of the residuals from the model for sampled and unsampled population units are 0, leading to an unbiased predictor of $\bar{y}$. See Rizzo [1990] for further details and a study of effects of misspecification of the form of $\mathcal{E}(y_i|\pi_i)$.

The actual predictive error of $\hat{\bar{Y}}_P - \bar{Y}$ under (3.1) will be

$$\hat{\bar{Y}}_P - \bar{Y} = \frac{1}{N}\sum_{i\notin s}\Pi_i(\Pi_s'\Pi_s)^{-1}\Pi_s' e_s - \frac{1}{N}\sum_{i\notin s} e_i \tag{3.3}$$

from (2.2), (2.3), (3.1). See also Rizzo [1990].

We can find both conditional and unconditional moment properties, i.e., we can find for example $\mathcal{E}(\hat{\bar{Y}}_P - \bar{Y} \mid \Pi_\infty, \underset{\sim}{\alpha})$ and $\mathcal{E}(\hat{\bar{Y}}_P - \bar{Y} \mid \Pi_\infty)$. In the former case we are conditioning on the sample selected, and in the latter case we are looking at the properties of $\hat{\bar{Y}}_P - \bar{Y}$ over all possible samples, i.e., the distribution of $\underset{\sim}{\alpha}$ given Π_∞. The latter case is essentially equivalent to a design-model or $P\xi$ unbiasedness operator. In this paper focus is on conditional properties. See Rizzo [1989] for a study of unconditional properties of $\hat{\bar{Y}}_P - \bar{Y}$.

When conditioning is made on $\underset{\sim}{\alpha}$ everything is fixed except the $\{e_i\}$. e_s consists of the sample residuals, and $\sum_{i\notin s} e_i$ is a sum of the residuals of the unsampled units. Both sets of e_i have expectation 0 by (3.2), thus $\hat{\bar{Y}}_P$ is an unbiased predictor of $\bar{Y}$ conditionally (and therefore unconditionally as well).

Computation of the conditional variance requires essentially specification of $Var(e_i|\Pi_\infty)$. Under similar assumptions as given in (1.3) this can be reduced to $Var(e_i|\pi_i)$, thus

$$Var(e_i|\pi_i, \alpha_i) = Var(e_i|\pi_i) = G_\pi(\pi_i) \tag{3.4}$$

for some function G_π of π_i. In the paper G_π will be assumed the constant function, i.e., $G_\pi(\pi_i) = \sigma_e^2$. See Rizzo [1990] or Rizzo [1989] for further details. Under this specification some straightforward algebra shows that

$$\begin{aligned} &\mathrm{var}(\hat{\bar{Y}}_P - \bar{Y}|\Pi_\infty, \alpha) = \\ &\sigma_e^2\left[\left(\frac{1}{N}\sum_{i\notin s}\Pi_i\right)(\Pi_s'\Pi_s)^{-1}\left(\frac{1}{N}\sum_{i\notin s}\Pi_i\right)' + \frac{N-n}{N^2}\right] \end{aligned} \tag{3.5}$$

See Rizzo [1989] for further details and the unconditional variance. Under (3.1) the predictive error of the classical estimator can be written as

$$\begin{aligned} \hat{\bar{Y}}_{HT} - \bar{Y} = &\frac{1}{N}\sum_{i\in s}\Pi_i\gamma\left(\frac{1}{\alpha_\pi\pi_i} - 1\right) - \frac{1}{N}\sum_{i\notin s}\Pi_i\gamma \\ &+ \frac{1}{N}\sum_{i\in s} e_i\left(\frac{1}{\alpha_\pi\pi_i} - 1\right) - \frac{1}{N}\sum_{i\notin s} e_i \quad . \end{aligned} \tag{3.6}$$

The conditional expectation is nonzero (the first two terms are fixed constants conditionally). However it is easy to show that the unconditional expectation is asymptotically zero under the usual linearization arguments. The conditional bias in fact tends to be small with most samples, as can be seen in the examples below. $\hat{\bar{Y}}_{NN}$ is designed to eliminate a good portion of this conditional bias. The variance of $\hat{\bar{Y}}_{HT} - \bar{Y}$ is also readily derivable and will not be given (see Rizzo [1990]).

