Estimating the Level of Underreporting of Expenditures among Expenditure Reporters: A Micro-Level Latent Class Analysis

Clyde Tucker, Bureau of Labor Statistics
Paul Biemer, Research Triangle Institute
Brian Meekins, Bureau of Labor Statistics
Jennifer Shields, Bureau of Labor Statistics

1. Introduction

This paper makes estimates of the level of underreporting of consumer expenditures. The paper examines reporting in particular commodity categories and attempts only to make estimates of underreporting among those that report at least one expenditure in the category. The measure of the level of underreporting in a category by a particular responding unit is based on latent class analysis using demographic characteristics as well as characteristics of the respondent's reporting behavior. It is assumed that the level of underreporting is similar within a particular subpopulation defined by these characteristics.

Latent class analysis, a theory for detecting unobserved variables, was developed by Paul Lazarsfeld (1950). According to Lazarsfeld, an unobservable variable (such as underreporting) could be constructed by taking into account the interrelationships between observed variables. The mathematics underlying this theory were extended by Lazarsfeld and Henry (1968) and Goodman (1974).

The paper begins with an introduction to the Consumer Expenditure Interview Survey (CEIS) sponsored by the Bureau of Labor Statistics (BLS) and conducted by the Census Bureau. Previous related work by the authors in this area is summarized and the design of this particular study is outlined. The following section presents the analytical results, and a final section is devoted to the discussion of the results and the consideration of additional avenues for research.

2. CEIS

In this section, we describe the data sets that will be analyzed in the study and some key operational definitions. The data used in this study consists of interviews collected in six years of the CEIS: 1996 through 2002. Each survey was designed to collect information 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 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 quarter of the year so that each quarter, 20 percent of the CU's are being interviewed for the first time. Only CU's completing and reporting an expense in wave 2 are used in this analysis, for a total of 14,877 respondents.

3. Previous Work

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.

Previous work by the authors used MLCA to make aggregate estimates of underreporting in a category only by respondents reporting no expenditures in that category. 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. Biemer and Tucker (2001) extended the earlier analysis using data from four consecutive quarters of the CEIS by considering CU's that were interviewed four consecutive times beginning in the first quarter of 1996 and ending in the last quarter of 1998. This allowed the authors to consider a wider-range of models including second-order Markov models. First order Markov models assume that

a purchase or non-purchase at quarter *q* is affected only by quarter *q*-1 purchases or non-purchases. A second order Markov model assumes that both quarters *q*-1 and *q*-2 affect purchasing behavior at quarter *q*. Their analysis provided evidence of second-order Markov effects and recommended that second-order terms be included in the models.

In Tucker, Biemer, and Vermunt (2002), model estimates with both unweighted and weighted data were compared. The results indicated that few differences were found between the two; therefore, given the ease of use, unweighted data were used in these analyses. A thorough examination of all explanatory variables considered in the previous studies was undertaken, and a reduced set of the most powerful ones was identified. A new diagnostic technique was developed and used to evaluate the validity of the models. Finally, methodology for estimating missing expenditures was outlined.

4. New Design

Unlike previous work, these authors chose a micro-level approach incorporating measures specific to a given interview. In essence, a latent variable that adequately accounted for the shared varaince among a set of observed response error indicators was created. The observed varaibles were based on information collected from each CU during the interview. The latent variable was believed to be a better measure of underreporting than any of the observed varaibles taken individually. Each CU then was assigned to a particular class of the latent variable representing its hypothesized level of expenditure underreporting based on the CU's values on the observed variables. See Tucker (1992) for an earlier empirical example.

We used only second wave data[1]. We examined reporters of expenditures and ignored nonreporters. We wished to develop a model separate from covariates with only indicators of the quality of response. We began with the simplest identifiable model composed of three indicators (each with three classes) and a latent variable with three classes. From this point we ran all possible combinations of three indicators for a three class latent variable. The analysis was further extended by examining restricted models based on the hypothetical relationship of some of the indicators with the latent variable, thus ordering the latent classes in what we believed to be an interpretable manner. These "restricted" models were compared to the unrestricted models to aid in interpretability and choices of model fit. Some of the indicators are dichotomous. These were entered into the best three variable models along with other combinations to create four indicator models. At this point we also allowed the latent variable to have four classes. Our goal was to develop a latent variable (preferably ordered) that indicated the quality of responses, such that poor reporters could be easily identified. The following indicators were explored:

1. Number of contacts the interviewer made to complete the interview
2. The ratio of respondents to total number of household members
3. The ratio of household members earning an income to the total number of household members
4. Whether the household was missing a response on the income question
5. The type and frequency of records used. This variable indicates whether a respondent used bills or their checkbook to answer questions, and how often they did so.
6. The percent of data requiring imputation or allocation.
7. The length of the interview
8. A ratio of expenditures reported for the last month of the 3 month reporting period to the total expenditures for the 3 months
9. And a combination of type of record used and the length of the interview.

5. Model Selection

Models were estimated using ŒEM. Model selection is based on a number of objective and subjective measures. We primarily used the Bayesian Information Criteria (BIC), the L^2 test statistic, and the dissimilarity index. However, for each model we examined the conditional probabilities of the latent variable given each values of each indicator. In this way we assessed the relative influence of each indicator and the degree to which an indicator effectively differentiated the respondents with respect to the

[1] Wave 2 data are used because wave 1 is a bounding interview.

classes of the latent variable (See *Table 1 for and example*).

Table 1: Probability of Indicator A Given Latent Variable X

A	*X*	*P(A|X)*
1	*1*	.0794
2	*1*	.9206
1	*2*	.8119
2	*2*	.1881
1	*3*	.1217
2	*3*	.8783

In addition to model fit, we used the ordered models as a guide. If the indicators of the unrestricted model were aligned to the latent classes in a similar manner as the restricted model, then the unrestricted model showed promise, at least in its interpretability. In cases where they did not and the objective fit statistics were not good, the models were immediately discarded. We were looking for a final model that was able to line up indicators with the latent classes in a logical way, while not sacrificing model fit.

Boundary values were also of considerable influence in selecting models, especially in the case of restricted models. The use of ordinal or inequality restrictions can lead to a number of boundary issues, especially if the models fit poorly. These models were discarded if it was obvious that the ordinal constraints were the source of the boundary problem, regardless of fit statistics (although they were usually quite poor). If this was not obvious, the models were re-run with a priori specified starting points on the conditional probabilities so that they were "pushed-off" of the boundary.

Table 2: The Probability of Latent Variable X Given Indicators A, B, and C

			X		
A3	*B3*	*C3*	*1*	*2*	*3*
1	*1*	*1*	0.7638	0.0013	0.2349
1	*1*	*2*	0.6370	0.1903	0.1727
1	*1*	*3*	0.1089	0.8232	0.0679
1	*2*	*1*	0.7020	0.0021	0.2959
1	*2*	*2*	0.5297	0.2735	0.1967
1	*2*	*3*	0.0671	0.8757	0.0573
1	*3*	*1*	0.5617	0.0014	0.4369
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
3	*3*	*3*	0.0233	0.6542	0.3225

Table 2 shows a portion of one of the tables for the purposes of example used to examine differentiation between latent classes. The first three columns show the values of the indicators. The final three columns show the probability of a respondent falling into that group of indicators given each value of the latent class, or X. For example, approximately 76% of respondents with a value of "1" on each of the three indicators are classified as a "1" on the latent variable.

If values are higher for one latent class than the other two, we can say that the indicators are able to differentiate between them, if the values are similar, then we would not be able to draw this conclusion.

6. Best Model

Using these methods a "best" model was selected. The model uses four indicators to define a three class latent variable. This model is not ordinal, and has adequate fit statistics:
BIC = -86.01, Dissimilarity index = 0.02, L2 = 233.96, p=0.00.
The indicators are:

1. Number of contacts
2. Missing income question
3. Type and frequency of records used
4. Length of interview

Each of these variables is thought to be related to the amount of effort expended by the respondent. After this final model was identified, we again used the probability of being in each latent class given a combination of indicators to assign each combination to a latent class, using the plurality rule. We then returned to the CE data and assigned each respondent to that latent class which corresponded to their characteristics. Expenditure means were then found for each "latent" class.

Latent classes aligned with expenditure means as expected. Those with lower expenditure means had higher levels of underreporting. For example, those in the low underreporting class had a total expenditure mean of 10,625, while those in the high underreporting class had a mean of 6,948 (See *Table 3)*. This may suggest that those in the high underreporting class failed to report more of their expenditures than did those in the other two classes.

