Properties of Alternative Sample Design and Estimation Methods for the Consumer Expenditure Surveys

John L. Eltinge and Jeffrey Gonzalez, Office of Survey Methods Research, Bureau of Labor Statistics
OSMR, PSB 1950, 2 Massachusetts Avenue NE, Washington, DC 20212

Eltinge.John@bls.gov Gonzalez.Jeffrey@bls.gov

Abstract:

The U.S. Consumer Expenditure Survey (CE) is a large-scale household survey conducted to collect information on expenditures at a relatively fine level of classification. Data currently are collected through quarterly interviews (the Consumer Expenditure Interview Survey) and weekly diaries (the Consumer Expenditure Diary Survey). Due to interest in reduction of respondent burden, improvement of data quality, or possible reduction in the cost of data collection, the Bureau of Labor Statistics is carrying out some preliminary exploration of alternatives to the current designs of the diary and interview surveys. This paper provides an overview of some of the alternatives and outlines several areas of research that will be required to evaluate these alternatives.

Key words: Adaptive sampling; Balanced repeated replication; Constraints; Cost weights; Efficient design of survey procedures; Generalized linear model; Global questions; Matrix sampling; Operational risk; Survey cost structures; Synthetic data; Two-phase sampling; Variance estimation.

1. Introduction

1.1 The Consumer Expenditure Survey

The U.S. Consumer Expenditure Survey is a large-scale household survey carried out to collect data on consumer expenditures in the U.S. civilian noninstitutionalized population. The survey is sponsored and managed primarily by the Bureau of Labor Statistics, and data collection is carried out by the Bureau of the Census. For some general background on the U.S. Consumer Expenditure Survey, see Bureau of Labor Statistics (2007).

For the current paper, four features of the Consumer Expenditure Survey are of special interest. First, the survey uses a stratified multistage sample design, with strata and primary sample units defined on the basis of geography. Second, the sample elements are known as “consumer units” or CUs, roughly equivalent to households. Third, sample consumer units are assigned to one of two instruments known respectively as the “Interview” and “Diary” components. Fourth, consumer units selected the Interview component are asked to participate in a total of five interviews in which a field representative collects very detailed information about expenditures by the consumer unit.

1.2 Alternatives to the Current Design of the Consumer Expenditure Survey

Due to the length of the interview, and the associated potential for concerns related to respondent burden and data quality, the Bureau of Labor Statistics is carrying out some initial exploration of alternatives to the current CE design. Several of these alternatives involve the use of *matrix sampling methods* to assign subsets of the current instrument to sample consumer units selected for the interview survey. This paper develops some notation and simple estimators that could be used for the Consumer Expenditure Survey under a matrix-sampling design; and outlines several areas of research that would be required to evaluate the potential practical value of matrix sampling designs for the CE, including weighting methodology; variance estimation; estimation efficiency; nonsampling errors; and cost issues.

2. Background, Notation and Simple Estimators for Matrix Sampling Designs

2.1 Matrix Sample Designs

To date, matrix sampling has been used extensively in the education-testing literature, and has also been used in some survey and administrative-data applications. For some general background on matrix sampling designs, see, e.g., Mislevy (1983), Hinkins (1983, 1984), Thomas and Gan (1997), zeger and Thomas (1997), Thomas et al. (2006), and references cited therein. In addition, Gonzalez and Eltinge (2007) provide a literature review and discuss the potential application of matrix sampling designs to the Consumer Expenditure Survey.

2.2 Some Simple Weighting and Point Estimation Methods

For the current Consumer Expenditure Quarterly Survey, in which each sample unit receives the full instrument, let Y_{hij} equal the m -dimensional row vector of data that one would collect under full response from consumer unit $j = 1,\ldots, J_{hi}$ in PSU $i = 1,\ldots I_h$, in stratum $h = 1,\ldots, H$. In addition, let U_h be the population of N_h consumer units within stratum h; let $U = \bigcup_{h=1}^{H} U_h$ be the overall population of $N = \sum_{h=1}^{H} N_h$ consumer units; let S_h be the full set of n_h sample consumer units within stratum h; let S be the full set of $n = \sum_{h=1}^{H} n_h$ sample consumer units; define the full-sample selection probabilities π_{hij} for each consumer unit $(h,i,j) \in U$; and define the corresponding simple inverse-probability weights $w_{hij} = \pi_{hij}^{-1}$. In addition, define the population total $Y = \sum_{h=1}^{H} Y_h$ where $Y_h = \sum_{(hij)\in U_h} Y_{hij}$; and define the corresponding population means $\bar{Y}_h = Y_h / N_h$ and $\bar{Y} = Y / N$. Also, to simplify notation, assume for this paper that selected sample units provide complete responses to all questions that they receive; i.e., that there is no “unplanned missing data.”

