Improving the Edit Process for the Public Libraries Survey While Migrating to a Web-based Format

Joanna Fane McLaughlin, Terri L. Craig, Economic Statistical Methods and Programming Division
Patricia O'Shea, Governments Division
U.S. Census Bureau, Washington, DC 20233-9100

Abstract

Increasing respondent burden and processing costs associated with editing Public Libraries Survey (PLS) data, along with the desire to improve data quality and ensure edit compatibility in the PLS' conversion to a web-based survey, drove our research on improving the PLS edit process. Using the old edit software, analysts followed up unnecessarily on a number of valid responses. Furthermore, some large errors escaped the edits altogether. With this in mind, our research focused on applying edit methodology in the appropriate places in the PLS data collection and processing phases to improve the overall edit process. We modified or removed malfunctioning edits, examined relationships between data items, researched competing edit methods, selected appropriate ratio edits, assigned the parameters for these edits, and determined which edits will go online.

Keywords: editing, edit parameters, macro-edits, micro-edits, ratio edits, resistant fences, web-based survey

1. Introduction

Data editing is an essential part of the survey process. In the past, editing typically followed data collection. However, due to technological advancements in computer hardware and software, editing now often follows data entry. Statisticians and programmers incorporate edits into computer program code in ways that capture and immediately report issues with data quality.

For establishment surveys in particular, a major objective is to continuously improve the edit process. This includes incorporating new and better editing methods, accounting for changes to the questionnaire, and updating existing parameters.

While edits are necessary in the survey process, they must be balanced with other concerns. In particular, they must be efficient and minimize respondent burden and survey processing costs. Edits should not flag items unnecessarily, resulting in the need for edit follow-up.

2. About the Survey

The PLS, sponsored by the National Center for Education Statistics (NCES) and conducted by the U.S. Census Bureau, is an annual census of public libraries and their public service outlets. The survey is administered to the more than 9,000 libraries in the 50 states, Washington, D.C., and the outlying areas of Guam, the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands. The survey collects identifying items such as name, address, and telephone number as well as items on population of the legal service area, full-time equivalent staff, service outlets, public service hours, library materials, operating income and expenditures, capital outlay, total circulation, circulation of children's materials, reference transactions, library visits, children's program attendance, interlibrary loans and several items on electronic services.

The PLS is administered via a Computerized Self-Administered Questionnaire (CSAQ). Starting with fiscal year (FY) 2005, the survey goes to a web-based format. Data is submitted on a voluntary basis. The Federal–State Cooperative System (FSCS) for Public Library Data collects the data. State Data Coordinators (SDCs) administer the FSCS. The SDCs collect the requested data from public libraries and submit the data to the NCES.

There are over 250 edits for 51 PLS data items. Data editing for the PLS is very thorough because the NCES displays individual library data on its website. The data are useful to federal, state, and local policymakers; library and

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

public policy researchers; and journalists and the public. NCES data products include state summary files and public-use data files.

The response rate for the PLS is generally high. For example, total non-response for FY 2003 was less than 3%! An incentive for responding is funding from federal, state, and local governments. SDCs overseeing data collection also help maintain a high response rate.

3. Current Editing Procedures

Prior to releasing the survey, programmers incorporate edits into the survey software, WinPLUS. Each SDC receives a copy of the survey software. After data entry, the SDC selects the software's "edit check" option, which generates an edit report. The SDC annotates and submits the edit report. Annotation includes verifying that the data flagged (i.e., in question) is in fact correct. If a library misreported data, the SDC has the opportunity to make changes.

Edit follow-up is the second stage of data editing. At this stage, NCES and Census Bureau analysts review the edit reports. The analysts also review data item aggregates, or may run edits that are not part of the software, also referred to as "internal edits." Analysts take any outstanding questions about a state's reported data to the SDC. Edit follow-up is often a lengthy process, because an SDC may have to re-contact libraries before responding to analyst questions.

4. Edits for Research

In general, edits are broadly classified as micro or macro edits. Micro editing occurs at the unit level, and macro editing occurs at the population or at a subset of the population level.

For this research, unit level edits were classified into one of three categories offered by Anderson et al. (2003): validation, consistency, or reasonableness. Validation edits include mandatory field, logic error, and range checks. Consistency edits include parts summing to totals, ratios of current year items, or current-to-prior year ratios that check that the data are consistent. Reasonableness edits include macro edits or comparisons of item totals.

Consistency edits – in particular, ratio edits – were a research priority because many parameters are no longer valid. For most surveys, consistency edits are not included in the survey instrument. Instead, they are run separately, as internal edits. This is not the case for the PLS. The SDCs are a sophisticated group of respondents, and data are released at the unit level. In large part, researching these edits for inclusion in the survey instrument is what separates our research from most.

