Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 12, NUMBER 2
DECEMBER 1986

Canadä

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1986

Published under the authority of
the Minister of Supply and
Services Canada

March 1987
8-3200-501

Price: Canada, $10.00, $20.00 a year
Other Countries, $11.50, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 12, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $10.00 per copy, $20.00 per year in Canada, $11.50 per copy, $23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. (Reduced prices are available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.)

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 12, Number 2, December 1986

CONTENTS

Special Section - Missing Data in Surveys*

* The June 1986 issue was entirely devoted to selected papers presented at the Symposium on Missing Data in Surveys. Due to space limitations in the June issue, some symposium papers are included here.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 107–120
Statistics Canada

Estimating a Monthly Index Based on Trimestrial Data

JOHN G. KOVAR[1]

ABSTRACT

A problem of estimating monthly movements in rents based on data collected every four months is explored. Five alternative composite estimators of the rent index are presented and justified, both from an intuitive as well as theoretical point of view. An empirical study testing and comparing the proposed methods is described and summarized. Recommendations are put forth.

KEY WORDS: Index numbers; Rotating samples; Composite estimation.

1. INTRODUCTION

The rent component of the Consumer Price Index is based on data collected on a six month rotating basis using a Labour Force Survey Supplement. Since changes in rents generally occur on an annual basis, the effective sample size of the Labour Force Survey design is reduced. Furthermore, special annual benchmarks, which are obtained by revisiting the June sample of dwellings one year later, indicate that the rent component can suffer from varying degrees of bias (Dolson 1982). To ameliorate the situation, several data collecting schemes were proposed in order to combine the monthly data with the yearly benchmarks in a continuous and timely fashion. One of these methods, which collects data every four months, was selected for practical application.

The proposed design consists of four sets of four rotation groups of rented dwellings, each set of which is to be surveyed in one of four consecutive months, on a rotating basis. Each month, one rotation group is surveyed for the first time and the other three are those that rotated in four, eight and twelve months ago respectively. Each group would thus be surveyed four times over a period of thirteen months,before rotating out of the sample. Every month, data on current rents, as well as matched rents collected four months ago, are available from exactly three rotation groups (the fourth group is new and thus has no matching "backrents"). Yearly benchmarks can be calculated monthly based on one rotation group. This paper discusses several methods of estimating a monthly index based on such trimestrial data.

In estimating the indices, the constraints of the Consumer Price Index publication policy must be kept in mind. In other words, it must be practically as well as technically possible to produce the indices on a monthly basis for each of the index cities. The estimates must be timely: produced no later than mid-month following the reference month. Furthermore, no revisions can be made once the indices are published. While not entirely essential, it would be desirable that any proposed estimator be able to reflect (real) sudden changes in trend very quickly. On the other hand, in order to remain credible,the indices must be relatively stable: volatile, saw-toothed indices are to be avoided.

[1] John G. Kovar, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

In Section 2, five estimators will be presented, justified, and compared on a theoretical basis. Some empirical adjustments to these indices will be discussed in Section 3. In order to compare the performance of these estimators over time and between locations, a simulation study involving eight cities with observations over a period of 48 months was performed. The results of the study are presented in Section 4. The conclusions and recommendations can be found in Section 5.

2. INDEX ESTIMATORS

In this paper, only matched indices will be considered. While relative changes could easily be derived by comparing independent (unmatched) estimates of rent levels at distinct time points, such estimates of levels would have to be very reliable, necessitating prohibitively large sample sizes. Moreover, past studies indicate that such direct estimators tend to be volatile, upwardly biased and generally not practical in use (Szulc 1983). In what follows, therefore, an estimate of relative change between two time points will be based only on those units that report rents for both of these time points.

We will denote by x_m the total rent paid, in the current month m, by a certain subset s of dwellings in a given city. Thus, more rigorously,

$$x_m = \sum_{i \in s} x_{mi}, \tag{2.1}$$

where x_{mi} denotes the rent paid by the i-th dwelling in month m. The rent index is customarily estimated by chaining one month relatives, that is, the ratios of average rents between two consecutive months denoted by r^m_{m-1}. In other words, the index in month m, I_m, over a base period zero, is estimated recursively by

$$\hat{I}_m = \hat{I}_{m-1} \times \hat{r}^m_{m-1} = 100 \times \hat{r}^1_0 \times \hat{r}^2_1 \times \ldots \times \hat{r}^{m-1}_{m-2} \times \hat{r}^m_{m-1}. \tag{2.2}$$

where 100 is the (abritrary) level of the index at time zero. The difficulty then rests only in estimating the relatives.

In general, consider the relative change in rent in month m over month 1, denoted by r^m_1. This "m over 1 relative" can be estimated by

$$\hat{r}^m_1 = x_m / x_1. \tag{2.3}$$

However, if one considers matched indices only, the only estimable relatives under the proposed design are the four-month relatives, in other words, those of the form r^m_{m-4j}, $j = 1, 2, 3$, because it is only in these cases that there are common units between the two months. These relatives are estimated by

$$\hat{r}^m_{m-4j} = x_m / x_{m-4j}, \tag{2.4}$$

where the set s of dwellings consists of only those units that report rents at both time m and $m-4j$. Unfortunately, the interest lies in estimating monthly relatives of the form r^m_{m-1}. On the positive side, the rotation scheme ensures that a four-month relative is available every month. It is also assumed that units rotating out of the sample are replaced by equivalent units rotating into the sample. As such, the set s of common dwellings in (2.1) depends on the time m only and any future reference to it, while implicitly retained, can thus be suppressed in what follows. For a rigorous discussion of these assumptions and the effect on the index if the assumptions fail, the reader is invited to consult Szulc (1983) and Kovar (1984).

In the following paragraphs, five methods of estimating monthly relatives from four-month relatives will be described. Each will be justified intuitively as well as theoretically, and its advantages and disadvantages will be pointed out. The first three methods are derived on a theoretical basis alone while the fourth attempts to exploit the rotation pattern of the survey. All four assume that at least a four month back history of data is available. The last approach takes advantage of prior empirical knowledge: that of high probability of observing one change in rent per year. Methods two and four have been discussed earlier by Kovar (1984).

2.1 Interpolated Index (Additive Index)

One way of estimating the relative r^m_{m-1} is to estimate the previous month's rent, x_{m-1}. This can be accomplished, among other methods, by linearly interpolating the observed rents at time m and $m - 4$, that is, by assuming that the rents increase (decrease) linearly over time. Note that this assumption does not require each individual rent to increase every month by a fixed amount, but merely that the sum of all the rents does. In general, to describe linear interpolation briefly, consider two measurements of the same quantity at two distinct time points, say y_t and y_{t-s}. Suppose that we wish to estimate the value of y at some point between the times $t - s$ and t, say at time $t - u$ $(u < s)$. Assuming that the measurements increase linearly in time, y_{t-u} can be estimates from y_t and y_{t-s} by

$$y_{t-u} = (1 - \frac{u}{s}) y_t + \frac{u}{s} y_{t-s} \tag{2.5}$$

or in the case at hand, where $s = 4$ and $u = 1$, by

$$y_{t-1} = (3/4) y_t + (1/4) y_{t-4}. \tag{2.6}$$

Thus the previous month's total rent can be estimated by

$$x_{m-1} = (1/4) x_{m-4} + (3/4) x_m \tag{2.7}$$

and consequently, the monthly relative for month m by

$$r^m_{m-1} = \frac{x_m}{x_{m-1}} = \frac{4x_m}{x_{m-4} + 3x_m}. \tag{2.8}$$

The index is then derived by chaining the relatives as in (2.2) above.

Provided that the rents follow the linear interpolation model, that is, provided that we can write the current month's rent as a recursive function of previous months' rents, namely, as

$$x_m = x_{m-1} + d = x_0 + md, \tag{2.9}$$

then it can be shown that the index at time m is given by $I_m = x_m/x_0$, as is desired. In other words, if the data follow the model in (2.9), the index will suffer no time lags. But, of course, if the model were true at all times, the index would be fixed for all time points, based on any two observations. Since this is clearly not the case, one can at best use (2.8) as an approximation over short periods of time only. In that case, however, if the relationship in (2.9) is not exact, the index at time m will depend on all the rents between time -4 and m. In other words, the index is then susceptible to accumulating various biases over time.

Note that the same index would be derived by assuming that the four-month increment, $x_m - x_{m-4}$, occurred in 4 equal additive steps: $(x_m - x_{m-4})/4$. Since then, the previous month's rent would be estimated by

$$x_{m-1} = x_m - (x_m - x_{m-4})/4, \tag{2.10}$$

which is the same as (2.7); hence the alias: additive index.

2.2 Geometric Index

In this section, in contrast to the above, we will attempt to estimate the relative directly. We first note that

$$\begin{aligned} r^m_{m-4} &= \frac{x_m}{x_{m-4}} = \frac{x_m}{x_{m-1}} \frac{x_{m-1}}{x_{m-2}} \frac{x_{m-2}}{x_{m-3}} \frac{x_{m-3}}{x_{m-4}} \\ &= r^m_{m-1}\, r^{m-1}_{m-2}\, r^{m-2}_{m-3}\, r^{m-3}_{m-4}. \end{aligned} \tag{2.11}$$

We then assume that the four relatives on the right hand side of (2.11) are equal, or equivalently, that the four-month movement is due to four equal movements which act multiplicatively (Kosary *et al.* 1982). Under this assumption, the relationship (2.11) can be written as

$$r^m_{m-1} = (r^m_{m-4})^{1/4}. \tag{2.12}$$

From (2.2) and (2.3), assuming that there are no sample changes or that units rotating out of the sample are replaced by equivalent units rotating into the sample, the index in month m over the base period zero becomes

$$\begin{aligned} I_m &= I_0 \times r^1_0 \times r^2_1 \times \ldots \times r^m_{m-1} \\ &= I_0 \times (r^1_{-3})^{1/4} \times (r^2_{-2})^{1/4} \times \ldots \times (r^m_{m-4})^{1/4} \\ &= I_0 \frac{(x_{m-3}\, x_{m-2}\, x_{m-1}\, x_m)^{1/4}}{(x_{-3}\, x_{-2}\, x_{-1}\, x_0)^{1/4}} \end{aligned} \tag{2.13}$$

In other words, the index is a ratio of two geometric averages; hence the name geometric index. We note that at any time, assuming the panels are stationary, the index depends on eight months worth of data only, and thus is independent of any movements between time 0 and $m-4$, though in practice matched sets contributing to each r^m_{m-4} are different, so the cancellation is only theoretical. By contrast the index suffers from one-month to three-month lags and will thus tend to dampen true sudden changes. These changes, however, will be reflected eventually, that is, the index will selfcorrect (Kovar 1984).

As a point of clarification, note also that the relatives in (2.12) can be rewritten as

$$\frac{x_m}{x_{m-1}} = \left[\frac{x_m}{x_{m-4}}\right]^{1/4}$$

or as

$$x_{m-1} = (x_{m-4})^{1/4}\,(x_m)^{3/4}$$

or finally as

$$\log(x_{m-1}) = (1/4)\log(x_{m-4}) + (3/4)\log(x_m). \tag{2.14}$$

The geometric index is therefore equivalent to an index derived by estimating the previous month's rent by linearly interpolating the logarithms of the observed rents at time m and $m - 4$. (See (2.6) with $y_m = \log x_m$.)

2.3 Incremental Index

Analogous to the above geometric index, here we assume that the four consecutive monthly relative net increments are equal and acting additively. More precisely, we can write r_1^mas

$$r_1^m = 1 + i_1^m$$

where $i_1^m$is the relative net increment in month m over month 1. To estimate r_{m-1}^m we need therefore i_{m-1}^m. Assuming that the available $i_{m-4}^m = 4\,i_{m-1}^m$, the relative r_{m-1}^m can be estimated. Namely, we will estimate i_{m-1}^m by

$$i_{m-1}^m = (1/4)\,i_{m-4}^m = (1/4)\,(r_{m-4}^m - 1) = (1/4)\left(\frac{x_m}{x_{m-4}} - 1\right), \tag{2.15}$$

and r_{m-1}^m by

$$r_{m-1}^m = 1 + i_{m-1}^m = \frac{x_m + 3x_{m-4}}{4x_{m-4}}. \tag{2.16}$$

We note that $r_{m-1}^m = x_m/x_{m-1}$ and thus (2.16) can be written as

$$\frac{x_m}{x_{m-1}} = \frac{x_m + 3x_{m-4}}{4x_{m-4}}$$

or as

$$\frac{1}{x_{m-1}} = (1/4)\frac{1}{x_{m-4}} + (3/4)\frac{1}{x_m}. \tag{2.17}$$

In other words, the incremental index corresponds to one which would be derived by estimating the previous month's rent by linearly interpolating the reciprocals of the observed rents at time m and $m - 4$. (See (2.6) with $y_m = x_m^{-1}$.)

As is the case with the interpolated index, the incremental index will be independent of the intermediate observations only under the restrictive condition that the interpolation model be followed. In this case, analogous to (2.9), the model is

$$\frac{1}{x_m} = \frac{1}{x_0} + md. \tag{2.18}$$

However, in most real situations, the chained incremental index will depend on all the data between times -4 and m and therefore will be susceptible to various accumulating biases.

Since all three indices discussed to this point can be described in terms of linear interpolation of various functions of the observed rents, it is also possible to compare them theoretically. It can in fact be shown that the three indices are ordered in magnitude, from smallest to largest in the order of their presentation. That is, in an inflationary situation the interpolated index will always be smaller in absolute value than the geometric index which in turn will always be dominated by the incremental index. The reverse holds true when the trend is downward, that is, when prices are decreasing. As one referee pointed out, this phenomenon can be explained by noting that "the interpolated, geometric and incremental relatives are respectively the weighted arithmetic, geometric, and harmonic means of rent quotations four months apart. The standard relationship between these means explains the behaviour of the estimates in inflationary or deflationary times".

2.4 Carried Index (Arithmetic Index)

The carried index is constructed by taking advantage of the rotating sample at hand. Noting that all units reappear periodically in the sample, we construct the index by simply carrying each unit's rent value forward until a new observation is recorded. In this way all units on the file have a matching previous month's rent and thus the monthly relative, r^m_{m-1}, can be constructed in a straightforward manner. The obvious drawback is that the rent increases (decreases) are not recorded until observed. However, since all changes are eventually recorded, the index will selfcorrect (Kovar 1984) but will suffer from a mixture of one to three-month lags. Just as for the geometric index, sudden (real) changes will be dampened but the carried index will reflect them eventually.

On the technical side, we note that in computing the carried index for any given month one quarter of the observations on the file reflect a four-month movement, whereas three quarters of the observations are carried for one to three months and reflect no change. In fact, in month m we observe x_m and carry x_{m-1}, x_{m-2} and x_{m-3}. Similarly, in month $m-1$ we observe x_{m-1} and carry x_{m-2}, x_{m-3} and x_{m-4}. The monthly relative is therefore given by

$$r^m_{m-1} = \frac{x_m + x_{m-1} + x_{m-2} + x_{m\text{-}3}}{x_{m-1} + x_{m-2} + x_{m-3} + x_{m-4}}. \tag{2.19}$$

Chaining the relatives as in (2.2), and assuming again that the samples are stationary, we obtain the index for month m over the base period zero as

$$I_m = I_0 \frac{x_{m-3} + x_{m-2} + x_{m-1} + x_m}{x_{-3} + x_{-2} + x_{-1} + x_0}. \tag{2.20}$$

In other words, the index is a ratio of two arithmetic averages. Analogous to the geometric index, the carried index depends on eight months worth of data only, and thus is independent of the movements between time 0 and $m - 4$. As mentioned above, it too suffers from one to three-month lags, and therefore dampens sudden changes.

2.5 Annual Index

Empirical observations suggest that most units change rent once a year. One could therefore argue that yearly relatives are more stable than monthly relatives, since the distribution of individual monthly relatives will necessarily demonstrate two spikes, one around the annual relative and the other at 1. The rotation pattern of the proposed rent pilot (Kovar 1984) ensures that an annual relative be estimable every month, that is that r^m_{m-12} be available. To compute the annual index on a monthly basis, we note that for any chained index the following relationships hold:

$$I_m = r^m_{m-1} I_{m-1} \tag{2.21}$$

and

$$I_m / I_{m-12} = r^m_{m-12}. \tag{2.22}$$

From these relationships we obtain an expression for a monthly relative r^m_{m-1} as

$$r^m_{m-1} = r^m_{m-12} I_{m-12} / I_{m-1}. \tag{2.23}$$

These relatives can then be chained as above to produce an index. Since such a relationship is recursive, we need 12 months worth of indices to be able to "start up". One possibility that exists, is to define the index for the first 12 months, by analogy to the geometric index, as

$$I_k = (r^k_{k-12})^{k/12}, k = 1, 2, \ldots, 12. \tag{2.24}$$

As defined, the annual index is independent of intermediate changes. On the other hand it will be saw-toothed unless individual monthly sample sizes are large. This is due to the fact that consecutive monthly estimates are totally independent. Moreover, it must be noted that the lagging problem will be at least as serious in the case at hand as it is for the indices presented earlier.

3. ADJUSTMENTS

In this section, two adjustment procedures for the above indices will be discussed. First, because the first four indices suffer from one to three month lags, they will smooth out true, sharp peaks. From prior data, it has been observed that rent indices do exhibit sharp rises, in certain cities, with some regularity. To "correct" the smoothed out index, an empirical adjustment will be proposed. By contrast, due to the volatility of the annual index, a smoothing adjustment will also be proposed.

3.1 Empirical Adjustments

It is known, for example, that most rents in Montreal change in July. The first four indices discussed in the previous section would distribute this July change over July, August, September and October. One could however adjust the index in July to reflect a larger change and counter adjust it in the following three months. More precisely, the index could be multiplied by r^* in the reference month and then by $(r^*)^{-1/3}$ in each of the following three months. Since all the proposed indices are chained indices, in the third month after the reference month the four multipliers will offset each other, leaving no trailing biases. As for the choice of r^*, this will depend on continued empirical observations in each particular city.

It is to be noted that such adjustments must be performed in rare situations only and with great care. It is imperative that the particular situation be monitored, for it is not uncommon for such aberrations to disappear suddenly.

3.2 Smoothing

As a last effort in redeeming a volatile, saw-toothed index, one could consider smoothing it. Like the above adjustments, smoothing should be considered in rare and extreme situations only: in cases where no other alternative exists. The smoothing procedure we consider here involves averaging the index at time m with a linear extrapolation to time m of the smoothed index from time $m - 1$ and $m - 2$. One possible choice of the smoothed index at time m, S_m, is then given by

$$\begin{aligned} S_m &= I_m/2 + (2S_{m-1} - S_{m-2})/2 \\ &= S_{m-1} + (I_m - S_{m-2})/2. \end{aligned} \tag{3.1}$$

Since the smoothing operation basically projects past data into the future, the smoothed index will extend past trends and therefore introduce some lags. Moreover, the method is recursive and consequently could also introduce unwanted biases. Other smoothing methods could be considered, although the utility of smoothing an index that suffers from serious lags is questionable.

4. EMPIRICAL STUDY

The study described in the following paragraphs was initiated in order to test the performance over time of the proposed indices and adjustments. The study provides quantitative information on the ability of the indices to track the true index accurately. It supports the mostly heuristic observations made above and reinforces the theoretical ones.

4.1 The Population

The population of rented dwellings used in this study was designed to duplicate the real situation as closely as possible. For this purpose, the cities,their sizes, and their sample sizes were selected to correspond to those used by the Rent Component of the CPI. Since all real data on rents is available for periods of six months only, the needed thirteen months of data had to be simulated. Eight cities were chosen for this purpose. Some are large, some are small, some have periodic jumps in their indices, but all are CPI index cities and have sufficient amount of rent data available. Moreover, while some of the indices in these cities are strictly increasing, others are both increasing and decreasing.

Only the initial rents of all units (those collected when the unit rotated in) on the CPI rent database for the years 1979 to 1984 inclusive, for the eight cities mentioned above, were

Table 1

Average Sample Sizes (Distinct Units) and the Index at 8401
for Eight Cities Based on the Simulated Population

City	Average Monthly Sample Size	Index at 8401 (8001 = 100)
Halifax	51	144.3
Montreal	268	136.6
Ottawa	35	130.0
Toronto	170	130.4
Winnipeg	105	132.0
Edmonton	112	125.2
Calgary	97	123.5
Vancouver	105	130.5

retrieved. For each unit, twelve additional months worth of data were then simulated using the observed parameters. (This approach is operationally easier then simulating seven months of data in addition to the existing six.) More precisely, for each unit, first a decision was made whether or not a change in rent will occur sometime in the next twelve months. The probability of this event was set to be equal to the observed probability of a rent change in that particular city and year. Then, given that a change was to occur,the appropriate month was selected proportional to the observed incidence of rent changes, again specific to the city and month at hand. The actual amount of the rent change was assumed to be distributed normally with a fixed mean and variance. Robust estimates of these two parameters were obtained from the existing data for each city and each month.

All programming was done in SAS (Statistical Analysis System). The random numbers were generated using the routines RANUNI and RANNOR. The resulting population consists of eight cities and four years of fully rotated data (that is,discarding start up months). The average monthly sample sizes and the value of the simulated index for January 1984 (with Jan 1980 = 100) can be seen for each city in Table 1. The indices, calculated for each of the cities, resemble very closely those observed originally. In the following comparisons, the indices of the simulated population were taken to be the true reference points to be reproduced.

4.2 Comparison of Indices

For the purpose of calculating the indices, it was assumed that of the 13 available observations for each unit, only those for months 1, 5, 9 and 13 were actually observed. All calculations were then based on this (4/13) subsample.The five indices described above were calculated for each city and compared to the true index. All indices are fixed at 100 in January 1980. The empirical adjustment was tested with the Montreal, Halifax and Winnipeg data, for the month of July, January and October respectively. While the results for all the possible combinations of cities and indices are too numerous to include herein, they are available from the author. Some selected highlights will be put forth in the following paragraphs. While not exhaustive, they are hoped to be representative as well as indicative of the situation at hand.

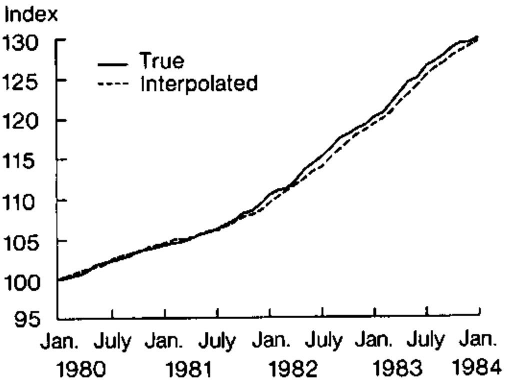


Figure 1. Plot of the True Index and the Interpolated Index for the City of Ottawa

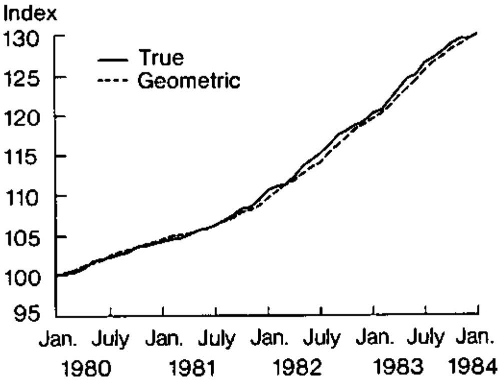


Figure 2. Plot of the True Index and the Geometric Index for the City of Ottawa

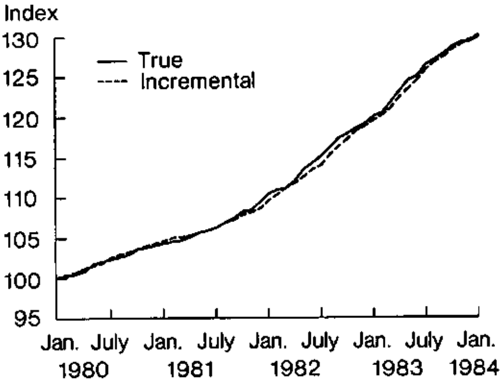


Figure 3. Plot of the True Index and the Incremental Index for the City of Ottawa

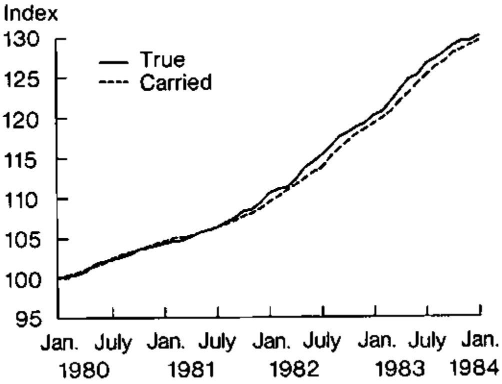


Figure 4. Plot of the True Index and the Carried Index for the City of Ottawa

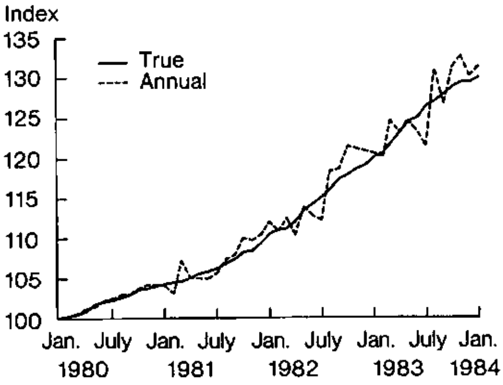


Figure 5. Plot of the True Index and the Annual Index for the City of Ottawa

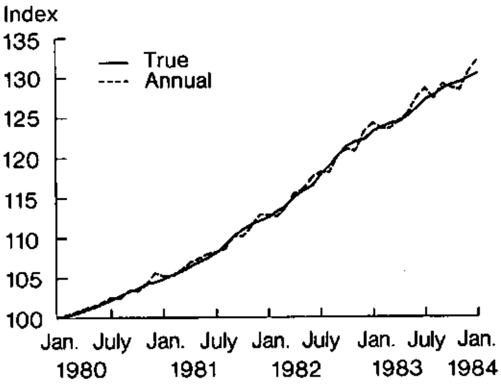


Figure 6. Plot of the True Index and the Annual Index for the City of Toronto

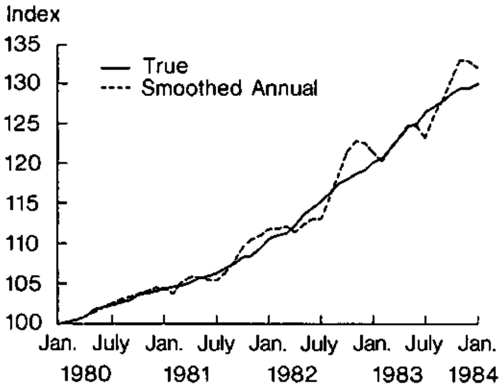


Figure 7. Plot of the True Index and the Smoothed Annual Index for the City of Ottawa

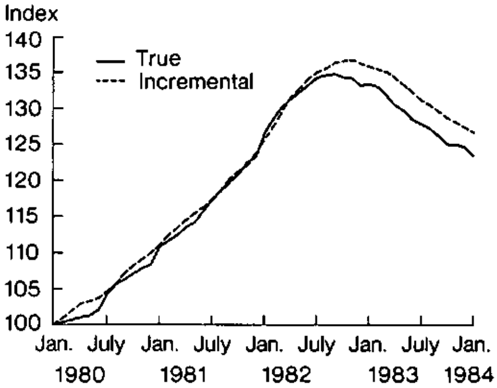


Figure 8. Plot of the True Index and the Incremental Index for the City of Calgary

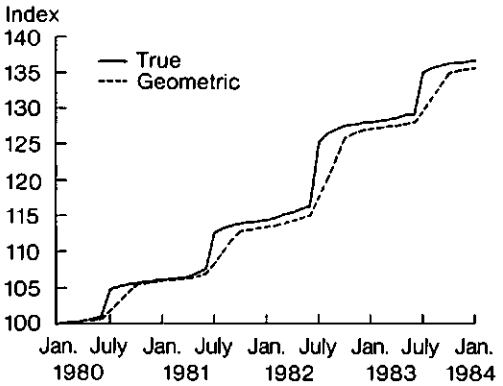


Figure 9. Plot of the True Index and the Geometric Index for the City of Montreal

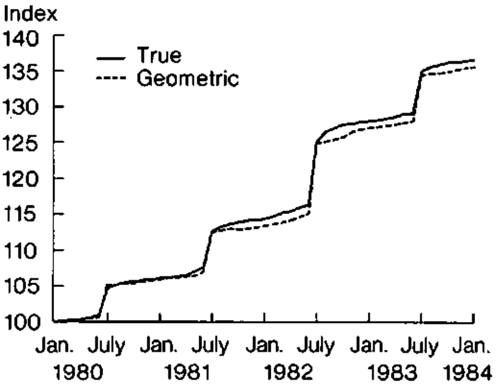


Figure 10. Plot of the True Index and the Adjusted Geometric Index for the City of Montreal

As can be seen in Figures 1-5, all five indices track the true index reasonably well, even in the case of small sample sizes such as in the city of Ottawa. As expected, the first four indices show some lags, those being more pronounced in the carried and interpolated index. (Note that the lagging problem could likely be accentuated by generating the population with exponentially increasing prices). Not surprisingly, the annual index is rather volatile. For cities with large sample sizes however,(e.g. Toronto), the annual index performs well (see Figure 6). While the smoothing adjustment of Section 3.2 does indeed smooth the index, the results are less than satisfactory as can be seen in Figure 7 (c.f. Figure 5). Perhaps a larger number of points should be used for the extrapolation but then the lagging problem would be even more pronounced. Figure 8 further demonstrates how sudden unexpected changes in trends are reported with a delay. However, expected jumps in the index (as in July in Montreal, Figure 9) can be adjusted successfully using the adjustment procedure of Section 3.1 (Figure 10).

Table 2

Mean Square Errors of Five Indices in Eight Cities

City	Interpolated		Geometric		Incremental		Carried		Annual	
Halifax	30*	(3)	19*	(2)	12*	(1)	48	(4)	74	(5)
Montreal	48*	(3)	24*	(2)	9*	(1)	160	(5)	82	(4)
Ottawa	17	(3)	12	(2)	8	(1)	22	(4)	95	(5)
Toronto	36	(4)	27	(3)	20	(2)	29	(5)	13	(1)
Winnipeg	27*	(3)	17*	(2)	10*	(1)	66	(5)	41	(4)
Edmonton	46	(1)	64	(4)	88	(5)	55	(3)	50	(2)
Calgary	56	(2)	81	(4)	121	(5)	64	(3)	46	(1)
Vancouver	70	(5)	53	(2)	39	(1)	64	(4)	60	(3)

Note: 1. Bracketed figures indicate ranking within cities.
2. Starred figures are results of adjusted indices as per Section 3.1.

Mean square errors of the five indices away from the true index have been calculated for each city (Table 2). The three interpolation based indices (interpolated, geometric and incremental) have been adjusted for the cities of Montreal, Halifax and Winnipeg. Table 2 also presents the rankings (from smallest to largest) of the mean square errors of the five indices within each city. The carried and the annual index tend to perform the worst. The three interpolation-based indices perform relatively alike. In general, in cities where the index is climbing consistently, the performance of these three indices worsens in the order: incremental, geometric, interpolated. The order is reversed in cities where sharp decreases in the index have been observed. It is unlikely, however, that the strategies could be interchanged based on observed behaviours only.

5. SUMMARY

Both the theoretical as well as the empirical observations suggest that the yearly index is too volatile in cities where sample sizes are not large enough. Smoothing, at least of the type described, has proven fruitless. For this reason the annual index should be reserved only for those rare cases where sample sizes permit. On the other hand, the annual index could be used in conjunction with one of the more stable four-month indices to produce a composite estimate analogous to that proposed by Kosary *et al.* (1982). However, empirical observations would be needed to determine the appropriate weights to be used in averaging the two indices.

By contrast, the carried, and to some degree, the interpolated index tend to be too smooth. That is they tend to smooth out all peaks in addition to demonstrating a one or two (index) point lag. While the incremental and geometric indices are not entirely free of these lags, they tend to track the true index a little more closely. The incremental index performs the best overall, however, because of the mathematical "cleanliness"of the geometric index (i.e. its theoretical independence of its history and its correspondence to the chaining structure), it is the latter that is recommended here. In other words, the geometric index does not retain terms that could cause biases in the long run.

It is also apparent that whenever possible, prior knowledge can be used to improve the index. Empirical adjustments as described in Section 3.1 can be useful, provided that they are well founded. If their use is contemplated, it is imperative that the empirical knowledge that leads to their application be monitored and its continued existence verified.

ACKNOWLEDGEMENT

Thanks are due to Mr. George Sampson for all of his help and to the referees and the editorial staff for their constructive comments.

REFERENCES

DOLSON, D.D. (1982). Rent status survey: Analysis. Technical Report, Statistics Canada.

KOSARY, C.L., BRANSCOME, J.M. and SOMMERS, J.P. (1982). Evaluating alternatives to the rent estimator. Technical Report, Bureau of Labor Statistics.

KOVAR, J. (1984). Note on calculating the rent index. Technical Report, Statistics Canada.

SZULC, B. (1983). Linking price index numbers. Technical Report, Statistics Canada.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 121–138
Statistics Canada

Regression Analysis Using Survey Data with Endogenous Design

ARIE TEN CATE[1]

ABSTRACT

This paper discusses the influence of the sampling design on the estimation of a linear regression model. Particularly, sampling designs will be discussed which are dependent on the values of the endogenous variable in the population: endogenous (or "informative") designs. A consistent estimator of the regression coefficients is given. Its variance is the sum of a sampling design component and a disturbance term component. Also, model-free regression is briefly discussed. The model-free regression estimator is the same as the model estimator in the case of an endogenous design.

KEY WORDS: Regression; Survey sampling; Endogenous design.

1. INTRODUCTION

The heart of any statistical model is the assumption that the value of one or more variables is generated by drawing from some probability distribution; for example, a regression model with normally distributed disturbances. In this paper a finite set of elements which behave according to such a model will be considered. This set is called the population. Next, a sample is drawn from this population, without replacement. The subject of this paper is the influence of the sampling design on the estimation of the parameters of the model. This influence depends mainly on whether the design is exogenous or endogenous with respect to the model. In the case of an endogenous (or "informative") design, the sampling probabilities depend on the value of the endogenous ("dependent") variables. Then, the design should not be ignored in the estimation of the model parameters. The nature of the problem is indicated in Figure 1, where a stratified sampling design is shown. There are 3 strata, defined in the endogenous variable of a regression model. The middle stratum has a higher sampling fraction than the other two. The diagram shows that the slope of the regression line estimated using the sampled data points only, is biased downwards if one ignores the design. This bias does not vanish in large samples. This can be seen in an intuitive manner by imagining that every white and black dot in Figure 1 denotes a large number of identical data points. Even if this large number tends to infinity, the slope of the estimated regression line will be biased downwards, because the shape of the scatter will remain the same.

There is a rapidly growing body of literature on the application of regression techniques in finite population sampling. This literature deals with a variety of problems. One problem is, how to use regression techniques in order to estimate a finite population total. Another problem concerns the estimation of population parameters such as $\Sigma xy / \Sigma x^2$, where the summation runs over all elements of the finite population. Reviews of the literature about these problems are given by Nathan (1981) and Smith (1981). A third problem is the estimation of the parameters of a regression model, using a sample from a finite population. This problem can be solved relatively easily in the case of a exogenous design. See Porter (1973, Section 1.2), DuMouchel and Duncan (1983), and textbooks such as Cramer (1971, p. 143). Texts

[1] Arie ten Cate, Central Planning Bureau, 2585 JR 's-Gravenhage, Van Stolkweg 14, The Hague, The Netherlands.

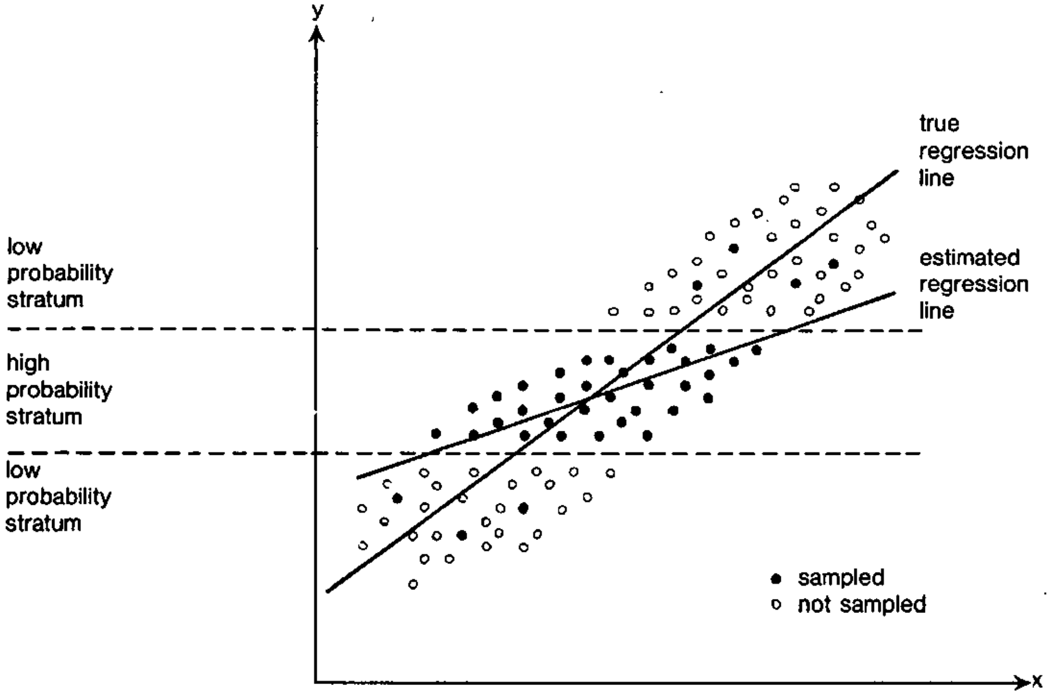


Figure 1. The Effect of Endogenous Stratification on the Estimated Regression Line

such as Kmenta (1971, Section 8.3) and Johnston (1972, Section 9.2) discuss the closely related topic of stochastic regressors. See also White (1980a) for non-linear regression. Our topic, regression analysis with endogenous design, is more complicated. Hausman and Wise (1981) discuss stratified endogenous designs in a very simple case: two strata and a regression model consisting of a constant term only. Jewell (1985) gives some iterative estimators for the case of endogenous stratification.

