GROSS FLOWS ESTIMATION FROM LONGITUDINAL ADMINISTRATIVE DATA WITH MISSING WAVES

Susana Rubin Bleuer, Statistics Canada
15-H, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada

Key Words: Non-response adjustments, Coverage history.

1. INTRODUCTION

Longitudinal Administrative Database (LAD) is a "panel" survey obtained from the T1 Family File. The T1 Family File (TIFF) is an yearly administrative file based on information contained in the annual income tax T1 forms. It is a list of the individual tax filers and their non-filing spouses for whom the Social Insurance Number (SIN) was reported, containing demographic and income-related data.

Most adult Canadians file an income tax return every year. This means that T1FF is a good frame for the adult Canadian population. When we compare estimates obtained from T1FF with Census or with Postcensal Estimates of Population we find that T1FF undercovers the Canadian population by about 6% every year. Then LAD will have at least 6% of records with missing data at every wave and an analysis involving two years of LAD data may have to deal with up to 12% of records with missing data. Thus estimates of gross flows may underestimate the parameters of interest; but the individuals not covered by T1FF in one year are not necessarily missing the next year; only a small percentage of the Canadian population is never covered. Therefore a method of compensation for the missing years in LAD may account for most of the population.

Two features make LAD different from other longitudinal surveys with respect to weight adjustments. One is that LAD is obtained from administrative data: the missing waves are a result of a cross-sectional coverage problem, and the "response" mechanisms are quite different from those in other surveys. The second feature is that a new panel is born every year in LAD. So a significant proportion of individuals have different patterns of missing data and we have to consider these patterns in the development of weights.

We are concerned with the calculation of simple longitudinal measures like gross flows and transition rates from LAD. The purpose of this study is to investigate if it is necessary to compensate for the missing data and in the process, to develop an appropriate method of weight adjustments.

The arguments for weight adjustments versus imputation in a longitudinal survey are well documented (Lepkowski 1989, etc.). The main disadvantage of a weighting system with respect to a realistic imputation method, is the requirement of the production, documentation and use of several weight sub-systems. The main advantages are the preservation of relationships and, in the case of longitudinal analyses requiring only two years of data, the speed of production and implementation. Both merits are important in the timely calculation of gross-flows.

The weights developed here for two-wave longitudinal analyses can be used until a good system of imputation is in place, and also they can be used for evaluation of the imputation method with respect to a possible distortion of relationships.

In the development of weights we use coverage history to predict coverage (response) and we follow the model of Little & David (1983); we modify this model to derive reasonably bounded weights.

2. MISSING PATTERNS

LAD is a longitudinal 1% Bernoulli sample obtained from TIFF for the years 1982 to 1992. Thus LAD collects information from personal income tax returns for the same individuals at different points in time or "waves". The records are selected randomly every year by Social Insurance Numbers (SIN), using a sampling scheme which guarantees that all individuals who were selected in the first year,will be selected in all subsequent years, provided that they belong to T1FF of that year (Demnati, 1992).

The sampling mechanism also ensures that every year, LAD is a representative sample of the T1FF population. If an individual starts filing for the first time in, say, 1988, then he or she has a 1% chance of being included in the LAD sample from 1988 on. Similarly, if an individual dies or leaves the country, he or she will stop being part of LAD after the individual's last return.

In our study we will consider individuals 19 years

old and over. We will also assume that the amount of false links is negligible and that each record represents a distinct individual.

Given the missing pattern distribution of the last five years of LAD, we have that 78% are complete respondents, 4% have patterns with missing waves only at the end of the 5-year sequence, 13% of the records have missing waves only at the beginning of the sequence (from 1988 on) and the residual 5% have missing waves in the middle of the sequence as well as at the ends (Bleuer & Bernier, 1996). The counts do not include individuals with patterns ending with zeros if in the last year of data available the death code indicates that the individual died. This distribution is quite different from the usual one-panel sample survey, where there are no births and most of the non-respondents quit the panel after the first interview.

