CONSTRUCTION OF ADJUSTMENT CELLS BASED ON SURROGATE ITEMS OR ESTIMATED RESPONSE PROPENSITIES

I. S. Yansaneh, ASA/NSF/BLS Fellow Program and J. L. Eltinge, Texas A&M University

I. S. Yansaneh, ASA/NSF/BLS Fellow Program, 2 Mass. Ave., NE, Washington, DC 20212

Abstract. Little (1986) discussed the construction of nonresponse-adjustment cells by grouping sample units according to their estimated response propensities or predicted item values. An important step in this cell-construction work is the use of available auxiliary data to fit response-propensity and predicted-item models. The present paper outlines such model-fitting work for a specific income nonresponse problem that occurred in the U.S. Consumer Expenditure Survey. A related paper by Eltinge and Yansaneh (1993) discusses in further detail the resulting adjustments obtained for the Consumer Expenditure Survey income nonresponse problem.

Key words. Consumer Expenditure Survey, incomplete data, missing data, nonresponse, weighting adjustment, weighting cell.

1. Introduction

1.1. Construction of nonresponse adjustment cells

In the adjustment of survey estimates to account for nonresponse, some common methods are based on the use of "adjustment cells." The main idea is to group sample units into "cells" that are approximately homogeneous with respect to response probabilities or survey items. The responding units within a given cell then receive an additional weighting factor equal to the inverse of the estimated mean response rate within the cell. The resulting weighted estimator of a population mean or total then has a nonresponse bias approximately equal to zero, provided the within-cell covariances between survey items and response probabilities are approximately equal to zero. In some cases, one may alternatively consider nonresponse adjustment through hot-deck imputation within specified adjustment cells.

Some previous nonresponse-adjustment work has defined adjustment cells formed through combinations of simple classificatory variables, e.g., age group, race, and sex. In recent years, however, Little (1986) and others have suggested that one form cells by grouping sample units according to their estimated response propensities or predicted item values. Czajka et al. (1992) presented a detailed case study of estimated-response-propensity-based adjustment cell methods applied to a problem with missing income-tax data.

1.2. Application to income nonresponse in the Consumer Expenditure Survey

We applied the estimated-propensity and predicted-item methods to construct adjustment cells appropriate for the problem of income nonresponse in the U.S. Consumer Expenditure Survey (CE). The present paper describes the specific response-propensity and income models used in this work. A longer paper by Eltinge and Yansaneh (1993) considers other aspects of this case study, including additional background on general adjustment-cell methods, an explanation of income nonresponse in the CE, and a detailed discussion of the adjusted mean income estimates obtained through the estimated-propensity- and predicted-item-based cell methods.

For a detailed explanation of the sampling and weighting methods currently used in the U.S. Consumer Expenditure Survey, see United States Bureau of Labor Statistics (1992) and Zieschang (1990). For the present discussion, it is useful to note that the interview component of the U.S. Consumer Expenditure Survey uses a stratified multistage rotation sampling design. Each selected sample consumer unit is asked to participate in a total of five interviews. Detailed income data are collected through a complex set of questions asked at the end of the second and fifth interviews. Based on the extent of response or nonresponse to the full set of income questions, the BLS classifies each second- or fifth-interview consumer unit as a *complete* or *incomplete* income reporter. See Garner and Blanciforti (1992) for a detailed explanation of the "complete income reporter" definition. In the present case study, incomplete income reporting was the nonresponse phenomenon of principal interest. For both the second-interview and fifth-interview datasets considered here, approximately 14 percent of interviewed consumer units were incomplete income reporters.

Current CE weighting methods account for *unit* nonresponse (e.g., noninterviews), but do not adjust for income nonresponse. The resulting weights (labeled FINLWT21 in BLS documents) were used in the model-fitting work in the present study. Due to the complexity of current CE weighting methods, the BLS uses variance estimation procedures based on pseudoreplication methods with 44 replicates. These pseudoreplication methods are approximately equivalent to standard balanced repeated replication (Wolter, 1985, Chapter 3). All standard errors reported in this paper are based on these pseudoreplication methods.

2. A Response Propensity Model

A logistic regression model for response (complete income reporting) propensities was fit using the LOGSITIC procedure of SAS (SAS Institute Inc., 1989). The standard CE FINLWT21 weights were used in calculation of point estimates, while the cor-

responding 44 sets of replicate weights were used in calculation of standard errors.

Previous work by Garner and Blanciforti (1992) developed a main-effects logistic regression model for complete-income-reporter probabilities for second-interview CE participants from 1987. The logistic regression models considered here are based on the Garner and Blanciforti (1992) models and explanatory variables, with the following modifications. First, the data used here involved the 5125 consumer units that had a second interview in 1990, and the 5093 consumer units that had a fifth interivew between the fourth quarter of 1990 and the third quarter of 1991. Second, separate models were fit for the second- and fifth-interview units. Third, the replicate-based parameter-estimate standard errors calculated here were somewhat larger than the standard errors reported by Garner and Blanciforti (1992); consequently, some explanatory variables included in the Garner and Blanciforti model (e.g., race and sex) were not included in the present model, and other variables (e.g., education of reference person) were used with coarser categories. Finally, the present model fitting work considered inclusion of a number of two-factor interactions.

