ASSESSING BIAS DUE TO CLASSIFICATION ERROR IN LABOUR FORCE SURVEY DATA

Jane F. Gentleman, Statistics Canada
Social and Economic Studies Division, Ottawa, Ontario, Canada, K1A 0T6

KEY WORDS: gross flows, labour force survey, bias

This paper discusses several assumptions commonly made in modelling labour force survey data. The implications of these assumptions are examined. In particular, it is shown that combining the assumption that observed stock levels are unbiased with certain other common assumptions implies that true stock levels can be determined as a function of classification error probabilities. The design of and data from the Canadian Labour Force Survey (LFS) are used to illustrate these points.

The Canadian Labour Force Survey is a sample survey of approximately 55,000 households. Its primary purpose is to generate a broad array of data on labour market conditions. Each month, approximately 125,000 individuals of age 15 or over are classified as being employed, unemployed, or not in the labour force. The totals in each class are disaggregated by age, sex, province, industry, occupation, etc. A multistage sampling design is used in which a household's probability of being selected depends on its geographical location. The survey has a rotating panel design; a given household remains in the sample for six consecutive months, and each month, one-sixth of the households enter and one-sixth leave the sample. Data are collected by Statistics Canada interviewers in personal or telephone interviews. A Reinterview Programme assesses the interview process and estimates the probability of non-sampling errors of classification. Reinterviews are conducted by senior interviewers for about one in forty-five households each month. Two-thirds of the reinterview data are subjected to a reconciliation process in which, if a discrepancy is found, the reinterviewer attempts to secure a true response. The reconciled sample can be used to estimate response bias, and the unreconciled sample to estimate response variance. Analytically, the reconciled data are sometimes treated as known, true responses, and the unreconciled one-third of the data are sometimes treated as an independent replication of the interview.

The term "stock levels" refers to the observed frequencies of R different labour force classifications. For the three-class system above, R=3. The term "gross flows" denotes the observed transition frequencies between R states at time t and R states at time t+1. While it is is not the primary purpose of labour force surveys to measure longitudinal effects, considerable interest exists in such statistics. The dynamics of the labour market and their impact on individual members of the population can be better understood if individual gross flows, not just net flows, are known.

Numerous problems exist in obtaining reliable estimates of gross flows. To consolidate ideas for handling them, a Conference on Gross Flows in Labor Force Statistics was sponsored in the summer of 1984 by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. Much of the ongoing concern with gross flows estimation arises because of the biases that would be introduced by using observed gross flows to estimate true gross flows in the presence of classification error. Statistical agencies have regularly published observed stock levels, but they have not always published observed gross flows. According to Evans (1985), although there are about 14 OECD countries which have regular household labour force surveys, Australia is the only country outside North America to publish gross flow data prominently and regularly. Typically, it has been assumed that bias can be considerable in the observed flows, especially between two different states, but that biases in the observed stock levels tend to cancel out (see, for example, Abowd and Zellner (1985, p. 45), Fuller and Chua (1985, p. 65), Fellegi (1979), Hogue (1985, p. 2), Veevers and Macredie (1983), Macredie (1983), and the quote in Evans (1985, p. 123) of the Australian Bureau of Statistics).

Assume there are R mutually exclusive exhaustive states into which a respondent can be classified. Let $Y = \{y_{ij}\}$ $(i=1,\ldots,R;\ j=1,\ldots,R)$ denote the observed gross flows between time t and time t+1; that is, y_{ij} is the observed number of respondents who were classified as being in state i at time t and in state j at time t+1. Assume that the y_{ij}'s have a multinomial distribution with sample size $N=\sum\sum y_{ij}$ and probabilities $P=\{p_{ij}\}$. The notation P(obs i$\longrightarrow$ obs j) will be used to represent p_{ij} when convenient to designate the probability of observing a flow from state i at time t to state j at time t+1. Similarly, let h_{ij} = P(true i$\longrightarrow$true j) be the true flow probabilities between state i at time t and state j at time t+1. When there are classification errors, these true flow probabilities are not equal to the observed flow probabilities. Specifically,

$$p_{ij} = \sum_{m=1}^{R} \sum_{n=1}^{R} g_{ijmn} h_{mn} , \qquad (1)$$

where g_{ijmn} = P(obs $i \rightarrow$ obs j | true $m \rightarrow$ true n). To reduce the number of unknowns in Eqn. 1 and to permit the use of available reinterview data, researchers and analysts have typically made the assumption of "independent errors":

$$g_{ijmn} = k_{im} l_{jn} , \qquad (2)$$

where k_{im} = P(obs i at time t | true m at time t) and l_{jn} = P(obs j at time $t+1$ | true n at time $t+1$). This reduces the number of needed estimates to $2R^2$. Reinterview data can be used to estimate classification error probabilities k_{im} and l_{jn}, if not g_{ijmn}. Note that a true error independence assumption would require only that g_{ijmn} is the product of P(obs i at time t | true $m \rightarrow$ true n) and P(obs j at time $t+1$ | true $m \rightarrow$ true n), which would reduce the number of needed estimates to only $2R^3$ and still would not permit available reinterview data to be used.