Records with zeros at the beginning of the sequence do not necessarily imply missing data: an individual appearing for the first time in LAD may have just become part of the population through birth or immigration to Canada.

Similarly, records with zeros at the end of the sequence may correspond to individuals who died or emigrated. In this case, we should not treat these observations with the methodology used for missing data, since the said individuals do not belong to the population of interest.

In a preliminary study on the characteristics of the individuals with missing data (Bleuer & Bernier, 1996), a comparison between LAD estimates and counts of Leaving Canadians from Vital Statistics, showed that very few records with missing waves in the middle correspond to people who left the country. Similarly, from each pattern with missing waves at the end of the sequence, Bleuer & Bernier estimated the number of individuals who were truly dead, and showed that most patterns with missing waves at the end correspond to individuals who still belong to the Canadian population. We also know that every year at least 500,000 Canadians, representing 2.5% of the LAD adult population, file after we receive the data; this will translate into a missing wave for the corresponding records.

Bleuer & Bernier (1996) also found that individuals who missed at least one wave are different from the rest of the population: they have lower earnings and they are usually younger and single. A very small group have higher incomes.

Thus, ignoring the missing data could induce a bias in the analysis performed with LAD.

3. DEFINITION OF THE SAMPLE

Suppose we want to estimate a parameter of the cross-sectional population in 1990.

The LAD records with a missing wave in 1990 are not in the 1990 T1FF either. The sample is drawn from the population of Social Insurance Numbers and an individual may belong to the sample and not to T1FF for some of the waves. The missing waves are due to undercoverage, though it could also be considered as a non-response problem, since it is caused by the individual not filing on time or not filing at all. From now on, we use the terms "coverage" and "response" to mean the same phenomenon. In order to compensate for the missing data, we use the same methods used for non-response and apply a "weighting" adjustment to the estimator.

The weighting adjustment depends on the coverage rate for 1990. The coverage rate is the ratio of the number of individuals who were covered in 1990 over all "sampled" Canadians and landed immigrants who were living in Canada in 1990.

Thus we need to know who is in the sample in 1990, that is, who should have been covered in 1990.

We can find some of the individuals who should have been covered by looking at the longitudinal records. If we do not see an individual in LAD in one wave, we may see him or her in another wave. We may also deduce if it is a "birth" in the wave we do observe the individual or a "death" in the last wave we observed, through the birth date, immigration and death codes. Thus we may reach a reasonable conclusion on wether the individual belongs or not to the population.

There is however an important proportion of individuals who were alive in the last observed wave, and for whom we do not have any data for many waves until the end. Are these individuals still in the population, and if we assume so, when do we cut off? We have to decide upon a rule to define the respective cross-sectional populations. We estimate the number of deaths among the records with missing waves at the end and define the population accordingly (Bleuer 1996).

If R_t denotes the set of individuals in LAD who responded and are alive in wave t, then for this study, we define the cross-sectional sample $cs(t)$ at

time t by the collection of all individuals who belong to R_t, R_{t-1}, R_{t-2} or R_{t-3}:

$$cs(t) = R_t \cup R_{t-1} \cup R_{t-2} \cup R_{t-3}.$$

For example, the 1990 cross-sectional sample is the collection of all individuals with data available in either 1990, 1989, 1988 or 1987, and were alive by 1990. If an individual responded in 1987 and did not respond since 1988 to 1991 we assume that the individual is dead in 1991. Given this assumption, the cross-sectional sample in 1991 consists of all individuals in the 1990 sample, minus those who died by 1991, plus those who responded for the first time (were born) in 1991:

$$cs(1991) = cs(1990) - \text{deaths}(1991) + \text{births}(1991).$$

We will also assume, for the sake of simplicity, that any person who appears for the first time in LAD is just born in the population, and that the zeros at the beginning of the sequence do not represent missing data. This assumption can later be relaxed by adjusting the rate of initial response (or initial coverage).

