ON AN ε-CONTAMINATION MODEL FOR MULTINOMIAL RESPONSE ERROR WITH APPLICATION TO LABOUR FLOWS

A.C. Singh and G.E. Lemaître
Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, CANADA

ABSTRACT

A model for the response error associated with reported categorical data is proposed. For each individual k, we assume that there is a small chance ε_k that his reported response would be prone to error which however, may or may not lead to a response error. In this case, the reported response follows a contamination distribution. On the other hand, with probability $1-\varepsilon_k$, his reported response would not be prone to error and follows the true underlying distribution. The average of ε_k's over all individuals defines the mixing parameter ε for the contamination model. We illustrate an application of the proposed model in correcting classification bias for the month to month labour flow data obtained from the Canadian Labour Force Survey. It does not require the use of reinterview data as is commonly the case for other adjustment methods. Instead, we use administrative sources, namely, Unemployment Insurance files to estimate the model parameter ε.

KEY WORDS: Measurement error; Right-wrong models; Unbiased response error; Stocks and flows; Unemployment Insurance claims.

1. INTRODUCTION

In this paper we present a model-based approach to the problem of gross flow estimation for the characteristic "labour force status". Labour market data are often obtained from multistage stratified clustered samples of households with rotating panel designs. In the case of the Canadian Labour Force Survey (LFS), approximately 5/6th of the households are common to two consecutive months and each sampled household is interviewed consecutively for six months before being dropped. For each sampled individual belonging to the civilian noninstitutional population 15 years of age or over, data are collected on labour force status for the week prior to the week in which the survey is conducted. Individuals are classified as either E(employed), U(unemployed) or N(not in the labour force).

The main problems in modeling panel survey data include the presence of nonresponse, response or classification error, inflows to or outflows from the population of interest, and inconsistency with external population counts. Little (1985), Fay (1986), Stasny and Fienberg (1985), and Stasny (1986) among others considered the first problem by modeling in the presence of nonrandom nonresponse, i.e. the distribution of variable of interest for nonrespondents was not assumed to be the same as that for respondents.

The second major problem in the estimation of gross flows (i.e. proportions of individuals making transitions between categories at one time point to another) is that due to classification errors. It is believed that biases in the stocks or marginal counts are negligible but the interior counts or flows are considerably biased. In particular, there may be serious upward biases in the off-diagonal cell flows (which happen to be of main interest) of the (E, U, N) by (E,U,N) gross-flow table because most of the individuals do not tend to change their status from one month to the next. Thus even with small chances of misclassification, the relative errors in the off-diagonal flows are expected to be quite high. The factors that could contribute to errors in the classification process include proxy response, a misunderstanding of questions asked in the interview, coding errors, misinterpretation by the interviewer of the classification criteria, frequent status changes, etc.

A popular model used for correcting classification errors is based on the availability of reinterview data for estimating error rates under the assumption of independent classification errors, see e.g. Abowd and Zellner (1985), Fuller and Chua (1985), Chua and Fuller (1987), Poterba and Summers (1986). These methods differ with respect to how the error rates are estimated from reinterview results. The independence assumption means that an individual's observed classification at time t depends stochastically on his true classification at time t but not on his true or observed classification at t-1. If the response errors were serially dependent, say, positively correlated, then they would tend to decrease the number of reported changes and increase the number of reported continuations of the previous state. In general, this means that the adjustments for response errors would be smaller than those under serially independent classification errors; see also Gentleman (1988) for implications of some commonly made assumptions.

We propose a model for correcting for classification errors in the multinomial distribution defined by the two way table of (E,U,N) by (E,U,N) for the time period t-1 to t. We assume that the marginal proportions (or stocks) are unbiased. Under the assumption of unbiased response error at a single point in time, the flows (or interior proportions of the table) can still be seriously biased, see e.g. Chua and Fuller (1987). In our model, we specify response errors by means of a contamination distribution which is estimated under the assumption of unbiased stock estimates and that responses for the two time points are independent only for those individuals who are prone to error. Unlike the usual right-wrong models (Fuller 1987, p. 273), which involve response probabilities conditional on the true classification, the contamination model is based on unconditional probabilities for the reported categories. The link with the true underlying distribution is provided by a mixing parameter, denoted here by ε. Moreover, instead of using reinterview data to estimate response probabilities conditional on the true classifications, we propose to use administrative data (from Unemployment Insurance files) to estimate the model parameter ε.