Regression analysis with endogenous design is related to the problem of endogenous non-response in regression analysis (see Heckman (1979)). However, we have a lesser problem here, since the probabilities involved in the sampling process are assumed to be known: they constitute the chosen design. On the other hand, as we shall see in Subsection 6.1, variance estimation with an endogenous design is in general rather difficult.

Regression analysis with endogenous design may be compared with logit analysis with endogenous design, also called logit analysis with choice based sampling or case-control sampling. See Manski and McFadden (1981, Chapters 1 and 2) and Breslow and Day (1980, Section 6.3).

The contents of the rest of the paper are as follows. In Sections 2 and 3 the main theorems are given. These theorems give a consistent estimator of the parameters of a linear regression model, using a sample with an endogenous design. Consistency is defined here in a similar way as in the discussion of the bias in the example above, though slightly more subtle: the x-values are replicated a large number of times and the y-values behave according to the regression model. In Sections 4 and 5 the variance of the estimator of the regression coefficients is studied. Section 6 discusses the estimation of this variance. Section 7 deals with model-free regression, Section 8 discusses the various motives for weighted regression and finally, Section 9 concludes the paper.

2. THE MODEL, THE SAMPLE AND A REGRESSION ESTIMATOR

In this section the asymptotic properties of an estimator of a regression model are studied within the framework of finite population sampling without replacement. Asymptotic theory for samples drawn without replacement from a finite population may seem a contradiction since such a sample must be bounded. This contradiction is solved by increasing both the population size and the sample size, without bound, at the same rate. The dependence between the inclusions of population elements in the sample constitutes another problem, especially in the case of complex sampling designs. Here we use an idea of Brewer (1979). In Brewer's system, limit theorems on sequences of independent variables can be used, while the results may still be applied to complex designs. Basically, this system consists of the replica idea already introduced informally above. This replica idea will be used extensively throughout the rest of this paper. For another approach, see Robinson (1982).

First, the structure of the population and the model are given. Consider a finite set of N_0 elements. Each element has r real-valued exogenous non-stochastic characteristics, together forming an $(N_0 \times r)$-matrix X_0. One of the fundamental assumptions of this paper is the following. The population consists of K replicas of this set of N_0 elements, having $N \equiv KN_0$ elements. Its matrix of exogenous variables is X, with

$$X = \iota_K \otimes X_0. \tag{1}$$

Here, ι_K is the K-vector with all elements equal to unity and $\otimes$ denotes the Kronecker matrix product. Aymptotic results will be derived by allowing K to tend to infinity.

The model assumptions describe the standard linear model. Each of the N elements of the population has a score on a stochastic, endogenous, variable. Together they form an N-vector y. It is assumed that

$$E_\xi(y) = X\beta \tag{2}$$

for some fixed, unknown r-vector β. E_ξ denotes the expectation over all $y \epsilon R^N$. Next we define

$$\varepsilon = y - X\beta. \tag{3}$$

It is assumed that the N elements of ε are i.i.d. It follows from (2) that all elements of ε have expectation zero. Their variance is σ^2, that is,

$$E_\xi(\varepsilon\varepsilon') = \sigma^2 I. \tag{4}$$

Sampling is done without replacement here, as is common practice. The sample is described by a diagonal $(N \times N)$-matrix T, such that

$$t_{ii} = \begin{cases} 1 \text{ if population element } i \text{ is in the sample} \\ 0 \text{ otherwise} \end{cases}$$

for all $i = 1, ..., N$. Obviously, T is idempotent. The sample space S is the set of all such matrices T. This set is finite. The sampling design is some probability distribution over the elements of the sample space S. The sampling design is endogenous here, meaning that it depends on y. Hence, the sampling design itself is stochastic. (A design which does not depend on y is called exogenous, or uninformative.) Let T be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let T_k be the k-th diagonal block, related to the k-th replica. Similarly, let y be partitioned in K N_0-vectors, such that $y' = (y_1', y_2', ..., y_k', ..., y_K')$. It is assumed that the sampling design depends on y in the following sense: the K pairs $(T_1, y_1), ..., (T_K, y_K)$ are i.i.d.

The expectation over all elements of S, conditional on y (or ε), plays an important role in this paper. It is denoted by E_p. Then we define

$$\Pi \equiv E_p(T). \tag{5}$$

It is assumed that Π is known. The diagonal elements of Π are called inclusion probabilities: the probabilities that the population elements are included in the sample. The matrix Π is partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. Let Π_k be the k-th diagonal block, related to the k-th replica. Note that each Π_k is stochastic because it depends on y_k. By the above assumption, the $\Pi_1, ..., \Pi_k$ are i.i.d. The dependence of the Π_k on y is denoted by a function F, such that

$$\Pi_k = F(y_k) \tag{6}$$

for all $k = 1, ..., K$. It is assumed that $F(y_k)$ is non-singular for every y_k. In other words, the inclusion probabilities are always positive.

This framework and Brewer's (1979) differ in somewhat. Brewer has no endogenous variables and therefore all his Π_k are nonstochastic and equal. One may also compare this approach with the idea of "constant in repeated samples" in the econometric literature; see e.g. Theil (1971, p. 364).

The stage is now set for the estimation of β. The stochastic properties of estimators will be considered over all pairs $(y, T) \in (R^N \times S)$. The corresponding expectation will be denoted by $E_\xi E_p$. We shall consider a generalized least square estimator of β, say $\hat{\beta}$, with weights equal to the square roots of the inclusion probabilities, as follows,

$$\begin{aligned} \hat{\beta} &\equiv [(\Pi^{-1/2}X)'T(\Pi^{-1/2}X)]^{-1}(\Pi^{-1/2}X)'T(\Pi^{-1/2}y) \\ &= (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty. \end{aligned} \tag{7}$$

Recall that the matrix Π is known. Note that X and y relate to the population, but T effectuates summation over the sampled elements. As an alternative to considering $\hat{\beta}$ as a generalized least squares estimator, assume that all elements of Π^{-1} are integer numbers. Then, if each observation i in the sample is copied π_{ii}^{-1} times, $\hat{\beta}$ is the ordinary least squares estimator applied to this inflated sample. In this view, no square roots of the probabilities are involved. See also Hausman and Wise (1981, p. 373). The main theorem of this paper is:

Theorem 1. Under the assumptions made above ((1), (2) and the distribution of ε and T), the generalized least squares estimator $\hat{\beta}$, defined in equation (7) is consistent for $K \to \infty$.

The rest of the section is devoted to the proof of this theorem. The following lemma will be used in this proof and the proof of subsequent theorems.

Lemma 1. Consider an N-vector z, such that $z = \iota_k \otimes z_0$, where z_0 is some fixed N_0-vector. Consider also an N-vector η, partitioned such that $\eta' = (\eta_1', \eta_2', \ldots, \eta_K')$. Each η_k has N_0 elements. Assume that each η_k is a function of X_0, β and ε_k, all functions being the same. Then

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}z'\Pi^{-1}T\eta\right) = z_0'E_\xi(\eta_0), \tag{8}$$

where $E_\xi(\eta_0)$ is the expectation of any η_k, being equal for all k.

Proof of lemma 1: Consider the expectation of $\Pi_k^{-1}T_k\eta_k$:

$$E_\xi E_p(\Pi_k^{-1}T_k\eta_k) = E_\xi[\Pi_k^{-1}E_p(T_k)\eta_k] = E_\xi(\eta_k), \tag{9}$$

for all k. Since the distribution of η_k is the same for each k, one may write

$$E_\xi E_p(\Pi_k^{-1}T_k\eta_k) = E_\xi(\eta_0) \tag{10}$$

for all k. Also, the K vectors $z_0'\Pi^{-1}T_k\eta_k$ are i.i.d. Thus, Khintchine's theorem applies as follows,

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}z'\Pi^{-1}T\eta\right) = \operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\sum_k z_0'\Pi_k^{-1}T_k\eta_k\right) = E_\xi E_p(z_0'\Pi_1^{-1}T_1\eta_1)$$

$$= z_0'E_\xi E_p(\Pi_1^{-1}T_1\eta_1). \tag{11}$$

Substitution of (10) in (11) gives the lemma. The proof of theorem 1 is now straightforward.

Proof of theorem 1: The generalized least squares estimator of the theorem can be written as

$$\hat\beta = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty = \beta + (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}T\varepsilon. \tag{12}$$

Thus,

$$\operatorname*{plim}_{K\to\infty}\hat\beta = \beta + \left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right)\right]^{-1}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}T\varepsilon\right)$$

$$= \beta + (X_0'X_0)^{-1}X_0'0 = \beta. \tag{13}$$

The expression $X_0'X_0$ is formed by repeated application of lemma 1, substituting the columns of X for both z and η. Notice that $E_\xi(X_0) = X_0$ since X_0 is a constant. The expression $X_0'0$ is formed by repeated application of lemma 1, substituting the columns of X for z and ε for η.

3. THE ESTIMATION OF THE DISTURBANCE VARIANCE

The regression model described in Section 2 has two parameters: β and σ^2. Theorem 1 considered estimation of β; in this section the estimation of σ^2 will be considered. The result of this section is given in the following theorem.

Theorem 2. The disturbance variance σ^2 is estimated consistently by the weighted sample variance of the residuals of y if these weights are equal to the inverse of the square root of the inclusion probabilities.

Proof: The variance estimator of the theorem is

$$\hat{\sigma}^2 = (\iota_N'\Pi^{-1}T\iota_N)^{-1}\tilde{e}'\tilde{e} \tag{14}$$

with

$$\tilde{e} \equiv \Pi^{-1/2}T(y - X\hat{\beta}). \tag{15}$$

Let

$$\tilde{y} \equiv \Pi^{-1/2}Ty, \tag{16}$$

$$\tilde{X} \equiv \Pi^{-1/2}TX, \tag{17}$$

and

$$\tilde{\varepsilon} \equiv \Pi^{-1/2}T\varepsilon. \tag{18}$$

Then

$$\tilde{e} = \tilde{y} - \tilde{X}\hat{\beta} = \tilde{y} - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \tag{19}$$

and

$$\begin{aligned}\tilde{e}'\tilde{e} &= \tilde{y}'[I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}']\tilde{y} = (\tilde{X}\beta + \tilde{\varepsilon})'[I_N - \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'](\tilde{X}\beta + \tilde{\varepsilon}) \\ &= \tilde{\varepsilon}'\tilde{\varepsilon} - \tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}.\end{aligned} \tag{20}$$

The first term in the right-hand side (RHS) of (20) converges in probability as follows

$$\begin{aligned}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{\varepsilon}'\tilde{\varepsilon}\right) &= \operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\varepsilon'\Pi^{-1}T\varepsilon\right) = \operatorname*{plim}_{K\to\infty}\left[\frac{1}{K}\iota_N'\Pi^{-1}T\,\mathrm{diag}(\varepsilon)\varepsilon\right] \\ &= \iota_{N_0}'(\sigma^2\iota_{N_0}) = N_0\sigma^2.\end{aligned} \tag{21}$$

Here, diag(ε) indicates the diagonal matrix with as the diagonal. Lemma 1 has been applied with ι_N substituted for z and diag(ε)ε for η, using model equation (4). Next, consider the second term in the RHS of (20).

$$\operatorname*{plim}_{K\to\infty}\left[\frac{1}{K}\tilde{\varepsilon}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\varepsilon}\right]$$

$$= \left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{\varepsilon}\right)\right]'\left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{X}\right)\right]^{-1}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{X}'\tilde{\varepsilon}\right)$$

$$= \left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}T\varepsilon\right)\right]'\left[\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right)\right]^{-1}\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}T\varepsilon\right)$$

$$= 0'(X_0'X_0)^{-1}0 = 0. \tag{22}$$

In the derivation of (22), use has been made of lemma 1 in the same manner as in the derivation of (13). The combination of (20), (21) and (22) gives

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\tilde{e}'\tilde{e}\right) = N_0\sigma^2. \tag{23}$$

Finally, lemma 1 is applied to the first factor in (14), with ι_N substituted both for z and η. This gives

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}\iota_N'\Pi^{-1}T\iota_N\right) = N_0. \tag{24}$$

With (23) and (24) we have

$$\operatorname*{plim}_{K\to\infty}(\hat{\sigma}^2) = \sigma^2, \tag{25}$$

which proves the theorem. Finally it may be useful to note, as a corollary of (23), that

$$\left(\frac{1}{N}\right)\tilde{e}'\tilde{e} \tag{26}$$

is also a consistent estimator of σ^2.

4. THE VARIANCE OF $\hat{\beta}$

In this section the asymptotic variance of the estimator $\hat{\beta}$ is given.

Theorem 3. The asymptotic variance of $\hat{\beta}$ is given by

$$\operatorname{Var}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \tag{27}$$

with

$$V \equiv E_\xi\left[\operatorname{diag}(\varepsilon)\Pi^{-1}P\Pi^{-1}\operatorname{diag}(\varepsilon)\right], \tag{28}$$

and

$$P \equiv E_p(T\iota\iota'T). \tag{29}$$

The elements of P are the so-called second order inclusion probabilities: the probability for any pair of elements of the population of being included in the sample. The diagonal of P is equal to the diagonal of Π. The rest of this section is devoted to a proof of this theorem.

Proof: Consider the asymptotic distribution for $K \to \infty$ of

$$\begin{aligned} K^{½}(\hat{\beta} - \beta) &= K^{½}[(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}Ty - \beta] \\ &= K^{½}(X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}T\varepsilon. \end{aligned} \tag{30}$$

Since

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right) = X_0'X_0, \tag{31}$$

the asymptotic distribution of $K^{½}(\hat{\beta} - \beta)$ is equal to the asymptotic distribution of δ, with

$$\delta \equiv K^{-½}(X_0'X_0)^{-1}X'\Pi^{-1}T\varepsilon = K^{-½}(X_0'X_0)^{-1}\sum_k X_0'\Pi_k^{-1}T_k\varepsilon_k = K^{-½}\sum_k \delta_k, \tag{32}$$

and

$$\delta_k \equiv (X_0'X_0)^{-1}X_0'\Pi_k^{-1}T_k\varepsilon_k, \tag{33}$$

for all $k = 1, ..., K$. (See e.g. Rao (1973), p. 122). Since the vector δ_k ($k = 1, ..., K$) are i.i.d. and also

$$\begin{aligned} E_\xi E_p(\delta_k) &= (X_0'X_0)^{-1}X_0'E_\xi E_p(\Pi_k^{-1}T_k\varepsilon_k) \\ &= (X_0'X_0)^{-1}X_0'E_\xi[\Pi_k^{-1}E_p(T_k)\varepsilon_k] \\ &= (X_0'X_0)^{-1}X_0'E_\xi(\varepsilon_k) = 0, \end{aligned} \tag{34}$$

the variance of δ, say Var(δ), is equal for all K and also equal to the variance of the asymptotic distribution of δ for $K \to \infty$. This variance can be written as

$$\operatorname{Var}(\delta) = E_\xi E_p(\delta_k\delta_k') \tag{35}$$

for any $k \in \{1, ..., K\}$. Since the vectors δ_k are i.i.d. this may be rewritten as

$$\begin{aligned}
\text{Var }(\delta) &= \frac{1}{K}\sum_k E_\xi E_p(\delta_k\delta_k')\\
&= \frac{1}{K}(X_0'X_0)^{-1}\left[E_\xi E_p\left(\sum_k X_0'\Pi_k^{-1}T_k\varepsilon_k\varepsilon_k'T_k\Pi_k^{-1}X_0\right)\right](X_0'X_0)^{-1}\\
&= K(X'X)^{-1}\,[E_\xi E_p(X'\Pi^{-1}T\varepsilon\varepsilon'T\Pi^{-1}X)\,]\,(X'X)^{-1}\\
&= K(X'X)^{-1}X'\,\{E_\xi E_p\,[\text{diag}\,(\varepsilon)\Pi^{-1}T\iota\iota'T\Pi^{-1}\text{diag}\,(\varepsilon)\,]\,\}X(X'X)^{-1}\\
&= K(X'X)^{-1}X'\,\{E_\xi\,[\text{diag}\,(\varepsilon)\Pi^{-1}E_p(T\iota\iota'T)\Pi^{-1}\text{diag}(\varepsilon)\,]\,\}X(X'X)^{-1}. \qquad (36)
\end{aligned}$$

Division of $\text{Var}(\delta)$ by K gives $\text{Var}(\hat{\beta})$ and completes the proof.

5. A DECOMPOSITION OF VAR $(\hat{\beta})$

The variance formula (27) can be rewritten as

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1} + (X'X)^{-1}X'V^*X(X'X)^{-1} \qquad (37)$$

with

$$V^* \equiv E_\xi\,[\text{diag}(\varepsilon)\,(\Pi^{-1}P\Pi^{-1} - \iota\iota')\,\text{diag}(\varepsilon)\,], \qquad (38)$$

using (4). The first term in the RHS of (37) might reasonably be called the ξ-component of the variance of $\hat{\beta}$. This component would contain all the variance of $\hat{\beta}$ if the whole population was sampled. It is entirely due to the variation in the disturbance ε and it is the familiar expression for that case. The second term in the RHS of (37) might be called the p-component of the variance of $\hat{\beta}$. This component contains the matrices Π and P, which describe the sampling design. This component looks like the variance formula of the estimator of a total or average of a finite population. The theory of such estimators will be discussed briefly in the rest of this section, as an aid in the interpretation of the p-component of $\text{Var}(\hat{\beta})$.

Consider a finite population of N elements. (No replica structure is assumed here). Each element of this population has a score on some real non-stochastic variable, collected in an N-vector x. From this population a sample without replacement is taken. The sample is described by the diagonal matrix T, as before. Also as before,

$$\Pi \equiv E_p(T) \qquad (39)$$

and

$$P \equiv E_p(T\iota\iota'T), \qquad (40)$$

the first order and second order inclusion probabilities, respectively. There is no regression model here, so Π and P are fixed known matrices. Horvitz and Thompson (1952) suggested to estimate the population total $X'\iota$ by

$$\hat{X} = x'\Pi^{-1}T\iota \qquad (41)$$

Obviously this is an unbiased estimator, in view of (39). The variance of $\hat{X}$ is

$$\begin{aligned}\operatorname{Var}(\hat{X}) &= E_p(\hat{X}^2) - [E_p(\hat{X})]^2 = E_p(x'\Pi^{-1}T\iota\iota'T\Pi^{-1}x) - x'\iota\iota'x \\ &= x'(\Pi^{-1}P\Pi^{-1} - \iota\iota')x. \end{aligned} \tag{42}$$

The last member of equation (42) is the variance formula of the Horvitz-Thompson estimator, which can be found in textbooks on sampling, such as Cochran (1977), though usually not in matrix format. The expression in parentheses in the last member of (42) is equal to the expression in parentheses in (38), the definition of V^*. The latter is contained in the formula of the p-component of Var($\hat{\beta}$). Thus, the diagonal elements of the p-component of the variance matrix Var($\hat{\beta}$) can be considered as the ξ-expectation of the p-variance of the Horvitz-Thompson estimator of the row totals of $(X'X)^{-1}X'$ diag (ε). These totals are the elements of the vector $(X'X)^{-1}X'\varepsilon$.

6. THE ESTIMATION OF VAR($\hat{\beta}$)

6.1 The General Case

In this section the estimation of the asymptotic variance Var($\hat{\beta}$) is considered. Consistent estimation of Var($\hat{\beta}$) is rather difficult, since this requires knowledge of the relationship F between y and the sampling design, as it appears in the matrix V. In practice, only the sampling design for the actual values of y will be known. In general, it is difficult to tell from this design only, what the design would be like if y took on different values. In a sense not only a regression model is involved, but also a model of the designer himself!

For the moment we assume that the function F is known, and therefore V is a known function of X and the parameters of the model. (See Subsection 6.2 for a special case). This is expressed as follows.

$$V = V(\beta,\sigma^2;X), \tag{43}$$

It is assumed that $V(\beta,\sigma^2;X)$ is a continuous function. For the sake of brevity, $\hat{V}$ is defined as

$$\hat{V} \equiv V(\hat{\beta},\hat{\sigma}^2;X), \tag{44}$$

where and $\hat{\beta}$ and $\hat{\sigma}^2$ are consistent estimators of β and σ^2 respectively. The rest of this subsection gives a theorem on consistent variance estimation, and its proof. Consistent estimation of Var($\hat{\beta}$) by vâr($\hat{\beta}$) is interpreted here as follows:

$$\operatorname*{plim}_{K\to\infty} K\,\widehat{\operatorname{var}}(\hat{\beta}) = \lim_{K\to\infty} K\operatorname{Var}(\hat{\beta}). \tag{45}$$

Theorem 4. Under the assumptions made above, the asymptotic variance Var($\hat{\beta}$) is estimated consistently by

$$\widehat{\operatorname{var}}(\hat{\beta}) = (X'\Pi^{-1}TX)^{-1}X'T\left(\frac{\hat{V}}{P}\right)TX(X'\Pi^{-1}TX)^{-1}, \tag{46}$$

where $(\hat{V}/P)$ denotes the matrix consisting of the elements of $\hat{V}$ divided by the corresponding elements of P.

Proof: First the structure of V will be considered. Let V be partitioned in a square $K \times K$ array of $(N_0 \times N_0)$ blocks. The (k, r)-th off-diagonal block of V is equal to

$$\begin{aligned} & E_\xi\,[\mathrm{diag}(\varepsilon_k)\,\Pi_k^{-1}E_p(T_k\iota\iota' T_r)\,\Pi_r^{-1}\mathrm{diag}(\varepsilon_r)\,] \\ &= E_\xi\,[\mathrm{diag}(\varepsilon_k)\,\Pi_k^{-1}E_p(T_k)\,\iota\iota' E_p(T_r)\,\Pi_r^{-1}\mathrm{diag}(\varepsilon_r)\,] \\ &= E_\xi\,(\varepsilon_k\varepsilon_r') \;=\; 0, \end{aligned} \tag{47}$$

using the assumed replica structure of the population and the sampling design. The diagonal blocks of V are identical and depend on X_0. Thus, $V(\beta,\ \sigma^2;\ X)$ can be written as

$$V(\beta,\sigma^2;X) \;=\; I_K \otimes V_0(\beta,\sigma^2;X_0), \tag{48}$$

where $V_0(\beta,\ \sigma^2;\ X_0)$ is an $N_0 \times N_0$ matrix function. Together with (1), equation (48) can be used to rewrite $K\mathrm{Var}(\hat{\beta})$ as follows.

$$K\mathrm{Var}(\hat{\beta}) \;=\; (X_0'X_0)^{-1}X_0'\,V_0X_0(X_0'X_0)^{-1}, \tag{49}$$

where V_0 denotes $V_0(\beta,\ \sigma^2;\ X_0)$. The RHS of (49) is independent of K and therefore equal to its limit as K tends to infinity. Next, the LHS of (45) is considered.

$$K\mathrm{v\hat{a}r}\,(\hat{\beta}) \;=\; \left(\frac{1}{K}X'\Pi^{-1}TX\right)^{-1}\left[\frac{1}{K}X'T\left(\frac{\hat{V}}{P}\right)TX\right]\left(\frac{1}{K}X'\Pi^{-1}TX\right)^{-1}. \tag{50}$$

Earlier, in the derivation of (13) and (22), use has already been made of

$$\operatorname*{plim}_{K\to\infty}\left(\frac{1}{K}X'\Pi^{-1}TX\right) \;=\; X_0'X_0. \tag{51}$$

It follows from the assumption that $V(\beta,\ \sigma^2;\ X)$ is a continuous function, that

$$\operatorname*{plim}_{K\to\infty}\,\hat{V}_0 \;=\; V_0, \tag{52}$$

where $\hat{V}_0$ denotes $V_0(\hat{\beta},\ \hat{\sigma}^2;\ X_0)$. Using (1), (48) and (52) gives

$$\begin{aligned} \operatorname*{plim}_{K\to\infty}\frac{1}{K}X'T\left(\frac{\hat{V}}{P}\right)TX &= \operatorname*{plim}_{K\to\infty}\frac{1}{K}\sum_k\left[X_0'T_k\left(\frac{\hat{V}_0}{P_0}\right)T_kX_0\right] \\ &= \operatorname*{plim}_{K\to\infty}\frac{1}{K}\sum_k\left[X_0'T_k\left(\frac{V_0}{P_0}\right)T_kX_0\right] \;=\; X_0'\,V_0X_0. \end{aligned} \tag{53}$$

Here P_0 denotes $E_p\ (T_k\iota\iota'\ T_k)$, which is the same for all $k = 1, ..., K$. The last equality sign results from the application of Khintchine's theorem, since the terms in the second summation over k in (53) are i.d.d. with p-expectation equal to $X_0'\,V_0X_0$. Finally, the combination of (50), (51) and (53) gives

$$\operatorname*{plim}_{K\to\infty} K\hat{\mathrm{var}}(\hat{\beta}) = (X_0'X_0)^{-1}X_0'\,V_0X_0\,(X_0'X_0)^{-1}, \tag{54}$$

which is the same expression as the RHS of (49).

6.2 Stratified Sampling

In this subsection the computation of the matrix $T(\hat{V}/P)T$ is given for a special case: (1) the disturbances are normally distributed, and (2) the sampling design is an endogenously stratified sampling design, such that the inclusion probability π_{ii} of element i of the population is a function f of only the i-th element of y, say $y_{(i)}$. Thus,

$$\pi_{ii} = f(y_{(i)}), \tag{55}$$

for $i = 1, ..., N$. As an example, consider the stratified sample which was shown in Figure 1. The design contains three strata there. The elements in the middle stratum have the highest inclusion probability. Figure 2 shows the corresponding function f.

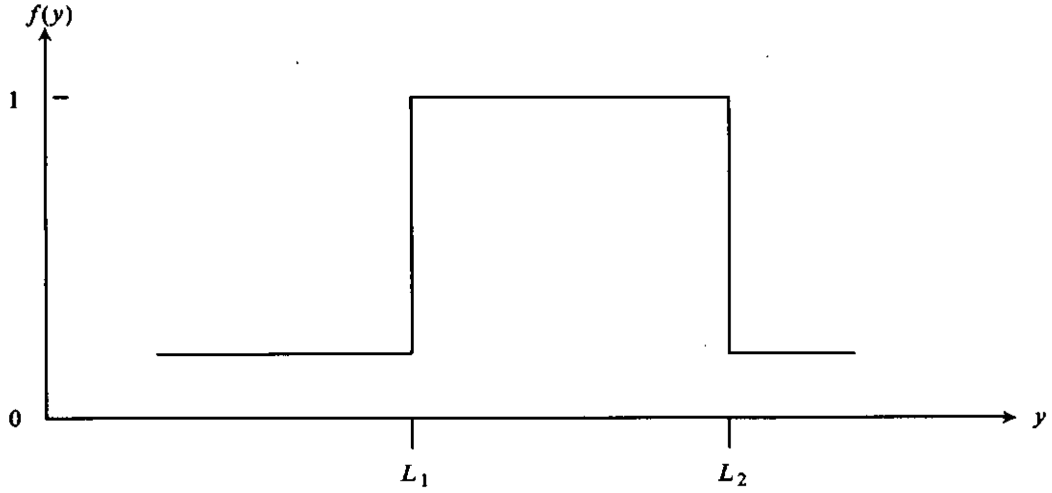


Figure 2. The Probability Function f Corresponding to Figure 1

In general, let there be H strata, indicated by $h = 1, ..., H$. Let the boundaries of these strata be $L_0, L_1, ..., L_H$. Typically, $L_0 = -\infty$ and $L_H = +\infty$. Let $\pi_{(h)}$ be the inclusion probability of the population elements in stratum h. More formally, the function $f(\cdot)$ is such that $f(y)$ equals $\pi_{(h)}$ if $L_{h-1} \le y < L_h$. The values of $\pi_{(h)}$ and L_h are usually known in practice, since the actual sampling design depends on their values.

In stratified sampling, the second order inclusion probability of any two population elements not in the same stratum equals the product of their respective first order inclusion probabilities: their inclusions in the sample are independent. For any two population elements in the same stratum this holds approximately. Thus, approximately the off-diagonal elements of P are equal to the off-diagonal elements of $\Pi\iota\iota'\Pi$. The diagonal of P is equal to the diagonal of Π, as before. Thus, approximately,

$$P = \Pi\iota\iota'\Pi - \Pi^2 + \Pi. \tag{56}$$

Then

$$\begin{aligned} V &= E_\xi [\operatorname{diag}(\varepsilon)(\iota\iota' - I + \Pi^{-1})\operatorname{diag}(\varepsilon)] \\ &= E_\xi [\varepsilon\varepsilon' - \operatorname{diag}^2(\varepsilon) + \operatorname{diag}^2(\varepsilon)\Pi^{-1}] = E_\xi [\operatorname{diag}^2(\varepsilon)\Pi^{-1}], \end{aligned} \tag{57}$$

in view of assumption (4). Thus V is a diagonal matrix here. Then

$$T\left(\frac{V}{P}\right)T = T\Pi^{-1}E_\xi [\operatorname{diag}^2(\varepsilon)\Pi^{-1}], \tag{58}$$

which is also a diagonal matrix. Now consider a population element i, which is included in the sample. Then, using (58) and assuming normally distributed disturbances,

$$\begin{aligned} \left[T\left(\frac{\hat{V}}{P}\right)T\right]_{ii} &= \frac{1}{\pi_{ii}} \sum_{h=1}^{H} \frac{1}{\pi_{(h)}} \int_{L_{h-1}-x_i'\hat{\beta}}^{L_h - x_i'\hat{\beta}} \varphi(\varepsilon_i;\hat{\sigma}^2)\varepsilon_i^2 d\varepsilon_i \\ &= \frac{\hat{\sigma}^2}{\pi_{ii}} \left\{ \frac{1}{\pi_{(H)}} + \sum_{h=1}^{H-1} \left(\frac{1}{\pi_{(h)}} - \frac{1}{\pi_{(h+1)}}\right) \Psi[(L_h - x_i'\hat{\beta})/\hat{\sigma}] \right\}. \end{aligned} \tag{59}$$

Here, $\phi(\cdot;\hat{\sigma}^2)$ indicates the normal density with mean zero and variance $\hat{\sigma}^2$. The function $\Psi(\cdot)$ is defined as

$$\Psi(x) \equiv \int_{-\infty}^{x} \varphi(\varepsilon;1)\varepsilon^2 d\varepsilon = \Phi(x) - x\varphi(x;1), \tag{60}$$

where $\Phi(\cdot)$ denotes the cumulative density function for the standard normal distribution. In the derivation of (59), use has been made of $\Psi(L_0) = 0$ and $\Psi(L_H) = 1$.

7. MODEL-FREE REGRESSION

7.1 Consistent Estimation

As a digression from the main theme of this paper, model-free regression will be considered in this section. Firstly, model-free regression can be usefully applied in the case of doubt about the validity of a linear model. See Fuller (1975), who studies model-free regression for some specific designs. Van Praag (1981, 1982) studies model-free regression in the case of repeated sampling from some probability distribution. See also DuMouchel and Duncan (1983). White (1980b, Section 3) studies related problems. Secondly, the so-called regression estimator of a population total uses model-free regression. See textbooks such as Cochran (1977), the review papers mentioned above by Nathan (1981) and Smith (1981) and Bethlehem and Keller (1983).

The purpose of model-free regression is the estimation of the population parameter vector

$$b \equiv (X'X)^{-1}X'y, \tag{61}$$

without assumptions about the probability distribution of y. In fact, both X and y are considered non-stochastic. Further, the same replica structure as in Section 2 is used, as follows.

$$X = \iota_K \otimes X_0, \tag{62}$$

and

$$y = \iota_K \otimes y_0, \tag{63}$$

where y_0 is some fixed N_0-vector. As before, the K diagonal matrices T_k $(k = 1, \ldots, K)$ are i.i.d. These matrices describe the sample as in Section 2. Together the matrices T_k form the matrix T. No additional assumptions are made concerning the distribution of T.

It is proved relatively easily, along the same lines as in Section 2, that the weighted estimator $\hat{\beta}$ defined before in (7), is a consistent estimator of b defined in (61). See also Jönrup and Rennermalm (1976), who indicates $\hat{\beta}$ as an "approximately unbiased" estimator of b, and Van Praag (1982, Section 4d), where "selectivity bias" with known inclusion probabilities is studied for the model-free case.

It follows in the same manner as in Section 4 that in the model-free case the asymptotic variance of $\hat{\beta}$, say $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$, equals

$$\mathrm{Var}_{\mathrm{MF}}(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}, \tag{64}$$

with

$$e \equiv y - Xb, \tag{65}$$

$$V = \mathrm{diag}(e)\Pi^{-1}P\Pi^{-1}\mathrm{diag}(e), \tag{66}$$

and with P defined as before in (29). Notice that V in (66) differs from V in (28) in the omission of the ξ-expectation and the substitution of e for ε.

It is interesting to rewrite $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$ in the same way as $\mathrm{Var}(\hat{\beta})$ was rewritten in Section 5. In doing so, use will be made of

$$X'e = 0, \tag{67}$$

which follows directly from (61) and (65). The $\mathrm{Var}_{\mathrm{MF}}(\hat{\beta})$ can be rewritten as

$$\begin{aligned}\mathrm{Var}_{\mathrm{MF}}(\hat{\beta}) &= (X'X)^{-1}X'\mathrm{diag}(e)(\Pi^{-1}P\Pi^{-1} - \iota\iota')\mathrm{diag}(e)X(X'X)^{-1} \\ &\quad + (X'X)^{-1}X'ee'X(X'X)^{-1} \\ &= (X'X)^{-1}X'\mathrm{diag}(e)(\Pi^{-1}P\Pi^{-1} - \iota\iota')\mathrm{diag}(e)X(X'X)^{-1}. \end{aligned} \tag{68}$$

The last member of (68) corresponds with the p-component of the decomposition of Var($\hat{\beta}$) in (37). It may be concluded from (68) that in model-free regression the variance of the estimator of the regression coefficients consists of the p-component, while the ξ-component vanishes.

Notice finally that, using the discussion at the end of Section 5, the last member of (68) can be written as

$$(X'X)^{-1}\Sigma(X'X)^{-1}, \tag{69}$$

where the matrix Σ is the p-variance-covariance of the row totals of X'diag(e). A similar result was reached by Binder (1983, Section 4), though along different lines.

8. DISCUSSION

In this section some practical considerations are given concerning the use of weights in regression analysis. Several motives for the use of weights are discussed shortly, related to the preceding technical sections of this paper.

First of all, it must be noted that the difference between weighted and unweighted regressions may be of some significance. An important example is the case where business firms are the unit of study - either farms, industrial enterprises of any other kind of business firms varying considerably in the number of employees. At the Netherlands Central Bureau of Statistics, for instance, the classification by number of employees is a standard stratification variable in sampling designs of business firms, giving a considerable range of inclusion probabilities - the large units chosen with relatively large probabilities. In studies with employment as the endogenous variable, such a sampling design is endogenous, which calls for weighted regression; the large units receiving small weights.

Secondly, in the case of units varying widely in size, a major problem with regression analysis is the heteroscedasticity of the error term. This calls for weighted regression, of the same sort as the weighting due to an endogenous design discussed in Section 2: large units receiving small weights.

Finally, there is a third motive for the weighting of sampled data: the notion of a model free regression, as discussed in Section 7 above. Again, the weights here are of the same sort as the weights in Section 2.

Summing up, there seems to be no reason not to incorporate the sampling design in regression analysis.

9. CONCLUSIONS

In this paper the estimation of a regression model with survey sample data has been studied. In particular, samples drawn with an endogenous design have been studied; for example, a sample stratified on the endogenous variable. It has been shown that for such a sample the weighting of the observations with the inverse of the square root of the sampling fractions gives a consistent estimator. The concept of consistency used here is a modification of Brewer (1979). The asymptotic variance of the estimator has been given, as well as a consistent estimator of this variance. The variance is the sum of a sampling component and a model component.

Also, model-free regression has been considered. Model-free regression requires the same weighting as endogenous stratification. The variance of the estimator of the model-free regression coefficients contains only the sampling component, and not the model component.

Finally, some practical considerations relative to the weighting of the data have been given.

ACKNOWLEDGEMENT

The author thanks Abby Israëls and Albert Verbeek and several anonymous referees for their comments on previous versions of this paper.

REFERENCES

BETHLEHEM, J.G., and KELLER, W.J. (1983). Weighting sample survey data using linear models. Internal Report, Department for Statistical Methods, Netherlands Central Bureau of Statistics, Voorburg.