5. Methodology

Our research goals included: determining which items are inadequately edited, adding new edits and updating edit parameters, determining which edits to place in a web instrument, and determining which edits to perform in a post-collection in-house review. Edit review, adding new edits, and updating edit parameters are discussed more in the following sections.

5.1 Edit Review

We reviewed every PLS edit for duplication and correctness. For each edit, we went through the following steps:

a. Classifying each as a validation, consistency, or reasonableness edit
b. Deciding to keep, or recommend deleting based on past performance, changes to the questionnaire, etc.
c. Considering modifications such as updating parameters or changing the edit conditions based on item definitional changes

The following example for the data item Library Visits illustrates this process. (Library Visits is the number of persons that entered the library during the year.)

These are three of the edits we reviewed for Library Visits:

a. Current year (CY) Library Visits is reported as 0 or filled in by the software as -2 (i.e., left blank)
b. Prior year (PY) Library Visits was reported as -1 (i.e., needs imputing) and CY Library Visits is reported as 0
c. The CY to PY ratio of Library Visits is outside the range [0.58, 1.90]

Edit "a" is a validation edit that we will keep. We see no need to modify the edit, because, by definition, the number of library visits should be greater than zero. Edit "b" is also a validation

edit. Subject matter specialists recommended deleting this edit, because the number of library visits should never be zero. Furthermore, cases that report a zero are caught with the previous edit. Edit "c", a ratio edit, falls under the category "consistency edit." We will keep this edit, but will update the parameters based on recent reporting patterns.

5.2 New Edits

For some items we decided to add one or more edits. Deciding to add a new edit was based on reasons such as respondent feedback, reporting patterns, and statistical research. Two of the new edit formats are given below:

- Item A is 0, and item B is greater than or equal to 0
- The CY ratio of item A to item B lies outside a predetermined range.

When developing an edit, the first step is to review item correlations. The second step is to confer with subject matter specialists about the plausibility of a new edits for a given data item.

Regarding the first step, the items we choose to compare should be highly correlated over at least three consecutive survey cycles. To achieve this, we looked at correlations over multiple survey cycles to make sure the items remain highly correlated over time. In some cases, like for State Government Revenue and Salaries, the data appear stabile over two cycles, but are clearly volatile across three or more survey cycles. For example, in fiscal year (FY) 2003 State Government Revenue was highly correlated with Salaries (.89), but in FY 2002 they were only mildly correlated (.59).

Regarding the second step, we need to know which relationships subject matter specialists deem most important. In other words, two items that are highly correlated are not necessarily a priority for review. Subject matter specialists often offer insight into the data that methodologists were not previously aware of.

As a quality assurance step, analysts test new ratio edits internally for at least one survey cycle before migrating them to the survey instrument. Of the records flagged internally, analysts might follow-up on the most unusual reported values.

5.3 Establishing Parameters

We examined item distributions to establish non-ratio edit parameters. For example, if a library reported having no librarians the previous year, we want to know a reasonable non-zero response to expect in the current year. We examined the current year item distribution for all libraries that reported no librarians in the prior year and determined, with the help of analysts, a reasonable cutoff.

For establishing ratio edit parameters, we want a method that is repeatable and works well for different distributions. For example, revenue and expenditure items will not have the same distributions as staffing and electronic resource items. A method that does not exclude "small" libraries is also important, because the data are published at the unit level. Equally important is finding a method that requires minimal parameter updating each year. We considered three methods for establishing ratio edit parameters: the current method, the Hidiroglou-Berthelot (HB) method, and a resistant fences method.

For each method, we compared the data after each stage of editing: pre-follow-up and post-follow-up. Using FY 2003 data, we generated a list of records that required analyst corrections, including item values before and after follow-up. We examined which of the three methods is best at flagging records that were actually changed, while not flagging more records than necessary. Ideally, we wanted to use "dirty" data. That is, we wanted to compare data before any editing took place (i.e., before any edit reports were generated) to data after edit follow-up, but we did not have access to entirely dirty data.

5.3.1 Current Method

Under the current method, data is used from previous survey cycles to calculate predetermined bounds:

$$\bar{r} \pm 2.5 * se(\bar{r}),$$

where $\bar{r}$ is the mean ratio, whether it be a current ratio or current-to-prior ratio. Values that fall outside the bounds are flagged as outliers.