The population estimates, derived from LAD with the increased cross-sectional samples, were compared with the corresponding Intercensal and Postcensal Estimates of Population (cross-sectional and longitudinal populations), to check that we did not include too many dead units in the population.

Table 3.1 shows estimates of the 1991 population, and the 1991-1992 longitudinal population. The relative differences between the LAD estimates and the Postcensal Estimates of Population are considerably reduced when we use the increased sample as defined above as opposed to the sample of 1991 "respondents". The sample counts were adjusted for the death counts as described in Bleuer (1996).

Table 3.1 Canadian population 19+

Estimator	1991	1991-1992
Unadjusted LAD	19,506,600	18,978,300
Relative Difference	-6.9%	-9.5%
Adjusted LAD	20,568,600	20,356,800
Relative Difference	-1.8%	-2%
Postcensal N_{91}, $N_{91} - D_{92}$	20,948,593	20,752,593

N_{91} = Postcensal Estimate of Population as of April 1st, 1992.
D_{92} = 1992 Death Count obtained from Vital Statistics.

4. ESTIMATION OF THE COVERAGE RATES

An individual may file on time or not in 1992 depending, for example, on whether he had low income in 1992 or wether he invested heavily in Registered Retirement Savings Plan (RRSP) in 1992. Since many of these variables are correlated with the corresponding ones for 1991, the coverage of an individual in 1992 may be predicted by the values observed for the individual in 1991. Moreover, the event of having filed late or not filed at all in 1991 may also indicate the filing behaviour (and thus the coverage) of the individual in 1992.

We have to formulate a coverage model that will permit us to estimate the different coverage rates required for our system. We will assume that the response (coverage) is missing at random within classes.

In a longitudinal survey we have information about previous response to help us predict response in the current wave. Most longitudinal surveys are one panel surveys and use the information obtained in the first wave to predict response in the current wave. In our case, LAD is a collection of many panels and we cannot ignore the births. We may also want to consider the different missing patterns and this translates into using coverage (response) history in order to predict coverage.

The approach we take to calculate the weighting adjustment is to define the coverage rate as the rate of coverage within the classes.

The difficulty lies in selecting the most relevant concomitant variables for the coverage model, since we have too much information.

A stepwise option in a logistic regression procedure was used to identify the most significant variables in predicting the coverage rate for the 1992 wave.

The following are some considerations that we took into account to develop the model.

Every year, the respondent provides the same type of demographic and financial information. It is reasonable to assume that if we account for the last available wave of data, the values corresponding to waves previous to this one add little information to the prediction of the response in the current year. In other words, if we want to predict response in 1991 based on data for 1990, the data provided for 1989 or 1988 does not improve the estimation of the response rates. This is a type of "Markov Chain" property that simplifies the response model: we do not have to include all missing patterns. We verified this assumption for some of the waves and for some key concomitant variables, through a logistic

regression model.

We define the coverage (response) history variables as follows. For the prediction of response in 1992 (r_{92} = 1) we consider the history of response up to 1991 inclusive:

88	89	90	91	92	
*	*	*	X	r_{92}	then hist(91)=91,
*	*	X	O	r_{92}	then hist(91)=90,
*	X	O	O	r_{92}	then hist(91)=89.

We set hist(91) = 92 for the units who appeared in LAD for the first time in 1992 and assign them a rate of response equal to 1. A value of hist(91) = 91 means that we can observe a vector of concomitant variables X_{91} from 1991 tax data. If hist(91) = 90 then the first set of auxiliary variables that we can observe is given by X_{90}. Thus, the covariate we use in the response model is

$$Z_{91} = (\text{hist}(91), X_{\text{hist}(91)}).$$

Similarly, we define hist(90) and Z_{90} to use in the response model for 1991.

