Searching for More Effective Variables to Use in a Household Survey's Nonresponse Adjustment Procedure

Lauren Vermeer[1], David Swanson[1], Sharon Krieger[1]
[1] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212

Abstract
Like most household surveys, the Consumer Expenditure Survey's response rate is decreasing. Also like most household surveys, it has a risk of bias, including nonresponse bias. The Consumer Expenditure Survey uses a nonresponse adjustment procedure to remove the bias generated by nonresponding households from its data. However, as the survey's response rate decreases, any imperfection in the nonresponse adjustment procedure increases, highlighting the importance of keeping the procedure current and accurate. The Consumer Expenditure Survey uses the traditional cell adjustment method for its nonresponse adjustments, which relies on having a set of cell-defining variables. In this paper we describe a technique we recently employed to systematically search for a more effective set of cell-defining variables.

Key Words: Non-response, non-response bias, consumer expenditure, R^2, effective variables

1. Introduction

The Consumer Expenditure Survey (CE) is a nationwide household survey that collects expenditure data from a representative sample of U.S. households to find out how they spend their money. The CE is conducted jointly by the U.S. Bureau of Labor Statistics and the U.S. Census Bureau, and it consists of two separate surveys – a quarterly Interview survey, and a two-week Diary survey. The quarterly Interview survey focuses on large and recurring expenditures, such as rent and car payments, while the two-week Diary survey focuses on smaller expenditures, such as food and apparel. A household selected to participate in either survey is asked to recall and record its expenditures for the time it is in the survey.

When a field representative first visits a selected household, it is not known whether the household will be a respondent or nonrespondent. Some nonrespondents verbally decline to participate in the survey but provide some basic demographic information, while other nonrespondents simply do not answer and provide no information. It is important to know something about who the nonrespondents are so they and their expenditures can be properly represented in the surveys' data. Not properly representing them can generate bias in the surveys' estimates.

Bias occurs when there is a systematic difference between the population's true parameters and the survey's estimates of them. Nonresponse bias is a specific type of bias that occurs when the respondents' data differs from the nonrespondents' data, and adequate steps are not taken to account for this difference. For example, in the CE, nonresponse bias is caused by the respondents and the nonrespondents having different expenditures. If the respondents are wealthier than the nonrespondents, then the surveys' expenditure estimates

may be too high, and they may over-emphasize the kinds of things wealthier households tend to buy, unless steps are taken to address the differences between the respondents and nonrespondents.

The CE considers these differences between respondents and nonrespondents when it adjusts the weights of the respondents to account for the nonrespondents. This adjustment removes bias generated by the nonrespondents, and this procedure is done using the traditional cell method. The procedure relies on having a set of cell-defining variables, so we explored the option of searching for more effective cell-defining variables. Our process allowed us to systematically search for these variables, calculate variances for them, and analyze the results.

2. CE's Nonresponse Adjustment Procedure

The CE, like other household surveys, begins the process of producing unbiased estimates of its variable of interest, household expenditures, by selecting an unbiased sample of households. However, some amount of bias is unavoidable, despite beginning with an unbiased sample of households. This bias occurs because not every household in the sample will be a respondent, and respondents and nonrespondents may have different expenditures. This unavoidable bias generated by nonrespondents gives the nonrespondents adjustment procedure its purpose – to adjust the data.

As mentioned above, the CE uses the traditional cell adjustment method to correct its data by adjusting the weights of the respondent households to account for the nonrespondent households. This method relies on using information that is known for both respondents and nonrespondents. (The available sources of this information will be described in **Section 4** of this report.) The general idea is to partition the sample households into disjoint subsets, or cells, with each cell containing households with similar probabilities of responding to the survey, and then increase the weights of the respondent households by multiplying them by an adjustment factor equal to the inverse of their cell's response rate.