2. ADJUSTMENT METHODS BASED ON REINTERVIEW DATA : A BRIEF REVIEW

We shall follow the description given in Fuller (1987, p. 274) in the section on right-wrong models for multinomial response error. The basic idea is to write down the relationship between expected observed gross flow proportions on one side and true gross flows along with the classification error (conditional response) probabilities on the other. This formula is then used to estimate the true gross flow proportions. We have

$$\pi^{(0)}_{k\ell}(t-1,t) = \sum_k \sum_\ell \phi_{ij|k\ell}(t-1,t)\ \pi^{(1)}_{k\ell}(t-1,t) \qquad (2.1)$$

for $k,\ell = E,U,N$ where $\pi^{(0)}_{ij}$ is the expected fraction in the ij^{th} cell of the observed gross flow table for $(t-1,t)$, $\pi^{(1)}_{k\ell}$ is the true proportion for $(k\ell)^{th}$ cell, and $\phi_{ij|k\ell}$ is the conditional probability of observing or reporting (i,j) classification when in fact the true classification in labour force status from t-1 to t is (k,ℓ).

Under the assumption of independent classification errors for time points t-1 and t, we have

$$\phi_{ij|k\ell}(t-1,t) = \beta_{i|k}(t-1)\ \beta_{j|\ell}(t), \qquad (2.2)$$

which implies that

$$\pi^{(0)}(k-1,t) = B(t-1)\ \pi^{(1)}(t-1,t)\ B(t)' \qquad (2.3a)$$

or alternatively it can be conveniently expressed as

$$\text{vec}\ \pi^{(0)} = (B(t) \times B(t-1))\ \text{vec}\ \pi^{(1)}, \qquad (2.3b)$$

where $B(t-1)$, $B(t)$ are matrices of conditional response probabilities $\beta_{i|k}(t-1)$, $\beta_{j|\ell}(t)$ respectively, vec $\pi^{(0)}$ is the column vector obtained by listing the columns of $\pi^{(0)}$ one below the other, and $\times$ denotes the Kronecker product. One can solve (2.3) to obtain

$$\text{vec}\ \pi^{(1)} = (B(t)^{-1} \times B(t-1)^{-1})\ \text{vec}\ \pi^{(0)} \qquad (2.4)$$

or

$$\pi^{(1)} = B(t-1)^{-1}\ \pi^{(0)}\ B(t)^{-1}.$$

Let $\hat{\pi}^{(0)}$ be the observed two way table of flow proportions, and $\hat{B}(t-1)$, $\hat{B}(t)$ the estimates of classification error rates from reinterview data at each time point. Then an estimator of $\pi^{(1)}$ is easily obtained from (2.4). For an estimate of the covariance matrix of $\hat{\pi}^{(1)}$, the assumption of multinomial sampling is generally made.

The various adjustment methods based on reinterview data differ with respect to how the B matrices are estimated. In the Abowd and Zellner approach (1985), for example, the reconciled reinterview sample data are used. It is assumed that the reconciled reinterview states are correct and quarterly aggregates of interview-reinterview tables are used to estimate B matrices. In the Chua and Fuller (1987) approach, on the other hand, the unreconciled reinterview data is used and model-based estimates of $B(t-1)$ and $B(t)$ are generated based again on some suitable aggregates of monthly interview-reinterview data. Poterba and Summers (1986) calculate two sets of error rates. The first are the usual rates from the reconciled sample used by Abowd and Zellner. They state that these rates are commonly regarded as downward-biased; indeed the rates of inconsistency from the reconciled sample before reconciliation are substantially smaller than those for the unreconciled sample. Poterba and Summers attempt to correct for this bias by calculating from the reconciled sample the probability of each "true" status, conditional upon first and second interview status. These probabilities are then used to synthetically estimate the number of individuals in the unreconciled sample with each "true" labour market status.