Tables 1 and 2 report the final logistic regression parameter estimates and standard errors for the second and fifth interviews, respectively. Standard errors are reported to three significant digits, and point estimates are reported to the corresponding number of digits. Note that all explanatory variables were classificatory; for each classificatory variable, the baseline class is indicated in parentheses, while the other classes are indicated in the subsequent indented list. See Garner and Blanciforti (1992) for detailed definitions of these explanatory variables. In addition, note that the same set of main effects were included in the second- and fifth-interview models, but that different two-factor interactions were included in the two models.

3. A Regression Model for Income

Linear regression models for Y = reported annual income were fit with the REG procedure of SAS (SAS Institute Inc., 1989). Separate models were estimated for the second and fifth interviews, using income and explanatory variables provided by the complete income reporters in the datasets described in Section 2. For point estimation, weights were set equal to the FINLWT21 weight divided by the unit's estimated response propensity as calculated from the model developed in Section 2. Similarly, replicate weights for calculation of standard errors were based on the 44 original sets of BLS replicate weights, each divided by the appropriate response propensity estimate, where response propensity estimates were based on the logistic regression model fits computed separately within each of the 44 replicates.

Tables 3 and 4 report parameter estimates and standard errors for the final regression model fits for the second and fifth interview reported incomes, respectively. The explanatory variables were similar to those used in the logistic regression models, with four exceptions. First, expenditures reported for the most recent quarter were used as a continuous variable, rather than a classificatory variable. Second, slightly different groupings were used for some classificatory variables. Third, the income regression model included some main effects (e.g., consumer unit structure and degree of urbanization) that were not used in the logistic regression propensity models. Finally, two common two-factor interaction terms were included in both the second- and fifth-interview income models.

4. Acknowledgements

The authors thank Richard Dietz, Thesia Garner, Paul Hsen, Eva Jacobs, Geoffrey Paulin, Stuart Scott, and Stephanie Shipp for many helpful discussions of the Consumer Expenditure Survey. This work was carried out while the authors were visiting the Bureau of Labor Statistics through the ASA/NSF/BLS Research Fellow Program, and was supported by a grant from the National Science Foundation (SES-9022443). Eltinge's research was also supported in part by a grant from the National Institutes of Health (CA 57030-04). The views expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of Labor Statistics.

5. References

Czajka, J. L., Hirabayashi, S. M., Little, R. J. A. and Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *J. Bus. Econ. Statist.* **10**, 117-131.

Eltinge, J. L. and Yansaneh, I. S. (1993). Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey. Technical Report #202, Department of Statistics, Texas A&M University. Submitted for publication.

Garner, T. I. and L. A. Blanciforti (1992), Household income report completeness: An analysis of U. S. Consumer Expenditure Survey data. Technical report, Division and Price and Index Number Research, Bureau of Labor Statistics, Washington, DC. Submitted for publication.

Little, R. J. A. (1986), Survey nonresponse adjustments for estimates of means, *Internat. Statist. Rev.* **54**, 139-157.

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, SAS Institute Inc., Cary, North Carolina.

United States Bureau of Labor Statistics (1992), *BLS Handbook of Methods*, Bulletin 2414, U. S. Department of Labor, Washington, DC.

Wolter, K. M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey, *J. Amer. Statist. Assoc.* **85**, 986-1001.

Table 1. Estimated Logistic Regression Propensity Model Coefficients, Second Interview

Independent Variable	Estimated Coefficient	Standard Error
Intercept	1.982	0.259
Age of Reference Person (Age 35 to 54 years)		
Age 34 years or less	0.498	0.147
Age 55 to 64 years	-0.026	0.157
Age over 65 years	0.759	0.273
Education of Reference Person (High School or Less)		
More Than High School	-0.092	0.125
Principal Occupation of Reference Person (Salaried Position)		
Laborer	-0.007	0.241
Craft	0.034	0.169
Sales	-0.124	0.177
Services	0.258	0.230
Self Employed	-0.912	0.181
Retired	-0.085	0.268
Not Working or other	-0.377	0.208
Marital Status (Not Married or other)		
Married	-0.451	0.128
Income Means Tested Program (Non-Participant)		
Participant	0.443	0.231
Housing Tenure (Owns)		
Rents	0.262	0.138
Region (South)		
Northeast	-0.317	0.175
Midwest	0.033	0.126
West	0.408	0.157
Expenditure Categories (Expenditures of $ 4500 to $5999)		
Expenditures of less than $ 1500	-0.976	0.258
Expenditures of $ 1500 to $ 2999	-0.612	0.198
Expenditures of $ 3000 to $ 4499	-0.568	0.183
Expenditures of $ 6000 to $ 7499	0.135	0.162
Expenditures of $ 7500 to $ 9999	0.464	0.173
Expenditures of $ 10000 or more	0.437	0.195
Married*(Age over 65 years)	0.487	0.273