The CE's cell adjustment method is performed in three steps. First, each survey's complete sample of households is partitioned into 192 cells based on their region of the country (one of four regions), their zip code's average income according to the IRS (one of three income classes), their household's size (one of four size classes), and their number of contact attempts (one of four contact attempt classes). Together they make 192 cells, 192 = 4 x 3 x 4 x 4. Households within each of these cells have similar probabilities of responding to either survey. Second, the response rate for each cell is calculated. And third, the weights of the respondent households in each cell are increased by multiplying them by the inverse of their cell's response rate, which is the adjustment factor that accounts for the nonrespondent households.

When any of the 192 cells have too few sample households to generate a credible response rate, it is collapsed together with one or more other cells to form a larger cell, until there are enough sample households in the collapsed cell to generate a credible response rate. Then the weights of the respondent households in the collapsed cell are increased by multiplying them by the inverse of the collapsed cell's response rate, which is the adjustment factor that accounts for the nonrespondent households in the collapsed cell. The

need to collapse is determined by whether the cell's nonresponse adjustment factor is above or below a certain threshold.

The CE's nonresponse adjustment procedure is designed to remove the bias generated by the nonrespondents, but any imperfection in the procedure gets magnified when the amount of nonresponse increases. The imperfections can be minimized by occasionally reviewing the procedure and adjusting as needed in order to keep it current and accurate. Our focus is on the systematic approach we took to find a more effective set of cell-defining variables, for this important procedure review.

3. Evaluating the Effectiveness of Cell-Defining Variables

A popular strategy for analyzing data is the within-cell variance method, which measures how well the sample households are clustered together based on their similarity to other households that are in the same cluster or "cell." It requires the data to be partitioned into "cells" before performing the variance calculation. [1] Using this idea, we produced an R^2 statistic to measure the effectiveness of each cell-defining variable.

The within-cell method measures the similarity of the households within the cells, with respect to their probability of responding to the survey. The goal is for the households in a cell to have similar probabilities of responding to the survey. In this paper we measured the effectiveness of the cell-defining variables by an R^2 statistic, using the following formula:

$$R^2 = 1 - \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(I_{hi} - RR_h)^2}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(I_{hi} - RR)^2} = 1 - \frac{\sum_{h=1}^{H} n_h RR_h(1 - RR_h)}{nRR(1 - RR)}$$

Here I_{hi} is a zero-one variable indicating whether the i-th sample household in the h-th cell responds to the survey ($I_{hi} = 1$ if it responds, and $I_{hi} = 0$ if it does not respond); where n_h and n are the number of eligible units in the h-th cell and in the whole survey; [2] $RR_h = \frac{1}{n_h}\sum_{i=1}^{n_h} I_{hi}$ is the response rate of the sample households in the h-th cell; and $RR = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} I_{hi}$ is the response rate of all sample households in the survey. The formula on the left can be rewritten as the formula on the right, which can be easier to program if one has already calculated the response rates. [3] The formula on the right looks like the variance formula for a binomial distribution.

---

[1] This popular method is widely used, so multiple versions of this formula exist. Our calculations use the formulas listed.

[2] "Eligible" units are the sample addresses with housing units that are occupied by people within the survey's target population. Unoccupied housing units, housing units that are occupied by people outside the survey's target population, and non-existent and non-residential building units are all "ineligible" units. In the CE, eligible units are categorized as either "completed interviews" or "Type A noninterviews," while ineligible units are categorized as either "Type B noninterviews" or "Type C noninterviews."

[3] The formula on the left can be rewritten as the formula on the right using algebra and the fact that $I_{hi}^2 = I_{hi}$. That is because I_{hi} is a zero-one variable, and $0^2 = 0$ *and* $1^2 = 1$. The formula on the right gives an alternative formula that looks a little simpler and is easier to program if one has already calculated the response rates.

A similar formula can be used to measure the similarity of the households within the cells with respect to the survey's variable of interest, which for the CE survey is household expenditures. The goal here is for each of the households within a cell to have similar expenditures. This method uses the following formula:

$$R^2 = 1 - \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(x_{hi} - \bar{x})^2}$$

Here x_{hi} is the reported expenditure of the i-th respondent household in the h-th cell on all expenditure categories added together; $\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ is the average reported expenditure from all respondent households in the h-th cell; and $\bar{x} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} x_{hi}$ is the average reported expenditure from all respondent households in the whole survey.