BINDER, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review,* 51, 279-292.

BRESLOW, N., and DAY, N.E. (1980). *Statistical Methods in Cancer Research, Volume 1: the Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

BREWER, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association,* 74, 911-915.

COCHRAN, W.G. (1977). *Sampling Techniques.* New York: Wiley.

CRAMER, J.S. (1971). *Empirical Econometrics.* Amsterdam: North-Holland.

DuMOUCHEL, W.H., and DUNCAN, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association,* 78, 535-543.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā,* C37, 117-132.

HAUSMAN, J.A., and WISE, D.A. (1981). Stratification on endogenous variables and estimation: the Gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications,* (Eds., C.F. Manski and D. McFadden), Cambridge: MIT Press.

HECKMAN, J.J. (1979). Sample selection bias as a specification error. *Econometrica,* 47, 153-161.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47, 663-685.

JEWELL, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika,* 72, 11-21.

JOHNSTON, J. (1972). *Econometric Methods.* Tokyo: McGraw-Hill Kogakusha.

JONRUP, H., and RENNERMALM, B. (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics,* 33-36.

KMENTA, J. (1978). *Elements of Econometrics.* New York: McMillan.

MANSKI, C.F., and McFADDEN, D. (eds.) (1981). *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

NATHAN, G. (1981). Notes on inference based on data from complex sample designs. *Survey Methodology,* 7, 110-129.

PORTER, R.D. (1973). On the use of survey sample weights in the linear model. *Annals of Economic and Social Measurement,* 2, 141-158.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: Wiley.

ROBINSON, P.M. (1982). On the convergence of the Horvitz-Thompson estimator. *Australian Journal of Statistics,* 24, 234-238.

SMITH, T.M.F. (1981). Regression analysis for complex surveys. In *Current Topics in Survey Sampling,* (Eds. D. Krewski, R. Platek, and J.N.K. Rao), New York: Academic Press, 267-292.

THEIL, H. (1971). *Principles of Econometrics.* New York: Wiley.

VAN PRAAG, B.M.S. (1981). Model-free regression. *Economics Letters,* 7, 139-144.

VAN PRAAG, B.M.S. (1982). The population-sample decomposition with an application to minimum distance estimators. Report 8218, Center for Research in Public Economics, Leyden University.

WHITE, H.,(1980a). Nonlinear regression on cross section data. *Econometrica,* 48, 721-746.

WHITE, H., (1980b). Using least squares to approximate unknown regression functions. *International Economic Review,* 12, 149-170.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 139–150
Statistics Canada

A Cluster Analysis of Activities of Daily Living From the Canadian Health and Disability Survey[1]

D.A. BINDER and G. LAZARUS[2]

ABSTRACT

The Canadian Health and Disability Survey, administered as a supplement to the Canadian Labour Force Survey in October 1983, collected data on potentially disabled persons by means of a screening questionnaire and a follow-up questionnaire for those screened-in. The data from the screening questionnaire, consisting of a set of activities of daily living, were used to group respondents according to identifiable characteristics. A description of the groups of respondents is provided along with an evaluation of the methods used in their determination. An incompletely ordered severity scale is proposed.

KEY WORDS: Disability scale; Discriminant analysis.

1. INTRODUCTION

Considerable efforts have been made to acquire a better understanding of the disabled population. These efforts have focussed on the development of a useful vehicle for capturing the potentially disabled population as well as the analysis of survey data for the purposes of gaining a better understanding of the various dimensions of disability and to develop useful measures of severity. Examples of papers which examine these issues are Dolson *et al.* (1984) and Raymond *et al.* (1981), among others. This paper chronicles the development of an exploratory technique in order to gain a better understanding of the disabled population in Canada. In particular, a cluster analysis based on results of several discriminant analyses was performed.

The next section presents information about the Canadian Health and Disability Survey. The third section describes the development of the clusters. Section 4 focusses on the characterization of the clusters. Some analysis of the behaviour of the derived clusters is given in Section 5. The paper concludes with some closing remarks.

2. BACKGROUND

In response to a need for data on disabled persons in Canada, Statistics Canada undertook a program to create a disability database. The Canadian Health and Disability Surveys (CHDS) were administered as supplements to the Canadian Labour Force Survey (LFS) in October 1983 and June 1984. In both cases, separate questionnaires were administered to children and to adults. In the October survey, the adult questionnaire was administered to everyone in the LFS sample (the frame includes about 97% of the Canadian population aged 15 or more). In June, the adult survey was restricted to those aged 15 to 64 from the six provinces with the smaller sample sizes in October (i.e. Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick, Manitoba and Saskatchewan). Children from all provinces were surveyed in both October and June.

[1] This is a revised version of the paper presented at ASA meetings, Social Statistics Section, Las Vegas, August 1985.

[2] D.A. Binder and G. Lazarus, Social Survey Methods Division, Informatics and Methodology Branch, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

This paper concentrates on work which utilized only the data from the adults questionnaire in October 1983. This survey obtained 92,945 adult respondents from approximately 47,000 households.

2.1 Questionnaire

2.1.1 Screening Section

The Labour Force Supplement included a screen which was used to identify respondents for a follow-up questionnaire. The screening section consisted of nineteen items – seventeen activities of daily living, an activity limitation item and an item about mental handicap. The activities of daily living (ADL's) are a set of activities which any person would perform during the course of his/her regular living pattern. The set used here was a modified version of those developed by the Organization for Economic and Co-operative Development (OECD) and has been utilized by several other countries.

The ADL's are presented in Table 1 with the questionnaire identification and the orientation of the specified activity. Two ADL's are related to hearing troubles, two to vision troubles, four to mobility troubles, one to speaking and being understood and the remaining eight to agility troubles.

Table 1

Activities of Daily Living

Questionnaire Item	Description	Orientation
A10	Walking 400 Metres	Mobility
A11	Walking up and down stairs	Mobility
A12	Carrying 5 kg. object for 10 metres	Mobility
A13	Moving from one room to another	Agility
A14	Standing for long periods	Mobility
A15	When standing, bending down to pick up object	Agility
A16	Dressing and undressing	Agility
A17	Getting in and out of bed	Agility
A18	Cutting own toenails	Agility
A19	Using fingers to grasp or handle	Agility
A20	Reaching	Agility
A21	Cutting own food	Agility
A22	Reading newsprint	Vision
A23	Seeing clearly a face across the room	Vision
A24	Hearing conversation with another person	Hearing
A25	Hearing conversation with two or more persons	Hearing
A26	Speaking and being understood	Speaking and being understood

An example of the wording of these questions in the screening section of the questionnaire is as follows: (A20) Does . . . . have any trouble reaching? The activity limitation item (A27) concerned limitation "in the kind or amount of activity he/she can do at home, at work or going to school because of a long-term physical condition or health problem". The final item in the screen section (A28) concerned mental handicap.

It should be noted that the survey was concerned with long-term conditions or health problems - those that had lasted or were expected to last more than six months (excluding pregnancy). An individual was screened in if he/she had trouble with at least one of the ADL's, the activity limitation item or had a mental handicap. (Proxy responses were required for mentally handicapped individuals).

2.1.2 Follow-up Section

The follow-up section of the questionnaire was completed for individuals selected by the screening section. This section included an item which sought to determine if the respondent was completely unable to perform the ADL('s) he/she had trouble with. Other segments of the follow-up questionnaire pertained to: nature of the disability (related to trouble seeing or reading, trouble hearing, trouble speaking and being understood, and mobility); problems related to the ability to work or the workplace itself; obstacles to education and availability of special educational facilities; problems related to local and long-distance travel; and problems in current residence and special facilities. The information in the follow-up questionnaire, given above, could be used to analyze the cluster characteristics, or to develop a severity index (see Lazarus; 1985a, 1985b).

3. CLUSTERS

This section presents a description of the procedures used in the development of the clusters. The clustering procedures employed were developed specifically for this application. Technical details concerning the methods used are given in Sections 3.2 and 3.3. All computations were performed using SAS.

3.1 Methodology

This section summarizes the methodology used to derive the final clusters. The clustering procedure consisted of two steps:

a) a divisive step, where the 12,907 individuals were sequentially partitioned using PROC CANDISC.

b) an agglomerative step, where the partition was collapsed.

For the divisive step, the following procedure was employed iteratively. First, the starting point put all the observations into a single cluster. Each step subdivided each of the current clusters into two groups. For each of the current clusters, a canonical correlation analysis was performed by taking each non-constant variable as a grouping variable and using all other non-constant variables as explanatory variables. The cluster was then split into two, based on the discriminant analysis with the largest *F*-value. In this way the determinant of the between-sums-of-squares matrix is maximized.

For the agglomerative step, subjective criteria were used, based on the magnitude of the *F*-value, the size of the groups and the plots of the points. Collapsing was accomplished in the reverse order of splitting, for the most part.

For the divisive step, data based on both unweighted and weighted covariances were used separately. The results were essentially the same. It was decided to continue without the sampling weights because of the added complexity which would be incurred by their inclusion. Furthermore, the weights were not expected to be important with respect to the characteristics of the clustered individuals. Inclusion of weights is necessary for evaluation and analysis.

3.2 Description

The cluster analysis was a procedure which grouped together those screened in respondents with similar but not necessarily identical "profiles". For our purposes, a respondent's profile consisted of the responses to the seventeen ADL's (yes, has trouble/no, does not have trouble), responses to the major activity limitation item (positive/negative), and the mental handicap item in the screening section of the questionnaire.

Table 2 details the final clusters. The symbols *U* and *Z* demonstrate how the groups are defined. The symbol *U* means that the group is defined through that variable being one, i.e. 100% by definition. The symbol *Z* is used when the defining screening section item is zero, i.e. 0% by definition. Note that six of the nineteen screening items are not used explicitly in the process of classifying respondents. These are A11, A13, A18, A20, A23 and A24.

4. CLUSTER CHARACTERIZATION

This section explores the ways and means of identifying the clusters. The concepts of "trouble orientation" and "umbrella" group are introduced and the clusters are ranked according to the severity of disability.

4.1 Trouble Orientation

Threshold values were established to assist in the cluster classification process. The values were chosen by ordering the clusters according to orientation and locating an obvious gap in the *E*(NADL) for the orientation, where *E*(NADL) referred to the average number of troubles among ADL's A10 - A26. In general, a cluster was recognized as having trouble with an activity orientation when the *E*(NADL) for a particular orientation exceeded the established threshold value. For example, for mobility orientation, *E*(NADL) was computed for activities A10, A11, A12 and A14. The *E*(NADL) for each cluster over each orientation may be found in Table 3.

Clusters were labelled as follows. If a cluster had trouble with an activity, the corresponding letter was included in the label. Two clusters, containing individuals who had trouble speaking and being understood or were mentally handicapped, were "special". Clusters which had neither mobility nor agility troubles exceeding the established values were so designated with an *N*. For example, HMA1 and HMA2 refer to clusters with a large proportion having hearing, mobility and agility problems, but no particular problem with vision. Alternatively, VN1 refers to a cluster with the exact opposite set of problems.

4.2 Umbrella Groups

Clusters with similar orientation patterns became members of specified "umbrella" groups, where they could be better compared using *E*(NADL) within the umbrella. Table 4 shows the clusters according to the "umbrella" groups to which they belong.

Table 2
Cluster Analysis Results

Cluster	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
1	*U*	92.7	79.9	59.7	89.8	85.5	*U*	62.7	86.8	60.1
2	*U*	77.0	63.1	16.0	77.0	55.6	*Z*	11.2	46.5	31.0
3	*U*	85.1	66.5	19.4	75.8	*U*	*Z*	15.8	49.6	26.5
4	*U*	65.6	36.7	6.4	55.9	*Z*	*Z*	4.5	21.5	17.7
5	*Z*	18.7	18.2	3.4	25.6	24.6	4.9	6.9	21.7	20.7
6	*Z*	36.3	23.2	4.8	49.5	*U*	11.8	16.6	28.4	21.1
7	*Z*	9.2	5.3	0.3	10.8	*Z*	1.1	0.9	4.4	7.1
8	*U*	94.7	88.6	67.3	93.9	89.0	*U*	74.7	94.7	84.0
9	*U*	92.9	82.1	55.4	89.3	91.1	*U*	58.9	87.5	30.4
10	*U*	95.7	81.0	55.7	91.9	93.8	*U*	*U*	85.2	33.3
11	*U*	92.2	71.7	21.1	83.7	74.1	*U*	*Z*	59.0	28.9
12	*U*	91.9	71.3	25.0	81.3	*U*	*Z*	16.9	58.1	31.9
13	*U*	61.0	48.8	4.3	55.5	*Z*	*Z*	4.9	32.3	14.0
14	*U*	91.3	*U*	23.6	81.4	*U*	*Z*	16.8	40.2	*U*
15	*U*	93.6	*U*	29.9	84.0	*U*	*Z*	19.3	56.1	*Z*
16	*U*	74.9	*Z*	10.9	65.7	*U*	*Z*	12.9	32.8	16.4
17	*U*	66.7	58.3	12.5	37.5	*Z*	*Z*	0.0	37.5	20.8
18	*U*	74.0	55.5	7.5	59.5	*Z*	*Z*	10.4	29.5	*U*
19	*U*	79.6	*U*	11.5	60.8	*Z*	*Z*	2.9	14.6	*Z*
20	*U*	59.0	*Z*	2.7	45.6	*Z*	*Z*	2.2	10.4	*Z*
21	*Z*	14.7	12.6	1.9	19.4	13.9	5.5	4.7	22.2	11.5
22	*Z*	26.5	40.9	7.0	41.4	59.1	*U*	32.1	47.4	35.8
23	*Z*	29.0	26.1	2.1	43.3	*U*	*Z*	13.0	19.0	13.5
24	*Z*	2.4	2.4	0.0	2.0	*Z*	*Z*	0.4	7.7	3.3
25	*Z*	35.6	*U*	2.4	32.9	*Z*	*Z*	3.1	8.5	18.0
26	*Z*	13.5	*Z*	0.3	16.8	*Z*	*Z*	1.8	4.2	9.1
27	*Z*	17.0	13.7	0.3	*U*	*Z*	*Z*	2.4	6.2	5.4
28	*Z*	10.3	6.9	0.0	*Z*	*Z*	*Z*	0.1	7.8	*U*
29	*Z*	38.7	26.3	0.6	*Z*	*Z*	*Z*	2.2	10.9	*Z*

Cluster	A20	A21	A22	A23	A24	A25	A26	A27	A28	Size
1	63.7	42.2	38.6	27.1	73.3	*U*	23.4	94.4	6.3	303
2	35.3	11.8	*U*	50.8	71.7	*U*	9.6	85.0	1.6	187
3	34.6	5.9	*Z*	3.4	63.7	*U*	2.5	88.7	1.1	355
4	16.4	1.9	*Z*	1.6	57.9	*U*	2.6	73.3	1.0	311
5	17.7	8.4	*U*	46.3	59.6	*U*	12.8	55.7	7.9	203
6	24.9	3.5	*Z*	1.4	50.9	*U*	4.2	71.3	1.0	289
7	4.6	0.6	*Z*	1.3	60.5	*U*	5.6	26.3	1.6	1,770
8	78.4	*U*	32.6	16.7	1.2	*Z*	32.2	96.3	9.8	245
9	50.0	*Z*	*U*	30.4	5.4	*Z*	10.7	100.0	5.4	56
10	55.2	*Z*	*Z*	0.5	0.0	*Z*	2.4	89.0	1.9	210
11	45.8	*Z*	*Z*	1.8	0.6	*Z*	3.0	90.4	0.6	166
12	39.4	7.5	*U*	45.6	4.4	*Z*	5.0	93.1	1.9	160
13	20.7	5.5	*U*	42.7	1.2	*Z*	6.7	78.0	4.3	164
14	34.4	1.5	*Z*	1.0	0.9	*Z*	1.3	89.4	1.2	187
15	66.3	16.6	*Z*	2.1	2.1	*Z*	5.9	92.0	1.6	677
16	20.7	0.7	*Z*	0.0	0.4	*Z*	2.0	82.3	0.4	458
17	16.7	20.8	*Z*	0.0	0.0	*Z*	*U*	91.7	33.3	24
18	29.5	12.1	*Z*	0.0	0.0	*Z*	*Z*	82.1	1.2	173
19	19.4	1.0	*Z*	0.5	0.2	*Z*	*Z*	73.5	1.0	582
20	8.0	0.0	*Z*	0.7	0.4	*Z*	*Z*	66.7	0.6	857
21	9.7	7.1	*U*	41.1	2.6	*Z*	8.7	55.3	9.2	618
22	41.9	19.5	*Z*	1.4	1.4	*Z*	7.0	76.3	4.7	215
23	18.1	1.9	*Z*	0.8	0.7	*Z*	1.2	66.6	0.4	1,164
24	0.8	2.0	*Z*	0.0	0.0	*Z*	27.2	62.2	*U*	246
25	23.7	1.4	*Z*	0.3	0.0	*Z*	1.4	*U*	*Z*	295
26	7.3	1.2	*Z*	0.7	0.5	*Z*	1.9	*U*	*Z*	1,923
27	2.4	0.3	*Z*	0.3	0.3	*Z*	0.3	*Z*	*Z*	371
28	11.8	8.3	*Z*	0.0	0.5	*Z*	0.5	*Z*	*Z*	204
29	18.0	1.6	*Z*	6.5	5.7	*Z*	8.5	*Z*	*Z*	494

Table 3

Average Number of Troubles by Orientation

Cluster	Hearing	Vision	Mobility	Agility	Total
1	1.733	0.657	3.624	5.841	11.855
2	1.717	1.508	3.171	2.170	8.566
3	1.637	0.034	3.274	2.543	7.488
4	1.579	0.016	2.582	0.710	4.887
5	1.596	1.463	0.625	1.211	4.895
6	1.509	0.014	1.091	2.152	4.766
7	1.605	0.013	0.253	0.246	2.117
8	0.012	0.493	3.772	7.203	11.480
9	0.054	1.304	3.643	4.480	9.841
10	0.000	0.005	3.686	5.256	8.947
11	0.006	0.018	3.476	3.319	6.819
12	0.044	1.456	3.445	2.838	7.783
13	0.012	1.427	2.653	0.884	4.976
14	0.009	0.010	3.727	3.178	6.924
15	0.021	0.021	3.776	2.941	6.759
16	0.004	0.000	2.406	1.964	4.374
17	0.000	0.000	2.625	2.083	4.708
18	0.000	0.000	2.890	1.890	4.780
19	0.002	0.005	3.404	0.494	3.905
20	0.004	0.007	2.046	0.233	2.290
21	0.026	1.411	0.467	0.852	2.756
22	0.014	0.014	1.088	3.498	4.614
23	0.007	0.008	0.984	1.688	2.687
24	0.000	0.000	0.068	0.352	0.482
25	0.000	0.003	1.685	0.587	2.273
26	0.005	0.007	0.303	0.258	0.573
27	0.003	0.003	0.310	1.170	1.486
28	0.005	0.000	0.172	1.285	1.462
29	0.057	0.065	0.650	0.418	1.190

4.3 Severity

One area of analytic interest is the development of an index of severity of disability. The notion has been considered previously by Raymond et al, among others.

The index of severity would be useful in as much as it would allow for simple comparisons of disability among the screened-in respondents. The use of *E*(NADL) to draw such comparisons presumes that the orientations are self-weighting, noting, for example, that two ADL's are devoted to hearing troubles while four are devoted to mobility troubles. Also, the multidimensional nature of severity of disability is hidden by a single score such as *E*(NADL).

Table 4

Ordering of Clusters by "Umbrella" Groups

Umbrella Group	Cluster	Sample Count	*E*(NADL)	ID
HV (Hearing/Vision)	2	187	8.566	HVMA1
	5	203	4.895	HVN1
H (Hearing)	1	303	11.855	HMA1
	3	355	7.488	HMA2
	4	311	4.829	HM1
	6	289	4.760	HA1
	7	1,770	2.120	HN1
V (Vision)	9	56	9.841	VMA1
	12	160	7.783	VMA2
	13	164	4.976	VM1
	21	618	2.756	VN1
S (Special)	17	24	4.708	SMA1
	24	246	0.482	SN1
MA (Mobility/Agility)	8	245	11.480	MA1
	10	210	8.947	MA2
	11	166	6.819	MA4
	14	187	6.924	MA3
	15	677	6.759	MA5
M (Mobility)	16	458	4.374	M2
	18	173	4.780	M1
	19	582	3.905	M3
	20	857	2.290	M4
A (Agility)	22	215	4.614	A1
N (Neither)	23	1,164	2.687	N1
	25	295	2.273	N2
	26	1,923	0.573	N6
	27	371	1.486	N3
	28	204	1.462	N4
	29	494	1.190	N5

Table 4 presents an ordering of clusters according to "severity" within umbrella groups. This within group ordering better reflects the notion that severity is multidimensional than would an overall ordering.

5. CLUSTER CHARACTERISTICS

The principal components technique was used to examine the behaviour of the resulting clusters. Raymond et al also employed principal components; the main difference being that analysis here is based upon group means rather than individuals.

5.1 Methodology

We considered a subset of screened in cases, where more information per case is available. In particular, we added the responses to questions of the form: (B101) Is . . . completely unable to walk 400 metres without resting? This line of questioning was used for each of the ADL'S, A10-A26. Thus, 11,412 of the original 12,907 individuals who were screened in were usable. The other 1,495 were dropped because of non-response problems. These "completely unable" items were coded with "1" when the individual indicated that he/she was completely unable to perform the specified ADL, otherwise , a "0" was coded.

The means were obtained for the nineteen screening items and seventeen follow-up items for each cluster. The means for the completely unable items were then multiplied by the ratio of the overall average number of ADL's to the overall average of completely unable items in order to scale them consistently and to avoid the scaling problems associated with principal components analysis.

Principal components were obtained using the nineteen screening section and seventeen follow-up item means as variables, using the "clusters" as observations and weighting according to cluster size. The clusters were then ordered according to each of the first four principal component scores.

The final stage involved the pooling of cluster cases according to "umbrella" group membership and finding the means of the first four principal component leadings for each of the eight "umbrella" groups, where the weights were the numbers of members in the "umbrella" groups.

5.2 Results

We present the results in two stages. In the first stage, we examine the principal components and attempt to label them according to the scores. We also explore the "umbrella" group construct in terms of the principal component means. In the second stage, we examine the ordering of the clusters according to the first four principal components.

5.2.1 Components

The first four principal components for the nineteen screening section items and the seventeen follow-up items explained just over seven-eighths of the total variance and appeared to be most useful for our purposes.

The loadings of the first principal component are positive on all but four items (A24, A25 and B241 are hearing oriented, A28 is mental handicap). The negative loadings are close to zero. This first component appears to be an overall measure of strength. The first principal component explained nearly 66% of the total variance and is denoted as "OVERALL".

There are negative loadings on A10, A11, A12, A14 and A15 of the second component. The loading for A15 is nearly zero, however. Loadings are positive for ADL's with an agility-trouble orientation as well as for hearing-trouble and vision-trouble orientations. It appears then that this component polarizes mobility trouble against agility, hearing and vision troubles. The second component is labelled "AHV/M".

The third principal component has positive loadings for mobility and hearing oriented ADL's and negative loadings for agility and vision oriented ADL's. This third component is denoted "MH/AV".

The fourth principal component has positive loadings for mobility and vision oriented ADL's and negative loadings for agility oriented ADL's. This fourth component is designated "MV/A".

5.2.2 Mean Loadings

Table 5 presents the average differences of the principal component scores from the mean scores over all 11,412 individuals, for each of the eight "umbrella" groups. We can

Table 5

Average Differences of Principal Component Scores from Mean Scores

		Differences			
Umbrella Group	Sample Count	PRIN1 (Overall)	PRIN2 (AVH/M)	PRIN3 (MH/AV)	PRIN4 (MV/A)
Hearing/Vision	346	0.68	1.26	0.61	1.06
Hearing	2741	−0.33	0.54	0.81	−0.25
Vision	888	0.30	0.69	−0.76	1.27
Special	151	−1.02	−0.04	−0.47	−0.06
Mobility/Agility	1311	3.31	−0.33	−0.21	−0.33
Mobility	1893	0.30	−0.80	0.18	0.33
Agility	195	−0.19	0.31	−0.80	−0.78
Neither	3887	−1.11	−0.16	−0.41	−0.22

now check to see if the incomplete ordering presented earlier is consistent with the results from the principal components analysis. We note the following observations are taken from Table 5.

i) The mobility/agility "umbrella" group has the highest difference on the first principal component "overall", while the "umbrella" group "neither" has the lowest difference. The difference for the hearing/vision group is positive as is the mean for the vision group. The hearing group difference is negative, however, evidence that individuals with hearing-oriented troubles tend not to have other disabilities. There may be an in–clination to draw the same kind of conclusion with respect to agility-oriented troubles. It is observed that the mobility/agility and mobility groups have positive differences while the agility "umbrella" group has a negative difference. However, in this case, the result is somewhat ambiguous because the agility-oriented ADL's included speaking trouble (A26), a so-called "special" trouble area and it is clear indeed that the special "umbrella" group has a negative difference for the first principal component.

ii) The second component set mobility-oriented troubles (-) against agility, hearing and vision-oriented troubles (+). Positive differences are recorded for the hearing/vision, hearing, vision and agility "umbrella" groups while negative differences are associated with the mobility/agility, mobility and neither groups, as expected. The difference for the special groups is nearly zero.

iii) The third component set mobility-oriented and hearing-oriented troubles (+)against agility-oriented and vision-oriented troubles (−). Again, the results are consistent.

iv) The fourth principal component set mobility and vision-oriented troubles (+) against agility-oriented troubles (−). The results are again consistent with the umbrella-group construct.

5.2.3 The Scales

Table 6 shows the ranks of the clusters according to the first four principal component scores and *E*(NADL). Recall that the component loadings are for 11,412 cases and utilize follow-up information as well as screening section information while the *E*(NADL) scale is based on 12,907 cases and uses screening information only.

The cluster ranking according to principal components was done as follows. The component representing overall strength (OVERALL) ranked clusters from highest to lowest scores. The ranking of clusters on AHV/M tended to put clusters with mobility-oriented troubles at the bottom end as opposed to clusters with agility, hearing or vision oriented troubles which were ranked higher up on this scale. The ranking of clusters on MH/AV tended to put clusters with mobility or hearing troubles at or near the bottom of the scale while clusters with agility or vision-oriented troubles were ranked higher. Finally clusters with agility-oriented troubles were ranked higher on MV/A than the others. Given the bipolar nature of components 2, 3 and 4, it was necessary to make an arbitrary decision as to a trouble orientation scale. As cluster 8 had shown itself to be highly severe according to the *E*(NADL) scale, it was determined that cluster 8 should be similarly ranked along the other scales.

For most clusters, the rankings fluctuate over a wide range. This reflects the nature of the criteria upon which the scales were based. The first principal component, which provides an overall measure of strength, may be the most suitable candidate for ranking the clusters. Firstly, it incorporates the screening section information used in the development of the *E*(NADL) measure. As a result, the rank orderings provided by the OVERALL and *E*(NADL) scales are quite similar. The additional follow-up information used in the construction of

Table 6

Cluster Rank According to Alternative Scales

Cluster	ID	PRIN1 (Overall)	PRIN2 (AHV/M)	PRIN3 (MH/AV)	PRIN4 (MV/A)	*E*(NADL)
2	HVMA1	9	4	27	28	5
5	HVN1	22	2	22	25	12
1	HMA1	3	3	24	6	1
3	HMA2	10	14	28	10	7
4	HM1	16	15	29	20	13
6	HA1	20	8	25	3	15
7	HN1	29	7	26	9	24
9	VMA1	2	6	4	23	3
12	VMA2	4	10	7	27	6
13	VM1	13	11	11	29	11
21	VN1	23	5	2	26	20
8	MA1	1	1	1	1	2
10	MA2	5	20	13	4	4
14	MA3	6	24	16	7	8
11	MA4	7	23	17	8	9
15	MA5	8	28	20	18	10
18	M1	14	26	19	21	14
16	M2	15	25	18	17	18
19	M3	11	29	23	24	19
20	M4	18	27	21	22	22
22	A1	17	9	3	2	17
23	N1	21	17	6	5	21
25	N2	19	22	10	16	23
27	N3	24	19	15	12	25
28	N4	28	12	9	11	26
29	N5	25	16	12	15	27
26	N6	26	18	8	14	28
17	SMA1	12	21	14	19	16
24	SN1	27	13	5	13	29

this component leads us to believe that OVERALL is better than other scales such as *E*(NADL). It is worth noting that the ranking was done on all 29 clusters and depicted in Table 6 on an "umbrella" group basis. The "umbrella" group information was not incorporated into the principal components analysis, however.

6 CLOSING REMARKS

A clustering technique was employed to group screened-in individuals according to similar screening section profiles. The clusters were then ordered according to the information contained in the screening section of the questionnaire (the incomplete ordering based on *E*(NADL) and presented in Table 4) and finally according to information contained in the screening and follow-up sections of the questionnaire (the OVERALL scale presented in Table 6). This last scale is deemed presently to be the most suitable of those considered here. However, it could be argued that no single index of severity exists and in fact the severity index should be defined as a 4-dimensional scale corresponding to our principal components.

ACKNOWLEDGEMENT

The authors would like to thank the assistant editors for constructive comments on earlier drafts.

REFERENCES

DOLSON, D., GILES, P., and MORIN, J.-P. (1984). A methodology for surveying disabled persons using a supplement to the Labour Force Survey. *Survey Methodology*, 10, 187-197.

LAZARUS, G. (1985a). Characteristics of potentially disabled individuals based on the cluster analysis of activities of daily living. Working Paper, Institutional and Agricultural Survey Methods Division, Statistics Canada.

LAZARUS, G. (1985b). An application of the results of the cluster analysis of activities of daily living. Working Paper, Institutional and Agricultural Survey Methods Division, Statistics Canada.

RAYMOND, L., CHRISTE, E., and CLEMENCE, A. (1981). Vers l'etablissement d'un score global d'incapacite fonctionelle sur la base des questions de l'OCDE, d'apres une enquete en Suisse. *Revue d'epidemiologie et sante publique*, 29, 451-459.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 151–160
Statistics Canada

Additive Versus Multiplicative Seasonal Adjustment When There Are Fast Changes in the Trend-Cycle[1]

GUY HUOT and NAZIRA GAIT[2]

ABSTRACT

The seasonal adjustment of a time series is not a straightforward procedure particularly when the level of a series nearly doubles in just one year. The 1981-82 recession had a very sudden great impact not only on the structure of the series but on the estimation of the trend- cycle and seasonal components at the end of the series. Serious seasonal adjustment problems can occur. For instance: the selection of the wrong decomposition model may produce underadjustment in the seasonally high months and overadjustment in the seasonally low months. The wrong decomposition model may also signal a false turning point. This article analyses these two aspects of the interplay between a severe recession and seasonal adjustment.

KEY WORDS: Decomposition models; ARIMA; Lead-lag relationship.

1. INTRODUCTION

1981 and 1982 were atypical years afflicted by a severe recession. This recession has profoundly affected the evolution and structure of economic time series, and consequently their seasonal adjustment. Seasonally adjusted time series are necessary to diagnose the socio-economic health of a country. In turn, social and economic policies founded on these data influence decisions in both the private and public sectors. Thus, this recession raises many questions. One can readily see that a prompt examination of seasonal adjustment is necessary.

The series under consideration here are: initial and renewal claims received (for unemployment benefits) and beneficiaries. It is difficult to see how their trend and cycle components evolve when they are contaminated by seasonal variation, namely intra-annual climatic and institutional factors. Seasonal adjustment permits a better detection of fundamental tendencies, such as turning points, and evaluation of the present performance of the economy.

This article analyses some aspects of the interplay between a severe recession and seasonal adjustment. In just one year, that is in 1981, this recession has nearly doubled the level of beneficiaries. Such a sudden large change prompts questions about the structure of the series, the choice of the X-11-ARIMA decomposition model, the determination of turning points at the end of the series, and the use of ARIMA forecasts for seasonal adjustment.

In section 2, we discuss two important consequences of using a wrong decomposition model, namely a systematic over- and under-adjustment of series and the possibility of having a false turning point at the end of the series. In section 3, we use the lead-lag relationship between the claims and beneficiaries series to help seasonally adjust the latter series.

The ARIMA forecasts generally help to reduce the revision to the seasonal factors and they can help to provide a more accurate recognition of the turning points at the end of the series. Section 4 considers this question.

[1] This paper was presented at the 145th Annual Meeting of the American Statistical Association, Las Vegas, Nevada, 1985.

[2] Guy Huot, Time Series Research and Analysis Division, Statistics Canada. N. Gait, University of Sao Paulo, Brazil, was visiting Statistics Canada when the paper was written.

2. DECOMPOSITION MODELS FOR SEASONAL ADJUSTMENT

Most of the claims and beneficiaries series have similar characteristics, so we have chosen to study one claims series and one beneficiaries series which can clearly illustrate some of the problems peculiar to seasonal adjustment during a severe recession. It should be noted that the results of our analysis are equally valid during a sudden strong expansion in the economy. It is the sudden large change in the level of the series caused by the recession or the expansion that is important.

The X-11-ARIMA program (Dagum 1980) will be used to seasonally adjust these series. The program is applied to the claims and beneficiaries series, using data from January 1973 and May 1975 respectively, up to February 1983.

The X-11-ARIMA program provides three decomposition models for the estimation of the time series components. The program assumes an additive relationship between the components

$$O_t = \mathrm{TC}_t + S_t + I_t \tag{2.1}$$

or a multiplicative one

$$O_t = \mathrm{TC}_t\, S_t\, I_t \tag{2.2}$$

or a log additive one

$$\log O_t = \log \mathrm{TC}_t + \log S_t + \log I_t \tag{2.3}$$

where O stands for the observed and unadjusted series; TC, the trend-cycle; S and I, the seasonal and irregular components; and t, the time.

Seasonal adjustment means removing the seasonal variations S_t from the raw data O_t, thus leaving a seasonally adjusted series consisting of TC_t and I_t. In order to know whether a certain series contains a significant amount of seasonality and if so, whether an additive or multiplicative model provides the better fit, one can perform a test for the presence of seasonality and a model test on the series (Higginson 1977). The first test shows that both series contain a very significant amount of seasonality. According to the second test, the multiplicative model fits the beneficiaries series better when tested from May 1975 to June 1981. When the series is extended to February 1983, taking into account the impact of the recession on the series, the additive model then fits better. On the other hand, the model test favours neither the additive nor the multiplicative model for the claims series.

One usually adjusts the series using only one model, however, figure 1 shows the beneficiaries series adjusted using the two models, both without using the ARIMA option. During 1980 and 1981, the difference between the additive and multiplicative adjustments was small compared with the difference observed in 1982.

The multiplicative model assumes that the seasonal variation is proportional to the level of the trend-cycle. During 1982, the seasonal amplitude did not increase in this way. Consequently, using the multiplicative model is likely to overestimate it from June to November, the seasonally low months. As figure 1 shows, in underestimating the number of seasonal beneficiaries, the multiplicative model has drastically overestimated the number of seasonally adjusted beneficiaries. The converse is also true.

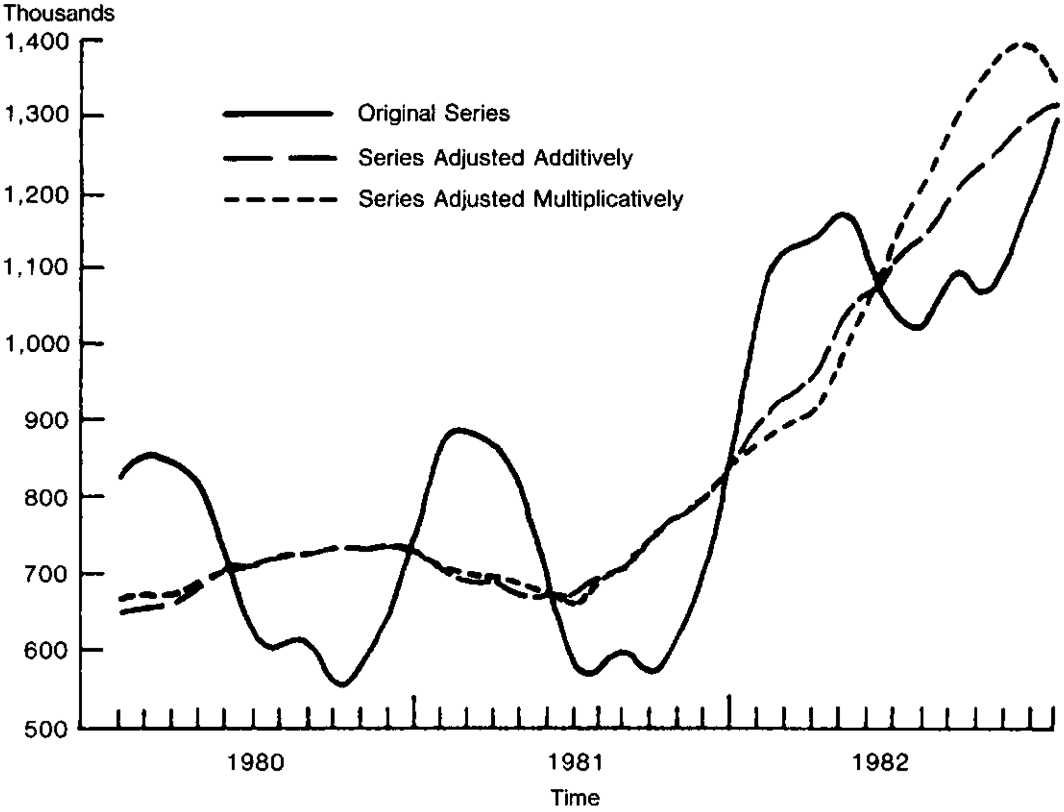


Figure 1. Beneficiaries

The additive model, on the other hand, does not assume that the components of the series evolve proportionately. Figure 1 confirms that the trend cycle increased while the seasonal amplitude remained constant. Thus, the additive model provides the better seasonal adjustment. It performs better in 1982 than the multiplicative model and is acceptable in 1980 and 1981.