IV. An Example Comparing $\hat{\bar{Y}}_{HT}$ and $\hat{\bar{Y}}_P$

The following example uses data from the National Educational Longitudinal Survey of 1988 (NELS88) sponsored by the U.S. Department of Education and carried out by the National Opinion Research Center (NORC). The survey is designed to track students in the eighth grade (in 1988) over time and collect longitudinal information. Schools are clusters, and are selected proportional to size (of the eighth grade class) with differing sampling fractions within sampling strata. In this paper only the school data for 1988 are studied, thus the clustering and longitudinal aspects do not come into play. The response variable used below is base salary of a beginning teacher with a B.S. The population of interest is restricted to four states in the Midwest (IL, IN, WI, and MI), and the target mean is the mean base salary for suburban, public schools. For more examples from the NELS88 data see Rizzo [1989].[1] The responding sample size is 52 (there were 10 nonrespondents: we assume the response mechanism is ignorable for simplicity). The subpopulation of interest is a domain crossing sampling strata, in this case state strata. The probabilities of selection are related to size of school and

state. The scatter plot of the response variable against the log of the probability is given as Figure 1 below.

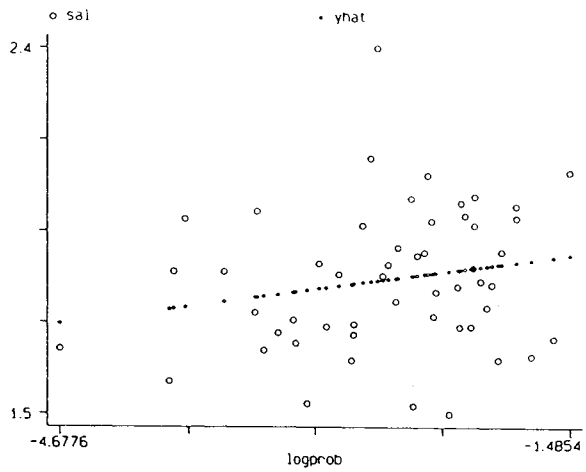

Figure 1. Base Salary against Log (Probability of Selection)

As can be seen linearity in the logarithm is a reasonable assumption, as is homoscedasticity. The estimator $\hat{\bar{Y}}_P$ based on this is $\hat{\bar{Y}}_P = 1.8078$, or \$18,078. The corresponding classical estimator is $\hat{\bar{Y}}_{HT} = 1.7972$. Assuming the predictive conditional expectation $\mathcal{E}(y_i|\pi_i)$ is of the form $\beta_0 + \beta_1(\ell n\ \pi_i)$ the following moments are estimated using the formulae of the previous section (and the obvious estimators of γ and σ_e^2):

	$\hat{\bar{Y}}_P$	$\hat{\bar{Y}}_{HT}$
Bias	0	-.007
Variance	8.11×10^{-4}	1.09×10^{-3}
MSE	8.11×10^{-4}	1.14×10^{-3}

The estimator $\bar{Y}_P$ has a mean squared error that is 71.1% of the MSE of $\bar{Y}_{HT}$ giving an efficiency gain of nearly 30%. See Rizzo [1989] for many other examples: generally $\bar{Y}_P$ gives superior efficiency conditionally and unconditionally in NELS88 examples. See Rizzo [1990] for further discussion of this example and a study of the robustness of the results against misspecification of $\mathcal{E}(y_i|\pi_i)$.

V. The Nearest Neighbor Estimator

As seen in the example $\hat{\bar{Y}}_{HT}$ can be biased conditionally: the squared bias as given with the point estimate of bias was approximately 5% of overall MSE. Where does this bias arise from? It can be understood in the following simple example: suppose there are 200 items in a population with 100 having a probability of selection of .1 and 100 having a probability of .2. Suppose $f(\pi_i) = 5$ for $\pi_i = .1$ and $f(\pi_i) = 10$ for $\pi_i = .2$. We expect 10 selected units from the $\pi = .1$ subgroup and 20 from the $\pi_i = .2$ subgroup unconditionally. Suppose in the sample selected we get 8 from the $\pi_i = .1$ group and 22 from the $\pi_i = .2$ group. The conditional Horvitz-Thompson estimator will weight the 8 in subgroup 1 by dividing by .1, and weight the 22 by dividing the .2. The $\pi_i = .2$ group wil be overrepressented conditionally- -in essence we will be predicting to a population with 80 in subgroup 1 and 220 in subgroup 2.

If the probabilities are *known* for the population by the analyst as they are in the NELS88 example then this conditional bias can be corrected readily, leading to a new version of the Horvitz-Thompson estimator with superior conditional properties.