3. THE MODEL

We can use the right-wrong model framework (Fuller 1987, p.273) to motivate the proposed response error model. The models described in the previous section belong to submodel I of right-wrong models in which every individual truly belongs to one of the 3^2 categories. Moreover, all individuals belonging to the same true category have a common matrix of conditional response probabilities. The proposed model can be viewed as a submodel II of the right-wrong models in which the conditional response probability matrix is allowed to vary for individuals within the same true category. In this case, these probabilities $(\phi_{ij|k\ell})$ would, however, be difficult to estimate. The response error is, therefore, specified indirectly through unconditional response probabilities as follows.

First we assume that individual k has a small chance ε_k of being susceptible to response error in the time period $(t-1,t)$ under consideration. With chance $1-\varepsilon_k$, the individual k responds with no error. We then assign an appropriate gross-flow distribution depending upon whether the individual k was or was not susceptible to error in the given time period. Thus the expected proportions $\pi^{(0)}$ of the observed flow tables are obtained as a result of a two-stage process. The chance ε_k would in general vary from individual to individual because of various factors (cognitive and situational). If an individual turns out not to be prone to error, then his response follows the true multinomial classification distribution $\pi^{(1)}$. If the individual k is prone to error, then he may or may not make an error; his response would then follow a different classification distribution, $\pi^{(2)}$ say, which we will refer to as the contamination distribution.

We next assume that if an individual is prone to error, then his response at time t does not depend on his response at time t-1, i.e. for each classification at time t-1, the response probabilities at time t are the same. In view of the factors enumerated in the introduction that could contribute to response error, this assumption may not be unreasonable because of the lapse of time between the two interviews. Thus, we write

$$\pi^{(2)}_{ij}(t-1,t) = \pi^{(2)}_{i+}(t-1)\ \pi^{(2)}_{+j}(t). \qquad (3.1)$$

This assumption is somewhat similar to the one of independent classification errors (2.2) except that this is assumed to hold only for a small fraction of individuals who were found to be susceptible to error. Also we remark that $\pi^{(2)}_{ij}$ denotes the probability of observing or reporting (i,j) classification when an individual is prone to error. Unlike the classification probability $\phi_{ij|k\ell}$ conditional on the true state (k,ℓ), it gives no indication of the chance of actually making an error.

For all individuals, we have assumed a true common multinomial classification distribution $\pi^{(1)}$ and a common contamination distribution $\pi^{(2)}$. Thus, the probability of observing (i,j) for the k^{th} individual is given by

$$\pi^{(0)}_{ij(k)} = (1-\varepsilon_k)\ \pi^{(1)}_{ij} + \varepsilon_k\ \pi^{(2)}_{ij}. \qquad (3.2)$$

However we are not interested in a particular individual k but in the aggregate proportion of individuals in the ij^{th} cell. Letting ε denote the average of ε_k's, we have that the expected fraction of individuals observed in the ij^{th} cell is given by

$$\Pi_{ij}^{(0)} = (1-\varepsilon)\,\Pi_{ij}^{(1)} + \varepsilon\,\Pi_{ij}^{(2)}. \tag{3.3}$$

This defines an ε-contamination model for multinomial response error, where ε is the mixing parameter and represents the rate of contamination. If $\varepsilon=0$, then the responses would be free from any error. A useful practical interpretation of ε can be given if we assume that the ε_k's are approximately the same for all individuals in the time period under consideration. Then ε approximates the fraction of individuals in the population who are prone to response errors in the given time period.

Now let $\hat{\Pi}_{ij}^{(0)}$ denote the observed flow proportions with expectations $\Pi_{ij}^{(0)}$. It can be seen that the response error bias b_{ij} in the estimator $\hat{\Pi}_{ij}^{(0)}$ for the true proportion $\Pi_{ij}^{(1)}$ is given by

$$b_{ij} = \Pi_{ij}^{(0)} - \Pi_{ij}^{(1)} = -\,\delta(\Pi_{ij}^{(0)} - \Pi_{ij}^{(2)}) \tag{3.4}$$

where δ is $\varepsilon/(1-\varepsilon)$. An estimate $\hat{b}_{ij}$ can be obtained by finding estimates of ε, $\Pi_{ij}^{(0)}$, and $\Pi_{ij}^{(2)}$ which would then provide the adjusted or corrected estimate of $\Pi_{ij}^{(1)}$ as

$$\hat{\Pi}_{ij}^{(c)} = \hat{\Pi}_{ij}^{(0)} - \hat{b}_{ij}\,. \tag{3.5}$$

We shall now assume as in Chua and Fuller (1987) that the marginals $\hat{\Pi}_{i+}$ and $\hat{\Pi}_{+j}$ have unbiased response errors. This implies that the stock estimates are unbiased, i.e.