Let $H=\{h_{ij}\}$, $K=\{k_{ij}\}$, and $L=\{l_{ij}\}$. Implicit in the definitions of these quantities are the properties that the elements of H add to one ($1_R'H1_R=1$, where 1_R is an Rx1 vector of ones), the columns of K each add to one ($1_R'K=1_R'$), and the columns of L each add to one ($L'1_R=1_R$). Under the independent errors assumption of Eqn. 2, Eqn. 1 becomes

$$P=KHL' . \qquad (3)$$

Assume that the model of Eqn. 3 holds. Then $E(Y/N) = P = KHL'$. Let b_{ij} be the bias of y_{ij}/N when used as an estimate of h_{ij}. Then

$$b_{ij} = \sum_m \sum_n k_{im} h_{mn} l_{jn} - h_{ij} . \qquad (4)$$

Obviously, the observed gross flows will all be unbiased if K and L are identity matrices, i.e., if there are no classification errors. Assuming that this is not (quite) true, the bias is sensitive to the values h_{ij}. Insight into the behaviour of b_{ij} and of the relative bias $r_{ij} = b_{ij}/h_{ij}$ can be gained by examining a simple approximation to Eqn. 4. Assume that $k_{ii} \cong l_{ii} \cong \delta$ ($i=1,\ldots,R$), where δ is less than but close to one. Assume also that $k_{ij} \cong l_{ij} \cong (1-\delta)/(R-1)$ ($i=1,\ldots,R$; $j=1,\ldots,R$; $i \neq j$). These simplifications imply that K and L are close to being identity matrices. Expanding the double summation to isolate h_{ij}, and assuming that $(1-\delta)^2/(R-1)^2 \cong 0$, an approximation to the relative bias is obtained:

$$r_{ij} \cong (1-\delta)\left[-1-\delta + \left(\frac{\delta}{R-1}\right)\left(\frac{s_{ij}}{h_{ij}}\right)\right] , \qquad (5)$$

where $s_{ij} = \sum_{m \neq i} h_{mj} + \sum_{n \neq j} h_{in}$.

The behaviour of r_{ij} depends importantly on whether or not $i=j$. Assume that the interval of time between t and $t+1$ is relatively small, as in the LFS, so that changing to or from state i is much less probable than staying in state i, and assume that R is small. Then s_{ii}/h_{ii} is fairly small, so r_{ii} will be negative, and r_{ij} will be positive and quite large. Accordingly, the hypothetical two-state example in Macredie (1983, pp. 2-3) produces relative errors of -.018 and -.006 for observed frequencies of flows between the same states, and relative errors of +.940 and +.297 for flows between different states.

If Y is multinomial, the marginal sums $Y1_R$ are multinomial, with $E(Y1_R/N) = KH1_R$, and the marginal sums $1_R'Y$ are multinomial with $E(1_R'Y/N) = 1_R'HL'$. Therefore, if the observed stock proportions of time t are used as estimates of the true stock probabilities at time t, their biases are as follows:

$$\text{Bias of } (Y1_R/N) = E(Y1_R/N)-H1_R = (K-I_R)H1_R ,$$

where I_R is an RxR identity matrix. Similarly, the biases of the column sums $1_R'Y/N$ are:

$$\text{Bias of } (1_R'Y/N) = 1_R'H(L'-I_R) .$$

Thus, the row sums of Y are unbiased if $K=I_R$, and the column sums are unbiased if $L=I_R$, i.e., if there are no classification errors. These conditions are sufficient, but not necessary, for the marginal sums of Y to be unbiased.

If $K \neq I_R$ (or $L \neq I_R$) and the row sums (column sums) are nevertheless assumed to be unbiased, the row sums (column sums) of H can be expressed explicitly in terms of only the elements of K (of L). That is, the common assumption that observed stock levels are unbiased is, in the presence of classification errors and the usual assumptions about them, sufficiently stringent that the true stock levels can be determined from the (assumed known) classification error probabilities alone. This relationship is derived as follows.