The numerator is a stratified variance of the eligible units' response rates (or the respondents' expenditures); and the denominator is an unstratified variance of them. One minus the ratio of variances is the percent of variance "explained" by the stratification variables.

Both formulas produce an R^2 statistic, which shows the variability between households' expenditures in the same "cell." Analysis of this statistic is straightforward. Its values always range from 0 to 1, with higher values indicating a better stratification of the households into cells with similar response rates (or similar expenditures), and lower values indicating a worse stratification. Generally, the higher the R^2, the better the model's fit.

4. Implementing the Analysis Using CE Data

4.1 Searching for Test Data

CE's nonresponse adjustment procedure currently uses four cell-defining variables (region, zip code's average income according to the IRS, household size, and number of contact attempts) as described in **Section 2**, however we want to know if a better set of variables exists. The information available for the nonrespondents is typically limited to just a few variables, mostly frame variables such as their geographic location. So, to widen our search, we looked outside our data and into the Census Planning Database.

The ideal cell-defining variable would need to have known values for both respondents and nonrespondents. This presents a challenge since such variables are rare – gathering information on respondents is easy, since they respond and provide the desired information, but gathering information on nonrespondents is hard, because by definition, they don't respond. Fortunately, we found a collection of variables within the Census Planning Database that can be tested.

The Census Planning Database (CPD) is a database published every year by the U.S. Census Bureau. It contains a wide range of housing, demographic, and socioeconomic information that can be used for census planning and for survey analysis in general. There are two versions of the database, a block-group version, and a tract-level version. We used the tract-level version from 2021. Tracts are small geographic areas defined by the U.S. Census Bureau that are designed to have about 4,000 people. They vary in size depending on population density, ranging in size from a few blocks in densely populated urban areas

to hundreds of square miles in sparsely populated rural areas. The U.S. has about 330 million people, each tract has about 4,000 people, so there are about 75,000 tracts.

Each tract in the CPD has about 600 variables: 100 variables from the 2010 census; 100 variables from American Community Survey (ACS), many of which have basically the same information as the 2010 census variables; 100 standard error estimates; and 300 percentages derived from the other variables.

Adding the CPD's tract-level demographic information to the CE's database was a simple matter because CE's database contains the tract number of every household in the sample, both respondents and nonrespondents, and that allowed the two databases to be merged. Then after adding the tract-level demographic information to the CE's database, the next step was to partition the sample households on the CE's database according to their tract-level demographic characteristics rather than the demographic characteristics of the individual households. That increased the number of potential cell-defining variables that were available for testing.

Most of the variables we selected from the CPD come from the ACS, which asks detailed questions about each household member and their housing unit. For example, the ACS asks questions about the household members' age, race, sex, and marital status. The ACS also asks questions about whether the housing unit is owned or rented, its type of structure, and its access to internet/telephone services. The ACS's data has a good reputation for accuracy, which is reflected in its large sample size and its high response rate.[4] In total, 45 variables were selected for testing from the CPD.

4.2 Preparing the Data

To begin, we wanted to evaluate the effectiveness of the current variables as a baseline. In total, 54 variables were tested, 45 variables from the CPD, and 9 variables from the CE's data, which include those currently used in CE's nonresponse adjustment procedure.

In the CE, the variables that are used to define cells for the nonresponse adjustment procedure are divided into three or four categories. Four regions of the country; three income categories; four household size classes; four number of contact attempts categories. In order to evaluate the potential new cell-defining variables the way they would probably be defined in production, we took the 45 CPD variables and divided each variable into four quartiles. For example, one of the CPD variables examined was the percent of people 5 years of age and older who speak English at home. About 80 percent of people 5 years of age and older speak English at home, but the percent varies from tract-to-tract with the distribution of tract-level percentages shown below:

[4] Over the ten-year period 2010-2019 the ACS averaged 2.2 million interviewed households per year, and it had a response rate of 94 percent. For details see https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/.