5.3.2 Hidiroglou-Berthelot Method

Hidiroglou and Berthelot (1986) describe another method for detecting outliers. For some data

item x at times t and t+1, calculate the ratio, r, for the i^{th} case.

$$r_i = x_i(t) / x_i(t+1), \; i = 1, \ldots, n.$$

After calculating r_i, transform the ratio.

$$s_i = \begin{cases} 1 - r_M / r_i, & \text{if } 0 < r_i < r_M \\ r_i / r_M - 1, & \text{if } r_i \geq r_M \end{cases},$$

where s_i is the transformed ratio, and r_M is the median of the ratios. Hidiroglou and Berthelot recommend transforming s_i, using a parameter, U, to provide "a control on the magnitude of the data." Specifically, the following transformation, E_i, ensures more "weight" is given to small change in a large unit than large change in a small unit.

$$E_i = s_i \left\{ \text{Max}\left(x_i(t), x_i(t+1)\right) \right\}^U,$$

where $0 \leq U \leq 1$. Next, calculate the quartile deviations d_{Q1} and d_{Q3}.

$$d_{Q1} = \text{Max}\left(E_M - E_{Q1}, \left|AE_M\right|\right).$$
$$d_{Q3} = \text{Max}\left(E_{Q3} - E_M, \left|AE_M\right|\right).$$

They suggest 0.05 for A. E_{Q1}, E_M, and E_{Q3} are the 1st quartile, the median, and the 3rd quartile, respectively.

Finally, outliers fall outside the interval (E_M-Cd_{Q1}, E_M+Cd_{Q3}), where C is some value larger than zero used to control the width of the confidence interval.

5.3.3 Resistant Fences Method

Thompson (2001) describes the resistant fences method for developing acceptable ranges.

Ideally, the data are symmetric. A log-transformation is recommended for a sample size $n \geq 50$ and moderately skewed data ($s_k > 6.75$). There is a variation on the method for asymmetric data from small samples, but here we address only the method for symmetric data, because we successfully used the natural logarithm to symmetrize the PLS data.

Given symmetric data, for a constant k, the resistant fences method flags a value if it is k interquartile ranges outside the 1st or 3rd quartiles. Defined values of k are 1.5 (inner fences), 2 (middle fences), and 3 (outer fences). The interval is:

$$(Q1-k*H, Q3+k*H),$$

where $Q1$ and $Q3$ are the 1st and 3rd quartiles, respectively, and H is the interquartile range, $Q3$-$Q1$.

6. Results

The Resistant Fences and HB methods each flagged fewer records than the current method, while still capturing the majority of the records that were actually changed by analysts. We tried the HB method at both national and state levels. Of the two, records were more often flagged incorrectly at the national level. The current method flags a number of unnecessary records.

Given that both the HB and resistant fences methods showed promising results, we examined the methods separately. Initially, we focused on the HB method because it uses incoming data to develop edit parameters; we do not have to establish parameters prior to releasing the survey. However, there is no guarantee that a SDC's computer is equipped with the statistical software required for this method. Moreover, the HB method focuses on questionable values that largely affect population totals, and we want to use a method that also focuses on editing at the unit level.

For editing at the unit level, the resistant fences method is a better tool for our immediate purpose. Moreover, it is a statistical method for setting tolerances that works well for different sets of economic data. It is likely to capture large value changes, regardless of an establishment's size. However, it does require that methodologists develop new parameters prior to the survey's release each year. Also, the method does not "work" nicely for all distributions. Specifically, we ran across four problematic scenarios:

a. A narrow interquartile range
b. A large frequency of reported zeros for a given data item
c. The same value of k does not work for each data item
d. No parameters look reasonable

6.1 Narrow Interquartile Range

The reported values for some data items are unlikely to change much from one year to the next. For example, it is quite possible a library keeps the same number of internet terminals from one year to the next. For such cases, *Q1* and *Q3* – unsymmetrized – both have a value of one, or close to one. It follows that the interquartile range is zero, or close to zero. This means we would flag all or almost all of the more than 9,000 records, which is clearly unreasonable. If we remove ratios of one from the calculations, the interquartile range becomes large enough that we flag a reasonable number of records.

6.2 Large Frequency of Reported Zeros for a Data Item

For some data items, a reported zero is not uncommon. For example, it is possible that some libraries do not have any librarians with a Masters degree. Therefore, for the prior year ratio of librarians with Masters degrees, it is likely that all values will fall outside the range calculated using the resistant fences method.

Instead of flagging all units that report having no librarians with a Masters degree, we have a separate edit for units that reported zero in either the current or prior year. For example, we examined the distribution of libraries that reported no librarians with a Masters degree in the prior year. Based on the distribution, we chose a reasonable cutoff for the current year value (as previously discussed in section 5.3). This method is subjective, but may work well if statisticians and subject matter specialists agree on an upper bound.