The estimator is constructed in this way: we predict a $\hat{y}_i$ for each $\pi_i, i \notin s$, by matching the $\pi_i, i \notin s$ to a 'nearest neighbor' in the sample, i.e., the π_j, , $j \in s$ such that $|\pi_j - \pi_i|$ is minimized. Assume for now the $\{\pi_j\}$, $j \in s$ are unique. Call this minimizing probability $\pi_{(i)}$. Let $y_{(i)}$ be the corresponding response. We predict y_i for $i \notin s$ by using $y_{(i)}$. This methodology is similar in spirit to 'hot deck' imputation for nonresponse and matching adjustments for treatment effect in observational studies (Rubin [1979]). The overall predictive mean estimator is

$$\hat{\bar{Y}}_{NN} = \frac{1}{N}\left[\sum_{i \notin s} y_{(i)} + \sum_{i \in s} y_i\right] \tag{5.1}$$

If there are two units π_j, $j \in s$, of equal distance to $\pi_i, i \notin s$ (i.e., on either side) choose one arbitrarily. One can also use $|\log \pi_j - \log \pi_i|$ as a distance criterion, or other possibilities (this distance was actually used in the example below).

The predictive error $\hat{\bar{Y}}_{NN} - \bar{Y}$ is as follows:

$$\hat{\bar{Y}}_{NN} - \bar{Y} = \frac{1}{N}\left[\sum_{i \notin s}[f(\pi_{(i)}) - f(\pi_i)]\right] + \frac{1}{N}\left[\sum_{i \notin s} e_{(i)} - \sum_{i \notin s} e_i\right] \tag{5.2}$$

Let n_i be the number of times y_i, $i \in s$, is used as a predictor. Then the conditional moments are easy to obtain:

$$\mathcal{E}(\hat{\bar{Y}}_{NN} - \bar{Y}) = \frac{1}{N}\sum_{i \notin s}[f(\pi_{(i)}) - f(\pi_i)] \tag{5.3}$$

$$\text{var}(\hat{\bar{Y}}_{NN} - \bar{Y}) = \frac{\sigma_e^2}{N^2}\sum_{i \in s} n_i^2 + \frac{N-n}{N^2}\sigma_e^2 \tag{5.4}$$

Now suppose the $\{\pi_i\}$, $i \in s$ are not unique. In this case use the sample mean at each unique π_i as a predictor for a neighboring π_i, $i \notin s$. The expectation and variance will be straightforward to calculate.

The bias term for $\hat{\bar{Y}}_{NN}$ ought to be small for most samples if f is sufficiently smooth. Suppose for example that $|\frac{\partial f}{\partial \pi}|$ exists over the interval [0,1] and is bounded above by M. Then

$$\begin{aligned} f_\pi(\pi_{(i)}) - f_\pi(\pi_i) &= (\pi_{(i)} - \pi_i)\frac{\partial f}{\partial \pi}|_{\pi_i^*} \\ &\qquad \pi_i \le \pi_i^* \le \pi_{(i)}, \quad \text{or} \\ &\qquad \pi_{(i)} \le \pi_i^* \le \pi_i \ . \\ &\le |\pi_{(i)} - \pi_i| M \end{aligned}$$

Thus

$$|\mathcal{E}(\hat{\bar{Y}}_{NN} - \bar{Y})| \le \max_{i \notin s} |\pi_{(i)} - \pi_i| M$$

For most large samples the π_i, $i \in s$, should be sufficiently dense among the π_i, $i \notin s$ such that this maximum will be small. If f is not smooth then the bias may not be small (the classical estimator will also suffer bias difficulties in this circumstance).

In the NELS88 example cited in example 5 the actual point estimate for $\hat{\bar{Y}}_{NN}$ was 1.7956. The conditional sampling properties show a bias of +.0002, which corresponds to a negligible squared bias, and a variance of 1.06×10^{-3}. Thus the overall MSE of $\hat{\bar{Y}}_{NN}$ is about 6% better than the corresponding MSE of $\hat{\bar{Y}}_{HT}$. The gain in efficiency is due to the utilization of the knowledge of the $\{\pi_i\}$'s in the population: knowledge certainly available to at least the person who selected the sample. $\hat{\bar{Y}}_{NN}$ can also be viewed as a nonparametric version of $\hat{\bar{Y}}_P$ that will be certainly robust against misspecifications of the form of $\mathcal{E}(y_i|\pi_i)$, unlike $\hat{\bar{Y}}_P$.