Assume that K is known, and suppose that the observed stock proportions $Y1_R/N$ at time t are unbiased estimates of the true stock proportions. Then $E(Y1_R/N)=H1_R$, and from Eqn. 3,

$E(Y1_R/N)=KH1_R$, so that $(K-I_R)H1_R = 0_R$ (where 0_R is a vector of R zeros). This is a system of R equations in the R unknowns $H1_R$, only R-1 of which are linearly independent (due to the restriction $1_R'K=1_R'$ on K). Replacing one of these equations, say the last, by the equation $1_R'H1_R = 1$, the system can be solved numerically or algebraically to obtain explicit values or formulas for the true stock probabilities $H1_R$ at time t. That is,

$$H1_R = \begin{bmatrix} A-I_{R-1} & B \\ 1'_{R-1} & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0_{R-1} \\ 1 \end{bmatrix} , \quad (6)$$

where A is an (R-1)x(R-1) matrix consisting of the first R-1 rows and first R-1 columns of K, and B is the first R-1 elements of the Rth column of K. Similarly, the true stock probabilities $1_R'H$ at time t+1 are a function of the l_{ij}'s if $1_R'Y/N$ is unbiased.

The formal hypothesis that an estimate of stock levels at time t is unbiased, e.g., that $E(Y1_R/N)=H1_R$, can be tested using a goodness-of-fit procedure. Under the hypothesis, and given K, $H1_R$ is completely determined, so the hypothesis is simple and the χ^2 test statistic will have the full R-1 degrees of freedom.

Once K, L, and the observed gross flows Y are all available, a formal goodness-of-fit test can be performed of the hypothesis that both sets of marginals of Y are unbiased, i.e., that $E(Y1_R/N) = H1_R$ and $E(1_R'Y/N) = 1_R'H$, simultaneously. This test uses the individual elements of Y as observed frequencies, rather than the marginals of Y. The hypothesis is composite; to compute expected frequencies for the χ^2 statistic, maximum likelihood estimates $\hat{H}$ of the individual elements of H are calculated by maximizing the likelihood function $L \propto \prod\prod p_{ij}^{y_{ij}}$, using the relationship of H to P of Eqn. 3. The marginals of $\hat{H}$ are fixed by the hypothesis, but there remain $(R-1)^2$ unfixed elements of $\hat{H}$. The χ^2 statistic will therefore have 2(R-1) degrees of freedom.

As an example, Table 1 provides observed gross flows (Y), and Tables 2A and 2B provide reconciled reinterview data (K and L) from 1983 Canadian Labour Force Survey data. These data were supplied by Georges Lemaitre, then of the Census and Household Surveys Division at Statistics Canada.

Table 3 gives the "corrected" estimate $K^{-1}YL'^{-1}/N$ of gross flow probabilities. Comparing Table 1 and Table 3, note that, consistent with the bias patterns described above (despite a fairly large off-diagonal element ((3,2)) in both K and L), all of the diagonal elements of Y/N are smaller than those of $K^{-1}YL'^{-1}/N$, and all of the off-diagonal elements have the reverse ordering. The relative differences are -.03, -.25, and -.03 for the diagonal elements, and they range from +.11 to +2.86 for the off-diagonal elements. The marginal sums of Y/N are close to those of $K^{-1}YL'^{-1}/N$, with absolute relative differences ranging from .005 to .05, giving support to their treatment as relatively reliable estimates even in the presence of classification errors.

Table 4A provides the maximum likelihood estimate $\hat{H}=K^{-1}\hat{P}L'^{-1}$, calculated under the composite hypothesis that the two sets of marginal sums of Y/N are unbiased estimates of the two sets of marginals of H. Table 4A is obtained assuming a multinomial distribution for Y, assuming independent errors (as in Eqn. 2), and treating K and L as true classification error probabilities. The row sums in Table 4A are true stock probabilities for time t, fixed under the simple hypothesis that the row sums of Y/N are unbiased estimates of the row sums of H; the column sums in Table 4 are true stock probabilities for time t+1, fixed under the simple hypothesis that the column sums of Y/N are unbiased estimates of the column sums of H.

Values of $\hat{P}$ are provided in Table 4B. To test the composite hypothesis that both the row and column sums of Y/N are unbiased estimates of the marginal sums of H, the χ^2 test statistic compares the observed frequencies Y in Table 2 to their expected frequencies $N\hat{P}$. A formal significance test is not required to reject the hypothesis, as there are clearly very large differences between Tables 1 and 4B. (The actual P-value of the significance test is near zero.)

In general, there are several ways to estimate stock levels and several ways to estimate gross flows. Despite the varying reliability and timeliness of these estimates, formal and informal comparisons among them can be used to assess the overall quality and consistency of interview data, reinterview data, and the models used to describe them. The above results indicate that combining what may seem, individually, to be reasonable (or at least convenient) assumptions yields a model which is not consistent with available data. The ability to estimate true gross flows would be enhanced by further research to provide more realistic, usable models, and by better and/or different data.

REFERENCES

Abowd, John M. and Zellner, Arnold (1985). Application of Adjustment Techniques to U.S. Gross Flow Data, Proc. Conf. on Gross Flows in Labor Force Statistics, U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, pp. 45-61.

