Estimation and Correction for Purchase Screening Errors in Expenditure Data: A Markov Latent Class Modeling Approach

Paul P. Biemer
Research Triangle Institute
P.O. Box 12194
Raleigh, NC 27709-2194 USA
ppb@rti.org

Clyde Tucker
Bureau of Labor Statistics
2 Massachusetts Avenue, N.E.
Washington, D.C. 20212 USA
tucker_c@bls.gov

Introduction

In the Consumer Expenditure Interview Survey (CEIS), respondents are asked a number of details about the expenditures they incurred during the past three months. These detailed questions are asked for a consumer item only if the respondent responds positively to a screening question that essentially determines if any purchases of the item have been made over the past three months. If the response to the screening item is negative, the detailed questions are skipped.

Due to the large number of items in the survey and the extensive information requested for some purchases, interviews can last between one and two hours or even longer and the burden on respondents is considerable. Although the potential for underreporting of expenditures as a result of false negative responses to the screening questions is apparent, there have been few studies to formally investigate the error in CEIS screening questions. One reason for the lack of information on this important source of error is the difficulty of evaluating the error.

Traditional methods for evaluating survey error require the use of a gold standard measurement that can serve as the truth for purposes of estimating reporting error. In the CEIS such gold standard measurements are often very difficult to obtain or are unavailable. For example to verify that the respondent had no automobile maintenance expense, a reinterview audit survey could be conducted where the respondent is asked to locate all receipts for automobile expenditures over the last three months. The interviewer could then check for the existence of auto maintenance receipts in this within household "census" of receipts. However, even then, the absence of automobile maintenance documentation is no assurance that expenditures of this type were not incurred over the reference period. Thus, gold standard studies of expenditure underreporting are not feasible.

Latent class analysis (LCA) is a fairly new methodology for evaluating the error in survey reports without the use of gold standard measures. In lieu of measurements which are accurate, the method assumes an error model for the available measurements and uses maximum likelihood estimation techniques to estimate the parameters of the error model. Thus, the validity of the LCA estimates hinges on the ability of the model to accurately represent the error-generating process. LCA usually requires at least two but preferably three replicate measurements of the same item (at the same point in time) as a condition of estimability of the error parameters.

For panel surveys such as the CEIS, a related statistical method referred to as Markov latent class analysis (MLCA) is available which essentially relaxes the requirement that the replicate measurements pertain to the same point. Thus, this method of analysis is feasible for analyzing repeated measurements of the same units at different time points available in panel surveys. MLCA requires a minimum of three measurements of the same units as would be the case for a panel survey where units are interviewed on three occasions. The MLCA model then specifies parameters for both the period to period changes in the status of the item as well as the measurement error associated with measuring those changes. In the automobile maintenance item example, the MLCA model would specify parameters associated with both the month to month change in the true status of the expenditure (i.e., transitions from no expense to some expense and vice versus) as well as the error in reporting this status in the CE.

Biemer (2000) applied the MLCA methodology to the CEIS in order to determine whether useful information on the magnitudes and correlates of screening question reporting error can be extracted directly from the CEIS panel data. His investigation illustrated how the MLCA approach can be applied to the CEIS for conducting exploratory data analysis of the correlates of screening error.

Biemer applied MLCA to three consecutive quarters of the CEIS in order to address a number of issues related to error in the screening questions for 19 selected consumer items. The results of his research suggests that MLCA analysis of CEIS data is plausible and the results of his multivariate analysis of 19 consumer items generally agreed with a priori expectations based upon a measurement error theory perspective.

However, Biemer identified a number of limitations of his analysis that need to be addressed in subsequent research. First, his study did not fully addressed the validity of the MLCA estimates. Although the MLCA estimates seemed plausible in many instances, other results seem puzzling and counterintuitive. In addition, because his investigation was preliminary and primarily pedagogical, the number of explanatory variables he considered in his models was small. He identified a number of other variables that are potentially related to screener reporting error that should also be investigated.

The purpose of the present paper is to continue the analysis of the CEIS along the lines investigated in Biemer (2000) and to address some of the areas of future research suggested in his paper. The present paper extends that work in the following ways:

- data from three years of the CEIS are used rather than only one year to control for seasonal effects,
- a more extensive list of variables are analyzed, and
- the validity of the first order Markov assumption is examined.