By mid-1982, it was not easy to tell which of the additive or multiplicative models would adjust the beneficiaries series better. Since this series was adjusted multiplicatively until June 1981, one would normally continue to do so in 1982. During 1982, were there some clues or pieces of evidence showing that the multiplicative model was no longer adequate?

The acceptance or rejection of model, given a sudden large change in the level of a series, clearly has to be based on a thorough analysis of the data. The set of quality control statistics included in the X-11-ARIMA program is not meant to detect that kind of problem in the model. In this experiment with the multiplicative model, none of the ten individual control statistics failed the guideline. However, the F test for the presence of moving seasonality showed the presence of increasing moving seasonality during 1982 in the final unmodified SI ratios.

Besides a systematic over and under-adjustment of the series, another consequence of using a wrong decomposition model is the possibility of having a false turning point at the end of the series.

Let us say that a cyclical turning point has occurred if the seasonally adjusted series shows a change in direction that persists for at least 5 months. Once the beneficiaries series has been seasonally adjusted multiplicatively, figure 1 shows the possible presence of a turning point around October 1982, where the upward trend has suddendly changed to a downward trend. This turning point seems to be confirmed when the series ending in December 1982 is extended by one month. The additively adjusted series, on the other hand, shows no turning point. The two results conflict. Thus, either the multiplicative model is signaling a false turn or the additive model is missing the turning point.

It is not that easy to show that the multiplicative model has signalled a false turn. The multiplicative model has created a turning point around October 1982. Table 1 shows that in the very short run, the updating of the series did not reverse this turning point.

Table 1

Multiplicatively Adjusted Beneficiaries Series
(in thousands, July 1982 - February 1983)

July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.
124	131	140					
124	130	140	142				
124	130	140	141	141			
124	130	140	142	138	131		
123	130	140	142	141	131	121	
123	129	139	142	141	134	121	123

3. LEAD-LAG RELATIONSHIP BETWEEN THE CLAIMS AND BENEFICIARIES SERIES

Leading indicators are sensitive to the evolution of the economic climate. They are measures of anticipations or new commitments, and as such they give an advance indication of changes expected in the trend-cycle of coincident and lagging indicators.

Figure 2 shows the claims series as a leading indicator for the beneficiaries series. The performance of the seasonally adjusted indicators can be tested using the criteria of Klein and Moore (1982). The two series satisfy these criteria. First, the correspondence between the series is one-to-one - the number of cycles is the same in each series. Second, there is uniformity in timing - the claims series always lead. Third, these are monthly series and they are current, or up-to-date. Thus, the claims series is likely to predict an upward or a downward change in the trend of the beneficiaries series.

The lead-lag relationship between the two series can help to seasonally adjust the beneficiaries series. It reduces the likelihood of mistaking an irregular turn for a cyclical turning point. Figure 2 shows September 1982 to be a turning point in the multiplicatively adjusted claims series. This is also true for the additive adjustment of the series. Since the cross-correlations between the two series shows a lead-lag relationship of 5 to 6 months, the September 1982 turning point in the claims series indicates that the multiplicative model applied to the beneficiaries series has signalled a false turn around October 1982. However, the leading indicator predicts a turning point around March 1983 in the beneficiaries series.

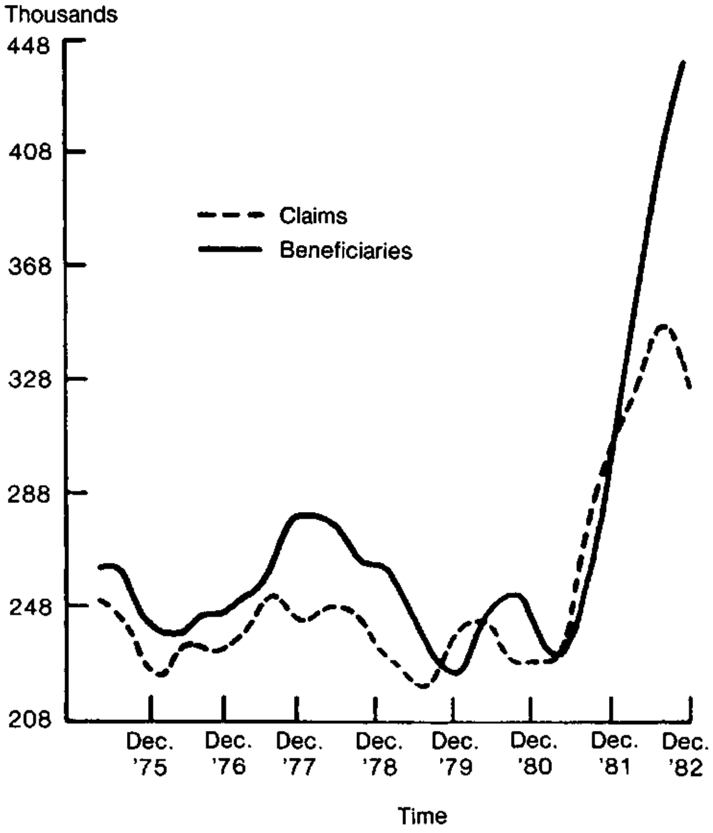


Figure 2. Claims and Beneficiaries. The Number of Beneficiaries has been Divided by 3 in Order to Make the Scale of Both Series Compatible.

4. ARIMA EXTRAPOLATIONS

An optimal seasonal adjustment procedure has to minimize the revision to the current seasonal factors and also has to produce reliable estimates of the trend-cycle, particularly of turning points, at the end of the series (Dagum 1979). The analysis carried on in the previous sections is based on seasonally adjusted data without using the ARIMA option. In this section, we shall focus on the use of the ARIMA forecasts as a variable that can provide an accurate recognition of the turning points.

The automatic X-11-ARIMA program proceeds as follows:

1. Three univariate ARIMA models of the general multiplicative form (p,d,q) $(P,D,Q)_s$ (Box and Jenkins 1970) are fitted to the monthly or quartely series that is to be seasonally adjusted. The models are

$$(0,1,1)\ (0,1,1)_s$$
$$(0,2,2)\ (0,1,1)_s$$
$$(2,1,2)\ (0,1,1)_s$$

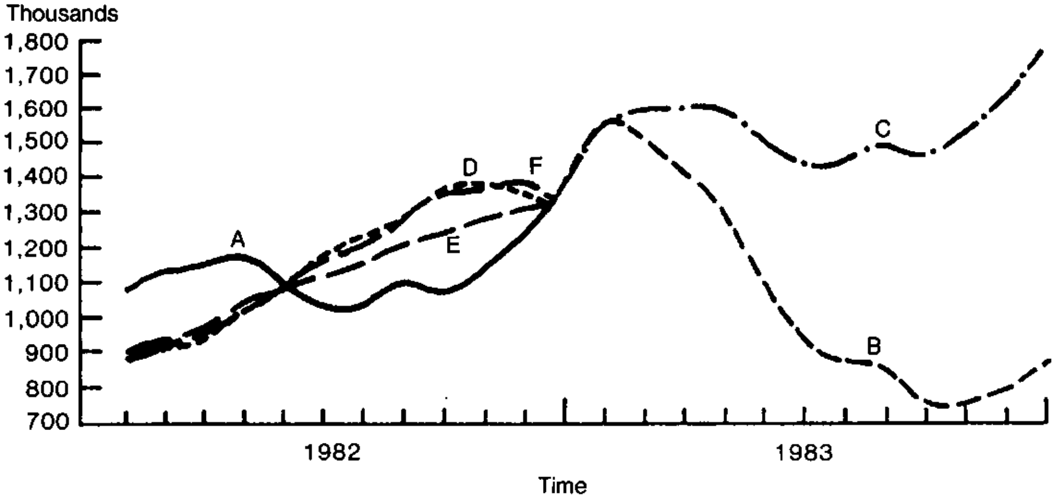
when the series is seasonally adjusted additively. For series adjusted multiplicatively, the same models are used and the log transform is applied to the data for the first two models.

2. The series is extrapolated one year in advance; and
3. provided the extrapolations are acceptable, the ordinary X-11 method is then applied to the series thus extended.

Figure 3 shows the beneficiaries series seasonally adjusted both additively and multiplicatively, using the automatic X-11-ARIMA options. The ARIMA models that best fit and forecast the series ending in December 1982 are (0,2,2) $(0,1,1)_{12}$ when the series is seasonally adjusted additively and log (0,2,2) $(0,1,1)_{12}$ when adjusted multiplicatively. The log (0,2,2) $(0,1,1)_{12}$ model has forecast a decrease in the series, while the (0,2,2) $(0,1,1)_{12}$ model has maintained the upward trend.

Figure 3. shows the multiplicative seasonal adjustment of the beneficiaries series using both the upward trend and the downward trend extrapolations. One can see from the comparison of figure 1 with figure 3 that ARIMA extrapolations did not modify the multiplicative estimates of the trend-cycle in the last year. The multiplicative model is still signalling a turning point around October (downward trend, log transform). The multiplicative model applied to either the non-extended beneficiaries series (figure 1) or to the extended series is questionable.

By the end of 1983, one could see that the true turning point has actually occurred around February 1983. Thus, the October or November 1982 turning point can hardly be corrected by extrapolation when it is due to the wrong selection of the decomposition model.



A. Original Series
B. Extrapolations Using Log (0,2,2)(0,1,1)
C. Extrapolations Using (0,2,2)(0,1,1)
D. Multiplicative Adjustment Using ARIMA Extrapolations with Log (0,2,2)(0,1,1)
E. Additive Adjustment Using ARIMA Extrapolations with (0,2,2)(0,1,1)
F. Multiplicative Adjustment Using ARIMA Extrapolations with (0,2,2)(0,1,1)

Figure 3. Beneficiaries Series Seasonally Adjusted Additively and Multiplicatively with Different ARIMA Extrapolations

Over and under-adjustment and problems of identifying the turning points occurred in other series as well. Figure 4 shows for instance, the series of "benefits paid" when seasonally adjusted multiplicatively with actual d ata available to the end of 1984. The seasonally adjusted series tends to oscillate systematically around the trend-cycle curve at the turning point, thus over- and underestimating the benefits paid. After the turning point, the oscillation decays to the trend-cycle curve; showing that the multiplicative model is doing poorly around the turning point. Note that this series has strong trading-day-variation which has also been removed.

5. SELECTION OF THE OPTIMAL SEASONAL ADJUSTMENT PROCEDURE

Figure 5 summarizes the criteria for seasonal adjustment that have been taken into account to overcome the problems due to the interplay between the 1981-82 recession and seasonal adjustment of the beneficiaries and claims series. The selection of the best seasonal adjustment procedure was primarily based on the first criterion.

In order to avoid over- and underestimation and false turning points in the seasonally adjusted figures, the appropriate decomposition model has to be selected. A thorough analysis of the data should be conducted by:

1. performing a model test on the series.
2. adjusting the series both additively and multiplicatively if the effort is justified. If the differenc e between the two adjustments becomes significant as in figure 1, one has to check for underadjustment in the seasonally high months and for overadjustment in the seasonally low months. One can also look in table D8 of the X-11-ARIMA program at the F tests for the presence of stable and moving seasona lity. The decomposition model that better adjusts the series will usually show the higher F value for stable seasonality and the lower F value for moving seasonality.
3. checking for turning points. For the claims series, both decomposition models have signaled a turn in August or September 1982. On the other hand, for the beneficiaries series, only the multiplicative model has signalled a turn in October 1982. Thus either the multiplicative model is signalling a false turn or the additive model is missing the turning point. The analysis has shown this turn to be a false one resulting from the drastic overestimation of the number of seasonally adjusted beneficiaries in the seasonally low months as shown in Figure 1.
4. using a bi- or multivariate approach to accurately estimate the turning points at the end of the series. The lead-lag relationship between the claims and beneficiaries series can help to seasonally adjust the beneficiaries series. It reduces the likelihood of mistaking an irregular turn for a cyclical turning point. Since the lead is about 5 to 6 months, the September 1982 turning point in the claims series confirms that the multiplicative model applied to the beneficiaries series has signalled a false turn in October 1982. However, the leading indicator is predicting a turning point around March 1983 in the beneficiaries series.
5. using the ARIMA option with concurrent seasonal factors. It usually gives smaller revisions to the seasonal factors wheth er an additive or a multiplicative seasonal adjustment is made. However, a false turning point can hardly be corrected b y extrapolations when it is due to the wrong selection of the decomposition model.

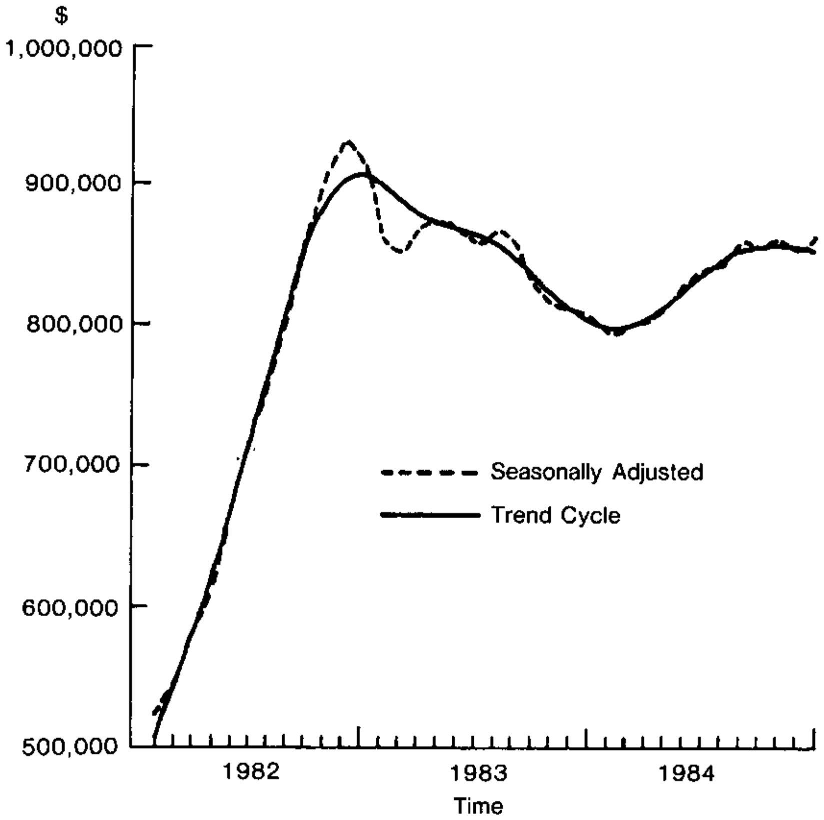


Figure 4. Benefit Paid (Seasonally Adjusted Multiplicatively)

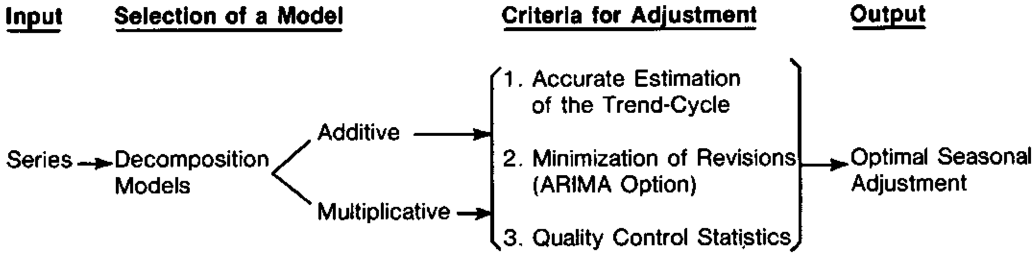

Input
Selection of a Model
Criteria for Adjustment
Output
Series
Decomposition Models
Additive
Multiplicative
1. Accurate Estimation of the Trend-Cycle
2. Minimization of Revisions (ARIMA Option)
3. Quality Control Statistics
Optimal Seasonal Adjustment

Figure 5. Optimal Seasonal Adjustment Procedure

6. checking both the raw and seasonally adjusted data. One cannot rely on tests only. For instance, the set of quality control statistics included in the X-11- ARIMA program is not meant to detect under- or overestimation of the series or false turning points.
7. all the above recommendations apply if the series is not strongly affected by trading-day-variation. If trading-day-variation is present, then it must be removed before the ARIMA option is used.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden Day.

DAGUM, E.B. (1979). Data extrapolation and smoothing with the X-11-ARIMA seasonal adjustment method. *Proceedings of the 12th Annual Symposium Interface Computer Science and Statistics,* (ed. Jane F. Gentleman), University of Waterloo, 195-202.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method.* Statistics Canada, Catalogue No. 12-564E.

HIGGINSON, J. (1977). *User Manual for the Decomposition Test.* Time Series Research and Analysis Division, Statistics Canada, Reference No. 77-01-001.

KLEIN, P.A., and MOORE, G.H. (1982). *The Leading Indicator Approach to Economic Forecasting Retrospect and Prospect.* Center for International Business Cycle Research, Rutgers University, Newark, N.J.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 161–180
Statistics Canada

Nonresponse Adjustment Procedures at the U.S. Bureau of the Census

DAVID W. CHAPMAN, LEROY BAILEY, and DANIEL KASPRZYK[1]

ABSTRACT

Nearly all surveys and censuses are subject to two types of nonresponse: unit (total) and item (partial). Several methods of compensating for nonresponse have been developed in an attempt to reduce the bias associated with nonresponse. This paper summarizes the nonresponse adjustment procedures used at the U.S. Census Bureau, focusing on unit nonresponse. Some discussion of current and future research in this area is also included.

KEYWORDS: Nonresponse adjustments; Imputation; Missing data; Weighting.

1. INTRODUCTION

The Bureau of the Census has long recognized the potential seriousness of measurement errors ascribed to survey nonresponse, and has consistently incorporated nonresponse adjustment or compensation procedures in the estimation methodologies for its numerous and varied surveys and censuses. The objectives of this paper are to provide an overview of procedures employed by the Census Bureau in compensating for nonresponse, primarily unit nonresponse. By unit nonresponse we mean that little or no information for the principal survey variables is obtained for the sample unit in question.

This presentation will include (1) a discussion of the general weighting scheme used for the demographic surveys; (2) a review of some of the distinct problems associated with nonresponse in the Survey of Income and Program Participation (SIPP); (3) a discussion of the handling of unit nonresponse for the economic surveys and censuses; and (4) a section on imputation for earnings for the Current Population Survey. In addition to providing descriptions of the various nonresponse compensation methods used by the Census Bureau, the authors will cite specific problems associated with those methods and note the Bureau's current nonresponse research activities and concerns.

2. NONRESPONSE IN DEMOGRAPHIC SAMPLE SURVEYS

At any given time, the Bureau of the Census may be involved with the conduct of 25-30 recurring or special demographic surveys. The concerns of these surveys include labor force participation, individual and family income, health care, transportation, leisure activities, crime, and other topics reflective of the current interests of the nation's people, governments, businesses, and institutions. Unit nonresponse rates for these surveys range from between three and four percent for the National Crime Survey to over 25 percent, which was recorded for the 1984 National Survey of Natural and Social Scientists and Engineers.

[1] David W. Chapman and Leroy Bailey are Principal Researchers, Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233. Daniel Kasprzyk is a Special Assistant, Office of the Chief, Population Division, U.S. Bureau of the Census, Washington D.C. 20233.

Weight adjustment within classes (Oh and Scheuren 1983), or cell balancing, is the predominant technique used to compensate for unit nonresponse in the Census Bureau's demographic surveys. There is variation among the surveys relative to the determination of adjustment classes within which weighting occurs. For some surveys, ancillary data available to define weighting classes are limited to basic geographic and survey design information, while for others a considerable amount of demographic and economic data are accessible.

The nonresponse adjustment factors for the Bureau's demographic surveys are usually the inverse of the survey's weighted or unweighted response rate. In a small number of the surveys this adjustment factor is modified slightly to reflect information gleaned from follow-up subsamples of the initial nonrespondents. Since the Census Bureau's general approach to survey nonresponse is essentially the same for all of its major demographic surveys, a general description will be given in Section 2.1 of the nonresponse adjustment procedure for the National Crime Survey (NCS), as the example of a "typical" Census Bureau application of weighting. Section 2.2 will consist of a discussion of alternative procedures and current unit nonresponse research in the demographic areas.

2.1 The National Crime Survey

The NCS sample is a national probability sample of about 72,000 households which is divided into six panels, each of which is interviewed in a given month and again at six-month intervals over three years. The survey focuses on measuring household crimes and the extent of victimization of household members age 12 and older by assault (including rape), burglary, larceny, auto theft, and robbery. [For a detailed description of the NCS, see U.S. Department of Commerce, Bureau of the Census (1977).]

Estimates for the NCS, which are produced quarterly, are derived by initially inflating the sample data by the inverse of the related selection probabilities. The noncontacts and refusals account for about three to four percent of the survey's occupied units in any given month. Adjustments for these units are made by applying adjustment factors to the weighted respondent data in weighting classes. An attempt is made to define these classes in such a way that the respondents and nonrespondents in each class have similar survey characteristics. In order to temper the impact of the nonresponse adjustment on the variance of the survey estimates, some of the smaller weighting classes generally have to be collapsed with other classes before a final nonresponse adjustment can be effected. Collapsing of classes also takes place if the weight adjustment factor becomes too large for one or more classes. [See Hanson (1978).] Collapsing is discussed further in Section 4.

Since the NCS employs a self-response method of interviewing, there is concern about the amount of within household nonresponse. Consequently, a separate set of weighting cells exists to compensate for within-household nonresponse. These cells or weighting classes, as well as those used for the household nonresponse adjustment, are indicated in Tables 1–3. The NCS household and within household nonresponse rates for 1984 are shown in Table 4.

To illustrate the NCS estimator of a total, there is a selection probability $\pi_i = 1, 2, \ldots, N$, associated with each of the N units in the population. It is assumed that among the n sample units, n_R are respondents. The NCS estimator for the population total, after adjusting for unit nonresponse, takes the following form:

$$\hat{Y}_{\text{NCS}} = \sum_{j=1}^{M} \sum_{k=1}^{P} (z_j u_k)^{-1} \sum_{\ell=1}^{n_{Rjk}} \frac{y_{jk\ell}}{\pi_{jk\ell}},$$

Table 1

NCS Noninterview Adjustment Cells for
Within Household Nonresponse

Household Relationship	Persons by Age, by Race of Head							
	Black				Non-black			
	12-24	25-44	45-64	65+	12-24	25-44	45-64	65+
Head of Household								
Wife of Head								
All other Persons								

Table 2

NCS Household Noninterview
Adjustment Cells for
Standard Metropolitan
Statistical Areas (SMSA's)

Race	Central City of SMSA	Balance of SMSA	
		Urban	Rural
White			
Not White			

Table 3

NCS
Household Noninterview
Adjustment Cells for
Non-SMSA's

Race	Urban	Rural	
		Non-farm	Farm
White			
Not White			

where for sample units in the k^{th} within household and j^{th} household weighting classes,

$y_{jk\ell}$ = value of the ℓth sample respondent,
n_{Rjk} = number of sample respondents,
n_{jk} = number of sample cases,
z_j = the estimated household response rate,
u_k = the estimated within household response rate,
$\pi_{jk\ell}$ = selection probability for the ℓth sample respondent,
P = total number of within household nonresponse weighting classes,
M = total number of household nonresponse weighting classes.

Implicit in the formation of the NCS nonresponse weighting classes, as well as those for other demographic surveys, are the following assumptions:

1. There is "significant" correlation between the major survey variables and the covariates used to define noninterview adjacent classes.
2. Within each household nonresponse weighting class, $E(\bar{y}_{Rj}) = E(\bar{y}_{\bar{R}j})$, where $\bar{y}_{Rj}$ and $\bar{y}_{\bar{R}j}$ are the means for the sample respondents and nonrespondents, respectively, in the j^{th} weighting class.
3. The weighting class means differ, that is, $E(\bar{y}_{Rj}) \neq E(\bar{y}_{Rj}')$, $j \neq j'$.

(Assumptions analogous to 2 and 3 above are also implicit for *within* household nonresponse adjustment classes.)

Table 4

NCS Noninterview Rates - 1984

	Average 1984	Jan.	Feb.	Mar.	Apr.	May	June
Household Noninterviews							
Total Interviewed HH's	11,769	11,916	11,925	11,743	11,809	11,918	9,482
Total	430	446	540	481	446	388	348
Rate	3.5	3.6	4.3	3.9	3.6	3.2	3.5
No one at home	0.9	0.8	1.1	0.9	0.9	0.7	1.0
Temporarily Absent	0.6	0.6	0.6	0.8	0.6	0.4	0.7
Refusal	1.9	2.1	2.6	2.2	2.2	2.0	1.9
Other	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Within Household Noninterviews							
Total	685	655	751	701	806	804	697
Rate	2.5	2.6	3.0	2.8	3.0	2.9	3.2
		July	Aug.	Sept.	Oct.	Nov.	Dec.
Household Noninterviews							
Total Interviewed HH's		9,869	9,446	9,895	9,350	9,692	9,410
Total		411	409	337	406	387	346
Rate		4.0	4.2	3.3	4.2	3.8	3.5
No one at home		0.9	0.9	0.6	1.0	1.2	1.0
Termporarily Absent		1.0	1.0	0.6	0.6	0.4	0.4
Refusal		2.1	2.3	2.0	2.4	2.1	2.1
Other		0.1	0.1	0.1	0.3	0.3	0.1
Within Household Noninterviews							
Total		709	678	666	728	735	803
Rate		3.1	3.1	2.9	3.4	3.3	3.7

The selection of weighting classes for this procedure is constrained by the requirement that measurements for the weighting class variables (covariates) must be available (either before or during the survey) for both the respondents and the nonrespondents. This essentially restricts the characteristics by which classes are defined to those associated with geography, race, urbanicity, housing unit characteristics, and design levels. The bias reduction capability of the procedure depends, in part, on the extent to which the NCS nonresponse weighting classes satisfy the three assumptions given above. No definitive results relating to this concern are currently available, but relevant research is underway and more empirical studies seem warranted.

2.2 Alternatives to Sample Weighting

There are a number of plausible alternatives to weighting to adjust for nonresponse. See, for example, Little (1986, Section 5). However, there are no definitive results which show that any of them offer appreciable advantages. Subsections 2.2.1 and 2.2.2 contain brief descriptions of two alternatives which are currently being investigated for application to demographic surveys.

2.2.1 Separate Estimates for Dissimilar Types of Nonresponse

In demographic surveys, nonrespondents can be placed into four categories: refusal (REF), not-at-home (NAH), other occupied unit (OTO), or a unit from which a response was not obtained due to extenuating circumstances. These are referred to as type A noninterviews. The NAH group can be divided into those households or individuals whose extended absence from their homes precludes an interview during the scheduled interview period (NAH_E), and the group which is expected to return home sometime during the survey period (NAH_S).

The authors are not aware of any data which show that the four nonresponse groups are generally similar. In fact, the Census Bureau's Current Population Survey and the Canadian Labour Force Survey suggest that the NAH_S households are likely to be smaller, younger, and have a larger proportion of employed people than the other groups. The NAH_E group is usually older with a relatively low employment rate. The interviewed group may be more reflective of the REF and OTO groups. [See Palmer and Jones (1967) and Paul and Lawes (1982).] It is conceivable that separate treatment of the four nonresponse groups could produce a better overall adjustment for nonresponse than is obtained from the current procedure. This option is being investigated by an NCS nonresponse adjustment research group.

2.2.2 Weighting With Response Probabilities

Several weighting techniques have been advanced which make use of the concept of response probabilities. Most of these techniques are based on concepts introduced by Politz and Simmons (1949) which group sample respondents according to estimates of their probabilities of responding. The factors with which the sample data in the resultant weighting groups are inflated are the inverses of the estimated response probabilities. The Politz-Simmons procedure has some serious limitations, such as its inapplicability to refusals. However, there have been a number of fairly recent extensions and applications of the procedure, including those presented by Anderson (1978), Thomsen and Sirling (1983). These methods may be applicable to recurring surveys for which extensive callbacks are made.

Research is in progress regarding the development of models which may be used to estimate response probabilities for several demographic surveys for units with similar values of the "independent variables." The feasibility and merits of computing nonresponse adjustment factors, as well as constructing weighting classes based on such models (sometimes referred to as response propensity stratification), are being examined. [See Rosenbaum and Rubin (1983) and Little and Samuhel (1983).] Moreover there are continued efforts to develop more objective methods of sample weighting for nonresponse, which are designed to control nonresponse-related errors.

3. THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

The Survey of Income and Program Participation (SIPP) is a new, ongoing national household survey program of the U.S. Bureau of the Census. The purpose of SIPP is to improve the measurement of information related to the economic situation of households and persons in the United States. It is the culmination of a large-scale development program, the Income Survey Development Program (ISDP), which examined concepts, procedures, questionnaires, recall periods, and the like. For a description of the ISDP, see Ycas and Lininger (1981). Data from SIPP are expected to be useful in studying the Federal transfer system, estimating program costs under changes in program eligibility rules, evaluating the effects of program changes on selected population subgroups, as well as studying changes to the tax system.

In October 1983 SIPP began as an ongoing survey program with one sample panel of approximately 21,000 occupied households eligible for interview in 174 Primary Sample Units (PSU's) selected to represent the noninstitutional population of the United States. (Beginning in 1985 a new panel is being introduced in February of each year; the 1985 panel consisted of 14,500 households eligible for interview.)

Each household is interviewed once every four months for approximately 2½ years to produce sufficient data for longitudinal analysis while providing a relatively short recall period for reporting monthly income. The reference period for the principal survey items is the 4 months preceding the interview. This design provides eight interviews per household, and allows cross-sectional estimates to be produced from more than one panel.

To facilitate field and processing operations, each sample panel is divided into four approximately equal subsamples, called rotation groups; one rotation group is interviewed in a given month. Thus, one cycle or "wave" of interviewing, using the same questionnaire, takes four consecutive months. Cumulative *household* noninterview rates are given in Table 5 for the 1984 SIPP panel.

At the time of the interviewer's visit, each person 15 years old or older who is present is asked to provide information about himself/herself; a proxy respondent is asked to provide information for those who are not available. An important design feature of SIPP is that all persons in a sample household at the time of the first interview remain in the sample even if they move to a new address during the next 2½ years. For cost and operational reasons, in-person interviews are only conducted at new addresses that are within 100 miles of a SIPP primary sampling unit. The geographic areas defined by these rules contain over 96% of the U.S. population. An attempt is made to conduct a telephone interview with those moving outside the 100-mile limit.

Table 5
Cumulative Household Noninterview Rates
for the 1984 SIPP Panels

Wave	Sample Loss
1	4.9%
2	9.4%
3	12.3%
4	15.4%
5	17.4%
6	19.4%
7	21.0%
8	22.0%
9	22.3%

After the first interview, the SIPP sample is a person-based sample, consisting of all individuals who were living in the sample unit at the time of the first interview. Individuals aged 15 and over who subsequently share living quarters with original sample people are also interviewed in order to provide the overall economic context of the original sample persons.

More detailed information concerning the SIPP design, content, and operations can be found in Nelson, McMillen, and Kasprzyk (1985).

3.1 Nonresponse Adjustments in SIPP

Data collected in SIPP can be viewed from two perspectives: cross-sectional or longitudinal. From the former point of view, each SIPP interview is treated as a separate cross-sectional survey, providing point-in-time estimates. For examples of these estimates, see U.S. Department of Commerce, Bureau of the Census (1984a). From the longitudinal point of view, data are collected at more than one point-in-time, and the survey record is viewed not as a set of unrelated observations, but as a set of variables with logical dependency between two or more points-in-time. Data processing operations, as well as statistical estimation, are treated from this point of view, and therefore, rely on the use of data collected at two or more interviews.

Since SIPP can be viewed from both the longitudinal and cross-sectional perspectives, SIPP's public-use microdata files include cross-sectional data files issued on a wave-by-wave basis as well as longitudinal files. This implies two distinct systems to treat survey nonresponse.

3.1.1 Cross-Sectional Unit Nonresponse Adjustments

The cross-sectional unit nonresponse adjustment in SIPP is similar to the way noninterview adjustments are made in other Census Bureau recurring surveys. The following variables were used to define household noninterview adjustment cells for the first interview wave of SIPP. See U.S. Department of Commerce, Bureau of the Census (1983 and 1984b).

1. Census Region - Northeast, Midwest, South, West.
2. Residence - Standard Metropolitan Statistical Area (SMSA), non-SMSA.
3. Place/not place - defined for units not in an SMSA,
   Central city/balance - defineds for units in SMSA's.
4. Race of reference person - black, non-black
5. Tenure - owner of home, renter.
6. Household size - 1, 2, 3, 4 or more.
7. Rotation group - 1, 2, 3, 4.

Two criteria must be met by each weighting class: (1) the weighting class must contain at least 30 unweighted units and (2) the noninterview adjustment factor for a weighting class must be less than or equal to 2.0. For a given rotation group, the collapsing procedure to satisfy these two criteria is applied independently for each of the four tenure by race combinations. (For the first wave, there was no *within*-household nonresponse adjustment factor.)

In subsequent waves of SIPP, the household nonresponse adjustment factor accounts for noninterviews associated with units which have moved and cannot be located or have moved more than 100 miles from a SIPP PSU and cannot be contacted by telephone as well as units which are refusals, etc. Adjustments are performed for each month of the reference period, as well as the interview month, to account for an increase in the number of noninterviews caused by splits of sample households. The procedure is similar to that described for determining the Wave 1 household nonresponse adjustment factor; however, the variables used to define the weighting classes differ. Those variables are:

1. Race (white, nonwhite) and Spanish-origin (Spanish, non-Spanish) of reference person: a) reference person is white and not Spanish, and b) others.
2. Household type - three categories: a) female householder, no husband present, with own children under 16, b) householder's age is sixty-five years or older, and c) others.
3. Education level of the reference person: a) less than 8 years, b) 8-11 years, c) 12-15 years, and d) 16 or more years.
4. Type of income received (using the most recently completed interview for members of the household) - two categories: a) households which received at least one of the following sources of income - Supplemental Security Income; Black Lung Payments; Aid to Families with Dependent Children; General Assistance, Indian, Cuban, or Refugee Assistance; foster child care payment; Women's, Infants', and Children's Nutrition program; Food Stamps; and Medicaid; and b) others.
5. Assets - two categories: a) households in which at least one member held an asset type other than a savings account or an interest-bearing checking account, and b) all others.
6. Tenure: a) owner of home and b) renter.
7. Public housing or rent subsidies--renters are identified as a) those living in public housing projects or receiving rent subsidies from the government; and b) those not living in public housing projects and not receiving rent subsidies from the government.
8. Household size: 1, 2, 3, 4 or more.

The variables used for household nonresponse adjustments for the second and subsequent SIPP interviews differ from the first wave variables because of additional data available after the first interview for use in nonresponse procedures for later interviews. Fifty-three weighting classes were created using these variables with tenure as the principal variable for partitioning the sample. [For a description of these weighting classes see U.S. Department of Commerce, Bureau of the Census (1984c).] Although a cell collapsing strategy has been defined which merges cases in cells exhibiting similar poverty-related characteristics, little collapsing takes place since the nonresponse adjustment factors are calculated for three rotation groups (the SIPP data processing cycle) rather than one rotation group, as in the first interview.

There is a within-household nonresponse compensation procedure for the second and subsequent waves. This procedure is to "hot deck" (i.e., duplicate) the entire record of a sample respondent who presumably has survey characteristics that are similar to those of the nonrespondent.

3.1.2 Longitudinal Nonresponse Adjustments

Since persons identified as living at the sample address at the time of the first interview constitute the SIPP sample for waves subsequent to the first, the most useful and logical way of describing the nature of the SIPP nonresponse problem from the longitudinal viewpoint is in terms of individuals or persons. Each individual's microdata record is an extended record containing variables which oftentimes reflect the same measure at different points in time. Thus, in a panel survey of n waves there exist 2^n possible noninterview patterns for a sample person. Noninterview patterns of the original sample persons for the first five interviews (waves) of the 1984 panel are given in Table 6, adapted from Kalton, McMillen, and Kasprzyk (1986).

Table 6

Interview patterns of the Original Sample Persons for the First Five Interviews of the 1984 SIPP Panel

Response Pattern	Percent
Response every interview (5 interviews)	
Pattern: XXXXX	79.1
Apparent attrition cases	13.8
Patterns: XXXXO	3.8
XXXOO	3.1
XXOOO	3.2
XOOOO	3.7
First and fifth interviews conducted, but one and more intervening interview missing	4.1
Patterns: XXXOX	1.6
XOXXX	0.6
XXOXX	1.2
XXOOX	0.1
XOXOX	0.1
XOOOX	0.3
XOOXX	0.2
Fifth interview missing and one or more intervening interviews missing	0.7
Patterns: XOXXO, XOXOO, XOOXO, XXOXO	
Left the universe (deceased, institutionalized, living in armed forces barracks, moved overseas)	2.3
Total	100.0 (25,128)

The first SIPP longitudinal microdata file will contain twelve months (three interviews) of data from the 1984 SIPP panel, with the individual as the principal analytic unit. The sample of cases to be weighted for this file will be only those persons with three completed interviews. Those sample persons with only one or two interviews will be treated as nonrespondents. Their reported data will help to define nonresponse adjustment classes.