We fitted the model

$$E(r_{92}/Z_{91}) = 1/(1 + \exp(-\beta' Z_{91}))$$

where r_{92} is the response or coverage indicator for the 1992 wave, β is the column vector of regression parameters, and Z_{91} is the vector of concomitant variables, including response history, up to 1991.

We examined many variables for inclusion in the model and we found that response history is the variable that most explains response. The variables given in the stepwise procedure (SAS, PROC LOGIST) were entered into the model with the FORWARD option and they entered in the following order: RESPONSE HISTORY, SINGLE STATUS, INCOME CLASS, QUEBEC, WEST and UIB. QUEBEC and WEST are the indicators of filing from Quebec and the Prairie provinces respectively and UIB is the indicator of unemployment insurance recipiency.

After accounting for these variables, most other variables were not significant in explaining response in 1992. Similar outcomes were obtained in the regression of the response indicator for 1991, versus Z_{90}, and when fitting r_{90} and r_{89} respectively. We then used response history and the six binary variables described above to define the response classes. The weighting adjustments were calculated as the inverse of the response rates within the response classes.

We further modified both sets of weights by an adjustment for the people included in the sample who did not belong to the population (deaths).

The distribution of the weights for 1991 had mean 1.233, standard deviation 0.384 and $1+CV^2=1.097$. The statistic $1+CV^2$ is an indicator of the increase in variance for the adjusted estimators using these weights. (Kish, 1992 and Rizzo et al, 1994). This means that the weight adjustments for cross-sectional analysis in 1991 result in an approximate increase in variance of 9.7%. The largest weights correspond to the class of individuals who responded in 1988 and not in 1989 and 1990: the proportion of individuals responding after missing two waves is very small.

The further back we go to include waves in the definition of the sample, the larger the variation of the resulting weights; and the size of response classes corresponding to individuals with previous missing waves is the most influential factor in the variation of the weighting adjustments. After accounting for this factor the statistic $1+CV^2$ is about the same magnitude for every year.

5. LONGITUDINAL WEIGHTS

The coverage (response) rates calculated in the previous section are estimates of conditional rates, since they were derived using the information available in the previous waves. The cross-sectional estimators that use conditional rates are approximately conditionally unbiased. Indeed, let

$$w_{92}^{-1} = pr(r_{92}=1 \mid Z_{91})$$

and

$$\hat{Y}_{92} = 100 \cdot \Sigma w_{92} \cdot y,$$

where the factor 100 refers to the LAD Bernoulli design weight, and the sum is over the sample of those covered in LAD 1992; we have

$$E(\hat{Y}_{92} / Z_{91}) = Y_{92},$$

and an estimate of the variance is given by the estimate of the conditional variance.

However, the longitudinal estimators are not so straight forward in their conditional properties.

In the estimation of longitudinal parameters we have to use longitudinal coverage (response) rates. We use the relationships between the cross-sectional and longitudinal response rates, conditional to the history; these result in obtaining all longitudinal

combinations as functions of the cross-sectional conditional rates (Little & David, 1983). For example,

$$pr(r_{91} = 1,\ r_{92} = 1 / Z_{90}) = pr(r_{92} = 1/r_{91} = 1) \cdot pr(r_{91} = 1 / Z_{90}),$$

due to the Markov property.

It follows that the longitudinal weight is

$$w_{91,92} = w_{91} \cdot w_{92}$$

and the longitudinal estimator of a gross flow $Y_{91,92}$ is

$$\hat{Y}_{91,92} = 100 \Sigma w_{91,92} \cdot y$$

where the sum is over the individuals responding in both 1991 and 1992; if E_{91} denotes the expectation conditional on Z_{91}, then

$$E_{91} = E(\hat{Y}_{91,92} / Z_{91}) \neq Y_{91,92},$$

however

$$E_{90}(E_{91} / Z_{90}) = Y_{91,92},$$

and the variance formula includes two steps:

$$var(\hat{Y}_{91,92}) = E_{90}(V_{91}) + V_{90}(E_{91}).$$

If we followed the same model, the longitudinal weights for analysis of the 1990 and 1992 longitudinal population would become too large (Little & David, 1983). In order to avoid this, we may use the response rate in 1992 and 1990, given the response history until 1990. That is, we collapse two levels of response history. The resulting weights are reasonably bounded and the corresponding estimator of the total is conditionally unbiased, when we condition on response history up to 1989 inclusive.