In the next section, we describe the data sets that will be analyzed in the study and the essential components of the MLCA methodology. Because of space limitations, we are not able to provide any of the results from our analyses. However, a full report of this investigation can be obtained by requesting it from one of the authors.

Data Sets and Models

The data used in this study consists of all interviews collected in three years of the CEIS: 1996, 1997, and 1998. Each survey was designed to collect information on data on up to 95 percent of total household expenditures. We define a consumer unit (CU) as members of a household who are related and/or pool their incomes to make to make joint expenditure decisions. In the CEIS, CU's are interviewed once every three months for five consecutive quarters to obtain the expenditures for 12 consecutive months. The initial interview for a CU is used as a bounding interview and these data are not used in the estimation. The survey is designed to collect data on major items of expense which respondents can be expected to recall for three months or longer. New panels are initiated every month of the year so that each month, 20 percent of the CU's are being interviewed for the first time.

For the years of the CEIS we are analyzing, approximately 7,000 sample units were contacted for an interview each quarter. Allowing for bounding interviews and nonresponse (including vacancies), the number of participating sample units per quarter was targeted at approximately 5,000.

As we will see in the discussion of the MLCA approach, a minimum of three consecutive panel survey observations on all the CU's in the analysis is required for the identifiability of the MLC models. However, a minimum of four observations is required for evaluating the first order Markov assumption which is a key assumption in the three observation per CU model. Since we are interested in both analyses, our analysis only considers CU's that were interviewed four consecutive times beginning in the first quarter of 1996 and ending in the last quarter of 1998. Thus, all CU's not completing all four interviews where deleted from the analysis.

An important objective of the present study is to assess measurement errors arising from the CEIS data collection operations. To this end, only unweighted data are used in the analysis since weighting the data could distort data on the error processes operating during the data collection operations. One drawback of using unweighted data is that inferences regarding the overall quality of the published CEIS estimates of expenditures cannot be made; however, this is not an important objective in the study. For purposes of this analysis, the sample will be treated essentially as a simple random sample from a superpopulation which is the CEIS data series for the current survey design.

MLC models assume that all the variables in the analysis are classification variables. For example, in our analysis we will consider the screening questions in the CEIS where the outcome variable is a dichotomous response taking the value 1 if the CU reports a purchase for a particular consumer item for the month and 2 if not.

Let the CE target population be divided into L groups or domains and let the variable G be the indicator for group membership. For example, G may be related to the administration of the survey - such as interview length, use of records, number of times previously interviewed, etc. - or may describe CU characteristics such as size, income, age of CU members, etc. Let, $G_i = 1$ if the ith population member is in group 1, $G_i = 2$ for group 2 and so on.

In the preliminary stages of our analysis, we considered models for describing the error in a dichotomous variable, say D_m, defined for a single consumer item (such as pet supplies) and a particular month, m, of the CE where $D_m = 1$ if the CU purchased the item during the month and $D_m = 2$, otherwise. Thus, for the nine months of data collection, we would define D_m, $m = 1, \ldots, 9$ for the expenditure pattern for the item over the entire nine-month period. However, there were a number of difficulties with this modeling approach. First, and most important, the MLC models provided a very poor fit to these data due primary, we believe, to the failure of the first order Markov assumption (described below) to hold. Second, the models were quite complex with many 0-cells that caused convergence problems in the EM algorithm used for maximum likelihood estimation. In addition, the model fitting process was quite tedious and time consuming since a single model run could require one hour on a Pentium III 450. Therefore, this modeling approach was abandoned in favor of the following simpler modeling approach.

Rather than specifying a variable for a monthly purchase, we define a summary variable for the frequency of purchases of the item for the months of a three-month interview reference period. Such a model would be much more likely to satisfy the Markov assumption and would run much more efficiently as a result of the reduction in the dimensionality of the problem.

Let the subscript combination (g, i) denote the i-th CU in group $G = g$ for $g = 1, \ldots, L$ and $i = 1, \ldots, n_g$. For a particular three-month interview period under investigation, we define the manifest variable A_{gi} as follows.

$$A_{gi} = \begin{cases} 1 & \text{if CU}(g,i) \text{ reports the item was purchased in all three months of the period} \\ 2 & \text{if CU}(g,i) \text{ reports the item was not purchased during the period} \\ 3 & \text{if CU}(g,i) \text{ reports the item was purchased in one or two months of the period} \end{cases}$$

For analyzing the observations for the same CU for three consecutive interviews, we define B_{gj}, and C_{gk} in analogy to $A_{gj.}$ Likewise, for analyzing the data from four consecutive interviews, we define B_{gj}, C_{gk}, and D_{gi} analogously, for the second, third, and fourth periods, respectively.