VI. Further Results on the Relative Properties of $\hat{\bar{Y}}_P$, $\hat{\bar{Y}}_{HT}$, and $\hat{\bar{Y}}_{NN}$

In this section the relative variances of the three predictive estimates will be studied. From (3.3) the variance of $\hat{\bar{Y}}_P$ can be written as

$$\begin{aligned}\mathrm{var}(\hat{\bar{Y}}_p - \bar{Y}) = \\ &\sigma_e^2 \left(\frac{N-n}{N}\right)^2 \\ &\times \left[\left(\frac{1}{N-n}\sum_{i \notin s} \Pi_i\right)(\Pi_s'\Pi_s)^{-1}\left(\frac{1}{N-n}\sum_{i \notin s} \Pi_i\right)\right. \\ &\left. + \frac{1}{N-n}\right] \qquad (6.1)\end{aligned}$$

The quantity $\frac{1}{N-n}\sum_{i \notin s} \Pi_i$ is the mean of the Π_i rows for the unsampled elements, which corresponds to a point in the design space. The first term in (6.1) is simply the leverage of the point $\frac{1}{N-n}\sum_{i \notin s} \Pi_i$ if it were a design point. Let $P_s'P_s$ be $\Pi_s'\Pi_s$ adjusted for sample means. Then assuming the intercept is included in the design matrix we can rewrite (6.1) as

$$\begin{aligned}\mathrm{var}(\hat{\bar{Y}}_P - \bar{Y}) = \\ &\sigma_e^2 \left(\frac{N-n}{N}\right)^2 \\ &\times \left[\frac{1}{n} + (\bar{P}_{ns} - \bar{P})(P_s'P_s)^{-1}(\bar{P}_{ns} - \bar{P}_s)\right. \\ &\left. + \frac{1}{N-n}\right] \qquad (6.2)\end{aligned}$$

where $\bar{P}_{ns} = \frac{1}{N-n}\sum_{i \notin s} \Pi_i$ and $\bar{P}_s = \frac{1}{n}\sum_{i \in s} \Pi_i$, the corresponding sample means of unsampled and sampled units (see Weisberg [1985], p. 113).

Suppose we look at the corresponding variance of $\hat{\bar{Y}}_{HT} - \bar{Y}$. From (4.4)

$$\begin{aligned}\mathrm{var}(\hat{\bar{Y}}_{HT} - \bar{Y}) &= \sigma_e^2 \left(\frac{N-n}{N}\right)^2 \\ &\times \left[\sum_{i \in s}\left(\frac{1}{N-n}\right)^2 \left(\frac{1}{\pi_i \alpha_\pi} - 1\right)^2 + \frac{1}{N-n}\right] \\ &= \sigma_e^2 \left(\frac{N-n}{N}\right)^2 \left[\sum_{i \in s} w_i^2 + \frac{1}{N-n}\right] \qquad (6.3)\end{aligned}$$

where $w_i = \frac{1}{N-n}\left(\frac{1}{\pi_i \alpha_\pi} - 1\right)$. The w_i's can be seen readily to add to 1. Thus the quantity $\sigma_e^2 \sum w_i^2$ is the variance of a weighted mean of residuals. The variance of $\hat{\bar{Y}}_{NN} - \bar{Y}$ can also be put in the form of (7.3) with a different set of w_i's that add to 1. What can now be said of relative variances? In the special case when $\bar{P}_{ns}$ and $\bar{P}_s$ are equal

$$\mathrm{var}(\hat{\bar{Y}}_P - \bar{Y}) \leq \mathrm{var}(\hat{\bar{Y}}_{HT} - \bar{Y})$$

with equality only if the w_i's are equal (which would correspond to equal π_i's). The special casse of equality of means is unlikely though as the unconditional expectations of $\bar{P}_{ns}$ and $\bar{P}_s$ are different. In general the variance of $\bar{Y}_P$ depends on the degree of extrapolation from $\bar{P}_s$ to $\bar{P}_{ns}$, and the variance of $\hat{\bar{Y}}_{HT}$ and $\hat{\bar{Y}}_{NN}$ depends on the effective weights w_i in both cases, and in general there is no domination result. There is a special case where domination is assured, however. The following example from NELS88 is consistent with a *null* model, i.e., a conditional expectation $\mathcal{E}(y_i|\pi_i)$ equal to a constant. In this case there will be a domination result. The data is base salary data from Catholic schools (from the same states as above).