Since the first microdata longitudinal file contains only persons responding to all three interviews, the nonresponse adjustment issue is virtually the same as for the cross-section case. There are, however, two nonresponse adjustment factors applied to the initial sampling weights. See Kobilarcik and Singh (1986). The first adjustment factor accounts for households classified as noninterviews in the first interview wave. The second factor accounts for persons who did not supply all three interviews.

For the first adjustment factor, only those household variables available at the first interview can be used. Adjustment factors are calculated separately within cells defined by the following variables:

a. Census Region
b. Residence (metropolitan, non-metropolitan)
c. Race of reference person
d. Tenure (own, rent)
e. Household size

The second set of adjustment factors is implemented on a person basis. The factors are calculated within cells defined by the following characteristics:

a. Monthly household income
b. Program participation status of the person's household
c. Labor force status
d. Race
e. Years of school completed
f. Type of assets of person's household

Cells are collapsed whenever they do not contain thirty sample persons or the nonresponse adjustment factor exceeds 2.

As the survey progresses, more sophisticated methods of adjusting for longitudinal nonresponse will be developed which make use of the data provided for partial respondents (i.e., for sample persons that provide some, but not all, of the interview waves requested). It is not obvious how to treat the partial response cases. Data gaps associated with persons who miss one or more interviews can be viewed as either person nonresponse, and typically handled by weighting adjustments, or as item nonresponse, usually handled by some type of imputation method. For example, one might consider an individual with a (R,NR,R) pattern as a case of item nonresponse since the missing interview is bounded on both sides by completed interviews; but one might consider an individual with an (NR,R,NR) pattern as total unit nonresponse, treating it the same as (NR,NR,NR). However, we need to recognize that even in the case of the response pattern (R,NR,R) for an individual, four kinds of response patterns are still possible at the item level. Thus, many options can be considered when developing nonresponse compensation procedures for the SIPP longitudinal data base. This issue is discussed by Kalton (1986) and by Kalton, Lepkowski, and Lin (1985).

3.2 SIPP Research Activities

There are two areas where work has recently begun which should aid future decisions concerning nonresponse adjustments. First, the SIPP questionnaire, beginning during the fourth interview, contains a "Missing Wave" section. This section uses a short series of questions on labor force participation, income sources, and asset ownership/nonownership for respondents in the current wave who did not respond in the preceding wave. Respondents who miss two or more consecutive interviews are not eligible to complete the "Missing Wave" section. By emphasizing data collection at the expense of minor reporting burden, the person nonresponse problem can be reduced to an item nonresponse problem. An evaluation of the quality of the retrospective data will be necessary prior to using these data.

The second area of work concerns general strategies in the treatment of person-wave nonresponse in the SIPP. Graham Kalton and his colleagues at the Survey Research Center will (1) compare longitudinal imputation and weighting strategies for handling person-wave nonresponse, (2) evaluate imputation and weighting models in terms of the analysis of change across waves and aggregation across waves, and (3) develop preliminary criteria for the choice of method for treating person-wave nonresponse. A discussion of these and other issues which will be studied can be found in Kalton (1986), and Kalton and Miller (1986).

Finally, there are several other research topics for which work is planned. These include: (1) quantifying the selection of variables used for determining weighting classes; (2) assessing the robustness of the survey estimates on the population and selected subgroups under different nonresponse compensation procedures, and different weighting class cell collapsing strategies; (3) investigating the potential for making separate nonresponse adjustments by type of noninterview; (4) investigating the effect of deleting reported survey data to simplify the nature of the SIPP missing data problem; and (5) evaluating the longitudinal nonresponse compensation procedures adopted for the first SIPP longitudinal research file.

4. UNIT NONRESPONSE PROCEDURES FOR ECONOMIC CENSUSES AND SURVEYS

The Bureau of the Census carries out six economic censuses every five years, the most recent ones covering 1982. These six economic censuses are identified by the following trade areas:

(1) Retail Trade
(2) Wholesale Trade
(3) Service Industries
(4) Manufactures
(5) Mineral Industries
(6) Construction

In addition to the economic censuses, the Census Bureau carries out the Census of Governments and the Census of Agriculture. Though not part of the economic censuses, they are conducted during the same years as the economic censuses for processing efficiencies and to allow for data linkage. In nearly all of these economic areas the Census Bureau also carries out a number of monthly, quarterly, and annual surveys.

Like the demographic areas, there is some unit nonresponse for all of the economic censuses and surveys. In most cases, missing data are imputed based on (a) previous responses provided by the nonrespondent, (b) data from administrative records, and (c) relationships established between various data items. Rather than reporting the percent of units not responding, the level of nonresponse for an economic census or survey is usually given as the percent of one or more item totals that are imputed. These percents will be referred to as imputation rates.

Explanations of the unit nonresponse methods used for five of the six economic censuses are given in Section 4.1. Section 4.2 addresses unit nonresponse procedures for three economic surveys, and Section 4.3 covers such procedures for the Census of Agriculture. Research and evaluation activities with regard to nonresponse procedures for economic censuses and surveys are discussed in Section 4.4.

More detailed explanations of the nonresponse procedures used in these censuses and several related surveys are given by Bailey, Chapman and Kasprzyk (1985).

4.1 The Economic Censuses

The frame for the economic censuses is the Standard Statistical Establishment List (SSEL), a computer file maintained by the Census Bureau. The SSEL is comprised of all employer establishments reported by multi-unit employer companies in the Census Bureau's Company Organization Survey (COS) and all single-unit companies that filed a tax form with IRS. The COS is an annual survey of multi-unit companies. Companies that have at least 50 employees are surveyed each year, while companies with fewer than 50 employees are surveyed every three years. Each company in the COS is sent a list of the establishments it reported most recently in the survey and asked to update the list. They are also asked to provide, for each establishment, employee counts for the first quarter of the previous year and total payroll for the previous year. For the economic censuses, each establishment on the SSEL, except small single-unit establishments, is sent a census questionnaire (via its company) designed for its standard industrial classification (SIC) code.

Although there are many similarities among the unit nonresponse procedures used in the six trade areas, some important differences exist. In the following description of the unit nonresponse adjustment procedures used for five of the economic censuses, the trade areas that use essentially the same procedure will be grouped together as follows:

(a) Retail trade, wholesale trade, services

(b) Manufactures, mineral industries

4.1.1. Retail Trade, Wholesale Trade, Service Industries

These three parts of the economic censuses are often referred to collectively as the business census. For these trade areas, data for the census year are collected on sales receipts, employment, and payroll. The imputation rate for sales/receipts varies from 10 to 15 percent for retail and wholesale trade and is about 20 percent for service industries.

For any establishment that does not provide the census data, responses are generally imputed using tax form information available from the Internal Revenue Service (IRS). For payroll information, the IRS has four quarters of data available for each employer identification (EI) number from tax forms. A company may have one or more EI numbers. Payroll data for a particular company are obtained by adding up the payroll figures for all EI numbers used by the company. First quarter employment counts are also available by EI number from IRS records and can be aggregated to the company level. For sales/receipts, various IRS forms are used depending on whether the nonresponding company is a sole proprietorship, partnership, or corporation.

The imputation procedure is complicated by the difference between the census enumeration unit and the IRS tax unit. For the business census, the unit of enumeration is the establishment (i.e., a single location). However, the tax unit for the IRS is an EI number. There may be one or more establishments reporting under the same EI number. If a nonresponding company has only one location (i.e., is a single-unit company), then it will have only one EI number and imputation is straightforward. However, for a multi-unit nonresponding company imputation is more complex since, in general, IRS data will not be available for each establishment. In such a case, the company structure is determined first by referring to the SSEL to obtain a list of all establishments contained in a company and all EI numbers used by the company. The total for a company for each data item is obtained by adding the item across all EI numbers used by the company, as discussed above. The company total is distributed to establishments by prorating the total based on the most recent data available for the company from an annual or monthly survey. If no data are available, an equal proration is used. If there is nonresponse for only a portion of the establishments in a multi-unit company, data for the nonresponding establishment are imputed based on prior year relationships.

4.1.2 Manufactures, Mineral Industries

In these two economic censuses, general information is obtained on the number of employees, hours worked, and on production levels by four-digit standard industrial classification (SIC) codes. Imputation rates vary from about 10 to 15 percent. The unit nonresponse procedure used depends on the type of company that did not respond (i.e., single-unit or multi-unit) and on whether or not a previous year's record is available. Thus, there are four types of nonresponse cases that occur. The method of treating nonresponse for these four cases follows:

(1) Single-unit company, previous year data are available from the Annual Survey of Manufactures.

In this case annual payroll is obtained from IRS tax forms and compared to the previous payroll total reported. The percent change from the previous period is determined. This percent change is applied to all data items in the previous record to obtain an imputed current record,execpt for employment and value of shipments whenever these are available from IRS.

(2) Single-unit company, no previous year data are available.

In this case, sets of ratios are developed between census items within each four-digit SIC, with payroll as the "seed." That is, the relationships are developed in such a way that all items can be imputed from these relationships either directly or indirectly if a payroll figure is obtained. The specific relationships are derived from historic data reported by the respondents in the same industry. Then the (seed) value of payroll is obtained from IRS tax records and all other items are imputed from the relationships derived.

(3) Establishment in a multi-unit company, previous year data are available for the establishment.

First, for each four-digit SIC, an aggregate growth factor between the previous and current period is developed from external sources for each of the following key items: payroll, employment, change in inventory, and change in capital expenditures. These four growth factors are applied to the appropriate prior year data items for each establishment to obtain imputed responses for the current period. These four imputed items are then used as "seeds" to impute other items.

(4) Establishment in a multi-unit company, no previous year data are available for the establishment.

In this case, basic data on payroll and employment are obtained for each establishment from the SSEL discussed earlier in Section 4.1. As indicated, the SSEL obtains data on employment and payroll obtained for all establishments included in the COS. Then, using the SSEL data as a base, the data record for each establishment is imputed from relationships developed between the SSEL data items and the other census items. This procedure is analogous to that used in case (2) above.

4.2 Economic Surveys

The Census Bureau conducts a large number of monthly, quarterly, and annual economic surveys in addition to the economic censuses. In particular, most of the six census trade areas have monthly or annual surveys. The unit nonresponse procedures used for the Monthly Retail Trade Survey and the Truck Inventory and Use Survey are described below. The unit nonresponse adjustment procedure used for the Annual Survey of Manufactures (ASM) is not described here since it is virtually the same as that used for the Census of Manufactures, described in Section 4.1.2. Imputation rates for the ASM vary from 5 to 10 percent.

4.2.1 Monthly Retail Trade

The Monthly Retail Trade Survey includes about 30,000 reporting units: about 3,000 selected with certainty and 27,000 selected on a probability basis. The certainty cases are surveyed each month, while a third of the noncertainty cases are surveyed each month. This provides a monthly mailing of about 12,000 reporting units. For a multi-unit company in the survey, a subsample of the establishments in the company is selected for inclusion. Monthly retail sales is the only item enumerated in the survey. The imputation rate for retail sales is about 11 percent.

If a single-unit certainty company or a sample establishment in a multi-unit certainty company does not report for a given month, a value for sales is imputed from the previous month's figure by multipling it by a "ratio of identicals." This adjustment ratio is derived by dividing the weighted sum of the current month sales by the weighted sum of the previous month sales for all establishments in the same adjustment cell for which sales were reported for both the current and previous months. Adjustment cells are generally defined by the first three digits (or four digits in a few cases) of the SIC code, by type of establishment (i.e., whether or not it belongs to a large multi-unit firm), and by sales size class. The weight used for each reporting unit used in computing the ratio of identicals is the inverse of the probability of selection of the reporting unit.

If a multi-unit certainty company does not report sales for any of its establishments, the sales values are imputed for each establishment and for the entire company as in the previous case: applying the ratio of identicals for the appropriate adjustment cell to the previous month sales figures. If such a company does report current monthly sales for the entire company, the imputed establishment responses are ratio adjusted to be consistent with the reported total for the entire company.

For noncertainty companies, imputation for missing sales data is carried out in a way similar to that used for certainty cases, except that an extra step is required since noncertainty companies report every three months. The first step is to impute the previous month's sales for a nonrespondent based on the response provided three months ago. This is done by multiplying the sales reported three months ago by a ratio of identicals based on the weighted sum of sales during the previous month and the weighted sum of sales three months ago (cell by cell). Once the previous month sales are imputed, the current month sales is generated from the imputed value for the previous month using the same method described for certainty cases.

If a nonrespondent is in the survey for the first time, the previous month's sales (if it's a certainty case) or the sales figure three months earlier (if it's a noncertainty case) is imputed from the sales reported in the most recent census, if available. If the nonrespondent was not in the most recent census, then it would be a birth case for which two months of sales data generally would have been provided at the time the company was added to the frame. This data would be seasonally adjusted and then inflated to an annual-based figure. The imputation would then be carried out as though a census sales figure had been available for the nonrespondent.

4.2.2 Truck Inventory and Use Survey (TIUS)

The TIUS is conducted every five years and provides data on the physical and operational characteristics of trucks nationwide. These characteristics include type of trailer (vehicle configuration), kinds of products carried, type of gasoline used, and annual miles driven. The universe for the survey consists of the truck registrations from all 50 states and the District of Columbia. The sample size is about 120,000 truck registrations. About 75 percent of the trucks selected for the survey respond.

Adjustments for unit nonresponse are made by "weighting up" the respondents to the total sample, separately within weighting classes. The weighting classes are taken to be the sample strata which consist of cross-classifications by state and body type (5 categories). The nonresponse weight adjustment is based on the number of trucks; within each class (stratum), the initial weight of each respondent is multiplied by the ratio of the number of trucks in the stratum to the sum of the initial weights of the respondents in the stratum.

Of the economic surveys investigated, the TIUS is the only one that uses a weight adjustment procedure to account for unit nonresponse. With other economic surveys, alternate sources of basic information are generally available to "build" a record for a nonrespondent.

4.3 Census of Agriculture

The census of agriculture provides data relating to the Nation's farming, ranching, and related activities. It is the leading source of agricultural statistics and the only source of consistent, comparable data about agriculture at the county, State, and national levels.

The task of nonresponse adjustment for the census of agriculture is made complex by the fact that the SSEL cannot be as effectively used as it is in the other economic areas. The agricultural census mailing list is constructed by combining several overlapping sources. The resultant frame may contain some duplication and always contains some nonfarm entities. Thus, the nonresponse methodology must first identify, or estimate, the extent to which an adjustment is needed before it can take place.

For the 1982 census, nonrespondents were designated as large or small based on whether their expected sales were above or below $100,000. A 100% telephone follow-up was conducted for all of the large nonrespondents. The small nonrespondents were then stratified based on other mail list characteristics. A sample of these units was followed up by mail and telephone to obtain estimates, by strata within states, of the percent of nonrespondents which were actually farms. These estimates were then used, along with data on in-scope percents of *respondents* by county, to make estimates of the number of nonrespondent farms at the county level for each stratum. The weights of a randomly selected sample of respondents by county, consistent with the estimated number of nonresponding farms, were then inflated by two. All other respondents retained their weight of one.

4.4 Research Activities for Nonresponse Adjustments in Economic Surveys

Probably the most important source of information for unit nonresponse imputation in economic surveys is IRS data from tax forms. Some differences between the IRS figures and those collected in the economic census may arise because of differences in definitions, forms, or the data collection procedures used. A study by Dyke (1984) compared administrative (IRS) data used to impute sales/receipts, payroll, and employment in the 1977 business census with corresponding responses obtained in a follow-up sample of nonrespondents. In general, he found that the survey values reported in the follow-up survey exceeded those obtained from administrative sources. The sizes of the differences varied by item. Also, the differences were more pronounced for multi-unit establishments. Additional comparisons of this type are needed. If systematic differences are identified, adjustment factors to apply to IRS figures may be developed.

For several of the censuses and surveys, a "ratio of identicals" is calculated and used to obtain a factor to apply to a previous-period figure to obtain an imputed value for the current period. It is possible that this ratio computed among *all* sample cases that reported in both periods may not apply very well to the nonrespondents for some items. Bailey (1986) looked at alternatives to using ratios of identicals for imputing missing values such as linear regression and quadratic regression, using various sets of independent variables.

With many of the economic unit nonresponse imputation methods, the sample cases - both respondents and nonrespondents - are placed into cells prior to computing (a) some type of ratio between current and prior periods for an item or (b) some type of relationship between the survey items and the basic items: payroll, employment, and receipts. A research project to investigate alternate choices of cell definition for the Monthly Retail Trade Survey was recently completed by Huang (1986). She found that for some SIC's an alternate procedure of defining cells reduces the mean square error (MSE) of estimated sales substantially. In addition, she compared the current method of imputing - using ratios of identicals - to three alternate methods with respect to bias and MSE. The current method was evaluated as the second best procedure. However, she concluded that the slight gains of the optimum procedure may not be worth the additional requirements associated with using it.

5. IMPUTATION FOR EARNINGS IN THE CURRENT POPULATION SURVEY

5.1 The Hierarchical Hot Deck

The Current Population Survey (CPS) is a Census Bureau ongoing monthly survey of about 60,000 U.S. households per month. The CPS, sponsored by the Bureau of Labor Statistics, primarily collects labor force and employment information. Each March, the CPS administers an income supplement as part of the survey questionnaire. About 11-12% of the sample members do not respond to the income questions. Therefore, a special procedure, referred to as the "hierarchical hot deck," has been developed to impute for missing responses.

With the hierarchical hot deck, missing earnings values are inserted from the response record of another sample unit - a donor. The goal in selecting a donor is to find one with survey characteristics similar to those of the item nonrespondent. The first step in the process of finding suitable donors is to partition the entire sample, excluding total noninterview cases, into cells based on multi-way classifications of a number of survey characteristics. Within each cell a list is made of the respondents and nonrespondents for a given item. Donors from the list of respondents are assigned to the nonrespondents systematically, with a random start. If there are more nonrespondents than there are respondents in a cell for a given item, the responses of some, or perhaps all, of the respondents in the cell will be used more than once. In some cells, there may be one or more nonrespondents but no respondents for an item.

To avoid the problem of having nonrespondents with no donors available, the process of defining cells and selecting donors for the item nonrespondents is carried out several times. At each stage, fewer cells are defined than were defined for the previous stage. For the final stage the number of cells defined is small enough so that it is certain that there will be donors available in each cell. The cells defined at successive stages are formed by collapsing the cells used at the previous stage. Each item nonrespondent will have one or more donors assigned. The donor used to obtain an imputed value will be the one identified at the earliest stage.

The major advantage of this hierarchical procedure is that a very large number of cells can be defined at the first stage, due to the backup stages used. Whenever a donor is found at the first stage, the item nonrespondent and donor will be matched on a large number of survey characteristics. In such cases there should be a good chance that an adequate imputation is made. In other cases the item nonrespondents and donors will be matched on fewer characteristics. This hierarchical procedure trys to pick donors in a way that maximizes the number of matched relevant survey characteristics.

For a more detailed description of this of this procedure, see Welniak and Coder (1980), Oh and Scheuren (1980a), or David, Little, Samuhel, and Triest (1986, Section 2).

5.2 Evaluation of the CPS Hierachical Hot Deck

There have been some evaluation studies of the CPS Hot Deck: Welniak and Coder (1980); Oh and Scheuren (1980a and 1980b); Lillard, Smith, and Welch (1982); and David *et al.* (1986). One of the weaknesses noted of the CPS hot deck is that donor values may be used repeatedly, resulting in variance increases. The procedure could be modified to avoid using donor values more that once or twice; however, this change has not been made. The CPS hot deck procedure is based on the assumption that the distribution of responses for a survey variable is the same for respondents and nonrespondents in the same cell - the ignorability assumption.

David *et al.* (1986) developed several model-based alternatives to the CPS hot deck and evaluated them and the CPS hot deck with respect to mean absolute and mean relative error. These evaluations were based on a CPS-IRS matched file. In creating this file, an attempt was made to match the March 1981 CPS file to the IRS tax records for 1980. Despite the hot deck's apparent limitations, the CPS hot deck had a lower mean absolute and mean relative error than did the model-based alternatives. However, the models were developed for only 10% of the full CPS sample used to develop the hot deck procedure.

6. SUMMARY AND AREAS OF FUTURE STUDY

In this paper an attempt has been made, primarily through examples, to describe the current approaches being taken to nonresponse adjustments in the U.S. Census Bureau's censuses and surveys. Emphasis has been placed on the need for additional empirical and theoretical studies in both the demographic and economic areas in order to provide more objective guidelines (a) to design nonresponse compensation procedures and (b) to measure the effects of nonresponse on survey results for a variety of survey conditions.

Some of the research called for in this paper is already underway but more will be needed. For example, to what extent can available ancillary data be used in conjunction with modeling and data analysis procedures to identify the key functional relationships needed to provide a "reasonably" accurate description of the response/nonresponse structure applicable to a given survey?

In general, adjusting for nonresponse is just one of several steps taken to reduce the variance and bias of survey results. The degree to which these other steps aid in reducing the impact of nonresponse is an area for further research. Moreover, there should be continued efforts in support of research on recurring issues such as the impact of unit nonresponse weights and item nonresponse imputation on complex variance estimators, model approaches to determining appropriate adjustment factors, and the effectiveness of combining various types of nonresponse adjustment techniques.

ACKNOWLEDGEMENTS

The authors are very appreciative of the useful information received on the nonresponse adjustment procedures used for many of the U.S. Census Bureau's Censuses and Surveys.

This information was provided by personnel in several Census Bureau divisions, including Agriculture Division, Business Division, Construction Statistics Division, Economic Surveys Division, Governments Division, Industry Division, Statistical Methods Division, and the Statistical Research Division.

The authors are grateful to Dr. Fritz Scheuren for reviewing the draft and providing many helpful comments.

The authors are also indebted to Hazel Beaton, Alice Bell, and Valerie Howard for the diligence and patience they displayed in typing the manuscript.

REFERENCES

ANDERSON, H. (1978). On nonresponse bias and response probabilities. *Scandinavian Journal of Statistics*, 6, 107-112.

BAILEY, L. (1986). A study of alternative imputation techniques for surveys in the Current Industrial Reports. Internal Census Bureau Report, December 24.

BAILEY, L., CHAPMAN, D.W., and KASPRZYK, D. (1985). Nonresponse adjustment procedures at the Census Bureau: A Review. *Proceedings of the Bureau of the Census First Annual Research Conference*, 421-444.

DAVID, M., LITTLE, R.J.A., SAMUHEL, M.E., and TRIEST, R.K. (1986). Methods for CPS income imputation. *Journal of the American Statistical Association*, 81, 29-41.

DYKE, T.C. (1984). Evaluation of the use of administrative record data for establishments which were non-respondents to the 1977 Census of Wholesale Trade, Retail Trade, or Selected Services. Internal report: Statistical Research Division Report Series, No. Census/SRD/RR-84/08, U.S. Bureau of the Census.

HANSON, R. (1978). The Current Population Survey: Design and Methodology. Technical Paper No. 40, Washington, D.C.: U.S. Bureau of the Census, pp. 55-59.

HUANG, E.T. (1986). Comparison of different imputation procedures in the Monthly Retail Trade Survey. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

KALTON, G., and LEPKOWSKI, J., and LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KALTON, G., McMILLEN, D. and KASPRZYK, D.(1986). Nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 147-164.

KALTON, G., and MILLER, M. (1986). Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

KOBILARCIK, E.L., and SINGH, R.P., (1986). SIPP: Longitudinal estimation for persons' characteristics. To appear in the *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

LILLARD, L., SMITH, J.P., and WELCH, F. (1982). What do we really know about wages: The importance of non-reporting and census imputation. *Journal of Political Economy,* 94, 489-506.

LITTLE, R.J.A. (1986). Missing data in Census Bureau Surveys. *Proceedings of the Bureau of the Census Second Annual Research Conference*, 442-454.

LITTLE, R.J.A., and SAMUHEL, M.E. (1983). Imputation models on the propensity to respond. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update 1. SIPP Working Paper Series No. 8401, U.S. Bureau of the Census.

OH, H.L., and SCHEUREN, F.J. (1980a). Estimating the variance impact of missing CPS income data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 408-415.

OH, H.L., and SCHEUREN, F.J. (1980b). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. *Proceeding of the Survey Research Methods Section, American Statistical Association*, 416-420.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, Vol. 2, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-184.

PALMER, S., and JONES, C. (1967). A look at alternate imputation procedures for CPS noninterview. *Proceedings of the Social Statistics Section, American Statistical Association*, 73-80.

PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and nonrespondent households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.

POLITZ, A., and SIMMONS, W. (1949). An attempt to get the 'Not-At-Homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44, 9-31.

ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

THOMSEN, I., and SIRLING, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In *Incomplete Data in Sample Surveys*, Vol. 3, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 25-59.

U.S. Department of Commerce, Bureau of the Census (1977). National Crime Survey, national sample, survey documentation. U.S. Bureau of the Census Report.

U.S. Department of Commerce, Bureau of the Census (1983). Cross-sectional weighting specifications for the first wave of the 1984 panel of the Survey of Income and Program Participation (SIPP). Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, November 25.

U.S. Department of Commerce, Bureau of the Census (1984a). Economic characteristics of households in the United States: Third Quarter 1983. *Current Population Reports, Series P-70, No. 1*, Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Commerce, Bureau of the Census (1984b). 1984 SIPP first wave weighting-first stage estimate factors and specifications for collapsing noninterview adjustment calls. Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, February 16.

U.S. Department of Commerce, Bureau of the Census (1984c). SIPP weighting: subsequent wave cross-sectional - revised. Internal U.S. Bureau of the Census Memorandum from C. Jones to T. Walsh, October 12.

WELNIAK, E.J., and CODER, J.F. (1980). A measure of the bias in the March CPS earnings imputation scheme. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 421-425.

YCAS, M., and LININGER, C. (1981). The income survey development program: Design features and initial findings. *Social Security Bulletin*, Vol. 44, No. 11, November.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 181–196
Statistics Canada

Hot Deck Imputation Procedure Applied to a Double Sampling Design

SUSAN HINKINS and FRITZ SCHEUREN[1]

ABSTRACT

From an annual sample of U.S. corporate tax returns, the U.S. Internal Revenue Service provides estimates of population and subpopulation totals for several hundred financial items. The basic sample design is highly stratified and fairly complex. Starting with the 1981 and 1982 samples, the design was altered to include a double sampling procedure. This was motivated by the need for better allocation of resources, in an environment of shrinking budgets. Items not observed in the subsample are predicted, using a modified hot deck imputation procedure. The present paper describes the design, estimation, and evaluation of the effects of the new procedure.

KEY WORDS: Double sampling; Hot deck; Imputation.

1. INTRODUCTION

When the U.S. Internal Revenue Service (IRS) is mentioned, the first words to cross one's mind may not be "sample surveys." But every April, those of you from the U.S. take part in at least one of our administrative "surveys" and file an individual income tax return. We sample this administrative data annually for statistical purposes. Another of our major programs is an annual sample of U.S. corporate tax returns; that is the sample survey discussed here.

The primary interest at a Symposium like this is in non-response or other undesirable missing data. Despite our extensive enforcement efforts, we at IRS also have such non-response problems. However, the present paper is concerned with a different type of missing data problem: missingness that is not unexpected, but is designed (see also, Strudler, Oh, and Scheuren 1986, for another example). We take the liberty of discussing these problems because we use techniques usually associated with non-response, e.g., hot deck imputation (Ford 1983). Our case allows an evaluation of the imputation procedure, since the underlying non-response mechanism is known.

Double sampling has been introduced in our corporate tax return sample in an effort to reduce costs with only a "tolerable" loss of information. Reweighting to account for the subsampling stage is a standard estimation approach in double sampling (e.g., Cochran 1977); however, in our application, we would have had to reweight almost on an item-by-item basis. This was judged unacceptable by our users, who require rectangular data sets. (For an analogous approach in a Canadian context, see Colledge *et al.* 1978.)

The imputation technique used – hot deck imputation – is procedurally simple. The need to discuss the application of such a relatively simple procedure may surprise theoreticians; but, as we will show, the problems of implementation within the setting of a large statistical operation are many.

[1] Susan Hinkins, Statistics of Income Division, Internal Revenue Service, P.O. Box 369, Bozeman, Montana 59771. Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.

In the remainder of the present paper, we describe in some detail the double sampling procedure and the imputation technique employed. Preliminary results on the impact of these procedures are also presented and the last section contains our conclusions and future plans. A brief theoretical discussion of the estimators we are using and their properties is given in an Appendix.

2. DESCRIPTION OF THE SAMPLING PROCEDURES

An annual sample of U.S. corporate tax returns is used by IRS to estimate National totals of both tax and economic variables. For example, approximately three million corporate tax returns will be filed for 1985, and the IRS sample will contain over 90,000 of these returns. (In Canada, there are two separate corporate tax return samples, each designed to meet narrower purposes. The Revenue Canada Taxation sample (e.g., Burpee and McGrath 1982) was developed for tax policy simulation purposes. The Statistics Canada sample (e.g., Ambrose 1985) is intended primarily to estimate economic aggregates. It is our belief that separate designs in the U.S., but not entirely separate processing systems, could lead to improvements in efficiency over the current procedures; however, the work done (Clickner *et al.* 1984) indicates that the problem is quite difficult and progress has been slow.)

The annual estimates obtained are for the entire corporate population and for subpopulations, usually defined by industrial activity and size. The underlying population is highly skewed. For most variables, a small proportion of the population accounts for a substantial fraction of the total dollar amount. Examples for 1982 corporations are given in Exhibit 1.

A highly stratified sample design is used; small corporations are selected with small probability and large corporations are selected with certainty (Jones and McMahon 1984). The strata are defined by industrial classification and the size of the corporation (i.e., in terms of assets and net income). Selection probabilities for each stratum are determined by employing a modified form of Neyman allocation. Almost all of the returns in the 100% strata (returns selected with certainty) have total assets of $50 million or more. A form of post-stratified raking ratio estimation is used to weight the sample results (Leszcz, Oh, and Scheuren 1983).

Retrieving the information from each sampled return is a time-consuming and expensive process. Over 600 items may be retrieved from a return, and these items are not simply extracted; they are also carefully checked and redistributed to compensate for taxpayer reporting variations. The complete process is referred to as "editing the return". The cost of "editing" varies by degree of complexity. It may take only twenty-five minutes to edit a fairly simple return but as long as a week to edit a really complicated one. The quality of the editing is vital to our estimates, as these checks reduce, but do not eliminate reporting inconsistencies.

Exhibit 1

Degree of Concentration of Selected Corporate Variables

Selected Items	Assets Under $50 Million	Assets $50 Million or more
Number of Returns	99.6%	0.4%
Total Assets	16.3	83.7
Total Receipts	39.3	60.7
Total Income Tax	25.9	74.1

Source: Internal Revenue Service, 1985.

Indeed, nonsampling error is a serious concern in the data "editing" process, particularly for the largest corporations. In order to spend proportionately more resources on reducing the nonsampling error for the large returns, we introduced stratified double sampling for the smaller returns; specifically, certain data items were retrieved on only a subsample of the returns (i.e., a subset of returns with assets under $50 million). Although this change would increase the error for some variables on the small returns, we expected that the procedure would have little adverse effect on the estimates of national totals, or on the subdomain estimates of primary interest to our major users. There were two main reasons for this conjecture:

- As already noted, corporate *returns* with total assets of $50 million or more were not subject to the extra sampling step.
- The information loss due to the subsampling was reduced by the choice of the *items* or variables to be subject to subsampling.

By and large, as will be shown, the results obtained so far confirm our expectations.

Items Selected for Subsampling

When certain miscellaneous items on a return are nonzero, the taxpayer must attach a schedule providing additional information. For example, if the item "Other Income" is nonzero, the corporation must describe what was included under this category. The schedules are attached on separate sheets of paper and have no standard form or length. The process of editing a schedule has several parts: finding the schedule, deciding whether the taxpayer included appropriate amounts in "Other Income", and making changes if there are errors.

Beginning with the tax year 1981 corporate program, the statistical editing of data from the tax return was done in stages, and certain items were initially transcribed for statistical use directly from the return. Employing automatic tests, items or schedules could then be "flagged" for abstraction or further scrutiny in later stages (Cys *et al.* 1982). This new strategy allowed us to:

- Retain original taxpayer information as reported so that the amount of editing change could be evaluated. Prior to the 1981 sample, we had no information regarding the extent of the adjustments being made by editing. The editors only recorded the final result. (See Powell and Stubbs 1981.)
- Decide whether or not to review a particular schedule based on the initial information transcribed. (Again, prior to the 1981 program, editors were, of course, required to completely edit all schedules.)

For the 1981 and 1982 corporate programs, seven items and their associated schedules were picked for subsampling: schedules for Other Income, Other Deductions, Other Costs of Goods Sold, Other Current Assets, Other (Noncurrent) Assets, Other Current Liabilities and Other (Noncurrent) Liabilities.

The reported amounts on a corporate return may be modified substantially as a result of the editing. For example, consider the "Other Income" schedule shown in Exhibit 2. The original amounts (in column 1) are observed initially for every return. The variables being subsampled are changes that would be made if the Other Income schedule were edited (column 2). In this hypothetical case, we have an original Other Income amount of $1,600, which, when examined by the editor, could be reclassified as including $900 from Business Receipts, $300 in Rents and $400 that really belongs in Other Income. The variables of interest are, of course, the final ("corrected") amounts for each item.

Before implementing the new processing system, an experiment was run comparing the amount of time it took to do the reduced, initial transcription and the amount of time it took to do the complete editing (reading all schedules). As expected, the reduced edit was

Exhibit 2

Illustration of Editing Other Income

Income Type	Original Amounts($)	Change Amount($)	Final Amounts($)
Other Income	1,600	− 1,200	400
Receipts	500	+ 900	1,400
Rents	0	+ 300	300
Interest	700	0	700

significantly faster (and therefore, cheaper). Considerable resources could be saved by subsampling. (Conservatively, we extrapolated 1981 cost savings of at least $300,000, assuming only limited use of the subsampling technique.)

Double Sampling

We are now ready to describe the basic two-dimensional stratification chosen for our double sampling. The returns are stratified into "crucial" returns (Group A) versus the remaining returns (Group B). "Crucial" returns include all returns with total assets of $50 million or more, thereby including the important "large" returns and most returns selected into the sample with certainty. In addition, crucial returns should include corporations of any size for which the likelihood of an editing change was high. What we want, obviously, is a subsampling plan that has us edit all schedules that have a high probability of a change (especially a large change) and lets us subsample the rest.

In an attempt to predict which schedules are likely to change, a record is included in Group A if the original amount in Other Income, to continue our illustration, is unusually large compared to the amount in Total Income.

Also, since we do not want to impute large amounts, cases where Other Income is above a certain dollar value should be included in Group A, as well. (Unfortunately, this was done only indirectly.) By inference, Group B is supposed to include only small returns which we believe are likely to have little or no change made as a result of editing. (See Barker *et al.* 1982, for details.)

For the crucial returns in Group A, all variables (items) are always completely observed. Only returns in Group B are subject to the subsampling of the seven schedules mentioned earlier. Even for Group B returns, the original amounts for all items are always recorded; therefore, some information is obtained for every item. The information not obtained for some records in Group B is the change due to editing a schedule. It is these changes that are being imputed using the procedure described in the next section. Not all variables are affected by the subsampling. For example, of the 600 items picked up for the 1981 corporation program, only 56 were in any way affected by the double sampling; however, of the approximately 100 major income and balance sheet items, nearly one half could be affected.

3. THE IMPUTATION PROCEDURE

The missing information (i.e., changes from editing) in Group B was imputed using a hot deck procedure within adjustment cells. A record with schedules to be imputed was matched to a donor record, in the same adjustment cell, with these same schedules edited. (The formation of adjustment cells is described later in this section.)

In 1981, the subsampling rate was 10% for the returns subjected to subsampling: one out of ten was selected systematically for editing (these were the hot deck "donors") and the other nine were left to be imputed. In 1982, the subsampling rate was kept at 10% for non-financial returns (trade, manufacturing, etc.) but was raised to 20% for financial returns (banks, insurance companies, etc.)

Within an adjustment cell, the number of returns, n', can be divided into the number of donors, n'', and the number of imputes, $n' - n''$. Because of the small subsampling rate, the number of donors is almost always smaller than the number of imputes. In particular, let $n' - n'' = rn'' + t$ where r and t are nonnegative integers and $0 \leq t < n''$. Then the hot deck procedure selects all n'' donors r times, and selects the remaining t units by simple random sampling without replacement.

To continue our illustration, recall that the item of interest is Z, the final "corrected" amount for Other Income; Z can be written as $Z = X - Y$, where X is the original taxpayer amount in Other Income and Y is the change made due to editing the Other Income schedule. It is only the change, Y, that is unobserved and must be estimated for a subset of the returns in Group B.