Associated with each of the three or four observed variables is a latent variable for the *true* quarterly purchase status of the CU for each time period. For periods 1, 2, and 3, let X_{gj},Y_{gi}, and Z_{gi} denote trichotomous variables with categories defined analogously to A_{gi}, B_{gi}, and C_{gi}, respectively, except that they represent the true rather than observe statuses of the CU's. Likewise, for four time periods we define the latent variables W_{gi}, X_{gj},Y_{gi}, and Z_{gi} corresponding to A_{gi}, B_{gi}, C_{gi}, and D_{gi}, respectively. For notational convenience, we will drop the subscripts (g,i), but retain the relationship of the unsubscripted variable to an individual unit within a group. Further, the term "true purchaser" will be used to describe CU's who purchase the item in all three months of a period, "true non-purchaser" for CU's not purchasing the item in any quarter, and "true mixed consumer" for CU's who purchase in the item in some but not all quarters. To simplify the description of our general approach, we will consider the analysis of three consecutive interviews in some detail.. Extensions of these ideas to the case of four consecutive interviews is fairly straightforward, however.

Let $\pi_{x,y,z|g}$ denote Pr(X=x, Y=y, Z=z | G=g), let $\pi_{y|g,x}$ denote Pr(Y = y | X = x, G = g) and let $\pi_{z|g,y,x}$ denote Pr(Z = z | Y = y, X = x, G = g). Then, the probability that an individual in group g is has purchase status x in time period 1, y in time period 2, and z in time period 3 is

$$\pi_{x,y,z|g} = \pi_{x|g}\pi_{y|g,x}\pi_{z|g,x,y}$$

Finally, under the first order Markov assumption (which is required for model identifiability; see Van de Pol and de Leeuw, 1986), we assume

$$\pi_{z|g,x,y} = \pi_{z|g,y},$$

i.e., at period 3, the true status of an individual does not depend on the period 1 status once the period 2 status is known. An alternate interpretation is that the period 3 purchase status given the period 2 status does not depend upon the period 1 to period 2 transition.

One can conceive of some situations where the Markov assumption may not hold for the CE interview survey. For example, for large and costly purchases such as automobiles, CU's who are non-purchasers in period 2 may be much more likely to be non-purchasers in period 3 if they were mixed consumers in period 1 than if they were non-purchasers in period 1. This is because buying an automobile occurs infrequently and CU's purchasing a car in the last six months are less likely to purchase another car in the next three months than CU's who have not purchased a car in the last six months. For other items, such as cable TV subscription, the Markov assumption may hold quite well since the purchases statuses are more stable over a three month period Thus, the fit of the MLCM's and the validity of the resulting estimates are expected to vary considerably by consumer item.

A method for assessing the validity of the first order Markov assumption for panel data was suggested by Van de Pol and de Leeuw (1986) and is based upon four waves of panel data. For four waves of data, we write

$$\pi_{wxyz|g} = \pi_{w|g}\pi_{x|gw}\pi_{y|gwx}\pi_{z|gwxy}$$

and assume the second order Markov property holds for the last term of this equation, viz.,

$$\pi_{z|gwxy} = \pi_{z|gxy}$$

in order to obtain an identifiable model. Thus, with four waves of panel data, it is possible to fit models which relax the first order Markov assumption.

Using an extension of the notation established above, we denote the response probabilities in each of these classifications as follows:

$$\pi_{a|g,x} = \Pr(A = a | X = x)$$
$$\pi_{b|g,y} = \Pr(B = b | Y = y)$$
$$\pi_{c|g,z} = \Pr(C = c | Z = z)$$

Thus, $\pi_{a=1|g,x=2}$ is the probability that the CE classifies a person in group g as a purchaser (A = 1) when the true status is non-purchaser (X = 2). Likewise, $\pi_{a=2|g,x=2}$ is the probability that the CE correctly classifies a person in group g as a non-purchaser.