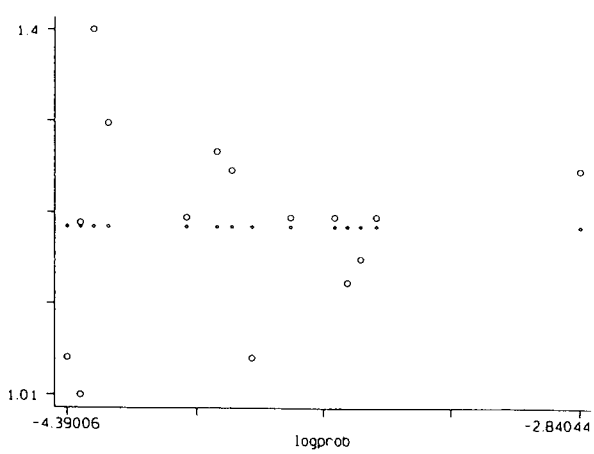

Figure 2. Adjusted Base Salary against $\log(\pi_i)$, Catholic Schools

Figure 2 shows a scatterplot of the response variable against the log of π_i. As can be seen the data is sparse but there appears to be no obvious relationship of y_i to π_i. There also appears to be heteroscedasticity which can be adjusted for (see Rizzo [1989]), but a homoscedastic model will continue to be assumed for simplicity.

In this situation $\hat{\bar{Y}}_P$ is the simple mean of the sample observations. $\hat{\bar{Y}}_{HT}$ and $\hat{\bar{Y}}_{NN}$ are both weighted means. All estimates have 0 bias under the null model, and $\hat{\bar{Y}}_P$ dominates both $\hat{\bar{Y}}_{HT}$ and $\hat{\bar{Y}}_{NN}$. The actual relative variances are 6.9×10^{-4}, 7.9×10^{-4}, and 1.08×10^{-3} for $\hat{\bar{Y}}_P$, $\hat{\bar{Y}}_{HT}$ and $\hat{\bar{Y}}_{NN}$ respectively. The efficiency gain of $\hat{\bar{Y}}_P$ is 13% in this situation over $\hat{\bar{Y}}_{HT}$. In a null model situation $\hat{\bar{Y}}_P$ must dominate in general.

An example on the other hand where $\hat{\bar{Y}}_{HT}$ will have a lower variance than $\hat{\bar{Y}}_P$ is from the urban public schools. A plot of y_i (the base salary) against $\log(\pi_i)$ is given in Figure 3 below with an estimated predictive conditional expectation (assumed quadratic).

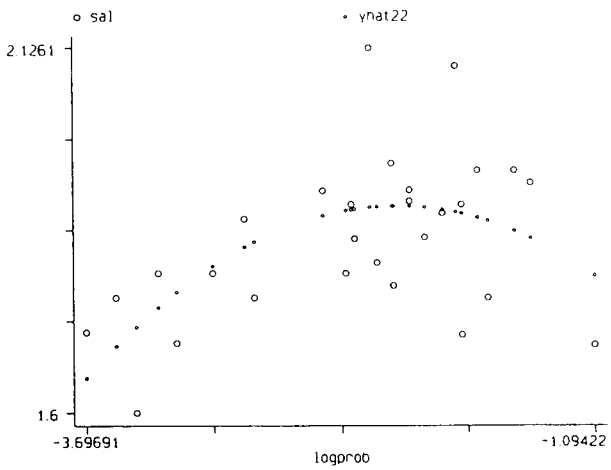

Figure 3. Adjusted Base Salary Against Log(π_i), Urban Public Schools

The point estimates and moment properties of the three estimators are given as follows:

	$\hat{\bar{Y}}_P$	$\hat{\bar{Y}}_{HT}$	$\hat{\bar{Y}}_{NN}$
Estimate	1.7256	1.8028	1.7885
Bias	0	+.0760	+.0906
Variance	1.49×10^{-3}	5.45×10^{-4}	1.30×10^{-3}
MSE	1.49×10^{-3}	5.80×10^{-3}	2.76×10^{-3}

(Note: the squared bias estimate that is part of the MSE estimate is a square of the bias estimate *minus* the variance of the bias estimate.) As can be seen the variance of $\hat{\bar{Y}}_P$ is much larger than the variance of $\hat{\bar{Y}}_{HT}$. Essentially there is an extrapolation problem: the smallest π_i in the sample was .0248, which was the 160^{th} smallest π_i of 670 in the population. The smallest population π_i was .0091 (In an independent sampling process there is approximately a 20% chance of failing to select 160 units in a row with $\pi_i = .01$, thus the occurrence is not incredible.) However it means we are extrapolating to these 160 population units with no data (this bad situation did not occur in the other two examples). The variance of $\hat{\bar{Y}}_{HT}$ is smaller simply because there are no 'small' π_i's in the sample, thus there are no larger weights w_i in the weighted average. $\hat{\bar{Y}}_{HT}$ essentially ignores the 160 units, but the effect of this shows in the large bias of +.076. The zero bias of $\hat{\bar{Y}}_P$ depends on the veracity of the quadratic model and the validity of the extrapolation to smaller π_i's than those in the sample, thus the zero squared bias is likely to be very optimistic. It would be fairer to say all three estimators have difficulties in this situation.