Evans, John M. (1985), Gross Flow Statistics Outside North America: Construction and Use. Proc. Conf. on Gross Flowsin Labor Force Statistics, U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, pp. 119-129.

Fellegi, I.P. (1979), Internal memorandum to Ian Macredie, Statistics Canada, Census and Household Surveys Field.

Fuller, Wayne A. and Chua, Tin Chiu (1985), Gross Change Estimation in the Presence of Response Error. Proc. Conf. on Gross Flows in Labor Force Statistics, U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, pp.

Hogue, Carma R. (1985), History of the Problems Encountered in Estimating of the Census and U.S. Bureau of Labor Statistics, pp. 1-7.

Lemaitre, Georges (1985). Private communication. Census and Household Surveys Division, Statistics Canada.

Macredie, Ian (1983), The Impact of Response Errors in the Estimation of Labour Market Flows, Internal publication, Statistics Canada, Economic

Veevers, Richard and Macredie, Ian (1983), Estimating Gross Flows from the Canadian Labour Force Survey, Internal publication, Statistics Canada, Economic Characteristics Division, 13 pages.

TABLE 1

Observed Gross Flows from Canadian Labour Force Survey: June to July 1983
Counts (Y) of individuals in sample, and proportions out of N = 92,192
Source: Lemaitre (1985)

JULY STATUS

JUNE STATUS	Employed	Unemployed	Not in L.F.	Total
Emp	50,191 (.5444)	1,073 (.0116)	1,351 (.0147)	: 52,615 (.5707)
Unemp	1,558 (.0169)	3,945 (.0428)	1,031 (.0112)	: 6,534 (.0709)
NILF	1,863 (.0202)	1,540 (.0167)	29,640 (.3215)	: 33,043 (.3584)
TOTAL	53,612 (.5815)	6,558 (.0711)	32,022 (.3474)	: 92,192 (1.0)

TABLE 2

Interview Data and Reconciled Reinterview Data from 1983 Canadian Labour Force Survey
Counts of individuals in sample, and proportions (K and L) out of column sums
Source: Lemaitre (1985)

A. Merged data (K) for January through June (excluding April)

RECONCILED REINTERVIEW STATUS

INTERVIEW STATUS	Employed	Unemployed	Not in L.F.	
Emp	2,475 (.9810)	8 (.0188)	19 (.0110)	
Unemp	12 (.0048)	360 (.8451)	24 (.0139)	
NILF	36 (.0143)	58 (.1362)	1,687 (.9751)	
TOTAL	2,523 (1.0)	426 (1.0)	1,730 (1.0)	: 4,679

TABLE 2 (Cont'd)

B. Merged Data (L) for July through November (excluding October)

RECONCILED REINTERVIEW STATUS

INTERVIEW STATUS	Employed	Unemployed	Not in L.F.	
Emp	3,291 (.9865)	12 (.0289)	17 (.0081)	
Unemp	12 (.0036)	368 (.8867)	16 (.0076)	
NILF	33 (.0099)	35 (.0843)	2,063 (.9843)	
TOTAL	3,336 (1.0)	415 (1.0)	2,096 (1.0)	: 5,847

TABLE 3

Estimates of True Gross flow Probabilities $(K^{-1}YL'^{-1})/N$ and of True Stock Probabilities After Correcting Observed Gross Flows for Classification Errors

JULY STATUS

JUNE STATUS	Employed	Unemployed	Not in L.F.		Total
Emp	.5618	.0098	.0048	:	.5764
Unemp	.0152	.0568	.0029	:	.0749
NILF	.0074	.0083	.3329	:	.3486
TOTAL	.5844	.0749	.3406	:	1.0

TABLE 4

Implications of the Hypothesis that Observed Stock Levels (Marginals of Table 2) are Unbiased Given Classification Error Probabilities K and L

A. Maximum Likelihood Estimate ($\hat{H}$) of True Gross Flow Probabilities, and True Stock Probabilities Fixed by Hypothesis

JULY STATUS

JUNE STATUS	Employed	Unemployed	Not in L.F.		Total
Emp	.3787	.0040	.0000	:	.3827
Unemp	.0201	.0389	.0027	:	.0617
NILF	.0233	.0061	.5262	:	.5556
TOTAL	.4221	.0489	.5290	:	1.0

B. Maximum Likelihood Estimate ($\hat{P}$) of Observed Gross Flow Probabilities, and True Stock Probabilities Fixed by Hypothesis

JULY STATUS

JUNE STATUS	Employed	Unemployed	Not in L.F.		Total
Emp	.3673	.0056	.0099	:	.3827
Unemp	.0199	.0294	.0124	:	.0617
NILF	.0349	.0140	.5067	:	.5556
TOTAL	.4221	.0489	.5290	:	1.0