Finally, we assume that

$$\Pr(A=a,B=b,C=c \mid G=g,X=x,Y=y,Z=z)$$
$$= \Pr(A=a \mid G=g,X=x)\Pr(B=b \mid G=g,Y=y)\Pr(C=c \mid G=g,Z=z)$$

and write

$$\pi_{a,b,c|x,y,z} = \pi_{a|gx}\pi_{b|gy}\pi_{c|gz}.$$

In the analysis that follows, less emphasis will be placed on the MLCA estimates for a particular item. Rather, we will be looking for trends in error rates across all the items in the study. Further, our analysis is exploratory in that we will be using MLCA to generate hypotheses regarding the causes of error that can be followed up and tested in other settings. For example, a finding that some consumer items are less subject to under-reporting than others could suggest a study that might be conducted with a small number of subjects in a laboratory setting to determine if the result can be verified and, if so, the reasons for the differential data quality. Thus, the manner in which MLCA will be used in the following is somewhat robust to failures of the model assumptions to hold.

With these assumptions, we can write the probability of classifying a CE sample member in cell (*g,a,b,c*) of the *GABC* table as follows:

$$\pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{a|g,x} \pi_{y|x,g} \pi_{b|y,g} \pi_{z|g,y} \pi_{c|g,z}$$

Under multinomial sampling, the likelihood function for the *GABC* table is

$$\text{Likelihood} = \Pr(GABC) = \text{constant} \times \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{g,a,b,c}}.$$

All the models we will consider are hierarchical models for which presence of an interaction term implies presence of all lower order interactions and main effect terms containing the sample letters. As an example, the hierarchical log-linear model containing the term *AXG* also contains the terms *A*, *X*, *G*, *AX*, *AG*, and *GX*. However, in specifying logistic log-linear models for conditional probabilities, all terms in the model must contain the dependent variable; thus, the model for $\beta_{a|xg}$ containing *AXG* contains the "main effect" terms *A*, *AX*, *AG* as well as the "first order interaction" term *AXG* only. This model is specified in shorthand notation as {*AXG*}.

Under the assumptions made previously, the model parameters are estimable using maximum likelihood estimation methods. Van de Pol and de Leeuw (1986) provides the formula for applying the E-M algorithm for estimating the parameters of this model and conditions for their estimability. These methods are implemented in the REM software which will be applied to the CE data set.

REFERENCES

Biemer, P. and Bushery, J. (to be published). "On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data" in *Survey Methodology*

Goodman, L. (1974). "Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part I: A Modified Latent Structure Approach," *American Journal of Sociology*, 79, 1179-1259.

Meyers, B. D. (1988). "Classification-Error Models and Labor-Market Dynamics," *Journal of Business & Economic Statistics*, Vol. 6, No. 3, pp. 385-390.

Poulsen, C.S. (1982). *Latent Structure Analysis with Choice Modeling Applications*, doctoral dissertation, Wharton School, University of Pennsylvania.

Singh, A.C. and Rao, J.N.K. (1995). "On the Adjustment of Gross Flow Estimates for Classification Error with Application to Data from the Canadian Labour Force Survey," Journal of the American Statistical Association, Vol. 90, No. 430, pp. 478-488.

Van de Pol, F. and de Leeuw, J. (1986). "A Latent Markov Model to Correct for Measurement Error," *Sociological Methods & Research*, Vol. 15, Nos. 1-2, pp 118-141.

Van de Pol, F. and Langeheine, R. (1997). "Separating Change and Measurement Error in Panel Surveys with an Application to Labor Market Data," in L. Lyberg, et al (eds.) *Survey Measurement and Process Quality*, John Wiley & Sons, NY.

Vermunt, J. (1997). *REM: A General Program for the Analysis of Categorical Data*, Tilburg, University.

Wiggins, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Elsevier S.P.C., Amsterdam.

RESUME

Previous analysis of the data from the U.S. Consumer Expenditure Interview Survey (CEIS) indicates that expenditure estimates are subject to substantial underreporting. In addition, missing data are often present for screener questions used to determine whether expenditures have occurred for particular commodities in the specified time period. One method for estimating and correcting for both types of errors is Markov Latent Class Analysis (MLCA). This method provides estimates of the probabilities of incorrect reporting of expenditures in different commodity categories. Using this procedure, imputation and/or the correction of responses to the screener questions may also be accomplished. This paper will describe the MLCA approach and present empirical results demonstrating the performance of the method for the CEIS.