Let $\hat{Y}_{90,92} = 100 \Sigma w_{90,92} \cdot y$ with

$w_{90,92} = w_{92/90} \cdot w_{90}$, and

$w_{92/90}^{-1} = pr(r_{92}=1,\ r_{90}=1 / Z_{90})$. Then

$$E_{90} = E(\hat{Y}_{90,92} / Z_{90}) \neq Y_{90,92}$$

and

$$E(E_{90} / Z_{89}) = Y_{90,92}.$$

With a variance increase for 1992 of 10%, the adjustment for the 1991–1992 longitudinal analysis results in an approximate variance increase of 21%:

$$(1 + CV_{91}^2) * (1 + CV_{92}^2) = (1.097) * (1.1) = 1.21.$$

The 21% is an overestimate anyways because the records which carry the largest weights in the 1991 and 1992 cross-sectional samples do not belong to the longitudinal sample.

The increase in variance corresponding to the longitudinal 1990–1992 weight is indicated by

$$(1 + CV_{90}^2) * (1 + CV_{92/90}) = (1.06)(1.02) = 1.08.$$

Here the statistic $1 + CV^2$ for the conditional weight $w_{92/90}$ is 1.02. The relatively small increase (2%) in the second level of variation is the result of collapsing levels in the response history class: the variance was reduced at the expense of some possible bias.

6. RESULTS

The relatively large increase in estimator variance implied by the statistic $1+CV^2$ can be controlled by considering alternative ratio estimators. The original estimator, $\hat{Y}_u$ (model U), unadjusted for missing data usually underestimates the quantities of interest, whereas $\hat{Y}_w$ (model W), the estimator that includes the weight adjustments, has a larger variability, but is unbiased under the model. We considered for comparison two other very simple estimators. If there is no known count of the population, we can control the variance and reduce some of the bias by considering an estimator developed under the assumption that the data is completely missing at random (model C.M.A.R.), $\hat{Y}_{C.M.A.R.}$. And finally we also look at the ratio-adjusted estimator $\hat{Y}_R$ (model R), which requires a known count of the population.

We first compared all four estimators by estimating some cross-sectional population characteristics for which we know the corresponding Postcensal Estimates of Population. Table 6.1 shows estimates of the male population, 19 years and older. The T1FF count in Table 6.1 is obtained from the entire T1 Family File. This quantity is affected only by undercoverage.

The unadjusted LAD estimates are almost equal to the total count for the T1FF. The small difference (e.g. 58,145 in 1991) can be attributed to sampling error in LAD. Thus the relative difference between the unadjusted LAD estimates and the Postcensal Estimates may be caused solely by the undercoverage.

The relative differences with the corresponding Postcensal Estimate of Population indicate that the adjustments have considerably improved the estimates and that the missing at random model is probably the correct one for this characteristic.

Table 6.1 Canadian population, 19+ male

Estimator	1991	1992
Postcensal Estimate of Population, N	10,074,228	10,373,780
T1FF Count	9,430,060	9,601,300
Unadjusted LAD, $\hat{Y}_u$	9,488,205	9,692,600
Rel. Diff. $(\hat{Y}_u - N)/N$	(−5.8%)	(−6.6%)
Adjusted Y_w	10,020,199	10,290,106
Rel. Diff.	(−0.5%)	(−0.8%)
Adjusted $\hat{Y}_{M.A.R.}$	10,004,773	10,248,095
Rel. Diff.	(−0.7%)	(−1.2%)
Adjusted−Ratio, $\hat{Y}_r$	10,205,318	10,420,915
Rel. Diff.	(1.3%)	(0.5%)

We compared the estimators under models U, W, C.M.A.R. and R, by looking at the estimates of some demographic, income and investment characteristics obtained from the four methods. We also calculated the adjusted and unadjusted proportions and tested for significance of the difference under the model W. This test suggests if there is any bias left after adjusting under C.M.A.R. conditions.