Percent of people 5 years of age or older who speak English at home		
Quartile	Percent	Number of tracts
1	0.00 - 71.44	18,770
2	71.45 - 88.16	18,096
3	88.17 - 95.40	18,077
4	95.41 - 100.00	18,136
Total		73,079

Any sample household in the CE that was in a tract whose percent of people speaking English was below 71.44 percent was placed in the first quartile; any household in the CE that was in a tract whose percent of people speaking English was between 71.45 percent and 88.16 percent was placed in the second quartile; and so on. That partitioned the sample households in the CE into four roughly equal groups.

4.3 Programming the Analysis

The testing procedure involves using SAS software to write and execute a program that efficiently calculates the R^2 statistic for each of the 54 tested variables. After running this SAS program, we were able to evaluate all the selected variables from the CPD database and CE data, efficiently calculating the desired R^2 statistics and outputting the results for all 54 variables. Using the generated output, the statistics for all variables were ranked and compared to determine which, if any, of the tested variables produced a high R^2 statistic.

The program uses the within-cell method and R^2 statistic to find effective variables that explain two things – the sample households' probabilities of responding to the survey, and the values of the survey's variable of interest, which for the CE surveys is a household's expenditures.

Also, the Interview and Diary surveys seek to collect different types of expenditures from households, so the most effective variable for one survey may not be the most effective variable for the other survey. Keeping this is mind, all 54 variables were tested for both surveys so that the best variables for each survey could be identified.

5. Results

The SAS program generated the output in **Table 1**. The program was run for the 54 variables mentioned above with 6 years of data (2015-2020). That generated 324 (= 54 x 6) R^2 statistics for the Interview survey's response probabilities; 324 R^2 statistics for the Diary survey's response probabilities; 324 R^2 statistics for the Interview survey's expenditures; and 324 R^2 statistics for the Diary survey's expenditures.

The most logical way to summarize the results for the comparison was to average all outputs across the tested years (2015 – 2020), therefore computing one average R^2 statistic per variable. This was done for CE's response rates and expenditures, and separately for CE's Interview and Diary surveys. Ultimately, each variable has four R^2 values, two for the Interview survey and two for the Diary survey. While R^2 statistics range from 0 to 1, the success of a tested variable might not have been as obvious. It is typical for R^2 statistics

in social sciences to be low, meaning we did not expect to get any values close to 1. To remedy this problem, criteria were formed to better analyze the results and determine which variable would give the most desired outcome.

To get a visual image of the results, the averaged R^2 statistics for response rates and expenditures were plotted on the same graph for comparison. This yielded two graphs. **Figure 1** is for the Interview survey and **Figure 2** is for the Diary survey. For each graph, the R^2 statistic for response rates is plotted along the x-axis, and the R^2 statistic for expenditures is plotted along the y-axis. In order to satisfy our requirements for the "most effective" variable, we are looking for any plotted points in the upper right-hand region of the graphs. Such variables do a good job of explaining both response rates and household expenditures. Unfortunately, as shown in Figure 1 and Figure 2, no variables appear in this region for neither the Interview survey nor the Diary survey.