$$\Pi_{i+}^{(0)} = \Pi_{i+}^{(1)},\ \Pi_{+j}^{(0)} = \Pi_{+j}^{(1)} \tag{3.6}$$

It now follows that the proportions $\Pi_{ij}^{(2)}$ must satisfy

$$\Pi_{i+}^{(2)} = \Pi_{i+}^{(1)},\ \Pi_{+j}^{(2)} = \Pi_{+j}^{(1)}, \tag{3.7}$$

and consequently,

$$b_{i+} = b_{+j} = 0. \tag{3.8}$$

This means that the biases for each row i and each column j cancel each other as one would expect under the assumption of unbiased stock estimates. Thus the ε-contamination model automatically satisfies the constraints (3.8) arising from the assumption of unbiased response error at a single point in time.

It follows from (3.6) and (3.7) that $\Pi_{i+}^{(2)}$, $\Pi_{+j}^{(2)}$ can be estimated unbiasedly by $\hat{\Pi}_{i+}^{(0)}$ and $\hat{\Pi}_{+j}^{(0)}$ respectively and so only ε remains to be estimated in order to obtain $\hat{\Pi}_{ij}^{(c)}$ of (3.5). For an interpretation of the ε - contamination model, consider the three matrices $\Pi^{(1)}$, $\Pi^{(0)}$, and $\Pi^{(2)}$. In the first matrix, diagonal proportions are high relative to the off-diagonal ones because most individuals tend to stay in the same status from one month to the next. For the second matrix, diagonal proportions are low relative to the first matrix due to the response error. Finally for the third matrix, diagonal proportions are expected to be further reduced due to independent responses at t-1 and t for error-prone individuals. The proposed model simply assumes that the difference between $\Pi^{(1)}$ and $\Pi^{(0)}$ is a constant fraction δ of the difference $\Pi^{(0)} - \Pi^{(2)}$. A natural generalization would be to consider different ε's for various post-strata. The $\hat{\Pi}^{(1)}$ would then be a weighted linear combination of the individual stratum estimates. This is, however, not considered in this paper.

4. ESTIMATION OF THE MIXING PARAMETER (ε)

For the Canadian labour market, unemployment insurance files can be used to provide an estimate of ε. The administrative data based on the number of new claims for unempoloyment insurance was also used by Abowd and Zellner (1985) for diagnostic purposes. Consider the subgroup of status changers, consisting of persons who in a given month became beneficiaries of the Unemployment Insurance program but were employed in the previous month. In order to qualify for benefits, a person must 1) have been a paid worker; 2) have worked usually 15 or more hours per week; 3) have worked at least 10 weeks; 4) not be a full-time student. In addition, the claimant must be available for work and unable to find a suitable employment. By means of various items on the Canadian Labour Force Survey (LFS) questionnaire, it is possible to identify persons in the sample satisfying the above criteria. The number of such persons suitably weighted will be referred to as the estimated beneficiary inflows.

It may be noted that the subgroup of beneficiary inflows in month t constitutes a composite category made up of parts of (E,U) and (E,N) categories for the time period (t-1,t). Let i = 1, 2, 3, 4, 5 denote one of the five sub-categories defined as 1 = Employed, 2 = Unemployed and not a beneficiary, 3 = Unemployed and a beneficiary, 4 = Not in the labour force and not a beneficiary, and 5 = Not in the labour force and a beneficiary. Then the new category of beneficiary inflows consists of cells (1,3) and (1,5) in the transition table corresponding to the five sub-categories. We now assume that the model (3.3) continues to hold for the above five sub-categories of the original three. To estimate ε from the combined cells (1,3) and (1,5), we need the true proportion of beneficiary inflows from unemployment insurance files. This can be obtained by identifying all persons who were not beneficiaries (with earnings in the first month) but were so in the second month. Now we can estimate δ(or $\varepsilon(1-\varepsilon)^{-1}$) using the expression given by (estimated proportion of beneficiary inflows from t-1 to t minus the true proportion) divided by [(proportion employed in month (t-1) $\times$ (proportion of beneficiaries in month t) - estimated proportion of beneficiary inflows].