Table 3: Mean Expenditure by Latent Class

Latent Class	*N.*	*Expenditure Mean*	*St. Dev*
Low Underreporting	11,507	10,625.18	112,8021
Moderate Underreporting	18,963	8,137.10	845,866
High Underreporting	12,525	6,948.97	904,372

After examining how each indicator related to the latent classes, we found that those in the latent class with the least underreporting:

1. Were not more or less likely to have a certain number of contacts,
2. Were more likely to use more records frequently
3. Were more likely to have longer interviews and
4. Were more likely not to have missing data on the income question

Expenditure means for respondents assigned to each latent class confirmed this finding (See *Table 4)*.

Table 4: Indicator Mean by Latent Class

Latent Class	*Indicator*	*Indicator Mean**
Low Underreporting	Contact	2.025
	Interview Length	1.069
	Type of Records	2.11
	Missing Income	2.80
Moderate Underreporting	Contact	1.91
	Interview Length	1.00
	Type of Records	1.53
	Missing Income	1.90
High Underreporting	Contact	2.08
	Interview Length	1.64
	Type of Records	1.25
	Missing Income	1.43

*Coding for indicators provided in Appendix A.

7. Discussion

This paper is ony a preliminary step toward a final measure of underreporting in the CEIS. A more thorough analysis of micro-level models will be completed, and a final model will be selected. Demographic analyses using the latent variable as a dependent variable will be examined to identify the charactersitics of members of each underreporting classs. Besides looking at mean total expenditure, the means for each latent class by commodity category will be compared to evaluate the discriminatory power of the latent variable for different types of expenditures.

At the final stage, the results from this micro-level analysis will be merged with the earlier aggregate level analysis to evaluate the overall underrepoting of expenditures in the CEIS. At that point, the contributions to underreporting from both nonreporters and incomplete reporters will be estimated for each commodity category and compared to external estimates of underreporting.

References

Bassi, Franesca, Jacques A. Hagenaars, Marcel A. Croon and Jeroen Vermunt. (2000). "Estimating True Changes When Categorical Panel Data Are Affected by Uncorrelated and Correlated Classification Errors: An Application to Unemploymnet Data." *Sociological Methods & Research* 29: 230-268.

Biemer, P.P. (2000). "An Application of Markov Latent Class Analysis for Evaluating the Screening Questions in the CE Interview Survey," Technical Report Submitted to BLS, June 6, 2000.

Biemer, P. P. and Tucker, C. (2001). "Estimation and Correction for Purchase Screening Errors in Expenditure Data: A Markov Latent Class Modeling Approach," *Proceedings of the International Statistics Institute*, Seoul, Korea.

Tucker, C. (1992). "The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Diary Survey." *Journal of Official Statistics.* 8: 41-61.

Tucker, C., Biemer, P., and Vermunt, J. (2002). "Estimation Error in Reports of Consumer Expenditures," Proceedings of the ASA, Survey Research Methods Section, New York, NY.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," Biometrika, 61, 215-231.

Langeheine, R. and Van der Pol, F. (2002). "Latent Markov Chains," in Hagenaars, J. and McCutcheon, A. (eds.) *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, UK

Lazarsfeld, P.F. (1950). "The Logical and Mathematical Foundation of Latent Structure Analysis." In S. Stauffer, E.A. Suchman, P.F. Lazarsfeld, S.A. Starr, and J. Clausen, Studies on Social Psychology in World War II, Vol. 4, Measurement and Prediction. Princeton: Princeton University Press.

Lazarsfeld, P.F. and Henry, N.W. (1968). Latent Structure Analysis. Boston: Houghton-Mifflin.

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). lEM: A General Program for the Analysis of Categorical Data, Tilburg, University

Appendix A

Coding for indicator variables used in final model:

Amount of contact

1	"0-2 contacts"
2	"3-5 contacts and missing"
3	"6 + contacts"

Interview Length

1	" < 45 minutes"
2	"45 <= minutes < 90"
3	">=90"

Type of records

1	"Almost never or never use of records"
2	"Single type of record and/or mostly or occasionally used records"
3	"Multiple types of records, almost always or always"

Missing on income

1	"Income not missing"
2	"Income missing"