Table 6.2 shows estimates of some longitudinal totals. The star means that there is a significant difference between the estimates adjusted under the missing completely at random and under model W. We see that when estimating the number of individuals who lived in the Atlantic Region in 1990 and moved out by 1992, there is no significant difference between the models, but in the case of estimating low income dynamics, not only there is a significant difference between the estimates under the first three models, but also the relative difference between the weight adjusted estimate and the unadjusted estimate has jumped up to 15%.

If we are interested in detecting late filers among a special and small domain of very high income individuals (1%), we may require to use a finer classification of income including a very high income level in the definition of the response classes.

The last entry in Table 6.2 refers to the total of individuals contributing to RRSP's for less than $500 in 1991 and over $1000 in 1992.

Even though the estimate under the model is larger than the unadjusted estimate by 26,000 individuals or 8% of the adjusted estimate, it does not represent a significant difference.

Nevertheless people who contribute to RRSP's usually do it with borrowed money and it is in their interest to file on time so they can receive their tax refunds as early as possible. We wanted to investigate if a 'finer' model would yield a different result. Thus, we developed special weights for this estimator, adding to the model an indicator of RRSP contributions. The resulting estimates were very similar to those in Table 6.2 and the difference between the unadjusted and the adjusted LAD totals was not reduced. This result may be an indication of the stability of our model; the weighting adjustments developed here provide a reasonable improvement in the estimators.

Table 6.2 Longitudinal Totals

Models	U	C.M.A.R.	W	R
Low Inc 90 & 92	1,809,404	1,971,035*	2,125,093	2,170,677
Atlantic filers 90, elsewhere 92	26,100	28,432	29,953	30,595
Low Inc 91, 92	2,206,084	2,393,122*	2,570,428	2,620,430
Low Inc 91,not in 92	1,221,997	1,325,601*	1,406,987	1,434,357
RRSP '91 < 500 and RRSP '92 > 1000	297,097	322,470	322,695	328,783

7. REFERENCES

Annual Demographic Statistics (1994) Statistics Canada Catalogue p.91-213.

Bleuer, S.-R. (1996) *"Non-Response Adjustments for Estimation of Gross Flows and Transitions from LAD"*. Internal Report, Small Area and Administrative Data Division, Statistics Canada.

Bleuer, S. and Bernier, J. (1996). *"Characteristics of the T1FF Undercoverage"*. Internal Report, SAAD, Statistics Canada.

Demnati, A. (1992). *"Données administratives longitudinales place de sondage"*. Internal Report, Social Survey Methods Division, Statistics Canada.

Kalton, G., and Miller, M. (1986). *"Effects of Adjustments for Wave Non-response on Panel Survey Estimates"*. Proc. Section on Survey Research Methods, ASA, p. 194-199.

Little, R. and David, M. (1983). *"Weighting Adjustments for Non-response in Panel Surveys"*. U.S. Bureau of the Census Working Paper.

Lepkowski, J.M. (1989). *"Non-response Adjustments for Wave Non-response"*. Panel Surveys.

Rizzo, L., Kalton, G., Brick, M. and Petroni, R. (1994). *"Adjusting for Panel Non-response in the Survey of Income and Program Participation"*. Proc. ASA, Survey Research Methods Section.

ACKNOWLEDGMENT

The author is grateful to Jon Rao and Barbara Bailar for their helpful comments.