Table 1: Summary of Tested Variables by R^2 Statistic and Survey

	Interview		Diary	
Variable Description	R^2 (Response Rates)	R^2 (Expenditures)	R^2 (Response Rates)	R^2 (Expenditures)
Number of contact attempts for weighting	0.0769	0.0077	0.0198	0.0013
Contact attempt description	0.0627	0.0092	0.0025	0.0017
Percent of mobile home housing units	0.0294	0.0446	0.0155	0.0116
Percent of population US citizens at birth	0.0291	0.0413	0.0158	0.0127
Percent of households w/ no Internet	0.0286	0.0971	0.0142	0.0193
Percent of non-Hispanic population; black only	0.0286	0.0460	0.0175	0.0139
Percent of households w/ no computer	0.0284	0.0851	0.0141	0.0175
Percent of non-Hispanic population; white only	0.0284	0.0463	0.0182	0.0150
Percent of non-Hispanic population; Asian only	0.0284	0.0558	0.0144	0.0123
Percent of population w/ college degree	0.0284	0.1006	0.0142	0.0156
Percent of population; married	0.0284	0.0750	0.0176	0.0206
Percent of households w/ Internet and computer	0.0283	0.1003	0.0142	0.0210
Number of returned Census forms from eligible households	0.0283	0.0565	0.0180	0.0186
Percent of households who rent their home	0.0283	0.0595	0.0167	0.0188
Median household income for tract	0.0282	0.1262	0.0149	0.0247
Percent of households who own their home	0.0282	0.0596	0.0167	0.0189
Percent of population; English speaking	0.0281	0.0417	0.0149	0.0126
Percent of population; Hispanic	0.0281	0.0430	0.0152	0.0120
Percent of population aged 25-44	0.0281	0.0378	0.0157	0.0123
Percent of population aged 65+	0.0281	0.0342	0.0154	0.0115

Percent of housing units with occupants who have moved in since 2010	0.0281	0.0383	0.0162	0.0130
Average household income for tract	0.0280	0.1274	0.0148	0.0232
Median of respondent's house value for tract	0.0280	0.0969	0.0143	0.0170
Percent of population; not a high school graduate	0.0279	0.0919	0.0146	0.0189
Percent of single-family housing units	0.0279	0.0535	0.0153	0.0185
Percent of population; female	0.0278	0.0349	0.0146	0.0115
Percent of population; male	0.0278	0.0350	0.0147	0.0115
Average number of people per households	0.0277	0.0375	0.0142	0.0106
Percent of households constructed 2010 or later	0.0277	0.0373	0.0145	0.0109
2015-2019 ACS self-response rate for tract	0.0277	0.0765	0.0175	0.0213
Percent of non-Hispanic population; other race	0.0277	0.0346	0.0145	0.0106
Percent of population who moved within past 1 year	0.0277	0.0371	0.0150	0.0133
Percent of households on public assistance	0.0277	0.0510	0.0146	0.0139
Average house value estimate for tract	0.0276	0.1131	0.0156	0.0229
Total number of eligible addresses for decennial Census	0.0276	0.0595	0.0142	0.0184
Percent of population; age under 5	0.0276	0.0352	0.0144	0.0109
Percent of housing units; large occupancy	0.0276	0.0428	0.0144	0.0148
Percent of population; aged 5-17	0.0275	0.0344	0.0144	0.0118
Percent of population; aged 18-24	0.0275	0.0448	0.0146	0.0116
Percent of population; aged 45-64	0.0275	0.0479	0.0145	0.0142
Percent of non-Hispanic population; American Indian or Alaskan Native only	0.0275	0.0338	0.0141	0.0107
Percent of non-Hispanic population; Native Hawaiian or Other Pacific Islander only	0.0275	0.0331	0.0141	0.0104
Percent of population; married w/ 1+ child	0.0275	0.0357	0.0144	0.0118
Percent of working civilians; aged 16+	0.0275	0.0530	0.0146	0.0142
Percent of working civilians; aged 16-24	0.0275	0.0352	0.0144	0.0104
Percent of working civilians; aged 25-44	0.0275	0.0456	0.0145	0.0120
Percent of working civilians; aged 45-64	0.0275	0.0404	0.0144	0.0118
Number of Consumer Unit (CU) members	0.0034	0.0788	0.0043	0.0137
Renter and owner quartiles by property value	0.0029	0.0640	0.0051	0.0188
Specifies if Consumer Unit (CU) is inside or outside CBSA	0.0021	0.0065	0.0028	0.0009
Region of selected household, BLS assigned groupings	0.0010	0.0064	0.0008	0.0041