Intuitively we might expect ε to be small, somewhere between .01 and .05. Although detailed calculations using unemployment insurance files and LFS have not been completed yet, some preliminary calculations similar to those given in Lemaître (1988) suggest δ or $\varepsilon(1-\varepsilon)^{-1}$ to be around .04. In the examples considered in section 6, we have used δ = .03 and .04 as working values.

5. VARIANCE OF $\hat{\Pi}^{(c)}$

The estimator $\hat{\Pi}^{(c)}$ can be expressed as

$$\hat{\Pi}^{(c)} = (1+\hat{\delta})\,\hat{\Pi}^{(0)} - \hat{\delta}\,\hat{\Pi}^{(0)}\,J\,\hat{\Pi}^{(0)} \qquad (5.1)$$

where J is a 3×3 matrix of ones. For large samples, $\hat{\Pi}^{(c)}$ would be approximately unbiased. For computing its asymptotic variance, we first assume δ known apriori for convenience.

5.1 δ assumed known, $\delta=\delta_0$

We have

$$\hat{\Pi}^{(c)} = (1+\delta_0)\hat{\Pi}^{(0)} - \delta_0\hat{\Pi}^{(0)}J\hat{\Pi}^{(0)}$$

$$\simeq (1+\delta_0)\hat{\Pi}^{(0)} - \delta_0(\Pi^{(0)}J\Pi^{(0)} + \Pi^{(0)}J(\hat{\Pi}^{(0)}-\Pi^{(0)}) + (\hat{\Pi}^{(0)}-\Pi^{(0)})\,J\Pi^{(0)}) \qquad (5.2)$$

So,

$$\mathrm{vec}(\hat{\Pi}^{(c)}-\Pi^{(c)}) \simeq \{(1+\delta_0)I_9 - \delta_0(I_3 \times \Pi^{(0)}J) - \delta_0(J\Pi^{(0)} \cdot I_3)\}\,\mathrm{vec}(\hat{\Pi}^{(0)}-\Pi^{(0)}) \qquad (5.3)$$

If the data were obtained under multinomial sampling, then

$$\hat{V}(\mathrm{vec}\,\hat{\Pi}^{(c)}) \simeq n^{-1}\,\hat{A}\{\mathrm{diag}(\mathrm{vec}\,\hat{\Pi}^{(0)}) - (\mathrm{vec}\,\hat{\Pi}^{(0)})(\mathrm{vec}\,\hat{\Pi}^{(0)})'\}\,\hat{A} \qquad (5.4)$$

where n is the sample size and the matrix $A_{9\times 9}$ is given by

$$A=(1+\delta_0)\,I_9 - \delta_0(I_3 \times \Pi^{(0)}J) - \pi_0(J\Pi^{(0)} \times I_3). \qquad (5.5)$$

5.2 δ unknown

In this case, $\hat{\delta}$ is used. If $\mathrm{Var}(\hat{\delta})$ is negligible, then expression (5.4) remains valid for the covariance matrix of $\mathrm{vec}\hat{\Pi}^{(c)}$ where δ_0 is replaced by $\hat{\delta}$. This situation may not be unreasonable because δ will usually be estimated by aggregating the data over several months/years as it is expected to be stationary over a long term. In general, we can account for the variability in $\hat{\delta}$ by modifying (5.4) as follows

$$V\{\mathrm{vec}\,\hat{\Pi}^{(c)}\} = E[V\{\mathrm{vec}\,\hat{\Pi}^{(c)} \mid \hat{\delta}=\delta_0\}] + V[E\{\mathrm{vec}\,\hat{\Pi}^{(c)} \mid \hat{\delta}=\delta_0\}]. \qquad (5.6)$$

The above expression can be easily evaluated if $\hat{\delta}$ and $\hat{\Pi}^{(0)}$ are uncorrelated. This would be so if the corresponding data sets are nonoverlapping.

It may be remarked that the standard error of $\hat{\Pi}^{(c)}$ so obtained under the multinomial assumption would be expected to be biased downward because of the clustered nature of the commonly used sample designs.