6.3 The Same Value of k Does Not 'Work' for Each Ratio

Initially, we tried k=3 for all of the current year ratios. This value of k gives the desired result in many cases, but does not always provide reasonable parameters. For example, using k=3 for the ratio Salaries to Total Staff gives an upper bound of $139,014. This is clearly unreasonable. The solution is to adjust the value of k, trying k=2 and k=1.5 as suggested by Thompson (2001).

Subject matter specialists are the best resource for determining whether or not parameters are reasonable. As Thompson (1999) notes, "Analysts who work with economic data develop an expert understanding of the distributions of ratios in a given industry." Subject matter specialists reviewed all parameters prior to approval.

6.4 No Parameters Look Reasonable

Sometimes no parameters look reasonable after trying different values of k. Perhaps the item was introduced to the survey recently and librarians still have questions about the item's definition. More likely is that the resistant fences method does not work well for the item's distribution – the method works well for most, but not all, distributions. Therefore, we may rely on subject matter specialists to advise us on reasonable edit parameters.

7. Conclusions

We are currently in the quality assurance stage of our research. This stage includes running all the new and modified edits with FY 2004 data and having analysts and methodologists review the output.

At this time, most edits will go into the survey instrument for FY 2005. Only edits for newer data items will remain out for internal testing. The only other in-house edits will be those requested at a later date by the survey sponsor.

Incorporating the edits into the web design is the final step. For the edit parameters, the hope is to establish a parameter file. A parameter file requires less programmer burden. Methodologists will update the edit parameters yearly, and programmers will read the file into previously written computer programs.

8. Topics for Future Research

As resources become available, there are topics we still want to address. One topic is: Would improving the questionnaire be one way to reduce edit failures? Another topic is: Do we effectively minimize the number of edit messages? A final thought is: Can we use the HB method for internal editing, or in the future as part of the web collection?

8.1 Improving the Questionnaire

If we think in terms of editing starting at the questionnaire design stage, we might consider adding resources to the survey design phase. Granquist (1995) notes that, "Form design problems are responsible for a significant number of respondent errors." Granquist goes on to note that, "Improving questionnaire design would improve the quality of incoming data." In this vein, it is important that the PLS survey goes through adequate pre-testing before it is released.

8.2 Generating Fewer Error Messages if an Item Fails Multiple Edits

Currently, if an item fails more than one edit, multiple error messages are shown. The SDC must account for each error message, often causing significant respondent burden. To reduce respondent burden, we do not want to flag an item more than once for failing multiple edits.

Error localization is a common method used to reduce respondent burden. An error localization program looks for the minimum number of fields to change to satisfy an edit, and makes the changes (Garcia 2003). Error localization will not necessarily solve the problem for PLS, because only respondents (including SDCs) may change reported values. Therefore, we need to research another way to minimize the number of error messages for a given data item.

8.3 Using the HB Method To Edit Internally

As previously noted, the NCES releases PLS data summary tables at the state level. For reported and imputed data, analysts review tables of item totals, comparing current and prior year values. For a more statistical approach, we want to test the HB method as a possible internal, macro-editing approach.

Acknowledgements

The authors would like to thank Cynthia Ramsey, Laura Hardesty, and Johnny Monaco for their help on this project. The authors would also like to thank Carma Hogue for her help on this project and for her many helpful comments on this paper.

References

Anderson, A., et al. (2003), "Changes to Editing Strategies when Establishment Survey Data Collection Moves to the Web," presented to the Federal Economic Statistical Advisory Committee, March 21, 2003, Washington, D.C., U.S. Bureau of Labor Statistics.

Garcia, M. (2003), "Error Localization and Implied Edit Generation for Ratio and Balancing Edits," Statistical Research Division, U.S. Census Bureau.

Granquist, L. (1995), "Improving the Traditional Editing Process," Cox, Binder, Chinnappa, Christianson, Colledge, and Kott (eds.), Business Survey Methods, John Wiley & Sons, pp. 385-401.

"Graphical Review of Economic Data: Final Report of the Graphical Analysis Working Group," Economic Statistical Methods and Programming Division, U.S. Bureau of the Census (1994).

Hidiroglou, M. and Berthelot, J. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," *Survey Methodology*, Vol. 12, pp. 73-83.

Thompson, K. J. (2001), "Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods," Statistical Policy Working Paper 29, a Federal Committee on Statistical Methodology Conference Paper available online at http://www.fcsm.gov/99papers/thompson.pdf.

Thompson, K. J. and Sigman, R. (1999), "Statistical Methods for Developing Ratio Edit Tolerances for Economic Data," *Journal of Official Statistics*, Vol. 15, No. 4, pp. 517-535.

WinPLUS User's Guide Version 2.5: Guide for Reporting Data for the Public Libraries Survey, FY2004 Using the Windows Public Library Universe System Software (November 2004).