Then simple estimators of the vectors of population totals at the stratum and full-population levels are, respectively,

$$(\hat{Y}_{Fh}, \hat{N}_{Fh}) = \sum_{(hij)\in S_h} w_{hij} (Y_{hij}, 1) \qquad \text{and} \qquad (\hat{Y}_F, \hat{N}_F) = \sum_{h=1}^{H} (\hat{Y}_{Fh}, \hat{N}_{Fh})$$

The corresponding estimators of means are $\hat{\bar{Y}}_{Fh} = \hat{Y}_{Fh} / \hat{N}_{Fh}$ and $\hat{\bar{Y}}_F = \hat{Y}_F / \hat{N}_F$.

Now consider a relatively simple matrix-sampling design under which some sample units (called “Group 0”) receive the full instrument; and all remaining sample units receive certain “core” questions as well as questions from one of K specialized sub-groups of questions. (In practice, one might consider more complex matrix sample designs, with a corresponding increase in the complexity of notation.). In parallel with the partition of the instrument, partition the vector of survey variables

$$Y = (Y_0, Y_1, \ldots, Y_K)$$

where Y_0 is the m_0-dimensional row vector corresponding to the "core" questions; for each $k=1,\ldots,K$, Y_k is the m_k -dimensional row vector corresponding to questions from section k of the instrument; and $m=\sum_{k=0}^{K} m_k$.

In addition, let $S_{(k)}$ be the subset of sample units that received questions from subgroup $k=1,\ldots,K$. Thus, the original full sample S is partitioned into $K+1$ subsamples: $S=\bigcup_{k=0}^{K} S_{(k)}$. In addition, for k = $0,1,\ldots,K$ define the selection indicators

α_{hijk} = { 1 if consumer unit *(h,i,j)* is contained in subsample k ; 0 otherwise }

Thus, consumer unit *(h,i,j)* receives section k of the instrument if and only if $\alpha_{hijk}=1$ or $\alpha_{hij0}=1$.

For the current discussion, assume that the assignment of sample consumer units to one of the subsamples $0,1,\ldots,K$ is based on a random process such that

$$p_{hijk}=P(\alpha_{hijk}=1)$$

where the probabilities p_{hijk} are assumed to be fixed. For $k>0$, define

$$p^*_{hijk}=p_{hij0}+p_{hijk} \;,$$

note that p^*_{hijk} is equal to the probability that consumer unit *(h,i,j)* receives section k of the instrument, and define the modified weights

$$w^*_{hijk}=w_{hij}\,/\,p^*_{hijk}$$

Then under the matrix-sample design described above, simple design-based estimators for population totals and means at the stratum and full-population levels, based only on data from units that received section k of the instrument, are

$$(\hat{Y}_{h(k)},\hat{N}_{h(k)})=\sum_{(hij)\in S_h} w^*_{hijk}(\alpha_{hij0}+\alpha_{hijk})(Y_{hijk},1) \quad \text{and}$$

$$(\hat{Y}_{(k)},\hat{N}_{(k)})=\sum_{h=1}^{H}(\hat{Y}_{h(k)},\hat{N}_{h(k)})$$

The corresponding estimators of means are $\hat{\bar{Y}}_{h(k)}=\hat{Y}_{h(k)}\,/\,\hat{N}_{h(k)}$ and $\hat{\bar{Y}}_{(k)}=\hat{Y}_{(k)}\,/\,\hat{N}_{(k)}$.

2.3 Some Simple Imputation-Based Point Estimation Methods

The estimators in Section 2.2 used only data collected directly from Section k of the instrument. One can develop alternative estimators based on imputation of data for Section k. For instance, define

$\tilde{Y}_{hijk}$ = {the observed Y_{hijk} if $\alpha_{hij0} = 1$ or $\alpha_{hijk} = 1$; an imputed vector otherwise }

and define the point estimators

$$\hat{Y}_{Ih(k)} = \sum_{(hij)\in S_h} w_{hij}\tilde{Y}_{hijk} \quad , \quad \hat{Y}_{I(k)} = \sum_{h=1}^{H}\hat{Y}_{Ih(k)} \quad , \quad \hat{\bar{Y}}_{Ih(k)} = \hat{Y}_{Ih(k)} / \hat{N}_{Fh} \quad \text{and} \quad \hat{\bar{Y}}_{I(k)} = \hat{Y}_{I(k)} / \hat{N}_{F} \quad .$$

Four examples of potentially applicable imputation methods are simple mean imputation within strata; mean imputation within more refined adjustment cells; regression or ratio-type imputation; or hot-deck imputation

3. Research Issues in Evaluation of Alternative Sample Designs and Estimation Methods

Detailed development and evaluation of the estimators considered in Section 2 would involve a wide range of theoretical and empirical issues, including the following.

3.1 Adjustment of Calibration Weighting Methods to Account for Subsampling

Current weighting for the CE involves several steps of nonresponse adjustment and incorporation of auxiliary information. For some general background on this methodology, see, e.g., Bureau of Labor Statistics (2007), Greenlees et al. (1982), Jayasuriya and Valliant (1996) and references cited therein. Extension of these methods to a matrix-sampling design would involve a number of adjustments. For example, one would need to adjust the weights to account for subsampling probabilities, as considered in the simple estimators in Section 2; and one would need to adjust the current CE weighting-cell structure to account for a potentially smaller number of sample units receiving a given instrument component k.