If we simply employ a conventional hot deck procedure and estimate the unobserved y_i value, on record i, with the observed value y_j from donor record j, then the resulting estimate of the final value z_i may not satisfy the edit checks. For example, assume the donor record had $30,000 originally as Other Income, and $15,000 was removed when the schedule was edited. Suppose that on the record to be imputed, the original amount in Other Income is $10,000, then the imputed change of $15,000 would result in a negative estimate for other income:

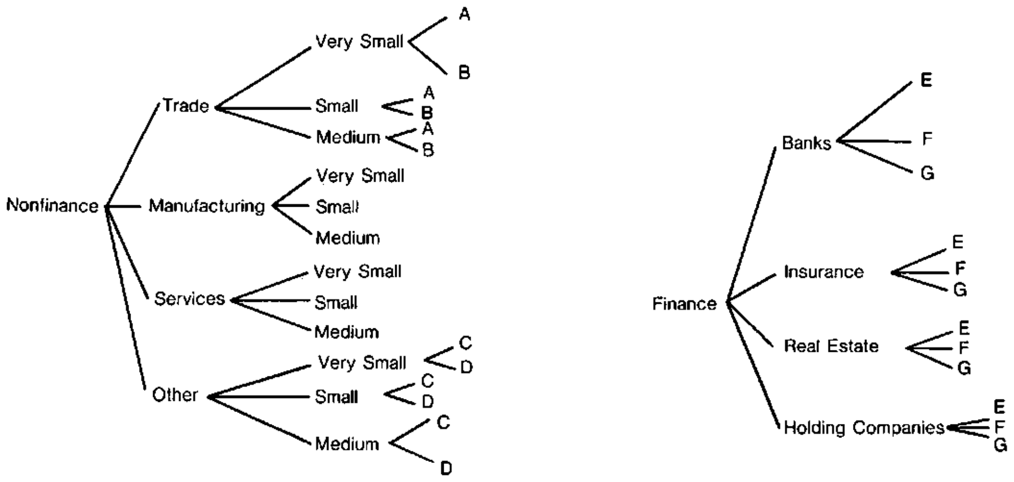
$$\hat{z}_i = x_i - \hat{y}_i = 10,000 - 15,000 = -5,000.$$

Since the amount for Other Income must be nonnegative, edit checks would fail and additional adjustments would have to be made to the record. (See Sande 1982, for a general discussion of this problem.) Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y. In this example, since the donor record had one half of the amount in Other Income removed after reading the schedule, then 1/2 should be removed on the imputed record. The estimated final amount in Other Income is then

$$\hat{z}_i = x_i - \hat{y}_i = 10,000 - (1/2)10,000 = +5,000.$$

In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however, the variance of the estimator is not analytically tractable and must be measured empirically. We have not yet verified in our corporation application the smaller variance that we conjecture; but simulation results do support the approach we have taken. However, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as we shall show.

The model associated with our imputation procedure is based on the definition of the double sampling strata being used and on the definition of the adjustment cells. Several constructive steps were taken to make the approach reasonable. In the initial stratification, an attempt was made to subsample only those records that were likely to have no changes or only small changes. Also, the adjustment cells were *subjectively* chosen to be homogeneous with respect to the magnitude of the relative editing change that might be made. In particular,



The coded tree branches above correspond to the following:

A = Retail, B = Wholesale, C = Transportation and Utilities, D = Other, E = Very Small,
F = Small, G = Medium.

Figure 1. Hierarchy of Ratio Hot Deck Adjustment Cells

the adjustment cells are defined in terms of industrial classification, corporation size and the pattern of items present on the return. There were thirty categories defined by various industrial and size criteria (see Figure 1). In addition, sixteen item patterns were treated separately, defined by the presence/absence of Other Income (2 classes), the presence/absence of either Other Deductions or Other Costs of Goods Sold (2 classes), Other Current Assets or Other Assets (2 classes) and, finally, Other Current Liabilities or Other Liabilities (2 classes). The maximum number of adjustment cells was 30 × 16 = 480.

For each item pattern, a hierarchical structure was developed so that collapsing could be done when there were an insufficient number of donors for use in the imputation (see Figure 1). The first division is into financial returns (banks, insurance companies, etc.) versus nonfinancial records; cells are not collapsed across this division. The next levels of the hierarchy separate cases according to fairly broad industrial classes and according to the size of the corporation, in terms of assets and net income. Recall that the largest corporations are not subject to subsampling and, so, should not need imputation; hence, broad industrial and size groups seemed sufficient.

The quality of our estimation depends on how much collapsing takes place. In 1981, we had 36,586 returns with at least one schedule to impute, and 3,989 donors. For the nonfinancial returns we never collapsed across the major industrial classification, and, in fact, we always had some size distinction. Many cells were not combined at all, but maintained the maximum detail possible. In contrast, for financial returns the size variable was often lost by combining all cells, and major industries were sometimes combined (Hinkins 1983). For one pattern, all financial returns were combined into the same cell.

Based on our 1981 experience, several changes were made in the 1982 double sampling design:

- Due to the extensive collapsing of cells for financial returns in 1981, the subsampling rate for small financial returns was doubled to improve the estimates (from 10% to 20%, as noted earlier).

Table 1

Selected Statistics on Hot Deck Ratio Imputation, 1981-1982

Item	Tax Year 1981		Tax Year 1982	
	Financial	Non-financial	Financial	Non-financial
NUMBER				
Donors	908	3,081	1,806	4,697
Imputes	7,912	28,674	10,719	43,477
Adjustment Cells	113	238	142	260
DONOR CELL SIZE				
Average	8	13	13	18
Maximum	68	58	126	98
Minimum	1	1	2	2
DONOR-TO-IMPUTE RATIOS				
Average	.11	.11	.17	.11
Maximum	1.00	.25	2.00	.28
Minimum	.05	.05	.05	.05

Note: For 1982, cell sizes of 2 donors each were required in order to make possible the calculation of the variance.

- In 1981, the double sampling procedure was not applied across the entire sample, but was restricted to certain processing centers. Other processing centers collected all information, as before. In 1982, the procedure was applied across the whole sample. The relative number of records in1982 with some items imputed was 63 percent, compared to 40 percent in 1981.
- In order to estimate the hot deck imputation variance (Oh and Scheuren 1980; Rubin and Schenker 1986), an additional restriction was imposed on the 1982 design, in that we required that there be at least two donors in each adjustment cell. (See Table 1.)

In 1982, there were 54,196 records to be imputed from 6,503 donors, and there was considerably less collapsing of adjustment cells (Hinkins 1984). In particular, for financial records, 94 percent of the records imputed in 1982 were in adjustment cells defined with some size distinction, compared to 75 percent in 1981. Table 1 provides a selection of other statistics on the operation of the 1981 and 1982 systems.

4. INITIAL EVALUATION OF BIAS

The evaluation of the 1982 double sampling system is still underway, but some initial results are available on the potential biasing effects of the imputation. Bias should be small if R, the ratio of the editing change to the original amount, is always small, or if R is constant within adjustment cells. We have taken the approach of looking for the "worst" cases of bias by looking for examples where R is neither small nor constant. We confine attention to only two variables: Other Income and Business Receipts.

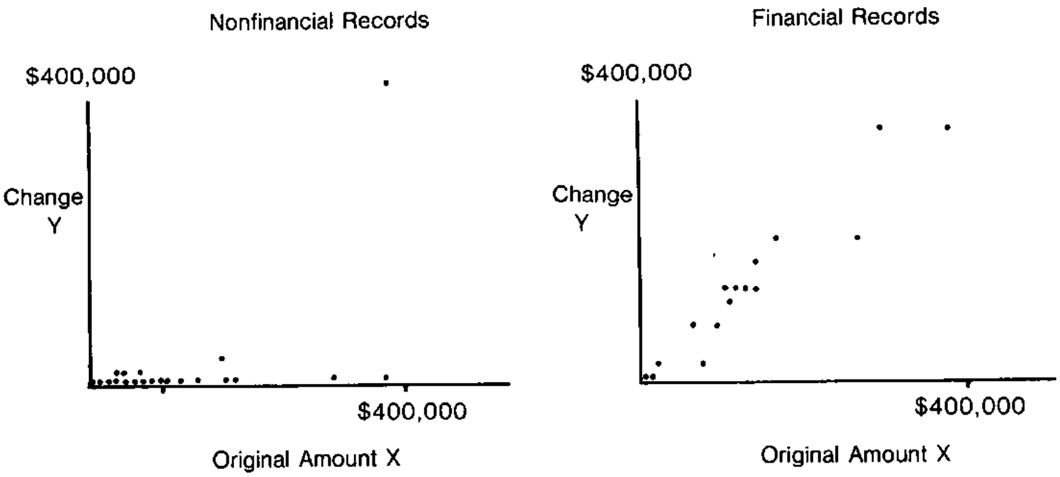


Figure 2. Changes in Other Income: Group B Donors only

Unbiased Model

The ratio bias in the hot deck imputation we are using would be zero if the relationship $Y = RX$ were to hold for all members of each adjustment cell chosen. An overall plot of the data might be useful, to look at the degree to which this model holds for Other Income. In Figure 2, therefore, we have plotted the Group B donors separately for financial and nonfinancial corporations. There is a distinct difference between these two categories. Nonfinancial returns are much less likely to change; in 1982, 14 percent of the nonfinancial donors had a change made to Other Income, compared to 59 percent of the financial records. Also, for financial returns at least, it looks as if the model $E(Y) = RX$ might be appropriate. Further work along these lines is intended, but the scatterplot encourages us to believe that, by and large, existing biases would be small.

Actual Bias Measures

Table 2 provides relative bias measures for selected worst case industries. These are shown for all returns in that industry and returns with assets under $25 million (i.e., for corporations likely to be most affected by the new procedures). Of the items changed in the double sampling the Other Income schedule showed some of the largest values of R and the most disperse distributions of R. The greatest change as a result of editing Other Income was made in the Business Receipt amount. It should be noted that the bias estimates in Table 2 are subject to considerable sampling error (Czajka 1986). Except for the very smallest amounts, however, it is conjectured that the estimates shown probably have the correct sign and are of the appropriate order of magnitude.

These examples indicate that within small subpopulations, there can be noticeable bias effects. However, even within a major industry, selected for its potential problems, the bias across all sizes is relatively small.

Table 2

Estimated Relative Biases for Business Receipts and Other Income by Selected Minor Industries, 1982

Selected Minor Industries	Business Receipts		Other Income	
	All Returns	Assets Under $25 Million	All Returns	Assets Under $25 Million
	(Biases as percent of applicable total)			
WHOLESALE TRADE				
Machinery, Equipment and Supplies	−1.40	−2.6	0.4	0.6
Miscellaneous Trade	−0.30	−0.5	−1.3	−2.4
RETAIL TRADE				
Auto Dealers and Service Stations	−0.30	−0.5	3.3	4.6
FINANCE AND INSURANCE				
Banking	−0.02	−0.7	0.1	2.4
Credit Agencies Except Banks	−0.50	−2.2	−0.9	−9.0
Insurance Agents	−0.60	−0.7	1.2	2.3

Note: All calculations are based on design-weighted estimates of the biases involved. The industries were selected to represent worst case examples.

Czajka's results (1986) indicate that for global estimates (across all industries), the bias effect of the imputation is small (less that 1% in all cases; considerably less than .05% in most cases).

There is no question that some of the biases in Table 2 appear large and warrant concern; however, it is important to realize that the overall effect on the root mean square error of the bias is small for all returns, generally 5% or less. These results give us strong evidence that the procedures employed did little or no harm to the data needed by our users; that, however, is not to say that major improvements, like those envisioned for 1985 and 1986, should not be made.

5. FUTURE PLANS AND SUMMARY

Double sampling and imputation were not used for the 1983 and 1984 samples because of processing constraints. They will be used again starting with the 1985 sample. As part of reinstituting the imputing process, we are planning to make several changes:

- It will no longer be necessary to initially transcribe certain items for statistical purposes before subjecting the records to double sampling. The fields needed are now being obtained directly from the IRS revenue processing system, so they are available before we begin reading and editing the tax return; thus, before editors first look at a return, we can designate whether or not they should review certain schedules. This makes the use of stratified double sampling even more appealing; the savings should increase.
- However, because of the new processing system, only three schedules are now available for subsampling. The schedules for 1985 are Other Income, Other Deductions and Other Costs of Goods Sold; the remaining four schedules used in 1981 and 1982 had to be dropped from the subsampling design.

- Despite the modest success of the 1981 and 1982 procedures, changes will be made for 1985 in the imputation methods. For example, the current definition of the adjustment cells could be improved, and separate imputation depending on the pattern of items represented needs to be reconsidered. The possible use of predictive mean matching within adjustment cells also bears examination (Little 1986). For 1986, refinements in the subsampling plan will need to be looked at too.
- Finally, we would like to base our estimates, in some way, on previous years' data, so as to be able to impute missing information earlier in the processing. In order to minimize the collapsing of adjustment cells, the 1981 and 1982 imputation processing had to wait for all records to be available. This delayed production by several weeks. We could avoid this problem by further increasing the number of donors; but, the editing of more records has the obvious disadvantage of increasing costs. On the other hand, by basing our approach in part on the previous year's data, we might not only improve the estimation, but also allow the imputation calculations to be done in the mainstream of processing.

Overall Summary

In this paper, we have described the reasons we had for making major changes in our statistical processing of corporate returns:

- The traditional complete data estimate was rejected in favor of double sampling because of cost considerations.
- The usual double sampling estimator (reweighting the complete data) was rejected because it did not result in a rectangular data set.
- A conventional hot deck approach was rejected because the resulting estimates could fail the edit checks.

Instead, the relative change was estimated using ratio hot deck imputation within adjustment cells.

We conjectured that because the double sampling procedure was restricted to a subset of the "small" corporations, the estimates of interest to our major users should be virtually unaffected; indeed, these estimates could even be improved, by better allocating our resources to validate and correct the records of the larger corporations. Our results so far largely vindicate these conjectures.

Compared to the traditional complete data estimator, the use of double sampling and hot deck imputation increased the mean square error of estimates in two ways; bias was introduced, and the variance of the estimator was increased. Our preliminary results indicate that there could be a significant bias effect for some estimates; however, the examples were chosen because they appeared to be cases where the hot deck ratio method would be weakest. Even so, the estimated overall effect of the procedure on the root mean square error appears relatively small. Looking at the increase in variance, the largest component is usually due to the decrease in sample size (double sampling). This increase in variance also turned out to be relatively small, since only one component of the final amount (the change) is imputed; the variance of the original values appears to dominate the variance of the changes.

In conclusion, while there are improvements to make, we feel encouraged to continue with our current double sample design and imputation technique. Perhaps at another Conference of this type we will be able to report on the further results of our research.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the considerable help they have received from the staff members in the SOI Division, whose day-to-day responsibilities are covered by the material presented here. We would also like to thank David W. Chapman and John L. Czajka for their many constructive comments, especially in clarifying our exposition. We, of course, accept full responsibility for any remaining obscurities or errors.

APPENDIX: SOME BASIC THEORY

This appendix provides some technical details on the double sampling procedure as applied in our particular situation. We contrast several potential estimators for the double sampling design we chose. An overall summary of the bias and variance expressions for these different approaches is found in Table A.

For this discussion, we ignore the underlying stratified sample design and act as if a simple random sample had been taken, or equivalently we consider estimates within a sampling stratum. To do otherwise would make the notation exceedingly complex, but would not change the main points we wish to make.

Let us again consider just one of the items subject to subsampling, namely Other Income as before. The variable of interest is Z, the final, corrected value of Other Income, and Z can be decomposed as

$$Z = X - Y,$$

where X = the original taxpayer (or revenue processing) value of Other Income,
Y = the change made to Other Income after reviewing the schedule.

The population values and parameters are indicated by upper-case letters and the sample statistics by lower case. The population parameters of interest are the finite population mean and variance, i.e.,

$$\bar{Z} = \sum Z_i / N = \bar{X} - \bar{Y},$$

$$S^2(Z) = \sum (Z_i - \bar{Z})^2 / (N - 1).$$

Complete Sample - Prior to the introduction of double sampling, the estimates were calculated from a complete sample of size n', and the unbiased estimator of $\bar{Z}$ was

$$\begin{aligned} \bar{z} &= \sum z_i / n' \\ &= \bar{x} - \bar{y}. \end{aligned}$$

Ignoring the finite population correction (N is large), the variance is

$$\mathrm{Var}(\bar{z}) = S^2(Z) / n'.$$

Table A

Selected Properties of Alternative Estimators

Estimator	Bias	Variance	Satisfy Edit?
Complete Sample	0	$\mathrm{Var}(\bar{z})$	Yes
Double Sample	0	$\mathrm{Var}(\bar{z}) + c_1 S_B^2(Y)$	Yes
Hot Deck			
Amount (Y)	0^a	$\mathrm{Var}(\bar{z}) + c_1(1 + c_2)S_B^2(Y)$	No
Ratio (R)	b_1	$\mathrm{Var}(\bar{z}) + V_1$	Yes
Combined Ratio	b_2	$\mathrm{Var}(\bar{z}) + V_2$	Yes

[a] In general, the basic hot deck procedure is unbiased only when it results in final values that satisfy the edit checks.

In Table A, we use the properties of $\bar{z}$ as a benchmark, to compare among alternative estimators.

Double Sampling Estimation – Using Cochran's notation (Cochran 1977, 12.2), the original sample of size n' has now been stratified into the two groups A and B, with $n_A{}'$ and $n_B{}'$ units respectively. A subsample of size n_B is selected from group B. The original taxpayer amount X is recorded for all $n' = n_A{}' + n_B{}'$ records. The changes due to editing Other Income, Y, will be recorded for all $n_A{}'$ units in group A and for the random subsample of n_B units in group B.

Since the double sampling procedure only applies to variable Y, within group B, the double sampling estimator of $\bar{Z}$ is

$$\bar{z}_d = \bar{x} - \bar{y}_d$$

$$= \bar{x} - \left(\sum y_{Ai} + (n_B'/n_B) \sum y_{Bj} \right)/n'$$

and $\bar{z}_d$ is unbiased.

Let N_B = number of population units falling in stratum B,
P_B = N_B/N, proportion of population falling in stratum B,
$\bar{Y}_B$ = population mean in stratum B,
$S_B^2(Y)$ = $\Sigma\,(Y_{Bi} - \bar{Y}_B)^2/(N_B - 1)$, $i = 1, 2, \ldots, N_B$,
$1/K$ = the subsampling proportion = n_B/n_B'.

If the sampling proportion, $1/K$, is assumed fixed (in our application, $1/K = .10$ or $.20$), it follows (Cochran 1977) that the unconditional variance of $\bar{z}_d$ is, ignoring the fpc,

$$\mathrm{Var}(\bar{z}_d) = \mathrm{Var}(\bar{z}) + c_1 S_B^2(Y),$$

$$= \left[S^2(Z) + P_B(K - 1)S_B^2(Y)\right]/\mathrm{n}',$$

where $c_1 = P_B(K - 1)/n'$.

Therefore the price paid for the reduction in cost due to not editing every schedule, is the increase in variance due to double sampling. This increase in variance looks potentially damaging because K is large. However, recall that $Z = X - Y$, and the increase in variance is a function only of the variance of Y within subpopulation B. We expect $S^2(X)$ to dominate $S^2(Y)$, which should further dominate $S_B^2(Y)$, i.e.

$$S^2(X) >> S^2(Y) >> S_B^2(Y).$$

This is because the size of the variance is related to the mean value, and Y should be small compared to X. (For most items, we expect the amount misclassified to be small, compared to the original amount). Therefore we expect $S_B^2(Y)$ to be so much smaller than $S^2(Z)$ that $P_B(K - 1)S_B^2(Y)$ will still be relatively small compared to $S^2(Z)$, and so the increase in variance due to subsampling will be relatively small. This is not guaranteed, but Czajka's results bear this out, for most items (Czajka 1986).

Hot Deck Imputation - Hot deck imputation was used, within adjustment cells, to reconstruct a rectangular data set. In particular, a return with schedules to be imputed was matched to a donor in group B, in the same adjustment cell, with these same schedules edited.

Imputing the missing values of y with a hot deck procedure, using simple random sampling, further increases the variance over using the double sampling estimate $(\bar{z}_d)$. However the additional increase in variance due to using hot deck imputation is small compared to the increase due to double sampling. This relative increase in variance due to imputing, denoted as c_2 in Table A, is bounded and in our case is small. (When $K \geq 2$, $c_2 \leq 0.125$. See, for example, Hansen, Hurwitz, and Madow 1953).

As discussed in the paper, there is a problem with using an ordinary hot deck approach. If we simply estimate the unobserved y_i value, on record i, with the observed value y_j from donor record j, then the resulting estimate of the final value z_i may not satisfy the edit checks. Additional corrections would have to be made to the record. Since the original amount is always observed, it seemed more reasonable to "hot deck" the relative change $R = Y/X$ rather than the actual change Y. In addition to satisfying the edit checks, we expected the ratio procedure to reduce the variance of our estimates relative to the basic hot deck approach; however the variance of our estimator is not analytically tractable and must be measured empirically. Also, by introducing the ratio, our estimators are now biased. We conjectured that the biases would be small and in fact they were, for the most part, as seen in Table 2. In practice, the hot deck imputation was done within adjustment cells, created by post-stratifying the records into what we hope are homogeneous cells. The effect of this post-stratification should be to reduce variance and bias effects, but that is dependent on our skill in defining the imputation cells (an area with ample room for additional work).

Ratio or Regression Estimation - We are also considering ratio (or regression) estimates within cells, instead of the hot deck estimates. For example, $\hat{z} = x_i - \hat{r}\, x_i$, where $\hat{r} = \bar{y}/\bar{x}$ is calculated within appropriate cells. Referring to Table A, the increase in variance, V_2, using the ratio estimator could be approximated using the formulas for the ratio estimator (e.g., Cochran 1977). However, these formulas are large sample approximations, and our sample sizes are almost always quite small. (In this case, the sample size is the number of donors, n_B, in an adjustment cell.) Therefore, empirical results are needed here.

Similarly, the bias, b_2, can be found using the results for ratio estimators. Unlike the hot deck ratio, the bias of the ratio estimator goes to zero as the sample size increases and in this sense the ratio estimator is more robust. In fact, the hot deck ratio estimator is unbiased only if the model $Y = \beta X$ is correct. (Of course, the bias of both estimators goes to zero as the fraction of missing data goes to zero). However, even if the model $Y = \beta X$ is incorrect, the ratio estimator is consistent.

There are of course many other options; multivariate regression models could be investigated. We are still in the early stages of this project and we certainly have our work cut out for us now and in the upcoming years.

REFERENCES

AMBROSE, P. (1985). Tax year 1985 business finance (T2) sample selection: detailed statement of requirement. Statistics Canada (Unpublished).

BARKER, D., HINKINS, S., and REHULA, V. (1982). 1981 corporation validation tests. Statistics of Income Division, Internal Revenue Service (Unpublished).

BURPEE, J., and McGRATH, A. (1982). Micro-model of corporation taxation sample design and estimates. Statistical Services Division, Revenue Canada Taxation (Unpublished).

CLICKNER, R.P., GALFOND, G.J., and THIBODEAU, L.A. (1984). Evaluation of the IRS corporate SOI sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 443-448.

COCHRAN, W.G. (1977). *Sampling Techniques,* (3rd ed.). New York: John Wiley and Sons, Inc.

COLLEDGE, M., JOHNSON, J., PARE, R., and SANDE, I.G. (1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 431-436. (See also the paper by S. Michaud in this issue.)

CYS, K., HINKINS, S., and REHULA, V. (1982). Automatic and manual edits for corporation income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 443-448.

CZAJKA, J. (1986). Imputation of selected items in corporate tax data: improving upon the earlier hot deck. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* (in publication).

FORD, B.L. (1983). An overview of hot deck procedures. In *Incomplete Data in Sample Surveys,* Volume 2 - Theory and Bibliographies (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 185-207.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory,* Vol. II. New York: John Wiley and Sons, Inc.

HINKINS, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 427-433.

HINKINS, S. (1984). Matrix sampling and the effects of using hot deck imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 415-420.

JONES, H., and McMAHON, P. (1984). Sampling corporation income tax returns for statistics of income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 437-442.

LESZCZ, M.R., OH, H.L., and SCHEUREN, F.J. (1983). Modified raking estimation in the Corporate SOI Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 434-438.

LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. Presented at the Second Annual Census Research Conference, March 1986. To appear in the *Journal of Business and Economic Statistics.*

OH, H.L., and SCHEUREN, F.J. (1980). Estimating the variance impact on missing CPS income data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 416-420.

POWELL, W.T., and STUBBS, J.R. (1981). Using business master file data for statistics of income purposes. *Statistical Uses of Administrative Records: Recent Research and Present Prospects,* Vol. 1., Washington, DC: Internal Revenue Service, 157-167. See, especially, the Appendix by Alan Freiden.

RUBIN, D., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association,* 81, 366-374.

SANDE, I.G., (1982). Imputation in surveys: coping with reality. *The American Statistician,* 36, 145-152.

STRUDLER, M., OH, H.L., and SCHEUREN, F.J. (1986). Protection of taxpayer confidentiality with respect to the IRS Individual Tax Model. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* (in publication).

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 197–206
Statistics Canada

Comparison of Weighting and Imputation Methods for Estimating Unsampled Data

SYLVIE MICHAUD[1]

ABSTRACT

The Canadian Census of Construction (COC) uses a complex plan for sampling small businesses (those having a gross income of less than $750,000). Stratified samples are drawn from overlapping frames. Two subsamples are selected independently from one of the samples, and more detailed information is collected on the businesses in the subsamples. There are two possible methods of estimating totals for the variables collected in the subsamples. The first approach is to determine weights based on sampling rates. A number of different weights must be used. The second approach is to impute values to the businesses included in the sample but not in the subsamples. This approach creates a complete "rectangular" sample file, and a single weight may then be used to produce estimates for the population. This "large-scale imputation" technique is presently applied for the Census of Construction. The purpose of the study is to compare the figures obtained using various estimation techniques with the estimates produced by means of large-scale imputation.

KEY WORDS: Weighting; Large-scale imputation; Unsampled.

1. INTRODUCTION

The Census of Construction (COC) is an annual survey which attempts to estimate expenses in the construction field. Although it is called a "census", in fact only businesses having a gross income exceeding $750,000 are surveyed. Various financial and non-financial data are collected by means of a long questionnaire mailed to these firms. For businesses with a gross income between $10,000 and $750,000, expenses are estimated from a sample of administrative data. First, two samples are selected independently from overlapping sample frames. Two subsamples are then drawn from one of the samples in order to obtain additional information.

Variables collected in the subsamples may be estimated in two different ways. The method currently used for the Census of Construction is to impute values for the businesses included in a sample, but not in a subsample. This creates a complete "rectangular" file, from which estimates for the overall population may be produced using only one weight. An alternative would be to calculate weights based on the probabilities of selection; these would have to be calculated separately for different subsets of data. The purpose of this study is to compare the estimates obtained by weighting with the estimates obtained by imputation.

The study was carried out on a population of unincorporated businesses only because, for fiscal year 1983, the sample selection strategies for unincorporated and incorporated businesses were different. The strategy used for corporations will be modified for fiscal 1984 to be equivalent to the strategy for unincorporated businesses. The strategy for unincorporated businesses was therefore examined. One hopes that the conclusions of this study will remain the same for incorporated businesses.

[1] S. Michaud, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

2. DESCRIPTION OF THE SAMPLING PLAN

As mentioned above, two independent samples are drawn from overlapping sample frames. The first is the prespecified sample selected for the Census of Construction; it is stratified by gross business income (GBI), province and 3-digit 1970 Standard Industrial Classification (SIC) code. The sample frame used is not completely up-to-date. It contains some "deaths", i.e. businesses which are no longer within the scope of the COC for various reasons (a firm which no longer exists, is no longer engaged in a construction activity, or whose gross income is below $10,000). Furthermore, the sample frame does not contain "births" or businesses which have changed activities and are now part of the construction industry. The second sample is a "cross-sectional" sample, selected independently by Revenue Canada from a complete database containing businesses in all SIC groups (not only construction). It is used to estimate "births". This sample is stratified by Gross Business Income ranges. Figure 1 below illustrates the situation.

Two independent subsamples are selected from the units of the prespecified sample: a financial subsample and a subsample of "other characteristics" (OC). The OC subsample is drawn directly from the prespecified sample, while the financial subsample is selected using data transcribed from the sample (and so "deaths" are not subsampled). Further details concerning the sampling plan may be found in Giles (1983).

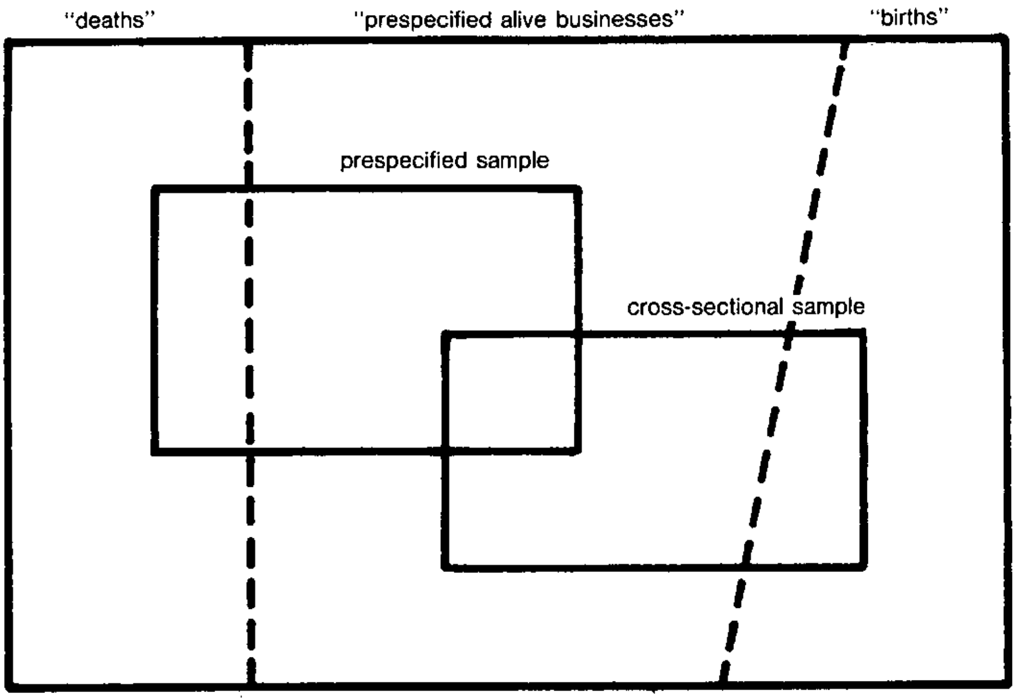


Figure 1. Representation of RC Sampling Plan

3. IMPUTATION TECHNIQUE

The COC uses a large-scale imputation technique to estimate the variables selected in a given subsample (i.e. values are imputed for each variable, for all records not selected in the subsamples). The imputation is carried out independently for each subsample. (The imputation is done in phases, and the imputation phases of the various subsamples are mutually independent and apply different techniques.) In each phase, the nearest neighbour is chosen from a subset of potential donor records, and is used to impute the variables which were not sampled.

The imputation is carried out differently for each subsample.

In the case of the financial subsample, the imputed value is the donor's value, adjusted by the ratio of an auxiliary variable which is available for both the donor and the candidate (the candidate being the record which is missing data to be imputed). (Note: The actual procedure is more complicated: the variables are imputed hierarchically and linear constraints are placed on the imputed values (the second variable is dependent on the value imputed to the first variable, etc.). Additional information on this procedure may be found in Philips and Emery (1976). A more detailed overview is also provided in Colledge *et al.* (1978)).

Suppose we use the following notation:

Y: the variable of interest (known for the donors, to be imputed for the candidate)
X: an auxiliary variable available for both the donor and the candidate
c: denotes the candidate
d: denotes the donor
I: denotes an imputed value.

For the financial subsample variables, the imputed value Y_c^I is defined to be:

$$Y_c^I = Y_d \frac{X_c}{X_d}$$

For the OC subsample variables, the imputed value is simply the value on the donor record:

$$Y_c^I = Y_d$$

The imputation procedure produces a complete rectangular file (the records of all the businesses that were selected in one of the samples contain values for all the variables of the samples/subsamples). Sampling weights may then be used to generate estimates for the overall population.

The weight assigned to a given record is the inverse of the probability of it being selected into at least one of the samples. If we use the following notation:

$P(\text{presp}_h)$: the probability of a record being selected in stratum h of the prespecified sample
$P(\text{cross}_k)$: the probability of a record being selected in stratum k of the cross-sectional sample
hk : cross-classification of records
h : denotes the stratum of the prespecified sample
k : denotes the stratum of the cross-sectional sample,

then the weight associated with each unit may be expressed as:

$$W_{hk}^{-1} = 1 - [1 - P(\text{presp}_h)]\ [1 - P(\text{cross}_k)]$$

Births and deaths cannot be cross-classified. Deaths have a zero weight $W_h = 0$ and the weight of a birth, W_k, is the inverse of the probability of being selected in stratum k of the cross-sectional sample. More details may be found in Bankier (1982).

Therefore, when the imputation technique is used, the estimator of the total is

$$\hat{Y} = \sum_{h,k} W_{hk} \sum_{j=1}^{n_{hk}} y^*_{jhk}$$

where $y^*_{jhk} = y_{jhk}$ if $j \in$ subsample

$= y^I_{jhk}$ if $j \notin$ subsample.

4. WEIGHTING TECHNIQUE

If a weighting technique were used to estimate subsample variables, there would be a number of possible estimators. The estimators are in the same form for both subsamples, but different weights are used.

The first estimator ($\hat{Y}_1$) would be based on the sampling plan used, adjusted for undercoverage of the population. In each of the SIC, PROV and GBI strata (Standard Industrial Classification, province, gross business income), a prespecified sample is selected. Once they have been transcribed (units sampled and still alive), the units are classified to two strata: "outside survey field" and "within survey field".The subsamples are chosen from the "within survey field" stratum. (We may assume that all the units in the "outside survey field" stratum have been subsampled and have a mean equal to zero.)The estimator contains a correction factor that compensates for undercoverage of the sample frame (calculated using information from the cross-sectional sample).

The second possible estimator ($\hat{Y}_2$) is a simplified version of the first estimator, $\hat{Y}_1$. Instead of assuming a double sampling to determine "within survey field" and "outside survey field" units, we could assume that a prespecified stratified sample is selected from "within survey field" units. A subsample is selected from the prespecified sample. The estimator must once again be adjusted to take undercoverage into account. If the differences between the first and second estimator turn out to be insignificant, the second would be a better choice because it is simpler.

The third possible estimator ($\hat{Y}_3$) is an estimator based on data from the cross-sectional sample only. We could assume that the units selected in both the subsample and the cross-sectional sample are selected from the cross-sectional sample. The reasoning behind such an estimator is that the cross-sectional sample is drawn from a complete sample frame. However, since the subsamples are selected from the prespecified sample, and not from the cross-sectional sample, the size of the subsamples in the cross-sectional sample will be small.

Finally, a fourth estimator ($\hat{Y}_4$) could be obtained by supposing that the subsample is selected from the complete sample (prespecified sample + cross-sectional sample), and that the complete sample comes from multiple frames. This fourth estimator is the one that most closely resembles the estimator obtained after large-scale imputation. Indeed, both of these estimators assume that births and new businesses "react" like the rest of the population. The imputation procedure does not make any special adjustment for such businesses, and the weighted estimator is not stratified in such a way as to distinguish these units. In addition, both estimators take into account the fact that the sample comes from a number of frames. The same sampling weight is therefore used in both cases to produce data up to the population level.

As mentioned above, the variables collected in the financial subsample are adjusted by the ratio of an auxiliary variable during the imputation.

We could therefore propose another type of estimator for the variables collected in the financial subsample: a ratio estimator. The auxiliary variable used would be the same one used for the imputation. As is the case for the simple weighting, different estimators could be calculated.

The various estimators and their variances are described in mathematical terms in the Appendix.

5. RESULTS

In the study, four of the seven variables in the financial subsample were considered.

As for the subsample of other characteristics, eight variables are collected for all businesses, while other variables are available for certain SIC groups only. The study was therefore limited to these eight variables.

The variables in the financial subsample presented in this report are "ADD" (additions to fixed assets) and "RM" (repair and maintenance). For the OC subsample, results are given for the variable "PCON" (percentage of construction in a specific field). However, the PCON variable is not published directly, but is multiplied by total expenses to obtain expenses in a specific field: PEXP. This second variable was the one studied.

As mentioned earlier, the variables in the OC subsample are not adjusted by a ratio during the imputation procedure. The ratio estimators will therefore not apply to these variables.

Tables 1, 2 and 3 provide values for the different estimators and estimates of their respective variances, based on 1983 tax data for unincorporated businesses.

In the first place, we see that there are no significant differences between the first two estimators. (According to the predetermined definitions, the second estimator is a simplified version of the first one.)The simplified version will therefore be retained.

Table 1

Estimated Values of PEXP (%EXP*EXPCONS) and Standard Deviation of PEXP

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_I$
Estimate (× 10^{11})	3.44	3.43	3.96	3.66	3.70
Standard deviation (× 10^9)	3.5	3.5	8.4	3.2	

Table 2

Estimated Values of ADD and Standard Deviation of ADD

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_I$
Estimate (× 10^8)	2.08	2.10	2.14	1.84	7.82	5.06	5.2	1.4
Standard deviation (× 10^7)	1.9	1.9	2.0	1.0	0.8	2.2	0.8	

Table 3

Estimated Values of RM and Standard Deviation of RM

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_1$
Estimate ($\times$ 10 ^{8})	1.5	1.5	1.43	1.55	0.9	1.63	1.67	1.75
Standard deviation ($\times$ 10^6)	6.9	6.9	8.9	5.3	3.1	11.0	4.3	

In general, for the variables in the financial subsample, the imputation technique appears to yield results similar to those produced by the weighting method ($\hat{Y}_4$). The estimator obtained by considering only units drawn from the cross-sectional sample ($\hat{Y}_3$) seems more variable than the other estimators. This variability could be explained by the smaller number of units used to calculate this estimator. It should be pointed out that these comparisons are based only on an observed sample, and so the conclusions are somewhat limited. However, owing to the nature of the data (often percentages and subdivisions of activity in the construction field), which is relatively stable in the strata (3-digit 1970 SIC, province and GBI), it was considered unnecessary to analyse these variables in greater depth.

