Evaluation of Race Edit Improvements in the Consumer Expenditure Survey

Brian Nix,[1] Sharon Krieger, and Barry Steinberg
Mathematical Statisticians, U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions, Statistical Methods Division
2 Massachusetts Avenue NE, Washington, DC 20212
Nix.Brian@bls.gov

There has been concern in recent years over the declining coverage rate of the Black/ African-American population in the Consumer Expenditure Survey (CE). People in this demographic group appear to be undercounted in the survey, and the problem seems to be getting worse. Nonrespondents for whom race is unknown or ambiguous are assigned a race, which is imputed by random as a last resort according to known probability distributions. However, previous research showed that too few nonrespondents were systematically placed into the Black/African-American category due to the usage of inadequate probability distributions. Several new procedures were implemented in 2012 to increase the accuracy of the imputations. This paper evaluates the results of those procedures to see how well they increased the accuracy of race imputations in the CE.

Key Words: Consumer Expenditures, Coverage Rate, Imputation, Race

1. Introduction

The Consumer Expenditure Survey (CE) is a nationwide household survey conducted by the U.S. Bureau of Labor Statistics (BLS) to find out how Americans spend their money. Its primary customer is the Consumer Price Index, which uses CE data to help establish its weights. Other customers include the Department of Defense which uses CE data to help determine cost-of-living allowances for military personnel, and the Department of Agriculture which uses CE data to calculate the annual cost of raising a child.

The CE consists of two sub-surveys, an Interview survey and a Diary survey, collected for the BLS by the U.S. Census Bureau. The purpose of the Diary survey is to collect detailed expenditure data on small, frequently purchased items such as food and apparel. The purpose of the Interview survey is to collect detailed expenditure data on large items such as property, automobiles, and major appliances, as well as on recurring expenses such as rent, utility bills, and insurance premiums. The data from the two surveys are then combined to provide a complete picture of consumer expenditures in the United States.

Households are selected to be in the survey by first drawing a stratified random sample of urban and rural areas across the country, and then drawing a systematic sample of households within those areas.[2]

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

[2] The CE's measurement unit for which expenditure reports are collected across the nation is called a consumer unit (CU), although informally the term "household" is used interchangeably, and will be used for the remainder of this paper. A CU is properly defined as all members of a household who are related by blood, marriage, or other legal arrangements, or two or more unrelated persons who live together and use their income to make joint expenditure decisions. On average, a CU consists of about 2.5 persons, with about 30 per cent of the CU's in our survey consisting of a single person.

Concern was raised in recent years over the declining coverage rate of the Black/African-American population in the CE. That demographic group seemed to be undercounted in the survey, and the problem was growing over time. As a result of the concern, research was conducted to fix the problem. The research showed a major source of the problem was the way nonrespondents were assigned to a race category. Some of the nonrespondents were randomly assigned to a race category, and the racial proportions used to assign them were skewed towards the Non-Black population.

Further research identified three ways to fix this problem: use the racial proportions observed in the latest Decennial Census instead of CE's respondents for the probabilities; use racial proportions at the county level instead of at the regional level; and add tenure information to the county-level proportions. These three methods were implemented into the CE in 2012. We now have several years of data to observe the results of the new methods.

This paper is a follow-up to a previous study of the data, conducted in 2013 by Krieger, Steinberg, and Swanson of the Bureau of Labor Statistics and published in the 2013 JSM Proceedings, using the data which were available at the time.

2. Data Description

The research presented in this paper is based on the Quarterly Interview survey data collected from 2005 through 2015. Each household in the survey is interviewed four times,[3] once per quarter, over a one-year period. Data are collected from approximately 7,000 households per quarter, for a total of approximately 28,000 households per year. Over this ten-year period, approximately 383,000 occupied households were contacted across the country, resulting in about 276,000 completed interviews (at an average 72% response rate.)

As mentioned above, there are two components to the CE, the interview and diary surveys. A household remains in the diary survey for two weeks, resulting in two individual one-week expenditure collections. As there is no time between the two weeks of the diary survey, over 95% of households who complete the first week complete both weeks, and therefore any information (such as race) which needs to be imputed in the first week is likely to be imputed in the second week as well. Therefore, only the interview survey was used in this study.