In addition, the simple estimators in section 2 above used only data collected directly for component k, or imputed for that component. One could develop an alternative set of point estimators for stratum and full-population means based on *pattern-mixture models.* For some general background on pattern-mixture models, see, e.g., Little (1993, 1994) and references cited therein.

3.2 Adjustment of Standard Balanced Repeated Replication Methods of Variance Estimation

At present, customary variance estimators for the Consumer Expenditure Survey area based on balanced repeated replication, in which the original sample is repeatedly partitioned into half-sample replicates, and each step in weighting adjustment and estimation is repeated for each half-sample replicate. In practice, under the current CE design, the weighting adjustment skips or simplifies some steps due to limitations on sample sizes within the half-sample replicate. Under matrix sampling, the issues with limited sample sizes are likely to become much more severe.

3.3 Two-Phase and Adaptive Forms of Matrix Sampling Designs

The development in Section 2 used a relatively simple form of matrix sampling in which the subsampling probabilities are fixed *a priori*. One could consider more complex forms of matrix sampling in which subsampling probabilities are determined adaptively, e.g., based on demographic data collected during the first CE interview. Under some conditions, the resulting design would lead to more efficient point estimators of expenditure means within specified sections.

3.4 Estimation of Population Covariance Matrices, Regression Coefficients, and Related Analytic Parameters

In addition to use in production of estimates of mean expenditures, data from the Consumer Expenditure Survey are also used in research on modeling of relationships among different types of expenditures, and on relationships between consumer-unit expenditures and incomes. One could continue to carry out this modeling research through use of data from sample units in Group 0. However, one could also consider modeling methods based on extensions of the literature on "synthetic datasets" e.g., Fienberg and McIntyre (2005), Reiter (2003, 2004, 2005a, 2005b) and references cited therein. The "synthetic datasets" literature developed originally to address issues with microdata disclosure risk, but much of the underlying mathematical structure would extend readily to the current matrix-sampling framework. In many of these cases, implementation would involve multiple imputation.

3.5 Evaluation of Burden, Total Survey Error and Costs

Finally, as noted in Section 1, consideration of matrix-sampling methods for the CE arose in part from concerns related to the perceived burden of the current interview component. Thus, evaluation of specific matrix-sampling options for the CE would require balanced consideration of empirical evidence related to perceived burden, as well as the impact of a matrix-sampling design on the total cost of data collection and the overall error components in the resulting data.

Evaluation of burden generally would involve both direct measures (e.g., total elapsed time for a given interview) and other measures (e.g., debriefings of field representatives or previous respondents). Evaluation of total survey error (e.g., sampling error, nonresponse and reporting error) would potentially involve classical sample-design approaches; laboratory and field testing of matrix-sampling instruments; and evaluation of current patterns of error components.

In addition, evaluation of costs would potentially include the relative costs of training field representatives on one or more forms of the instrument; survey initiation in wave 1; scheduling and re-contact in each of waves 2 through 5; interviewing for incremental minutes during waves 2 through 5; and conversion of reluctant respondents and other work related to caseload management, attrition and perceived burden.

Acknowledgements

The authors thank Jennifer Edgar, Karen Goldenberg, Sioux Groves, Mike Hidiroglou, Bill Mockovak and Clyde Tucker for helpful discussions of the application of matrix sampling and other design methods to the Consumer Expenditure Survey. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Bureau of Labor Statistics.

References

Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income. http://www.bls.gov/opub/hom/pdf/homch16.pdf (visited June 21, 2007).

Fienberg, S.E. and J. McIntyre (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics* **21**, 309-323.

Gonzalez, J.M. and J.L. Eltinge (2007). Multiple matrix sampling: A review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Greenlees, J.S., W.S. Reece and K.D. Zieschang (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* **77**, 251-261.

Hinkins, S.M. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 427-433.

Hinkins, S.M. (1984). Matrix sampling and the effects of using hot deck imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.

Jayasuriya, B.R. and R. Valliant (1996). An application of restricted regression estimation in a household survey. *Survey Methodology* **22**, 127-137.

Little, Roderick, J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125-134

Little, Roderick J.A. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471-483.

Little, Roderick, J.A. and Yougxiao Wang (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* **52**, 98-111

Mislevy, R.J. (1983). Item response models for grouped data. *Journal of Educational Statistics* **8**, 271-288.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181-188.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235 - 242.

Reiter, J.P. (2005a). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365 - 377.

Reiter, J.P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185 - 205.

Renssen, R.H. and N.J. Nieuwenbroek (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association* **92**, 368-374.

Thomas, N. and Gan, N. (1997). Generating Multiple Imputations for Matrix Sampling Data Analyzed with Item Response Models. *Journal of Educational and Behavioral Statistics*, **22**, 425-445.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. and Johnson, C.L. (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey. *Survey Methodology*, **32**, 217-231.

Zeger, L.M. and N. Thomas (1997). Efficient matrix sampling instruments for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association* **92**, 416-425.