Race of Consumer Unit (CU) used in weighting	0.0008	0.0101	0.0043	0.0030
Tenure (owner/renter) of household, used in weighting	0.0004	0.0426	0.0041	0.0132
IRS income of household, used for weighting	0.0003	0.0523	0.0011	0.0023

Figure 1: Percentage of Variance Explained by Tested Variables (Interview)

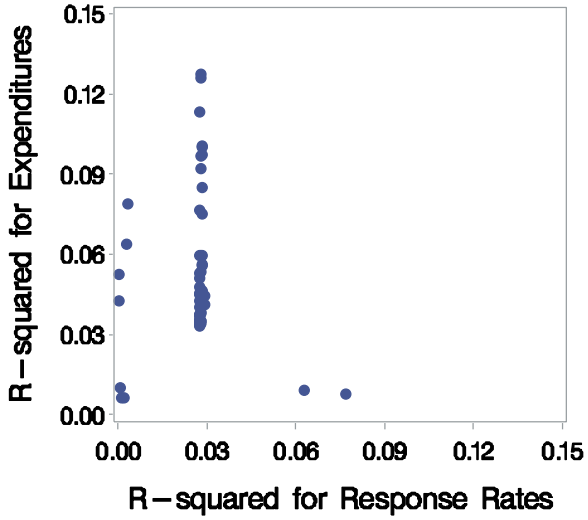
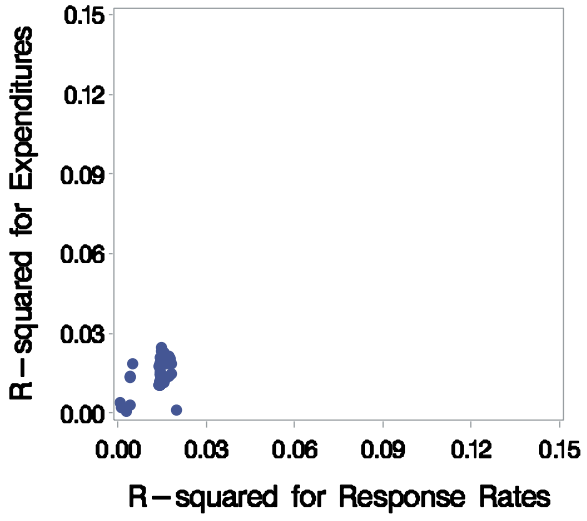


Figure 2: Percentage of Variance Explained by Tested Variables (Diary)



6. Conclusion

Our technique allowed us to evaluate the effectiveness of 54 potential cell-defining variables, which included variables currently used in the CE in addition to variables poised for incorporation into practice, should they be ideal. Using an R^2 statistic measures the percent of variance explained by each of the 54 variables, which translates to the percent of variance of the sample households' probability of responding to the survey, and the percent of variance of the respondent households' variable of interest.

Unfortunately, the results did not yield a new variable to be incorporated into practice, but they did show that some variables were better than others. The best variable for explaining a sample household's probability of responding to the survey was the number of contact attempts, which explained 7.69 percent of the variance in the CE Interview survey and 1.98 percent of the variance in the CE Diary survey. The best variable for explaining a respondent household's expenditures was the average household income in the respondent household's tract, which explained 12.74 percent of the variance in the CE Interview survey and 2.32 percent of the variance in the CE Diary survey. Although none of the tested variables were effective at explaining both variances, this outcome follows typical trends of R^2 statistics in social sciences. So, these results, although disappointing, were not unusual.

Our method provides a simple and objective way to measure the effectiveness of current and potential cell-defining variables for a nonresponse adjustment procedure, using the traditional cell adjustment method. Additionally, the R^2 statistic provides results which are easy to understand, interpret, and program.

This method can be employed again in the future, when new variables are introduced. Also, since it is important to have current data and current conclusions, future variable testing is likely to occur. An additional option for this future search may be to also include variables from other databases, to allow a slightly wider breadth of testing categories. While our variable of interest was expenditures, this same procedure could be applied to a different set of data of nonrespondents, and a different variable of interest.

7. Disclaimer

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.