VII. Conclusions

The methodology arising from the reducing assumptions in section 2 and 3 generates a new series of estimators that are essentially model-based, but highly robust against model misspecification because of the partial reliance on the randomization distribution induced by the sampling design. Only the relationship $\mathcal{E}(Y|\Pi_\infty)$ needs to be specified, which follows directly from the concept of ignorability. With assumptions 1 through 3 one can justify the use of the simple predictive regression estimator $\hat{\bar{Y}}_P$, which will be more efficient in general than the classical estimators with little loss in robustness. The main source of lack of robustness is likely to be in the specification of the form of the predictive conditional expectation and variance, but this can be partially alleviated by use of nonparametric fitting methods. $\hat{\bar{Y}}_{NN}$ can also be used as an alternative to give extra robustness against misspecification.

This technique is used in this paper for descriptive mean estimation, but it can be used as a replacement for p-weighting (weighting by the inverse of π_i) wherever p-weighting is used in either descriptive or analytic estimation. It is likely that when the weak assumptions of this paper are satisfied a gain in efficiency will result over p-weighting without the loss of robustness often endured with straight model-based estimators.

VIII. Acknowledgements

A portion of this work was part of the author's thesis at the University of Chicago. The author wishes to acknowledge his thesis advisor, David Wallace, for his input and to acknowledge a pre-doctoral fellowship from the IBM Corporation. Also thanks to NORC for providing access to their data files.

[1]The base year data is available in tape form from NORC, 1140 E. 60th St., Chicago, IL 60637

IX. References

Cassel, C., C. E. Särndal, and J. H. Wretman (1977), *Foundations of Inference in Survey Sampling*, New York: John Wiley and Sons.

Craven, P. and G. Wahba (1979), "Smoothing Noisy Data with Spline Functions," *Numerische Mathematik*, **31**, 377–403.

David, M. H., R. Little, M. Samuhel, and R. Triest (1983), "Imputation Models Based on the Propensity to Respond," *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 168–173.

Hanson, M. H., W. G. Madow, B. J. Tepping (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys," *Journal of the American Statistical Association*, **78**, 776–807.

Little, R. J. A. (1982), "Models for Non-response in Sample Surveys," *Journal of the American Statistical Association*, **77**, 237–250.

Little, R. J. A. (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review*, **54**, 139–157.

Little, R. J. A. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley and Sons.

Rizzo, L. (1989), "Predictive Regression Estimators of the Finite Population Mean Using Functionss of the Probability of Selection," unpublished Ph.D. dissertation, University of Chicago, Statistics Department.

Rizzo, L. (1990), "Predictive Regression Estimators of the Finite Population Mean Using Functions of the Probability of Selections," Technical Note No. 130, University of Iowa, Dept. of Statistics and Actuarial Science.

Royall, R. M. (1976), "The Linear Least-Squares Prediction Approach to Two-Stage Sampling," *Journal American Statistical Association*, **71**, 657–664.

Rubin, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, **74**, 318–328.

Rubin, D. B. (1985), "The Case of Propensity Scores in Applied Bayesian Inference," in *Bayesian Statistics 2*, eds. I. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Elsevier Science Publishers (North- Holland).

Scott, A. J. (1977), "On the Problem of Randomization in Survey Sampling," *Sankhya*, **C39**, 1–9.

Scott, A. J. and T. M. F. Smith (1973), "Survey Design, Symmetry, and Posterior Distributions," *Journal of the Royal Statistical Society, B*, **35**, 57–60.

Silverman (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society, B*, **47**, 1–52.

Sugden, R. A. and T. M. F. Smith (1984), "Ignorable and Informative Designs in Survey Sampling Inference," *Biometrika*, **71**, 495–506.

Weisberg, S. (1985), *Applied Linear Regression*, New York: John Wiley and Sons.