The dataset used in this research includes, from the interview survey, both respondents (households with completed interviews) and nonrespondents (households that could not be contacted or refused to give interviews).

3. How CE Gets the Race of Nonrespondents

Race is one of the characteristics that the Consumer Expenditure Survey requires of all households in its sample. The CE annually publishes tables of expenditures by different

[3] The survey actually had five interviews during the period of this study, but the first interview was a "bounding" interview which provided baseline data, and was not used to compute the survey's published expenditure estimates. Prior to 2010, the bounding interview did not have missing race values imputed, so the data from the first interview is only used in this study from that point forward. The bounding interview was eliminated in a sample redesign process in 2015.

categories, including race. Additionally, race is one of several demographics used in the weighting process to produce the nationwide expenditure estimates.

The race of respondents is almost always known, because it is reported by someone in the household. However, for nonrespondents, field representatives must determine their race in one of several ways.

Sometimes nonrespondents report their own race, because they reported it in a previous interview, and the information is carried forward to subsequent interviews. Also, many of the interviews which are classified as nonrespondents are actually partial interviews. The partial interviews are not considered complete enough for their expenditure data to be used, but they may include race information. Approximately 40% of CE's nonrespondents are assigned to a race category in one of these two ways. Two additional approaches are for the survey's field representatives to observe the information directly while trying to persuade the household to participate in the survey, and to ask a neighbor for the nonrespondents' race. These methods account for another 40% of the race assignments for nonrespondents.

For the 20% of nonrespondents who still have an unknown race category, a race category must be randomly assigned to them. Their race is imputed using probabilities generated from various racial distributions.

4. Summary of CE's Imputation Process for Race

CE's imputation process for race occurs at two stages. The first stage edits race at the household member level, and the second stage edits race at the household level.

The goal of the first stage (member race edit) is to assign every household member to one of the six[4] race categories used officially by the federal government: White, Black/African-American, Native American, Asian, Pacific Islander, or Multi-race.[5] In this stage, the following edit occurs for any household member having a missing race value:

1. If the member is Hispanic, then their race is set equal to White.
2. Otherwise, if the member is the reference person,[6] then their race is set equal to the race of another household member whose race is known.
3. Otherwise, their race is randomly chosen from one of the six race categories, by known probability distributions.

The goal of the second stage (household race edit) is to assign every household to a single race category of either Black/African-American or Non-Black. This is done for all households, both respondents and nonrespondents, in order to properly adjust their sampling weights. If the race of the reference person is Black/African-American, then the race of the household is set to Black/African-American. If the race of the reference person

[4] There are currently nine race categories (ten including "Other.") Beginning in 2011, the Pacific Islander race category was split into four separate categories: Native Hawaiian, Guamanian/Chamorro, Samoan, and [other] Pacific Islander. However, this study only focuses on whether or not an individual or household is Black/African-American.

[5] The way respondents are asked about their race allows them to identify each member of the household by more than one race category. For example, a person can say they are both White and Black/African American. When this happens, they are assigned to the Multi-race category during CE's member edit process. A person can also choose an "Other" category, with space to identify what they consider their race to be.

[6] The reference person is the first person mentioned by the respondent when asked to "Start with the name of the person or one of the persons who owns or rents this home."

is White, Native American, Asian, Pacific Islander, or Multi-racial, then the race of the household is set as Non-Black[7]. Otherwise, the following edit occurs:

1. If no member records are available for the household, then the household's race is set equal to Black/African-American or Non-Black based on the individual race category reported by field representatives.
2. Otherwise, the household's race is randomly chosen as Black/African-American or Non-Black.

5. Three Actions Taken to Improve Race Imputation

As mentioned in the introduction, three actions were implemented in 2012, based on previous research, to improve the random assignment process.

The first action was to use information from the 2000 Census to randomly assign the race of nonrespondents for which race was unknown. Previously, the racial distribution of CE's respondents, for whom race was known, was used to randomly assign them to a race category. The problem with this was that the known race for respondents had a nationwide distribution of 12.2% Black/African-American, but the 2000 Census showed the actual distribution to be 14.4%. Black/African-Americans have a lower response rate as a group, so the census probabilities are now used instead.