For the variables in the financial subsample, it was found that the estimators adjusted by the ratio do not always seem applicable (for example, the ADD variable). The estimates which they produce are extremely biased. One possible explanation is that the ADD variable and the auxiliary variable used have a high frequency of zero values. A "bad" sample in certain strata can thus inflate the estimates inordinately.

Some problems were also encountered with the imputation system (data imputed when they should not have been, data not imputed), which in certain instances may have affected the estimates obtained by the imputation method. Since the results were based on an observed sample only, and because it was difficult to estimate the impact of the system-related problems, it was decided that a simulation would be done.

6. SIMULATION

The simulation was carried out using a data subset, namely those businesses that had been selected in the financial subsample (all of the variables studied are present for this data subset). Then an attempt was made to apply a simplified version of the technique used by the Census of Construction. A stratified sample was selected, using sampling rates similar to those of the survey. The variables of the financial subsample, for the data not selected in the sample, were considered as missing, and then imputed by the system. The sample selection process and the imputation were repeated thirty times.

Estimates were produced, allowing us to compare the results obtained by summing the non-imputed and imputed data with the estimates produced using sampling weights equal to the inverse of the sampling rate. Since the value for the population is known, the bias and the variance of the estimates were calculated. The results for the ADD and RM variables are shown in Tables 4 and 5.

For the ADD variable, the value produced by ratio estimation differs significantly from the estimates obtained by imputation or by weighting. The bias of the estimate is also significantly not null. For the RM variable, all the estimators are equivalent (equal variances, bias not significant at a 5% level, estimates not significantly different).

Table 4

ADD Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate (× 10^7)	1.41	1.43	1.24	1.41
Standard deviation (× 10^5)		1.11	.85	1.15
Bias (× 10^5)		.22	−1.73	−0.07

Table 5

RM Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate (× 10^7)	1.06	1.06	1.07	1.04
Standard deviation (× 10^5)		4.52	4.11	4.87
Bias (× 10^5)		−0.07	−0.95	−1.38

7. CONCLUSIONS

According to the study results, there do not appear to be significant differences between the large-scale imputation technique and the weighting technique, for the variables in the other characteristics subsample. This was foreseeable, inasmuch as the variables studied seem to be relatively stable within each stratum.

The conclusions for the variables in the financial subsample are based on the results of the simulation. These seem to indicate that the estimates obtained by weighting by the inverse of the probability of selection are comparable to the estimates obtained from large-scale imputation.

The ratio estimator does not appear appropriate for the ADD variable (or for the other variables analysed, but not discussed in this report). Continuation of the study will try to determine whether a regression estimator would be more appropriate, and to evaluate the impact of the imputation on the variable correlation structure.

ACKNOWLEDGEMENT

The author would like to thank P. Giles for his comments and suggestions given during this study, and also the referees for their helpful comments.

APPENDIX

The following notation may be used for the proposed estimators:

h : stratum of the prespecified sample

k : stratum of the cross-sectional sample

N_h : size of the "prespecified" population in stratum h

$\hat{N}_{1h}$: size of the "prespecified" population with "alive businesses (within the scope of the survey) in stratum h (estimated)

$\hat{N}_{2h}$: size of the "prespecified" population with businesses "outside the scope of the survey" in stratum h (estimated)

$\hat{N}_k$: size of the population in stratum k, estimated using information from the cross-sectional sample

$\hat{N}_k'$: size of the population in stratum k, estimated using information from foth samples (multiple frames)

n_h : number of units sampled in stratum h of the prespecified sample

$\hat{n}_{1h}$: number of units sampled and transcribed in stratum h of the prespecified sample

$\hat{n}_k'$: number of units sampled and transcribed in stratum k

$\hat{m}_{1h}$: number of units subsampled from among "alive" businesses in stratum h

y : variable of one of the subsamples

x : auxiliary variable available for all units of the samples

s_{yh}^2 : estimate of the variance of y for the units of the subsample in stratum h

s_{xh}^2 : estimatee of the variance of x for the units of the subsample in stratum h

s_{yxh} : estimate of the covariance of x and y in stratum h.

i)

$$\hat{Y}_1 = \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec}}}\right) \sum_h \frac{N_h}{n_h} \frac{\hat{n}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} Y_{hj}$$

$$V(\hat{Y}_1) \simeq \left(\frac{\hat{N}_{1\text{ pre-spec}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec}}}\right)^2 \sum_h N_h n_h \left(\frac{N_h - 1}{n_h - 1}\right)$$

$$\times \left[W_{1h} S_h^2 \left(\frac{1}{\gamma_h} - \frac{1}{N_h}\right) + \frac{G_h}{n_h} S_h^2 \left(\frac{W_{1h}}{N_h} - \frac{1}{\gamma_h}\right) + \frac{G_h}{n_h} W_{1h}(1 - W_{1h})^2 \bar{y}_h^2 \right]$$

where $$G_h = \left(\frac{N_h - n_h}{N_h - 1}\right),\ \gamma_h = n_h \frac{\hat{m}_{1h}}{\hat{n}_{1h}}, \text{ and } W_{1h} = \frac{\hat{n}_{1h}}{n_h}.$$

ii)

$$\hat{Y}_2 = \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec.}}}\right) \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} y_{hj}$$

$$V(\hat{Y}_2) \simeq \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec.}}}\right)^2 \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \left(\hat{N}_{1h} - \hat{m}_{1h}\right) s_{yh}^2$$

iii)

$$\hat{Y}_3 = \sum_k \frac{\hat{N}_k}{\hat{m}_{1k}} \sum_j Y_{kj}$$

$$V(\hat{Y}_3) = \sum_k \hat{N}_k \left(\frac{\hat{N}_k - \hat{m}_{1k}}{\hat{m}_{1k}}\right) S_{yk}^2$$

iv) $$\hat{Y}_4 = \sum_k \frac{\hat{N}'_k}{\hat{m}_{1k}} \sum_{j=1}^{\hat{m}_{1k}} y_{kj}$$

$$V(\hat{Y}_4) = \sum_k \hat{N}'_\kappa \left(\frac{\hat{N}'_k - \hat{m}_{1k}}{\hat{m}_{1k}}\right) s^2_{yk}.$$

Ratio estimators may be calculated and, like simple estimators, they may take on different forms, depending on the hypotheses postulated. For example, the ratio estimator corresponding to estimator 4 would be:

$$\hat{Y}_{Q4} = \sum_k \hat{N}'_k \bar{Y}_{\text{sub}_k} \frac{\bar{X}_{\text{samp}_k}}{\bar{X}_{\text{sub}_k}}$$

where $\bar{X}_{\text{samp}_k}$ is the mean of variable X for the units selected in the complete sample, which are in stratum k

$\bar{X}_{\text{sub}_k}$ is the mean of X for the units selected in the subsample, which are in stratum k

$\bar{Y}_{\text{sub}_k}$ is the mean of variable Y in stratum k of the subsample.

$$V(\hat{Y}_{Q4}) = \sum_k (\hat{N}'_k)^2 \left(\frac{1}{\hat{m}_{1k}} - \frac{1}{\hat{n}'_{1k}}\right) \left[s^2_{yk} + \hat{R}^2_k s^2{}_{x_k} - 2\hat{R}_k s_{yx_k} + \left(\frac{1}{\hat{n}'_{1k}} - \frac{1}{\hat{N}'_k}\right) s^2_{yk}\right]$$

where $\hat{R}_k = \dfrac{\bar{Y}_{\text{sub}_k}}{\bar{X}_{\text{sub}_k}}.$

REFERENCES

BANKIER, M., (1982). Variance formula for an estimator based on any number of independant stratified samples of which some are Poisson samples. Technical document, Business Survey Methods Division, Statistics Canada.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

COLLEDGE, M.L., JOHNSTON, J.H., PARÉ, R., and SANDE, I.G.(1978). Large scale imputation of survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 721-726.

GILES, P. (1983). Construction division: Census of Construction. Technical document, Business Survey Methods Division, Statistics Canada.

PHILIPS, J.L., and EMERY, D. (1976), FIBCOC documentation. Technical document, Systems Development Division, Statistics Canada.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 207–216
Statistics Canada

A Regression Approach to Estimation in the Presence of Nonresponse

CARL ERIK SÄRNDAL[1]

ABSTRACT

In the presence of unit nonresponse, two types of variables can sometimes be observed for units in the "intended" sample s, namely, (a) variables used to estimate the response mechanism (the response probabilities), (b) variables (here called co-variates) that explain the variable of interest, in the usual regression theory sense. This paper, based on Särndal and Swensson (1985 a, b), discusses nonresponse adjusted estimators with and without explicit involvement of co-variates. We conclude that the presence of strong co-variates in an estimator induces several favourable properties. Among other things, estimators making use of co-variates are considerably more resistant to nonresponse bias. We discuss the calculation of standard error and valid confidence intervals for estimators involving co-variates. The structure of the standard error is examined and discussed.

KEY WORDS: Response mechanism; Adjustment group method; Co-variate; Robustness.

1. INTRODUCTION

We consider a finite population $U = \{1, ..., k, ..., N\}$ from which a sample s of size n is drawn with a sampling design under which the k-th unit has the (strictly positive) probability π_k of being selected. The sampling weight associated with the k-th unit is thus π_k^{-1}. We may admit a complex sampling design, not necessarily self-weighting, for example, a three-stage design with stratified selection of primary units. The probability under the design of jointly including the units k and l is denoted π_{kl} ($\pi_{kl} > 0$ for all $k \neq l$, and π_{kk} is interpreted as equal to π_k).

Given s, a certain unit nonresponse is assumed to occur. The responding subset of s is denoted by r, its size by m. The variable of interest, y, is observed for $k \in r$ only. To counteract the biasing effects of the nonresponse, we assume for the purpose of this paper that the widely used adjustment group method is employed: the sample s is subdivided into H groups $s_1, ..., s_h, ...; s_H$ of respective sizes $n_1, ..., n_h, ..., n_H$. The response set r is correspondingly divided into the subsets $r_1, ..., r_h, ..., r_H$, of respective sizes $m_1, ..., m_h, ..., m_H$. The response rate in group h is denoted $f_h = m_h/n_h$. The method calls for attaching (in addition to the sampling weight) the "adjustment weight" f_h^{-1} to an observation coming from group h. (The sizes and the composition of the adjustment groups at the population level are here assumed unknown.) We have:

$$n = \sum_{h=1}^{H} n_h; \; m = \sum_{h=1}^{H} m_h.$$

[1] Carl Erik Särndal, Department of Mathematics and Statistics, University of Montreal, Montreal, Quebec, Canada, H3C 3J7.

Let $t = \Sigma_U y_k$ be the unknown population total to be estimated. (If A is an arbitrary set of units, we shall systematically write $\Sigma_A y_k$ for $\Sigma_{k \epsilon A} y_k$.). The usual adjustment class estimator of t then becomes

$$\hat{t} = \sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}. \tag{1.1}$$

The adjustment group method is motivated theoretically by an assumption that units within the same group respond with the same (unknown) response probability. (More formally, this is expressed as Model A in Section 3 below.) The method clearly requires that group identity can be determined for each unit $k \epsilon s$. The (categorical) variables that permit this grouping can thus be regarded as variables used for the estimation of an underlying response mechanism.

A different category of variable may be observable for each $k \epsilon s$, namely, variables that explain y, in the ordinary regression theory sense. These variables will be termed co-variates. When incorporated in the estimator, such variables will not only reduce variance but also make the estimator more resistent to nonresponse bias. (They are not auxiliary variables in the usual sense of this term, since they are available not for the entire population U but only for the intended sample s.)

We shall thus keep a firm distinction in this paper between two types of variables observed for $k \epsilon s$, those that are used to estimate the response mechanism, and those that explain the target variable y. Little (1983), in presenting a general framework for data with nonresponse, distinguishes several types of variables. One attempt to describe our situation in terms of Little's setup would be to say that the set of complete item variables in Little's terminology are, in our case, further subdivided into one subset of variables used to model the nonresponse mechanism, and another subset (the co-variates) serving as explanatory variables for the incomplete item variable y. Our approach to inference is that of "quasi-randomization" (Oh and Scheuren 1983), where "quasi" refers to the fact that the nonresponse selection phase must be modelled, whereas the sample selection phase is controlled by the sampler.

2. SOME SIMPLE NONRESPONSE ADJUSTED ESTIMATORS OF THE POPULATION TOTAL

A slight development of the often seen formula (1.1) leads to a (generally somewhat "better") alternative in which the sampling weights π_k^{-1} can be said to be more fully used:

$$\hat{t}_{\text{EXP}} = \left(\sum_s \frac{1}{\pi_k}\right) \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}.$$

The formula (which becomes identical to (1.1) for a self-weighting design) can be written as an expansion of the response set mean:

$$\hat{t}_{\text{EXP}} = \hat{N}\tilde{y}_r,$$

namely, if we let the expansion factor be $\hat{N} = \Sigma_s \, 1/\pi_k$, and

$$\tilde{y}_r = \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{1}{\pi_k}}. \tag{2.1}$$

The symbol tilde will be used to indicate a properly weighted mean statistic. The "tilde mean" $\tilde{y}_r$, being a response set mean, is calculated by attaching to the k-th unit the multiplicative weight:

$$\text{sample weight} \times \text{non response adjustment weight} = \pi_k^{-1} f_h^{-1}$$

for each unit k in the h-th adjustment group.

The expansion estimator $\hat{t}_{\text{EXP}}$ is appropriate for the nonresponse situation: it takes into account the sampling design and it makes an effort to adjust for nonresponse. However, $\hat{t}_{\text{EXP}}$ can be improved upon if more information is at hand. Suppose that a single (and always positive) co-variate x is also observed for $k \epsilon s$. In the image of the classical ratio estimator, we can then construct

$$\hat{t}_{\text{RA}} = \left(\sum_s \frac{x_k}{\pi_k}\right) \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} \frac{x_k}{\pi_k}} = \hat{N} \tilde{x}_s \frac{\tilde{y}_r}{\tilde{x}_r},$$

say, where the tilde mean $\tilde{x}_r$ is formed according to (2.1) with x_k instead of y_k, and

$$\tilde{x}_s = \frac{\sum_s \frac{x_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}}.$$

The tilde mean $\tilde{x}_s$, being formed at the level of the intended sample s, employs sample weights only. (This type of mean can be calculated for the x-variable, which is observed for all $k \epsilon s$, but obviously not for y-variable, which is observed for $k \epsilon r$ only.)

The classical regresson estimator formula corresponds, in our context, to

$$\hat{t}_{\text{REG}} = \hat{N}\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}$$

with

$$b = \frac{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} (y_k - \tilde{y}_r)(x_k - \tilde{x}_r)/\pi_k}{\sum_{h=1}^{H} f_h^{-1} \sum_{r_h} (x_k - \tilde{x}_r)^2/\pi_k}.$$

(Note: sample weighting as well as nonresponse weighting is used in b too.)

In summary, we have a series of three estimators

$$\hat{t}_{\text{EXP}} = \hat{N}\tilde{y}_r, \tag{2.2a}$$

$$\hat{t}_{\text{RA}} \doteq \hat{N}\tilde{x}_s\frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.2b}$$

$$\hat{t}_{\text{REG}} = \hat{N}\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}. \tag{2.2c}$$

All three are properly sample weighted and nonresponse weighted. The obvious differences have to do with the co-variate: $\hat{t}_{\text{EXP}}$ uses no co-variate, whereas $\hat{t}_{\text{RA}}$ and $\hat{t}_{\text{REG}}$ do. It is also clear that $\hat{t}_{\text{RA}}$ appeals to an underlying relationship between y and the co-variate x in the form of a line through the origin, the slope of which is estimated by $\tilde{y}_r/\tilde{x}_r$. In the case of $\hat{t}_{\text{REG}}$, the relationship is a regression with a non-zero intercept. We shall further explore the role of the co-variate.

If the population size N is known, it is in general better to replace $\hat{N}$ by N in (2.2a) to (2.2c), yielding

$$\hat{t}^*_{\text{EXP}} = N\tilde{y}_r, \tag{2.3a}$$

$$\hat{t}^*_{\text{RA}} = N\tilde{x}_s\frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.3b}$$

$$\hat{t}^*_{\text{REG}} = N\{\tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r)\}. \tag{2.3c}$$

For estimating the population total, N must be known in these three estimators, which may not be the case. However, for estimating the population mean $\bar{Y}$, they lead, by dividing by N, to the convenient expressions

$$\hat{\bar{Y}}_{\text{EXP}} = \tilde{y}_r, \tag{2.4a}$$

$$\hat{\bar{Y}}_{\text{RA}} \doteq \tilde{x}_s\frac{\tilde{y}_r}{\tilde{x}_r}, \tag{2.4b}$$

$$\hat{\bar{Y}}_{\text{REG}} \doteq \tilde{y}_r + b(\tilde{x}_s - \tilde{x}_r). \tag{2.4c}$$

The three series of estimators (2.2), (2.3), and (2.4) are easy to accept on intuitive grounds since all that is involved are elementary weighting principles, plus standard ratio feature or regression feature. Somewhat less elementary is to draw the proper consequences for variance estimation and the construction of valid confidence intervals. These questions are discussed in Section 4. (Contrary to what the rather informal presentation of the estimators (2.2) to (2.4) may suggest, the formulas are not "ad hoc" but the result of a formalized general estimation procedure (with a multivariate regression) for two phases of selection; see Särndal and Swensson (1985a). Most importantly, the variance estimators and confidence intervals follow directly from this theory.)

3. RESPONSE MODELS

The nonresponse weights in the estimators seen in Section 2 can be justified through a response mechanism model involving individual response probabilities that are constant for each unit in a given group. More formally, consider the response mechanism:

MODEL A:

(1) The probability of response is constant (and equal to an unknown constant Θ_h) for all units $k \epsilon s_h$; $h = 1, ..., H$.

(2) The units respond independently of each other.

The theoretical response probabilities Θ_h may vary considerably between groups. (An indication that large differences in response propensity may exist between different subsets is, of course, an incentive to set up adjustment groups, and to weight accordingly.)

Consider a fixed sample realization, s. The group frequencies $n_1, ..., n_h, ..., n_H$ are then fixed. Let us also consider a fixed value of the vector of group response frequencies $\underset{\sim}{m} = (m_1, ..., m_h, ..., m_H)$. With s and $\underset{\sim}{n}$ fixed, the "selection" under Model A of a response set r_h can be shown to conform to a simple random selection of m_h from n_h. The conditional response probability of a unit k in the h-th group is therefore

$$\pi_{k|s,\underset{\sim}{m}} = \frac{m_h}{n_h} = f_h, \text{ all } k \epsilon s_h. \tag{3.1}$$

(This consideration underlies the weight f_h^{-1} used in the estimators.) Similarly one can show that given s and $\underset{\sim}{m}$, the probability under Model A that units k and l respond is

$$\pi_{kl|s,\underset{\sim}{m}} = \begin{cases} f_h & \text{if } k = l \\ \dfrac{f_h(m_h - 1)}{n_h - 1} & \text{if } k \neq l \epsilon s_h \\ f_h f_{h'} & \text{if } k \epsilon s_h \text{ ; } l \epsilon s_{h'} \ (h \neq h') \end{cases} \tag{3.2}$$

($\pi_{kk|s,\underset{\sim}{m}}$ is by definition equal to $\pi_{k|s,\underset{\sim}{m}}$.) These quantities (which remind us of stratified random sampling with m_h units chosen from n_h in the h-th stratum) are important for the calculation of variance estimates and standard errors; see below.

In practice, the analyst decides how to set up his groups s_h. The decision is crucial, for it will determine the adjustment weights f_h^{-1}, and thus the numerical value of the estimate of t, the variance estimate, and the confidence interval. Two different groupings may lead to widely different point estimates and confidence intervals.

The analyst is not so naive as to think that response probabilities exist that are exactly equal within the group that he has identified. He does, however, believe (and usually with good reason) that more valid point estimates and confidence intervals will result with these groups (and thereby the weights f_h^{-1}) than without them. The adjustment group approach is a sound and firmly established practice.

On closer scrutiny, several things may be wrong with a response model such as Model A: the response probability is perhaps not constant within groups. And, even if it were, the particular groups postulated by the model are perhaps wrongly defined; there should have been more groups than assumed, etc. Two cases must therefore be distinguished for the continued discussion:

(a) The assumed response mechanism (ARM; here in the form of Model A) is true. In practice, this is unlikely to be exactly the case.

(b) The ARM is more or less false. This is the unpleasant truth in the majority of all practical situations, and it leads to nonresponse bias. In the case of Model A, the groups may be formed more or less incorrectly.

As is usual in statistics, the statistical analyst will formulate the model corresponding to the best of his judgement; accordingly, he will draw certain inferences (confidence statements, for example). Then he will wonder about the robustness of these conclusions, that is, how well do they hold up if the model is false? In the same order of things, let us consider these questions in our particular situation.

4. VARIANCE ESTIMATORS BASED ON A CERTAIN ASSUMED RESPONSE MECHANISM

Model A, with a specified set of groups, is assumed to hold. The response rates, $f_h = m_h/n_h$, $h = 1, ..., H$, have been established. With this as a starting point, let us examine the variance estimators needed to construct a confidence interval at a specified $100(1 - \alpha)\%$ level. If $\hat{t}$ is one of the estimators in Section 2, and Model A really holds, we have:

(a) $\hat{t}$ is unbiased (except for a usually unimportant technical bias)

(b) an approximately $100(1 - \alpha)\%$ confidence interval for t is:

$$\hat{t} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{t})},$$

where the constant $z_{1-\alpha/2}$ is exceeded with probability $\alpha/2$ by the unit normal variate.

Under repeated draws of samples s and, for each fixed s, repeated realizations (obeying the assumed Model A) of response sets r, the interval will contain the true population total $100(1 - \alpha)\%$ of the time.

The variance and the estimated variance will be determined by two sets of selection probabilities:

1. π_k and π_{kl}, the probabilities of inclusion (first and second order) that accompany the sampling phase;
2. $\pi_{k|s,\underline{m}}$, $\pi_{kl|s,\underline{m}}$ the conditional response probabilities (first and second order) associated with the response Model A ("the nonresponse phase").

In our case, as a consequence of Model A, $\pi_{k|s,\underline{m}}$, and $\pi_{kl|s,\underline{m}}$, are given, respectively, by (3.1) and (3.2). As for π_k and π_{kl}, full generality is assumed; any design may be used for the sampling phase.

A detailed analysis will show that the total variance of any one of the estimators $\hat{t}$ seen in Section 2 can be broken down into two components:

$$V(\hat{t}) = V_1(\hat{t}) + V_2(\hat{t})$$

where $V_1(\hat{t})$ may be termed the sampling variance and $V_2(\hat{t})$ the nonresponse variance. The exact formulas given in Särndal and Swensson (1985a) are not reproduced here, but one notes that the components have some reasonable properties:

1. $V_1(\hat{t}) = 0$ if the whole population U is observed (a census rather than a sample survey);

2. $V_2(\hat{t}) = 0$ if the response is complete $(r = s)$;
3. $V_2(\hat{t})$ is greatly reduced in the presence of a strong co-variate, but $V_1(\hat{t})$ is not affected by the co-variate (naturally enough, since it is observed for $k \epsilon s$ only).

Let us examine somewhat more closely the variance estimators. If $\hat{V}_i(\hat{t})$ denotes the estimator of $V_i(\hat{t})$, $i = 1, 2$, the total variance $V(\hat{t})$ will be estimated by an expression of the form

$$\hat{V}(\hat{t}) = \hat{V}_1(\hat{t}) + \hat{V}_2(\hat{t}).$$

Here, the estimated sampling variance component is

$$\hat{V}_1(\hat{t}) = \sum_{k \epsilon r} \sum_{l \epsilon r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) \frac{1}{\pi_{kl|s,\underline{m}}} u_k u_l,$$

where $\pi_{kl|s,\underline{m}}$ is given by (3.2), and π_k, π_{kl} are the inclusion probabilities of the sampling design. The estimated nonresponse variance component is

$$\hat{V}_2(\hat{t}) = \sum_{h=1}^{H} n_h^2 \left(\frac{1}{m_h} - \frac{1}{n_h} \right) S_{wr_h}^2$$

with

$$S_{wr_h}^2 = \frac{1}{m_h - 1} \sum\nolimits_{r_h} (w_k - \bar{w}_{r_h})^2$$

The quantities u_k and w_k differ from one estimator $\hat{t}$ to another. Let us look first at the estimated nonresponse variance, $\hat{V}_2(\hat{t})$. This component is of "stratified form": the factor $n_h^2(1/m_h - 1/n_h)$ is characteristic of a stratified simple random selection with m_h units chosen from n_h in the h-th stratum. The reason for this structure lies in the conditional response probabilities $\pi_{kl|s,\underline{m}}$ given by (3.2).

The quantities w_h have the following appearance:

$$\text{For } \hat{t}_{\text{EXP}} \text{ and } \hat{t}^*_{\text{EXP}}\text{: } w_k = \frac{y_k - \tilde{y}_r}{\pi_k},$$

$$\text{For } \hat{t}_{\text{RA}} \text{ and } \hat{t}^*_{\text{RA}}\text{: } w_k = \frac{y_k - (\tilde{y}_r/\tilde{x}_r)x_k}{\pi_k},$$

$$\text{For } \hat{t}_{\text{REG}} \text{ and } \hat{t}^*_{\text{REG}}\text{: } w_k = \frac{y_k - \tilde{y}_r - b(x_k - \tilde{x}_r)}{\pi_k}.$$

The expressions for w_k are sample weighted regression residuals. Consequently, if x_k is a powerful explanatory variable for y_k, one will ordinarily have that the variance of the w_k (and thus $\hat{V}_2(\hat{t})$) is smaller for the RA and REG estimators than for the EXP estimator, where the quantity w_k is just a deviation of y_k from the response set mean $\tilde{y}_r$. Consequently, in fortunate circumstances, the part of the standard error that is due to the nonresponse will be reduced to near-zero levels, namely, when x and y have near perfect correlation.

The estimated sampling variance component $\hat{V}_1(\hat{t})$ is of less interest in this discussion, since it is not directly influenced by the co-variate. It should be mentioned, however, that the

u_k are determined as follows: $\hat{t}_{\mathrm{EXP}}$, $\hat{t}_{\mathrm{RA}}$, and $\hat{t}_{\mathrm{REG}}$, $u_k = y_k$, while for the "starred" series of estimators $\hat{t}^*_{\mathrm{EXP}}$, $\hat{t}^*_{\mathrm{RA}}$, and $\hat{t}^*_{\mathrm{REG}}$, $u_k = y_k - \bar{\hat{y}}_s$, where $\bar{\hat{y}}_s = (\Sigma_s \hat{y}_k/\pi_k)/(\Sigma_s 1/\pi_k)$ is the mean of the predicted values from the regression fit, so that for $\hat{t}^*_{\mathrm{EXP}}$, $\hat{y}_k = \tilde{y}_r$ for all k; for $\hat{t}^*_{\mathrm{RA}}$, $\hat{y}_k = (\tilde{y}_r/\tilde{x}_r)x_k$; and for $\hat{t}^*_{\mathrm{REG}}$, $\hat{y}_k = \tilde{y}_r - b(x_k - \tilde{x}_r)$.

A special case arises when $m_h = n_h$ for all h (that is, no nonresponse). Then $\hat{V}_2(\hat{t}) = 0$ (as is reasonable), and $\pi_{kl|s,\underline{m}} = 1$ for all k and l, leaving the non-zero component

$$\hat{V}_1(\hat{t}) = \sum_{k\epsilon r} \sum_{l\epsilon r} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}}\right) u_k u_l$$

which is the well-known variance estimator for the case of full response.

5. ROBUSTNESS PROPERTIES WHEN THE ASSUMED RESPONSE MECHANISM IS FALSE

Unbiased estimates and valid confidence intervals can be obtained with the aforementioned estimators, provided the ARM (given by Model A) holds. The presence of a strong co-variate brings about a reduction of the nonresponse component of the variance.

More interesting in a real-life situation is the case where the ARM breaks down. This case must be considered, because even the most careful judgement in setting up adjustment groups is bound to be less than perfect. The extent of the departure of the true response behaviour from that of the ARM will now determine behaviour of the various estimators. The statistical properties (bias, coverage rate achieved by confidence intervals, etc.) are in other words functions of the extent of model breakdown.

In Särndal and Swensson (1985a), a small scale Monte Carlo experiment was carried out to study the impact of certain types of breakdown in Model A. For purposes of illustration, we cite a few results from this study.

The true ARM in the experiment had $H = 4$ adjustment groups, with different response probabilities between groups (but constant response probability for all units in the same group). 1,000 simple random samples were drawn, and each sample was exposed to simulated nonresponse according to the true ARM (which is taken as known, since this is a controlled experiment).

As expected from theory, when the ARM underlying $\hat{t}_{\mathrm{EXP}}$ and $\hat{t}_{\mathrm{RA}}$ was true, there is essentially no bias, and the empirical coverage rates of the confidence intervals agree essentially with the nominal 95% rate. The advantage of $\hat{t}_{\mathrm{RA}}$ lies in a smaller component of variance due to nonresponse. (See "ARM is true" in Table 1.)

False ARM's were created by joining together groups of the true ARM. The estimator and the confidence interval (based on the false ARM) will then be calculated on the basis of fewer groups than ought to be the case. The case "ARM is false" in Table 1 represents the extreme situation where all four groups of the true ARM were joined into one, meaning that one acts in the estimation process as if all units throughout the population had the same (unknown, but estimated) response probability. The table shows that the co-variate estimator, $\hat{t}_{\mathrm{RA}}$, when compared to the no-co-variate estimator, $\hat{t}_{\mathrm{EXP}}$, has the following (not unexpected) advantages: (a) strong resistance to nonresponse bias (1.26 versus 4.85); (b) much better preservation of the nominal 95% confidence coefficient (92.6% versus 46.3% empirical coverage rate). In addition, $\hat{t}_{\mathrm{RA}}$ has a variance advantage, and therefore shorter confidence intervals on the average.

Table 1

Comparison of $\hat{t}_{\text{EXP}}$ and $\hat{t}_{\text{RA}}$

	Estimator	Absolute bias	Mean of the variance component $\hat{V}_2$	Empirical coverage rate (95% nominal)
ARM is true	$\hat{t}_{\text{EXP}}$	0.00	1.99	95.2%
	$\hat{t}_{\text{RA}}$	−0.01	0.78	95.5%
ARM is false	$\hat{t}_{\text{EXP}}$	4.85	2.55	46.3%
	$\hat{t}_{\text{RA}}$	1.26	0.78	92.6%

6. CONCLUSION

In summary, we have argued in this paper that two different categories of variables (observed for k in the intended sample s) are of importance:

(a) variables suitable for estimating the response mechanism (in the case of Model A, these variables allow the construction of the adjustment groups);

(b) variables (here called co-variates) that are powerful predictors of the y-variable; when used in the estimator formula, they reduce variance and improve the robustness properties.

Whenever possible, one should thus be on the outlook for suitable co-variates. One should also note that when several y-totals are to be estimated,the appropriate co-variates may differ from one y-variable to the other, whereas the weighting classes would probably be set up to apply uniformly for all variables of interest.

REFERENCES

LITTLE, R.J.A. (1983). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

SÄRNDAL, C.E., and SWENSSON, B. (1985a). A general view of estimation for two phases of selection. Part I: Randomized subsample selection (Two-phase sampling). Part II: Nonrandomized subsample selection (Nonresponse). Promemorior fran P/STM no. 20, Statistics Sweden.

SÄRNDAL, C.E., and SWENSSON, B. (1985b). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute* (45th session), 51:3, 15.2.1-16.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit non-response. In *Incomplete Data in Sample Surveys*, Vol. 2, (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 143-183.

Survey Methodology, December 1986
Vol. 12, No. 2, pp. 217–230
Statistics Canada

Ratio Estimation with Subsampling the Nonrespondents

PODURI S.R.S. RAO[1]

ABSTRACT

The procedure of subsampling the nonrespondents suggested by Hansen and Hurwitz (1946) is considered. Post-stratification prior to the subsampling is examined. For the mean of a characteristic of interest, ratio estimators suitable for different practical situations are proposed and their merits are examined. Suitable ratio estimators are also suggested for the situations in which the Hard-Core are present.

KEY WORDS: Auxiliary information; Post-stratification; Biases; Mean square errors; Linear model; Hard-Core.

1. INTRODUCTION

Consider a finite population of size N and a random sample of size n drawn without replacement. In surveys on human populations, frequently n_1 units respond on the items under examination, but the remaining $(n-n_1)$ units do not provide any response. The initial survey may be conducted through the mail or telephone calls, perhaps computer-aided.

In Sections 2, 3 and 4, we consider Hansen and Hurwitz's (1946) procedure of subsampling a portion of the $(n-n_1)$ nonrespondents. In this procedure the population is supposed to be consisting of the response stratum of size N_1 and the nonresponse stratum of size $N_2 = (N-N_1)$.

In Section 2, we discuss two procedures for post-stratifying the sampled units, prior to the subsampling of the nonrespondents.

Two ratio estimators for the mean of an item are considered in Section 3. Biases and Mean Square errors of these estimators are compared in Sections 3 and 4. In Section 4, two more ratio estimators, which may be suitable for some practical situations, are proposed and their relative merits are examined.

The Hard-Core problem is considered in Section 5. Six different estimators for this situation are proposed. Optimum conditions suitable for each one of the estimators are briefly described.

2. HANSEN AND HURWITZ'S ESTIMATOR AND POST-STRATIFICATION

Consider a characteristic of interest y_i, $i =$ $(1, 2, \ldots, N)$. Let $\bar{Y} = (\Sigma_1^N y_i)/N$ and $S^2 = \Sigma_1^N (y_i - \bar{Y})^2/(N-1)$ denote the mean and variance of the population. Let $\bar{Y}_1 = (\Sigma_1^{N_1} y_i)/N_1$ and $S_1^2 = \Sigma_1^{N_1}(y_i - \bar{Y}_1)^2/(N_1 - 1)$ denote the mean and variance of the response group. Similarly, let $\bar{Y}_2 = (\Sigma_1^{N_2} y_i) / N_2$ and $S_2^2 = \Sigma_1^{N_2}(y_i - \bar{Y}_2)^2/(N_2 - 1)$ denote the mean and variance of the nonresponse group. The population

[1] P.S.R.S. Rao, Department of Statistics, University of Rochester, Rochester, NY 14627, U.S.A.

mean can be written as $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$, where $W_1 = (N_1/N)$ and $W_2 = (N_2/N)$. The sample mean $\bar{y}_1 = (\Sigma_1^{n_1} y_i)/n_1$ is unbiased for $\bar{Y}_1$, but has a bias equal to $W_2(\bar{Y}_1 - \bar{Y}_2)$ in estimating $\bar{Y}$.

2.1 Subsampling the Nonrespondents

Hansen and Hurwitz (1946) suggest drawing a subsample of size $m = n_2/k$, $k \geq 1$, from the n_2 nonrespondents and assume that responses are available from all of them. The sample mean $\bar{y}_{2m} = (\Sigma_1^m y_i)/m$ is unbiased for the mean $\bar{y}_2$ of the n_2 units. The estimator for $\bar{Y}$ suggested by the above authors is

$$\hat{\bar{Y}}_{HH} = w_1\bar{y}_1 + w_2\bar{y}_{2m}, \tag{2.1}$$

where $w_1 = (n_1/n)$ and $w_2 = (n_2/n)$.

For a given set of n_1 respondents and n_2 nonrespondents, this estimator is unbiased for $\bar{y} = w_1\bar{y} + w_2\bar{y}_2 = (\Sigma_1^n y_i)/n$. Thus, it is unbiased for $\bar{Y}$.

The variance of this estimator is

$$V(\hat{\bar{Y}}_{HH}) = \frac{(1-f)}{n}S^2 + W_2\frac{(k-1)}{n}S_2^2, \tag{2.2}$$

where $f = (n/N)$; see Cochran (1977, p. 371).

Let $s_1^2 = \Sigma_1^{n_1}(y_i - \bar{y}_1)^2/(n_1 - 1)$ and $s_{2m}^2 = \Sigma_1^m(y_i - \bar{y}_{2m})^2/(m-1)$ denote the variances of the n_1 responses and the m subsampled units. An unbiased estimator of the variance is

$$\begin{aligned} v(\hat{\bar{Y}}_{HH}) = {} & \frac{(1-f)}{n}\left[\frac{(n_1-1)s_1^2 + (n_2-k)s_{2m}^2}{n-1}\right] \\ & + \frac{(1-f)}{n}\left[\frac{n_1(\bar{y}_1 - \hat{\bar{Y}}_{HH})^2 + n_2(\bar{y}_{2m} - \hat{\bar{Y}}_{HH})^2}{n-1}\right] \\ & + \frac{(N-1)w_2(k-1)\,s_{2m}^2}{N(n-1)}. \end{aligned} \tag{2.3}$$

This expression can also be obtained from the variance estimators for double sampling and stratification derived by Cochran (1977, p. 333) and Rao (1973); see also Rao (1983).