6. EXAMPLES

We consider two examples, one for Canadian LFS data, the other for United States CPS (Current Population Survey) data taken from Chua and Fuller (1987). We illustrate the adjustments under the ε- contamination model for two choices of δ or $\varepsilon(1-\varepsilon)^{-1}$ namely, .03 and .04. For LFS data, these adjustments are compared with those obtained under the Abowd and Zellner (1985) method. For CPS data, we compare the proposed method with Chua and Fuller (1987) method where we have assumed tentatively that the rate of contamination (i.e. the mixing parameter) ε for the Canadian and U.S. surveys is the same.

6.1 LFS data (October-November 1985).

The 3×3 table of weighted observed flow proportions obtained from matching the common respondents in the months of October and November, 1985, is given by

Table 6.1 (The matrix $\hat{\Pi}^{(0)}$ for LFS)

t-1 \ t	E	U	N	Row Total	
E	.0112	.0123	.5932	.5932	= $\hat{\Pi}^{(0)}_{1+}$
U	.0101	.0400	.0099	.06	= $\hat{\Pi}^{(0)}_{2+}$
N	.0099	.0109	.3260	.3468	= $\hat{\Pi}^{(0)}_{3+}$
Column Total	.5897	.0621	.3482	1	

In Abowd and Zellner's method, we need to estimate matrices of classification error rates B(t-1) and B(t) from interview-reinterview data. Using aggregate data over the period of 12 months (May '85 - April '86) and assuming B(t) and B(t-1) to be equal, we obtain

$$\hat{B}(t-1) = \hat{B}(t) = \begin{bmatrix} .9910 & .0178 & .0043 \\ .0017 & .9056 & .0061 \\ .0073 & .0766 & .9896 \end{bmatrix}, \qquad (6.1)$$

and

$$\hat{B}^{-1} = \begin{bmatrix} 1.0092 & -.0195 & -.0043 \\ -.0018 & 1.1048 & -.0068 \\ -.0073 & -.0854 & 1.0111 \end{bmatrix}. \qquad (6.2)$$

Now, $\hat{\Pi}^{(1)}$ can be obtained by using (2.4). The adjusted values are shown in Table 6.2.

Under the proposed model, we need to compute the matrix $\hat{\Sigma}$ whose $(i,j)^{th}$ element is $\hat{\Pi}^{(0)}_{ij} - \hat{\Pi}^{(0)}_{i+}\hat{\Pi}^{(0)}_{+j}$, and then use the formula (3.5) to compute the adjusted flows. The matrix $\hat{\Sigma}$ is given by

$$\hat{\Sigma} = \begin{bmatrix} .2198 & -.0256 & -.1942 \\ -.0252 & .0362 & -.0110 \\ -.1946 & -.0106 & .2052 \end{bmatrix} \qquad (6.3)$$

The adjusted flows under the ε-contamination model for δ = .03, .04 along with those for Abowd-Zellner method are given in Table 6.2.

Table 6.2 Adjusted flow proportions (LFS) $\hat{\Pi}^{(c)}$

t-1 \ t	E			U			N		
	A-Z	ε-CON		A-Z	ε-CON		A-Z	ε-CON	
		δ = .04	δ = .03		δ = .04	δ = .03		δ = .04	δ = .03
E	.5796	(.5697) .5785	.5763	.0104	(.0112) .0102	.0104	.0058	(.0123) .0045	.0065
U	.0092	(.0101) .0091	.0093	.0486	(.0400) .0414	.0411	.0049	(0.0099) .0095	.0096
N	.0034	(.0099) .0021	.0041	.0061	(.0109) .0105	.0106	.3316	(.3260) .3342	.3321

Note: Proportions in parentheses indicate observed or unadjusted flow proportions.

6.2 CPS Data (January-February 1979)

This example is taken from Chua and Fuller (1987). Table 6.3 represents observed flow proportions based on 3,198 interviews. The number 3,198 indicates the number of individuals who were also reinterviewed in February 1979.