The second action was to change the level of geographic detail used to create racial proportions. Previously, distributions had been devised at the level of the four regions of the country (Northeast, Midwest, South, and West). However, it was known that the concentration of the Black/African-American population could vary widely within region. For example, in the South, the Black/African-American percentage by state ranged in the 2010 Census from 37.0% in Mississippi to 3.4% in West Virginia. And within states, the percentage could vary even more so. Maryland, another state in the South, in the 2010 Census had a total Black/African-American population of 29.5%, but this ranged from 64.5% in Prince George's County, MD to 1.0% in Garrett County, MD. Therefore, beginning in 2012, county-level Decennial Census data have been used.

The third action was to add household tenure as a factor in the distribution. Tenure means whether someone owns or rents their home. The 2010 Census showed that 68.0% of Non-Black people owned their home, compared with only 44.3% of Black people. Thus, it seemed that a greater degree of accuracy could be achieved by using racial distributions at the county/tenure level rather than at the county level alone.

6. Evaluating the New Methods

In order to determine the effectiveness of the new changes in race imputation methods, this study needed to look at individuals and households in our survey whose race was unknown and therefore imputed in at least one interview, but then in a subsequent interview the race was reported. The non-imputed values in the subsequent interviews are regarded as the true

[7] A respondent who selects more than one race is classified as multi-race, and then assigned in the household race edit as Non-Black, even if Black/African American was one of the races originally selected. However, a 2003 study by the National Center for Health Statistics indicated that if a respondent were to choose multiple races including Black or African American, and were then instructed to choose a single primary race, then the respondent identified him- or herself as Black / African American 68.2 % of the time. This indicates that our classification procedures may need updating.

values. However, as the following tables illustrate, the number of such households is rather small.

Table 1 shows the race edit imputations at the individual member level. Only those values which were imputed in one interview, for which a value was reported in a later interview, are shown in the total numbers. The assumption is that the reported value in the most recent interview is the true value, and so the listed successes are those imputed values that match at later non-imputed value. 2015 data was therefore used to provide future values for 2014 imputed race values, with the imputed values analyzed ranging from 2005 to 2014.

TABLE 1 – Race Edit Success Rates at Member Level

Year	Correct Imputations	Incorrect Imputations	Success Rate	Indeterminable Imputations
2005	49	17	.7424	634
2006	29	24	.5472	518
2007	227	139	.6202	1,623
2008	50	30	.6250	878
2009	28	16	.6364	600
2010	63	38	.6238	716
2011	60	21	.7407	787
2012	95	59	.6169	4,093[8]
2013	156	62	.7156	4,956
2014	101	56	.6433	5,035

At the first (member-level) stage of imputation, over the 10 years in question, there were only 1,320 households that meet these criteria – 791 from 2005 to 2011 under the old method, and 529 from 2012 to 2014 under the new method. The respective numbers of successful imputations – that is, those that were later determined to have been correct – were 506 from 2005 to 2011, and 352 from 2012 to 2014. The rate of success does go up between the two groups (from 506/791 = 0.6397 to 322/484 = 0.6654), but is it significant?

The standard method for comparing two rates for statistical significance is to find the standard error as follows:

$$SE = \sqrt{\frac{p_1 * (1 - p_1)}{n_1} + \frac{p_2 * (1 - p_2)}{n_2}}$$

In this formula, p_1 and p_2 are the two proportions (or rates) out of their respective totals n_1 and n_2.

The standard error for the difference in proportions from the group (2005-2011) to the group (2012-2014) can be calculated, using p_1=0.6397, n_1=791, p_2=0.6654, and n_2=529. The standard error here is 0.0267. A 95% confidence interval for the difference, using this standard error, gives us 0 +/- 1.96*0.0267 = +/- 0.0523. But the difference between the two

[8] In 2012, the CE changed the way in which some values for race were coded. These included values which were carried forward from previous interviews, which are now classified as imputed, resulting in the abrupt increase in total number of imputations beginning in that year.

proportions, 0.6654-0.6397, is only 0.0257. This is within the confidence interval, so the difference between the two proportions is not statistically significant.

Table 2 shows the race edit imputations at the household level. As before, only those imputed values which can later be verified by non-imputed values are included, with a successful imputation defined as one that matches the later value.