Post-stratification and subsampling

The $(n - n_1)$ nonrespondents may be classified into $(L - 1)$ strata of sizes $(n_2, n_3, \ldots, n_L)$ according to an auxiliary characteristic, or for convenience in sampling at the next phase. Subsamples of size $m_h = (n_h/k_h)$, $k_h \geq 1$, provide the means $\bar{y}_{hm} = \Sigma_1^{m_h} y_{hi}/m_h$ and variances $s_{hm}^2 = \Sigma_1^{m_h}(y_{hi} - \bar{y}_{hm})^2/(m_h - 1)$.

The unbiased estimator for $\bar{Y}$ now is

$$\hat{\bar{Y}} = \sum_1^L w_h\bar{y}_{hm}, \tag{2.4}$$

where $w_h = (n_h/n)$ and $\bar{y}_{1m} = \bar{y}_1$.

The variance of the above estimator is

$$V(\hat{\bar{Y}}) = \frac{(1-f)}{n} S^2 + \sum_{2}^{L} \frac{W_h(k_h-1)}{n} S_h^2 \qquad (2.6)$$

where $S_h{}^2 = \Sigma_1^{N_h}(y_{hi} - \hat{\bar{Y}}_h)^2/(N_h-1)$. The estimator for the variance is

$$v(\hat{\bar{Y}}) = \frac{(1-f)}{n} \sum_{1}^{L} \frac{(n_h - k_h)s_{hm}^2}{(n-1)} + \frac{(1-f)}{n} \sum_{1}^{L} \frac{n_h(\bar{y}_{hm} - \hat{\bar{Y}})^2}{(n-1)}$$

$$+ \frac{(N-1)}{N(n-1)} \sum_{2}^{L} w_h(k_h-1)\, s_{hm}^2, \qquad (2.7)$$

where $k_h = 1$, $\bar{y}_{1m} = \bar{y}_1$, and $s_{1m}^2 = s_1^2$ as defined earlier.

Other types of post-stratification may be considered. For instance, the n units, respondents as well as the nonrespondents, may be post-stratified into L strata according to an auxiliary variable. The h-th stratum will now have n_{h1} respondents $(\Sigma_1^L n_{h1} = n_1)$ with mean $\bar{y}_{h1}$ and n_{h2} nonrespondents $(\Sigma_1^L n_{h2} = n_2)$. A subsample of size $m_{h2} = (n_{h2}/k_h)$ from the n_{h2} *units* will provide the mean $\bar{y}_{h2m}$. An unbiased estimator for the mean $\bar{Y}_h$ of the h-th stratum now is

$$\hat{\bar{Y}}_h = \frac{n_{h1}\bar{y}_{h1} + n_{h2}\bar{y}_{h2m}}{n_h} \qquad (2.8)$$

where $n_h = (n_{h1} + n_{h2})$, and the unbiased estimator for $\bar{Y}$ is

$$\hat{\bar{Y}} = \sum_{1}^{L} \frac{n_h}{n} \hat{\bar{Y}}_h = \sum_{1}^{L} \frac{n_{h1}\bar{y}_{h1} + n_{h2}\bar{y}_{h2m}}{n}. \qquad (2.9)$$

The variance of this estimator and its estimate can be found as in the above case.

The estimator in (2.4) is preferable if there is much difference among the means of the response and nonresponse strata. The estimator in (2.9) should be preferred if the means of the respondents and nonrespondents differ in each stratum, and if there is much difference among the means of the strata.

Sarndal and Swensson (1985) consider unequal probabilities of selection at the first phase and subsampling the nonrespondents after post-stratification.

3. RATIO ESTIMATORS

Let x_i, $i = (1, 2,..., N)$, denote an auxiliary characteristic with population mean $\bar{X} = (\Sigma_1^N x_i)/N$. Let $\bar{X}_1$ and $\bar{X}_2$ denote the means of the response and nonresponse groups. Let $\bar{x} = (\Sigma_1^n x_i)/n$ denote the mean of all the n units. Let $\bar{x}_1 = (\Sigma_1^{n_1} x_i)/n_1$ and $\bar{x}_2 = (\Sigma_1^{n_2} x_i)/n_2$denote the means of the n_1 responding units and the n_2 nonresponding units. Further, let $\bar{x}_{2m} = (\Sigma_1^m x_i)/m$ denote the mean of the $m = (n_2/k)$ subsampled units.

The population variances of x and y are denoted by S_x^2 and S_y^2, and the population covariance by S_{xy}. The correlation coefficient is $\rho_{xy} = (S_{xy}/S_xS_y)$. The sample variances are denoted by s_x^2 and s_y^2. As before, the subscripts 1 and 2 denote the response and nonresponse groups.

3.1 The Convential Estimator for the Mean

The ratio estimator for $\bar{Y}$ is

$$t_1 = \frac{\bar{y}^*}{\bar{x}^*}\bar{X} = r^*\bar{X} \tag{3.1}$$

where $\bar{y}^*$ is the same as $\bar{Y}_{HH}$ in (2.1), $\bar{x}^* = (w_1\bar{x}_1 + w_2\bar{x}_{2m})$, and $r^* = (\bar{y}^*/\bar{x}^*)$; see Cochran (1977, p. 374). Now,

$$t_1 - \bar{Y} = \frac{(\bar{y}^* - R\bar{x}^*)\bar{X}}{\bar{x}^*} \doteq (\bar{y}^* - R\bar{x}^*)\ (1 - \frac{\bar{x}^* - \bar{X}}{\bar{X}}) \tag{3.2}$$

where $R = (\bar{Y}/\bar{X})$. The approximation in (3.2) is obtained by expressing $(1/\bar{x}^*)$ in Taylor's series, and it is valid for large values of the sample sizes n and m. From (3.2) the bias of t_1 is

$$B_1 = E(t_1 - \bar{Y}) \doteq \frac{(1-f)}{n\bar{X}}\,(RS_x^2 - S_{xy}) + \frac{W_2(k-1)}{n\bar{X}}\,(RS_{x2}^2 - S_{xy2}). \tag{3.3}$$

The bias vanishes only if (a) the regression of y on x goes through the origin for both the response and nonresponse strata and (b) the slopes of both the regressions are equal to R. The first condition is needed for the ratio estimator to be the optimum estimator for $\bar{Y}$. For the second condition to be satisfied, $R_2 = (\bar{Y}_2/\bar{X}_2)$ should not differ much from $R_1 = (\bar{Y}_1/\bar{X}_1)$.

From (3.2), a large sample approximation to the Mean Square Error (MSE) of t_1 is

$$M_1 = E(t_1 - \bar{Y})^2 \doteq \frac{(1-f)}{n}\,S_d^2 + W_2\,\frac{(k-1)}{n}\,S_{d2}^2 \tag{3.4}$$

$$= \frac{(1-f)}{n}\sum_{1}^{2}\frac{(NW_h - 1)}{(N-1)}\,S_{dh}^2 + W_2\frac{(k-1)}{n}\,S_{d2}^2 \tag{3.4a}$$

where $S_d^2 = \Sigma_1^N(y_i - Rx_i)^2/(N-1)$ and $S_{dh}^2 = \Sigma_1^{N_h}(y_{hi} - Rx_{hi})^2/(N_h - 1)$ for $h = 1, 2$. The expression in (3.4) is briefly indicated by Cochran (1977).

An estimator for this MSE is obtained by replacing S_d^2 in (3.4a) by $s_{d1}^2 = \Sigma_1^{n_1}(y_i - r^*x_i)^2/(n_1 - 1)$, S_{d2}^2 by $s_{d2}^2 = \Sigma_1^m(y_i - r^*x_i)^2/(m-1)$ and W_h by w_h. It is possible to suggest alternative estimators for the above MSE.

3.2 An Alternative Estimator for the Mean

In some situations, there may not be any nonresponse on the auxiliary characteristic. Family size, years of education, years of employment, and the like, are the above type of auxiliary variables.

The subsample provides the means $\bar{x}_{2m}$ and $\bar{y}_{2m}$. However, since $\bar{x} = (\Sigma_1^n x_i)/n$ is available, for $\bar{Y}$ we may consider

$$t_2 = \frac{\bar{y}^*}{\bar{x}}\bar{X} = \frac{w_1\bar{y}_1 + w_2\bar{y}_{2m}}{\bar{x}}\bar{X}. \tag{3.5}$$

Since the expectation of $\bar{y}^*$ conditional on the first sample is equal to $\bar{y}$, the bias in t_2 is the same as the one in $\hat{\bar{Y}}_R = (\bar{y}/\bar{x})\bar{X}$. We note that $\hat{\bar{Y}}_R$ is the ratio estimator for the case of complete response. This result can also be derived from the expression

$$t_2 - \bar{Y} = \frac{\bar{y} - R\bar{x}}{\bar{x}}\bar{X} + \frac{\bar{y}^* - \bar{y}}{\bar{x}}\bar{X}. \tag{3.6}$$

Since the conditional mean of $\bar{y}^*$ is equal to $\bar{y}$, the bias of t_2 is

$$B_2 = E(t_2 - \bar{Y}) \doteq \frac{(1-f)}{n\bar{X}}(RS_x^2 - S_{xy}). \tag{3.7}$$

If the regression of y on x for the entire population goes through the origin, the bias of t_2 in (3.7) vanishes. If the regression for the second stratum also goes through the origin, the bias of t_1 in (3.3) would be small only when $R_2 = (\bar{Y}_2/\bar{X}_2)$ is close to R.

From (3.6), the MSE of t_2 is

$$M_2 = E(t_2 - \bar{Y})^2 \doteq \frac{(1-f)}{n}S_d^2 + \frac{W_2(k-1)}{n}S_{y2}^2 \tag{3.8}$$

$$= \frac{(1-f)}{n}\frac{\Sigma\,(NW_h - 1)S_{dh}^2}{N-1} + W_2\frac{(k-1)}{n}S_{y2}^2. \tag{3.8a}$$

Note that $S_d^2 = S_y^2 + R^2S_x^2 - 2RS_{xy}$. An estimator of this MSE is obtained by replacing S_{d1}^2, S_{d2}^2, S_{y2}^2, and W_h by s_{d1}^2, s_{d2}^2, s_{y2}^2, and w_h respectively, where

$$s_{d1}^2 = \sum_1^{n_1}(y_i - r^{**}x_i)^2/(n_1 - 1),$$

$$s_{d2}^2 = \sum_1^{m}(y_i - r^{**}x_i)^2/(m - 1),$$

$$s_{y2}^2 = \sum_1^{m}(y_i - \bar{y}_{2m})^2/(m - 1).$$

In theses expressions, $r^{**} = (\bar{y}^*/\bar{x})$.

Comparing the approximate expressions in (3.4) and (3.8), we find that when $R_1 = (\bar{Y}_1/\bar{X}_1)$ does not differ much from $R_2 = (\bar{Y}_2/\bar{X}_2)$, t_2 will have smaller MSE than t_1 provided the correlation ρ_2 in the nonresponse stratum is not too high. Secondly, if R_1 differs much from R_2, t_2 may have smaller MSE than t_1 even when ρ_2 is high. The following Section contains further comparisons between these two estimators.

3.3 Further Comparisons

In this Section, we compare t_1 and t_2 through the linear model. For the two groups, we consider the models

$$y_{1i} = \alpha_1 + \beta x_i + e_{1i},\ i = (1, 2, \ldots, N_1) \tag{3.9a}$$

and

$$y_{2i} = \alpha_2 + \beta x_i + e_{2i},\ i = (1, 2, \ldots, N_2), \tag{3.9b}$$

with the following assumptions:

$$E(e_{1i} \mid x_i) = 0,\ E(e_{1i}\, e_{1i'}) = 0,\ V(e_{1i}|x_i) = v_1 x_i^{\ell};$$

$$E(e_{2i} \mid x_i) = 0,\ E(e_{2i}\, e_{2i'}) = 0,\ V(e_{2i}|x_i) = v_2 x_i^{\ell}.$$

We note that $(i \neq i')$ and in practice ℓ may lie between zero and 2. Further e_{1i} and e_{2i} are assumed to be uncorrelated. Biases and MSE's of t_1 and t_2 are obtained in the Appendix with the assumption that the response group of size N_1 and the nonresponse group of size N_2 are samples from the super-populations represented by the above models.

Comparisons of the biases

Let I denote the observations from the first initial sample. Since $E[(1/\bar{x}^*)\,|I] \geq (1/\bar{x})$ and $E(1/\bar{x}) \geq (1/\bar{X})$, from (A.2) and (A.3) we find that both t_1 and t_2 overestimate $\bar{Y}$. Further the bias B_1 of t_1 is larger than the bias B_2 of t_2. From (A.6) and (A.7),

$$B_1 - B_2 = \frac{\alpha_W W_2 (k-1) S_{x2}^2}{n\bar{X}^2} \tag{3.10}$$

This difference in the biases increases with the size of the nonresponse stratum and decreases with an increase in the size of the subsample.

Comparison of the MSE's

From (A.9) and (A.20), the difference in the MSE's of t_1 and t_2 is

$$M_1 - M_2 = (A_1 - A_2) - C_2 + (D_1 - D_2). \tag{3.11}$$

From (A.10), (A.21), and (A.22),

$$(A_1 - A_2) - C_2 = [3V(\alpha_w) + \alpha_W^2 - \beta^2\bar{X}^2]\,\frac{W_2(k-1)}{n\,\bar{x}^2}\,S_{x2}^2. \tag{3.12}$$

We note that

$$\begin{aligned} V(\alpha_w) &= \alpha_1^2 V(w_1) + \alpha_2^2 V(w_2) + 2\alpha_1\alpha_2 \mathrm{Cov}(w_1, w_2) \\ &= \frac{N-n}{(N-1)n}\,(\alpha_1 - \alpha_2)^2 W_1 W_2. \end{aligned} \tag{3.13}$$

The difference in (3.12) becomes large as α_1 and α_2 differ much from each other. A sufficient condition for the right side of (3.12) to be nonnegative is that $\alpha_W > \beta\bar{X}$. Further analysis of this result shows that the above difference becomes large if $C_x = (S_x/\bar{X})$ becomes larger than $C_y = (S_y/\bar{Y})$ as the correlation $\rho_{xy} = (S_{xy}/S_x S_y)$ increases.

From (A.12) and (A.24),

$$D_1 - D_2 = E\{[2(\delta - \delta^*) + 3(\delta^{*2} - \delta^2)]\bar{e}^{*2}\}$$
$$+ 2E[\delta^* - \delta - \delta^{*2} + \delta^2)\,\bar{E}\,\bar{e}^*]. \tag{3.14}$$

We note that $(\delta^* - \delta) = (\bar{x}^* - \bar{x})/\bar{X} = w_2(\bar{x}_{2m} - \bar{x}_2)/\bar{X}$. Further, $E(\delta^* - \delta) = 0$.

When $\ell = 0$, from (3.14) and the results in (A.14) and (A.17), to $0(n^{-2})$,

$$\begin{aligned} D_1 - D_2 &= 3E[(\delta^{*2} - \delta^2)\bar{e}^{*2}] - 2E[(\delta^{*2} - \delta^2)\bar{E}\,\bar{e}^*] \\ &= \frac{3W_2(k-1)S_{x2}^2}{n^2\bar{X}^2}(W_1v_1 + kW_2v_2) - \frac{2W_2(k-1)S_{x2}^2}{Nn\bar{X}^2}(W_1v_1 + W_2v_2) \\ &= \{[2(1-f) + 1](W_1v_1 + W_2v_2) + 3(k-1)W_2v_2\}\,\frac{W_2(k-1)}{n^2\bar{X}^2}\,S_{x2}^2. \end{aligned} \tag{3.15}$$

This expression clearly is nonnegative.

When $\ell = 1$, from (3.14), (A.15) and (A.16), to $0(n^{-1})$

$$\begin{aligned} D_1 - D_2 &= 2E\left[(\delta - \delta^*)\frac{(w_1v_1\bar{x}_1 + w_2kv_2\bar{x}_{2m})}{n}\right] \\ &\quad + 2E\left[(\delta^* - \delta)\frac{(w_1v_1\bar{x}_1 + w_2v_2\bar{x}_{2m})}{N}\right]. \end{aligned} \tag{3.16}$$

Noting that $E[(\delta^* - \delta)\,\bar{x}_1|I] = 0$, from (3.16),

$$\begin{aligned} D_1 - D_2 &= -(2/n)\,E[kw_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 + (2/N)\,E[w_2^2\bar{x}_{2m}(\bar{x}_{2m} - \bar{x}_2)]v_2 \\ &= -(2/n)kE[w_2^2V(\bar{x}_{2m}|I)]v_2 + (2/N)\,E[w_2^2V(\bar{x}_{2m}|I)]v_2 \\ &= -\frac{2(Nk - n)\,W_2(k-1)S_{x2}^2}{Nn^2\bar{X}^2}\,v_2. \end{aligned} \tag{3.17}$$

Thus, when $\ell = 1$, $D_2 > D_1$. However, the difference in (3.17) becomes negligible when n is large.

The above results suggest that when $\ell = 0$, t_1 has larger MSE than t_2 if α is larger than $\beta\bar{X}$. When $\ell = 1$, t_1 will have larger MSE than t_2 if α is considerably larger than $\beta\bar{X}$.

4. SEPARATE RATIO ESTIMATORS

4.1 The First Estimator

If $(\bar X_1, \bar X_2)$ are known, the separate ratio estimator for $\bar Y$ that can be suggested is

$$\hat{\bar Y}_S = w_1 r_1 \bar X_1 + w_2 r_2 \bar X_2, \tag{4.1}$$

where $r_1 = (\bar y_1/\bar x_1)$ and $r_2 = (\bar y_2/\bar x_2)$. However, $(\bar X_1, \bar X_2)$ can be estimated by $(\bar x_1, \bar x_2)$ and $(\bar y_{2m}/\bar x_{2m})$ is an estimator of r_2. With these estimates, an estimator for $\bar Y$ is

$$t_3 = w_1 \bar y_1 + w_2 \frac{\bar y_{2m}}{\bar x_{2m}} \bar x_2. \tag{4.2}$$

This estimator can be used if $\bar x_2$ is available but $\bar X$ is not; however, it does not make use of $\bar x_1$.

From (4.2)

$$t_3 - \bar Y = (\bar y - \bar Y) + w_2 (\bar x_2/\bar x_{2m})(\bar y_{2m} - r_2 \bar x_{2m}). \tag{4.3}$$

If m is large, from (4.3) the bias in t_3 is

$$B_3 = E(t_3 - \bar Y) \doteq \frac{(k-1)}{n\bar X_2} (R_2 S_{x2}^2 - S_{xy2}). \tag{4.4}$$

The MSE of t_3 is

$$M_3 = E(t_3 - \bar Y)^2 \doteq \frac{(1-f)}{n} S_y^2 + \frac{W_2(k-1)}{n} S_{r2d2}^2 \tag{4.5}$$

where $S_{r2d2}^2 = \Sigma_1^{N_2} (y_i - R_2 x_i)^2/(N_2 - 1)$.

An estimator for this MSE is obtained by replacing the first term on the right of (4.5) by $v(\bar y) = (1-f)s_y^2/n$, S_{r2d2}^2 by $s_{r2d2}^2 = \Sigma_1^m (y_i - r_{2m} x_i)^2/(m-1)$, where $r_{2m} = (\bar y_{2m}/\bar x_{2m})$, and W_2 by w_2.

4.2 The Second Estimator

An estimator that utilizes $\bar X$ and $\bar x$ is

$$t_4 = t_3\left(\frac{\bar X}{\bar x}\right) = \left(w_1 \bar y_1 + w_2 \frac{\bar y_{2m}}{\bar x_{2m}} \bar x_2\right)\left(\frac{\bar X}{\bar x}\right). \tag{4.6}$$

It may be beneficial to consider this estimator since the conditional mean of t_3 for large m is equal to $\bar y$, and hence the conditional expectation of t_4 becomes equal to $(\bar y/\bar x)\bar X$.

From (4.6),

$$t_4 - \bar Y = \left(\frac{\bar y}{\bar x}\bar X - \bar Y\right) + w_2\left(\frac{\bar x_2}{\bar x_{2m}}\right)(\bar y_{2m} - r_2 \bar x_{2m})\left(\frac{\bar X}{\bar x}\right). \tag{4.7}$$

If n and m are large, the bias of t_4 is

$$B_4 = E(t_4 - \bar Y) \doteq \frac{(1-f)}{n\bar X}(RS_x^2 - S_{xy}) + \frac{(k-1)}{n\bar X_2}(R_2 S_{x2}^2 - S_{xy2}). \tag{4.8}$$

The MSE of t_4 is

$$M_4 = E(t_4 - \bar{Y})^2 \doteq \frac{(1-f)}{n} S_d^2 + W_2 \frac{(k-1)}{n} S_{r2d2}^2$$
$$= \frac{(1-f)}{n} \frac{\Sigma(NW_h - 1)S_{dh}^2}{N-1} + \frac{W_2(k-1)}{n} S_{r2d2}^2. \qquad (4.9)$$

An estimator of M_4 is obtained by replacing S_{d1}^2, S_{d2}^2, S_{r2d2}^2, and W_2 by s_{d1}^2, s_{d2}^2, s_{r2d2}^2, and w_2 respectively, where

$$s_{d1}^2 = \sum_1^{n_1} (y_i - r^* x_i)^2/(n_1 - 1),$$
$$s_{d2}^2 = \sum_1^{m} (y_i - r^* x_i)^2/(m - 1),$$
$$s_{r2d2}^2 = \sum_1^{m} (y_i - r_{2m} x_i)^2/(m - 1).$$

We note that $r^* = (\bar{y}^*/\bar{x}^*)$ as defined in Section (3.1).

Comparing (4.5) and (4.9), we find that t_4 will have smaller MSE than t_3 if the population correlation between x and y is high.

Further investigation is needed to evaluate the merits of the above two separate estimators relative to the estimators in the previous Section.

5. RATIO ESTIMATORS IN THE PRESENCE OF THE HARDCORE

It is becoming increasingly apparent that in spite of subsampling the nonrespondents and a number of call-backs, a significant proportion of the sampled units, the hard-core, do not respond to the items in the survey.

For this situation, we consider the population to be composed of three groups of sizes (N_1, N_2, N_3), $N = \Sigma_1^3 N_i$, with means $(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ and variances $(S_{y1}^2, S_{y2}^2, S_{y3}^2)$. The means and variances for the auxiliary characteristic are $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ and $(S_{x1}^2, S_{x2}^2, S_{x3}^2)$. The population means of these two items are $\bar{Y} = (W_1\bar{Y}_1 + W_2\bar{Y}_2 + W_3\bar{Y}_3)$ and $\bar{X} = (W_1\bar{X}_1 + W_2\bar{X}_2 + W_3\bar{X}_3)$, where $\Sigma_1^3 W_i = 1$. Let $R_1 = (\bar{Y}_1/\bar{X}_1)$, $R_2 = (\bar{Y}_2/\bar{X}_2)$ and $R_3 = (\bar{Y}_3/\bar{X}_3)$.

In the initial sample of size n, only n_1 units respond and provide the means $(\bar{x}_1, \bar{y}_1)$. The number of units (n_2, n_3) in the last two groups are not known, but their sum $(n_2 + n_3) = (n - n_1)$ is known. The means $(\bar{x}_2, \bar{x}_3)$ of the auxiliary characteristic may be known, but $(\bar{y}_2, \bar{y}_3)$ for the item of interest are not observed.

We consider the situation where in the subsample of size $m = (n - n_1)/k$, only m_2 units respond and provide the means $(\bar{x}_{2m}, \bar{y}_{2m})$. The remaining $m_3 = (m - m_2)$ units, the "hard-core", do not respond. Note that m_1 is not defined.

In Rao and Jackson (1984), a number of estimators for $\bar{Y}$ for the above situation are examined, without utilizing the auxiliary information. In this Section, we suggest the following six estimators that utilize the additional information. We briefly present the conditions for which these estimators may be the optimum ones. For the sake of space, we have not presented the derivations for these estimators.

(I). The difference between R_1, R_2 and R_3 is negligible. The m_3 units of the third group, the hard-core, is a random subsample of the m_2 respondents at the second phase. In this case,

$$\hat{Y}_{H1} = \frac{n_1\bar{y}_1 + (n - n_1)\bar{y}_{2m}}{n_1\bar{x}_1 + (n - n_1)\bar{x}_{2m}}\bar{X}. \tag{5.1}$$

(II). Same conditions as in I, but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H2} = \frac{n_1\bar{y}_1 + (n - n_1)\bar{y}_{2m}}{n\bar{x}}\bar{X}. \tag{5.2}$$

(III). $\bar{X}_3 = (N_1\bar{X}_1 + N_2\bar{X}_2)/(N_1 + N_2)$ and $\bar{Y}_3 = (N_1\bar{Y}_1 + N_2\bar{Y}_2)/(N_1 + N_2)$, and (R_1, R_2, R_3) do not differ much from each other. Under these conditions,

$$\hat{Y}_{H3} = \frac{n_1\bar{y}_1 + km_2\bar{y}_{2m}}{n_1\bar{x}_1 + km_2\bar{x}_{2m}}\bar{X}. \tag{5.3}$$

Note that, since $E(m_2/m) = n_2/(n - n_1)$, an unbiased estimator of n_2 is $[(n - n_1)/m]\, m_2 = km_2$.

(IV). Same conditions as in (III), but poor correlation in the second and third groups. For this case,

$$\hat{Y}_{H4} = \frac{n_1\bar{y}_1 + km_2\bar{y}_{2m}}{(n_1 + km_2)\bar{x}}\bar{X}. \tag{5.4}$$

(V). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the n_2 units of the second group. In this case,

$$\hat{Y}_{H5} = \left[\frac{n_1}{n}\bar{y}_1 + \frac{(n - n_1)}{n}\frac{\bar{y}_{2m}}{\bar{x}_{2m}}\bar{x}_2\right]\left(\frac{\bar{X}}{\bar{x}}\right). \tag{5.5}$$

(VI). The three ratios differ from one another. The n_3 units of the third group are a random subsample from the $(n_1 + n_2)$ units of the first two groups. Under these conditions,

$$\hat{Y}_{H6} = \left(\frac{n_1}{n_1 + km_2}\bar{y}_1 + \frac{km_2}{n_1 + km_2}\frac{\bar{y}_{2m}}{\bar{x}_{2m}}\bar{x}_2\right)\left(\frac{\bar{X}}{\bar{x}}\right). \tag{5.6}$$

While we expect the above conditions to be satisfactory, further research is needed to evaluate the performances of the above six estimators.

ACKNOWLEDGMENTS

The author would like to thank Dr. J.N.K. Rao and Dr. M.P. Singh for their interest in this topic. Thanks also to the referee for making suggestions towards improving the presentation of the material.

APPENDIX: BIASES AND MSE'S UNDER THE SUPER POPULATION MODEL

Let $\alpha_W = W_1\alpha_1 + W_2\alpha_2$, $\alpha_w = w_1\alpha_1 + w_2\alpha_2$,

$$\bar{E} = \sum_1^N e_i/N,\ \bar{e}_1 = \sum_1^{n_1} e_i/n_1,\ \bar{e}_{2m} = \sum_1^m e_i/m \text{ and } \bar{e}^* = w_1\bar{e}_1 + w_2\bar{e}_{2m}.$$

Now

$$\bar{Y} = \alpha_W + \beta\bar{X} + \bar{E}, \tag{A.1}$$

$$t_1 - \bar{Y} = \frac{\bar{X}}{\bar{x}^*}\alpha_w - \alpha_W + \frac{\bar{e}^*}{\bar{x}^*}\bar{X} - \bar{E}, \tag{A.2}$$

and

$$t_2 - \bar{Y} = \frac{\bar{X}}{\bar{x}}\alpha_w - \alpha_W + \beta\left(\frac{\bar{x}^*}{\bar{x}} - 1\right)\bar{X} + \frac{\bar{e}^*}{\bar{x}}\bar{X} - \bar{E}. \tag{A.3}$$

1. Biases

Let $\delta^* = (\bar{x}^* - \bar{X})/\bar{X}$ and $\delta = (\bar{x} - \bar{X})/\bar{X}$. Taylor's expansion about $\bar{X}$ gives

$$\frac{\bar{X}}{\bar{x}^*} = 1 - \delta^* + \delta^{*2}.... \tag{A.4}$$

and

$$\frac{\bar{X}}{\bar{x}} = 1 - \delta + \delta^2 .... \tag{A.5}$$

With these expansions, from (A.2) and (A.3), to $0(n^{-1})$ the biases of t_1 and t_2 are

$$B_1 = \frac{V(\bar{x}^*)}{\bar{X}^2}\alpha_W = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2 + \frac{W_2(k-1)}{n\bar{X}^2}S_{x2}^2\right]\alpha_W \tag{A.6}$$

and

$$B_2 = \frac{V(\bar{x})}{\bar{X}^2}\alpha_W = \left[\frac{(1-f)}{n\bar{X}^2}S_x^2\right]\alpha_W. \tag{A.7}$$

2. Mean Square Error of t_1

From the expansion in (A.4),

$$\left(\frac{\bar{X}}{\bar{x}^*}\right)^2 \doteq 1 - 2\delta^* + 3\delta^{*2}. \tag{A.8}$$

From (A.2), the MSE of t_1 can be written as

$$M_1 = E(t_1 - \bar{Y})^2 = A_1 + D_1, \tag{A.9}$$

where

$$\begin{aligned} A_1 &= E\left(\frac{\bar X}{\bar x^*}\alpha_w - \alpha_W\right)^2 \\ &\doteq E\left[(1 - 2\delta^* + 3\delta^{*2})\alpha_w^2\right] + \alpha_W^2 - 2E\left[(1 - \delta^* + \delta^{*2})\alpha_w\right] \\ &= V(\alpha_w) + \left[3E(\alpha_w^2) - 2\alpha_W^2\right]\left[V(\bar x^*)/\bar X^2\right] \\ &= V(\alpha_w) + \left[3V(\alpha_w) + \alpha_W^2\right]\left[V(\bar x^*)/\bar X^2\right] \end{aligned} \tag{A.10}$$

and

$$D_1 = E\left(\frac{\bar e^*}{\bar x}\bar X - \bar E\right)^2. \tag{A.11}$$

With the expansions in (A.4) and (A.8)

$$\begin{aligned} \left(\frac{\bar e^*}{\bar x^*}\bar X - \bar E\right)^2 &= \left(\frac{\bar X}{\bar x^*}\right)^2\bar e^{*2} + \bar E^2 - 2\left(\frac{\bar X}{\bar x^*}\right)\bar E\bar e^* \\ &\doteq (1 - 2\delta^* + 3\delta^{*2})\bar e^{*2} + \bar E^2 - 2(1 - \delta^* + \delta^{*2})\bar E\bar e^* \\ &= (\bar e^* - \bar E)^2 - (2\delta^* - 3\delta^{*2})\bar e^{*2} + 2(\delta^* - \delta^{*2})\bar E\bar e^*. \end{aligned} \tag{A.12}$$

Now,

$$\bar e^{*2} = w_1^2\bar e_1^2 + w_2^2\bar e_{2m}^2 + 2w_1w_2\bar e_1\bar e_{2m}. \tag{A.13}$$

Thus, conditional on n_1 and n_2, when $\ell = 0$,

$$E(\bar e^{*2}|n_1, n_2) = \frac{w_1^2}{n_1}v_1 + \frac{kw_2^2}{n_2}v_2 = \frac{w_1v_1 + kw_2v_2}{n}. \tag{A.14}$$

Similarly, when $\ell = 1$,

$$\begin{aligned} E(\bar e^{*2}|n_1, n_2) &= \frac{w_1^2}{n_1^2}v_1\left(\sum_1^{n_1} x_i\right) + \frac{(kw_2)^2}{n_2^2}v_2\left(\sum_1^{m} x_i\right) \\ &= \frac{1}{n}(w_1v_1\bar x_1 + w_2kv_2\bar x_{2m}). \end{aligned} \tag{A.15}$$

Further,

$$\bar E\bar e^* = \frac{1}{N}\left[\sum_1^{n_1} e_i + \sum_1^{m} e_i + \sum_1^{N-(n_1+m)} e_i\right](w_1\bar e_1 + w_2\bar e_{2m}) \tag{A.16}$$

From (A.16), when $t = 0$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 + w_2 v_2). \tag{A.17}$$

Similarly, when $\ell = 1$,

$$E(\bar{E}\bar{e}^*) = \frac{1}{N}(w_1 v_1 \bar{x}_1 + w_2 v_2 \bar{x}_{2m}). \tag{A.18}$$

3. Mean Square Error of t_2

From the expansion in (A.5)

$$\left(\frac{\bar{X}}{\bar{x}}\right)^2 \doteq 1 - 2\delta + 3\delta^2. \tag{A.19}$$

From (A.10), the MSE of t_2 can be written as

$$M_2 = E(t_2 - \bar{Y})^2 = A_2 + C_2 + D_2. \tag{A.20}$$

With the expansions in (A.5) and (A.19)

$$\begin{aligned} A_2 &= E\left(\frac{\bar{X}}{\bar{x}}\alpha_w - \alpha_W\right)^2 \\ &\doteq E\Big[(1 - 2\delta + 3\delta^2)\alpha_w^2\Big] + \alpha_W^2 - 2E\Big[(1 - \delta + \delta^2)\alpha_w\Big]\alpha_W \\ &= V(\alpha_w) + \Big[3E(\alpha_w^2) - 2\alpha_W^2\Big]\Big[V(\bar{x})/\bar{X}^2\Big] \\ &= V(\alpha_w) + \Big[3V(\alpha_w) + \alpha_W^2\Big]\Big[V(\bar{x})/\bar{X}^2\Big], \end{aligned} \tag{A.21}$$

$$\begin{aligned} C_2 &= \beta^2 E\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)^2 \bar{X}^2 \\ &\doteq \beta^2 E(\bar{x}^* - \bar{x})^2 = \beta^2 E\Big[w_2^2(\bar{x}_{2m} - \bar{x}_2)^2\Big] \\ &= \beta^2 W_2 \frac{(k-1)}{n} S_{x2}^2, \end{aligned} \tag{A.22}$$

and

$$D_2 = E\left(\frac{\bar{e}^*}{\bar{x}}\bar{X} - \bar{E}\right)^2. \tag{A.23}$$

With the expansions in (A.5) and (A.19)

$$\left(\frac{\bar{X}}{\bar{x}}\bar{e}^* - \bar{E}\right)^2 = (\bar{e}^* - \bar{E})^2 - (2\delta - 3\delta^2)\bar{e}^{*2} + 2(\delta - \delta^2)\bar{E}\bar{e}^*. \tag{A.24}$$

We note that

$$E\left[\frac{\bar{X}}{\bar{x}}(\alpha_w - \alpha_W)\left(\frac{\bar{x}^* - \bar{x}}{\bar{x}}\right)\bar{X}\right] \doteq E\Big[(\bar{x}^* - \bar{x})(\alpha_w - \alpha_W)\Big] = 0. \tag{A.25}$$

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques.* New York: John Wiley and Sons, Inc.

HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

RAO, P.S.R.S. (1983). Randomization approach. In *Incomplete Data in Sample Surveys*, Vol. 2; (Eds. W.G. Madow, I. Olkin, and D.B. Rubin), New York: Academic Press, 33-44.

RAO, P.S.R.S., and JACKSON, J.E. (1984). Estimation through the procedure of subsampling the nonrespondents. Presented at the American Statistical Association Meetings, Philadelphia.

SÄRNDAL, C.E., and SWENSSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Proceedings of the 45th Session of the International Statistical Institute*, Section 15.2.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees during 1986. An asterisk indicates that the person served more than once during the year.

J. Armstrong, *Statistics Canada*

A. Baldwin, *Statistics Canada*

D.A. Binder, *Statistics Canada*

L.S. Cahoon, *U.S. Bureau of the Census*

D.W. Chapman, *U.S. Bureau of the Census*

N. Chinnappa, *Statistics Canada*

*H. Choudhry, *Statistics Canada*

S.G. Currie, *Statistics Canada*

J.-P. Dion, *University of Quebec*

P. Foy, *Statistics Canada*

J. Gambino, *Statistics Canada*

P. Giles, *Statistics Canada*

G.J. Goldmann, *Statistics Canada*

J.G. Graham, *Carleton University*

G.B. Gray, *Statistics Canada*

R. Groves, *University of Michigan*

*M. Hidiroglou, *Statistics Canada*

G. Kalton, *University of Michigan*

T.S. Kheoh, *University of Western Ontario*

*H. Lee, *Statistics Canada*

G. Lemaître, *Statistics Canada*

R.J. Lowe, *Statistics Canada*

R. Platek, *Statistics Canada*

S. Presser, *National Science Foundation*

*J.N.K. Rao, *Carleton University*

D. Royce, *Statistics Canada*

D.B. Rubin, *Harvard University*

G. Sande, *Statistics Canada*

C.E. Särndal, *University of Montreal*

F. Scheuren, *Internal Revenue Service*

A.J. Scott, *University of Auckland*

R.H. Shumway, *University of California*

K.P. Srinath, *Statistics Canada*

R. Valliant, *U.S. Bureau of Labour Statistics*

Acknowledgements are also due to those who assisted during production of the 1986 issues of the Journal. These people include to mention only a few names, J. Dufour, S. Hupé, D. Lemire, P. Létourneau, A. McGuire, P. Pariseau, L. Quinn, P. Tessier.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

I. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.