Table 2. Estimated Logistic Regression Propensity Model Coefficients, Fifth Interview

Independent Variable	Estimated Coefficient	Standard Error
Intercept	1.538	0.252
Age of Reference Person (Age 35 to 54 years)		
Age 34 years or less	0.480	0.170
Age 55 to 64 years	-0.261	0.157
Age over 65 years	0.441	0.205
Education of Reference Person (High School or Less)		
More Than High School	-0.1789	0.0894
Principal Occupation of Reference Person (Salaried Position)		
Laborer	0.166	0.169
Craft	0.214	0.222
Sales	0.466	0.171
Services	0.386	0.197
Self Employed	-0.593	0.231
Retired	0.586	0.290
Not Working or other	-0.290	0.218
Marital Status (Not Married or other)		
Married	-0.307	0.103
Income Means Tested Program (Non-Participant)		
Participant	0.316	0.175
Housing Tenure (Owns)		
Rents	0.258	0.149
Region (South)		
Northeast	-0.207	0.228
Midwest	0.231	0.142
West	0.495	0.188
Expenditure Categories (Expenditures of $ 4500 to $5999)		
Expenditures of less than $ 1500	-0.917	0.308
Expenditures of $ 1500 to $ 2999	-0.619	0.159
Expenditures of $ 3000 to $ 4499	-0.111	0.177
Expenditures of $ 6000 to $ 7499	0.277	0.146
Expenditures of $ 7500 to $ 9999	0.793	0.188
Expenitures of $ 10000 or more	0.667	0.173
(Expenditure > $ 10000)*(Age 34 years or less)	1.253	0.580
(Expenditure > $ 10000)*(Age 55 to 64 years)	1.441	0.708
(Expenditure > $ 10000)*(Not working or Other)	1.476	0.399

Table 3. Estimated Income Regression Model Coefficients, Second Interview

Independent Variable	Estimated Coefficient	Standard Error
Intercept	12780	2350
Expenditure	3.069	0.311
Age of Reference Person (Age over 34 years)		
Age 34 years or less	-336	735
Education of Reference Person		
(High School graduate but did not complete college)		
Elementary	-2225	755
Did not complete high school	-868	905
College graduate	3610	1420
Postgraduate	8550	2680
Principal Occupation of Reference Person (Salaried Position)		
Laborer	-4660	1580
Craft	-3580	1760
Sales	-5170	1260
Services	-7770	1180
Self Employed	-15060	2130
Retired	-2050	2620
Not Working or other	-9680	2010
Consumer unit structure (Single or Single Parent)		
Husband and wife only	4640	1050
Husband, wife and children under 18	3840	1520
Husband, wife and other	6620	1850
Other family combinations	4260	1150
Income Means Tested Program (Non-Participant)		
Participant	-2979	918
Housing Tenure (Owns)		
Rents	-4187	657
Degree of Urbanization (City)		
Suburb	-1880	911
Rural	-3573	909
Region (South, Northeast or Midwest)		
West	2120	803
Postgraduate * Age 34 years or less	5470	2990
Expenditure * Retired	-1.854	0.490

Table 4. Estimated Income Regression Model Coefficients, Fifth Interview

Independent Variable	Estimated Coefficient	Standard Error
Intercept	6460	3460
Expenditure	3.826	0.438
Age of Reference Person (Age over 34 years)		
Age 34 years or less	660	1140
Education of Reference Person		
(High School graduate but did not complete college)		
Elementary	-2890	1180
Did not complete high school	-1653	850
College graduate	1670	1550
Postgraduate	8720	2520
Principal Occupation of Reference Person (Salaried Position)		
Laborer	-4010	1810
Craft	-3350	1750
Sales	-210	1580
Services	-4960	1850
Self Employed	-8250	2310
Retired	2290	4070
Not Working or other	-5620	3570
Consumer unit structure (Single or Single Parent)		
Husband and wife only	4980	1440
Husband, wife and children under 18	2930	1650
Husband, wife and other	4410	1860
Other family combinations	1220	1290
Income Means Tested Program (Non-Participant)		
Participant	-1586	910
Housing Tenure (Owns)		
Rents	-3112	843
Degree of Urbanization (City)		
Suburb	-1530	1360
Rural	-2131	999
Region (South, Northeast or Midwest)		
West	3310	1080
Postgraduate * Age 34 years or less	-7950	2920
Expenditure * Retired	-2.087	0.591