TABLE 2 – Race Edit Success Rates at Household Level

Year	Correct Imputations	Incorrect Imputations	Success Rate	Indeterminable Imputations
2005	136	55	.7120	763
2006	106	30	.7794	762
2007	117	41	.7405	936
2008	172	39	.8152	902
2009	145	68	.6808	856
2010[9]	298	92	.7641	1,268
2011[10]	327	112	.7449	1,491
2012	458	165	.7352	1,638
2013	553	164	.7713	1,861
2014	491	159	.7554	2,170

At this second (household level) stage of imputation, there were 1,301 successes out of 1,738 total imputations that were later assigned true values from 2005 to 2011, for a success rate of 0.7486, and 1,502 successes out of 1,990 total imputations from 2012 to 2014, for a success rate of 0.7548. The success rate again increased; however, also not by a statistically significant amount. The increase was only 0.0062, well inside the 95% confidence interval of +/- 0.0278, calculated in the same manner as before.

7. Results

In addition to looking at the accuracy of the imputations for individual persons and households in the CE, the overall distributions were also examined. Table 3 shows the percentage of Black/African Americans for CE's respondents and nonrespondents who reported their own race. Almost all respondents report their own race, with 10.6% of them reporting their race to be Black/African-American. Many nonrespondents also reported their race. The table also shows the percentage of Black/African Americans who resided in the 461 counties that were part of CE's sample from 2005 to 2015, according to the 2000 census. The percentages in this table show that the distribution of CE's nonrespondents is closer to that of the Decennial Census than to CE's respondents, which was the reason for using racial information from the Decennial Census for the probability distributions used in randomly assigning race values.

[9] An abrupt increase in the number of imputations beginning in 2010 is visible in both tables; this was due to the imputation of race values for the first (bounding) interview, which previously retained a missing value if race was unknown. (There is also a generally increasing trend in the number of imputations each year also reflects a steadily declining response rate in the CE.)

[10] The 2011 figures do not include data from April. In that one month, a classification error caused an abnormally high number of values to be flagged as imputed. This error was later corrected, but it is not possible to determine which values were actually imputed, so the data for that month has been omitted. Including those observations increases the total number of imputations from 2011 to 2,647, with 2,088 regarded as successes, for a success rate of 78.9% for that year.

TABLE 3 – Percentage of Black/African Americans in CE When their Race Is Known

Census Region	CE Respondents	CE Non-Respondents	2000 Census
Northeast	11.9 %	14.1 %	15.3 %
Midwest	9.7	12.0	15.8
South	20.4	19.2	20.7
West	4.4	5.9	6.4
Total U.S.	10.6	13.3	14.4

Table 4 shows the percentage of Black/African-American race values which were imputed for nonrespondents, before and after the three changes in race edit methods. The more recent imputed percentages are much closer to the 2000 Census values, indicating that the new methods are improvements to the survey.

TABLE 4 – Percentage of Imputed Black/African-American Nonrespondents in CE, Before and After Changes in Imputation Methods

Region	Nonrespondents, 2005-2011	Nonrespondents, 2012-2015	2000 Census
Northeast	10.5 %	15.1 %	15.3 %
Midwest	9.4	13.3	15.8
South	18.9	21.8	20.7
West	4.5	5.5	6.4
Total U.S.	11.7	14.8	14.4

8. Conclusions

In conclusion, it cannot yet be proven that the changes to the CE’s race edit imputation methodologies are increasing the accuracy of those imputations. However, the changes do seem to be improving the number of imputations which can be later confirmed. Additionally, the overall proportions of Black/African-American nonrespondents is getting closer to the true values known from the census.

We remain confident that the changes are improvements. Hopefully after examination of more years of data in the future, the change in accuracy will become more visible.

9. References

Krieger, S., Steinberg, B., and Swanson, D. 2013. “Improving the Race Edit in the Consumer Expenditure Survey.” Paper presented at the Joint Statistical Meetings, Montreal, QC, Canada.

Bureau of Labor Statistics, Washington, DC. Handbook of Methods: Consumer Expenditures and Income, 2016 edition. http://www.bls.gov/opub/hom/cex/home.htm

National Center for Health Statistics, 2003. “United States Census 2000 Population with Bridged Race Categories.” Published in Vital and Health Statistics, Series 2, Number 135.

American Factfinder – Census Bureau, https://factfinder.census.gov/ (used for 2000 and 2010 census data)