Table 6.3 (The matrix $\hat{\Pi}^{(0)}$ for CPS)

t-1 \ t	E	U	N	Row Total
E	.5316	.0081	.0188	.5585
U	.0094	.0147	.0066	.0307
N	.0172	.0097	.3839	.4108
Column Total	.5582	.0325	.4093	1

In Chua and Fuller's method, the matrices B(t-1) and B(t) are estimated using a certain model under the constraints that marginals are unbiased and were found to be

$$\hat{B}(t-1) = \begin{bmatrix} .9837 & .0552 & .0186 \\ .0030 & .8415 & .0077 \\ .0133 & .1033 & .9742 \end{bmatrix} \tag{6.4}$$

and

$$B(\hat{t}) = \begin{bmatrix} .9835 & .0550 & .0181 \\ .0032 & .8422 & .0082 \\ .0133 & .1028 & .9737 \end{bmatrix} \tag{6.5}$$

From formula (2.4), the adjusted estimates $\hat{\Pi}^{(c)}$ can be obtained as in Abowd and Zellner. These values are shown in Table 6.4.

We next consider the adjustment under the proposed model for the same example using the same values of δ as before for LFS data. First the matrix $\hat{\Sigma}$ is computed.

$$\hat{\Sigma} = \begin{bmatrix} .2198 & -.0101 & -.2097 \\ -.0077 & .0137 & -.006 \\ -.2121 & -.0036 & .2157 \end{bmatrix} \tag{6.6}$$

The adjusted flows $\hat{\Pi}^{(c)}$ are also shown in Table 6.4.

Table 6.4 Adjusted flow proportions (CPS) $\hat{\Pi}^{(c)}$

t-1 \ t	E			U			N		
	C-F	ε-CON		C-F	ε-CON		C-F	ε-CON	
		δ = .04	δ = .03		δ = .04	δ = .03		δ = .04	δ = .03
E	.5484	(.5316) .5404	.5382	.0063	(.0081) .0077	.0078	.0038	(.0188) .0103	.0125
U	.0081	(.0094) .0091	.0092	.0206	(.0147) .0152	.0151	.0020	(.0066) .0064	.0064
N	.0017	(.0172) .0087	.0108	.0056	(.0097) .0096	.0096	.4035	(.3839) .3925	.3904

Note: The marginals of the above table match those for the observed flow proportions.

7. CONCLUDING REMARKS

It was seen that the proposed ε-contamination model could provide a quick and simple procedure for correcting classification error bias if a suitable estimate of ε is available apriori. It may be noted that ε is expected to be stationary over a long term and so different estimates of ε for several consecutive periods would not be required. In the examples considered it was found that the ε-contamination model (with the working values of δ or $\varepsilon(1-\varepsilon)^{-1}$ as .03 and .04) generally tends to correct less than those under alternative methods. Clearly, it would be important to evaluate the performance of the proposed method for various cases after a more precise estimate of ε is determined.

ACKNOWLEDGEMENTS

The first author's research was supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University as an Adjunct Research Professor. Thanks are due to John Armstrong for helpful discussions and to Johane Dufour for assisting in computations for example 6.1. We would also like to thank Judy Clarke and Dula Edirisinghe for their efficient manuscript processing.

REFERENCES

Abowd, J.M., and Zellner, A. (1985). Estimating gross labor flows, *Journal of Business and Economic Statistics, 3, 254-283*

Chua, T.C., and Fuller, W.A. (1987). A model for multinomial response error applied to labor flows, *Journal of the American Statistical Association*, 82, 46-51.

Fay, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

Fuller, W.A. (1987). *Measurement Error Models, New York, John Wiley.*

Fuller, W.A., and Chua, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the confrence on Gross Flows in Labor Force Statistics*, Washington, D.C., 65-80.

Gentleman, J.F. (1988). Assessing bias due to classification error in labour force survey data. *American Statistical Association, Proc. Sec. Surv. Res. Meth.*

Lemaître, G.E. (1988). The measurement and analysis of gross flows. Working Paper No. SSMD-88-1E. *Statistics Canada.*

Little, R.J.A. (1985). Nonresponse adjustment in longitudinal surveys: models for categorical data. Bulletin of the International Statistical Institute, 51, 15.1.1-17.

Poterba, J.M., and Summers, L.H. (1986). Report errors and labour market dynamics. *Econometrica*, 54, 1319-38.

Stasny, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.

Stasny, E.A., and Fienberg, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse: *Proceedings of the conference on Gross Flows in Labor Force Statistics*, Washington, D.C., 25-43.