A Comparison of Two Ratio Edit Methods for the Annual Survey of Government Finances

Elizabeth Cornett, Joanna Fane McLaughlin, Carma R. Hogue,
Office of Statistical Methods and Research for Economic Programs,
Stephen D. Owens, Governments Division
U.S. Census Bureau[1], Washington, D.C. 20233-9100

ABSTRACT

The Annual Survey of Government Finances (ASGF) is conducted yearly to collect data on revenue, expenditures, debt, and assets of state and local governments. The questionnaire was redesigned for 2005. As a result, some data items were consolidated and others were split into multiple data items. These changes required a redesign of the ratio edits and the methods for establishing the ratio edit bounds. Historical edits were no longer possible. New current year edits had to be determined.

We researched two methods for editing ASGF data at the unit level: resistant fences and Hidiroglou-Berthelot (HB). The first is an editing technique that is designed to work well with different distributions. The second technique focuses on small changes in large units versus large changes in small units. In this paper, we compare these methods for the ASGF.

KEY WORDS: ratio edits, resistant fences, asymmetric fences, Hidiroglou-Berthelot, selective editing, Annual Survey of Government Finances

1. Introduction

The questionnaire for the Annual Survey of Government Finances (ASGF), which is conducted by the Census Bureau, was redesigned for 2005 in an effort to reduce respondent burden and collect better quality data. The new questionnaire is available online at http://www.census.gov/govs/www/surveyforms.html. After changing the data collection instruments, it was also necessary to redesign the edit, imputation, and estimation processing systems. Our main focus was on the edit system, which is used to reduce nonsampling error. In effort to reduce the processing time and improve the quality of the data, we first examined the logical edits that were present in the existing system before turning our attention to the ratio edits. Most of the ratio edits in the prior system were current to prior period comparisons. There were no ratio edits to check the data integrity within the questionnaire. A team was organized to establish new logical edits and current year ratio edits with their tolerances for 2005.

After finalizing the logical edits and the current year ratio edits, the last phase of research for the annual survey will be to examine the tolerances for the current to prior year ratios. Since the questionnaire for 2005 is new, ratios of 2005 to 2004 are not valid. Consequently, historical ratio edits cannot be introduced until the 2006 survey year.

The ASGF collects data on revenues, expenditures, debt, and assets from a probability proportional to size sample of state and local governments each year. Roughly 11,400 governments were selected from a population of about 74,000. In years ending in 2 and 7, the same set of questionnaires is used to collect data from all state and local governments. Some of the collected data are quite volatile. For example, debt can be incurred to build a new facility, or capital outlay can rise rapidly as a new building is constructed. Likewise, intergovernmental revenue can be volatile from year to year as federal, state, and local governments decide to change their level of funding of other governments. Other variables like taxes, charges, and non-construction expenditures are quite stable over time.

The team consisted of analysts and mathematical statisticians. During the first few months of the project, the analysts focused on the logical edit processes while the mathematical statisticians studied various papers on selective editing, the

[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.

Hidiroglou-Berthelot (HB) method of detecting outliers, and resistant fences methods for setting edit tolerances in an effort to find the best method(s) to use for the ratio edits. After looking at correlations, the analysts selected current-year ratios of interest to them. The team initially focused its attention on a ratio of Annual Payroll Expenditures from the ASGF to a similar annualized payroll variable from the Survey of Public Employment and Payroll. The HB method and the resistant fences methods were analyzed for this ratio. After examining this ratio in detail, other variables were examined.

Selective editing was thought to be a possibly viable method for editing special districts (governments that generally have one purpose, for example, an airport authority) and data from supplements (agencies that supply additional data for cities, counties, and towns if the parent government cannot supply that information). These units typically answer far fewer data items than a general purpose government. General purpose governments (cities, counties, and townships) have multiple purposes and can be expected to supply over a hundred items. The only way to use selective editing with these governments would be to use the procedure to assign editing priorities after the ratio edit failures have been determined. Selective editing can also be used to determine the very worst cases for analyst examination prior to running the estimation programs. Our experiences with selective editing will not be discussed in this paper. Likewise, our experience with the construction of logical edits will not be discussed.

In section 2 of this paper, we discuss the approach that we took in our research of ratio edit parameters. We discuss the resistant fences approach to determining ratio edit bounds in section 3. Our experiences using the HB approach with the ASGF data in determining outliers are discussed in section 4. Our combined conclusions and a discussion of the methods chosen by our analysts are presented in section 5. Topics for future edit research are presented in section 6.

2. Research Approach

The research study had to be divided into several parts. We started our research with separate examinations of general purpose and special district governments on a national level. As mentioned in the previous section, general purpose governments are quite different from special districts. They report revenues, debts, expenditures, and assets for many more functions than do the special district governments. For example, while a questionnaire from a county will report financial data for taxes, charges, intergovernmental revenue, and expenditures (highway, hospital, airport, housing, welfare, etc.), a special district generally will report a small subset of that information because they are usually operating a government with a single purpose. The magnitude of total revenue, etc., would usually be much smaller than the totals for general purpose governments. Therefore, it was decided that the special districts would have to be examined separately from the other governments. When it is feasible (like in the 2007 Census), cities, counties, and towns are also examined separately. In order to use the HB and the resistant fences methods, as in Thompson (2000), we decided that a cell must have at least 20 responses before we would run the comparison.

To start our study, we decided to examine national by type (cities, counties, towns, special districts) of government edits. From past experience, we know that eventually we will want to look at state (or census division or region) by type of government edits. We used 2002 reported, unedited census data for our research. Since we are developing the parameters based on units in the 2004 sample, we matched to the 2004 Government Identification File and used only those matching sample units along with their weights in the first phase of the study. As we get closer to the 2007 census, we will examine the parameters that are yielded by using the entire census.

The data is edited in phases. For a general purpose government, the data are often collected from a number of sources. If the government has schools that are dependent on the government, these data will be submitted using a different questionnaire. Larger governments often fail to supply data from dependent agencies, so we get supplemental data using the special district government questionnaire. A government with all of its dependent schools and supplements is called a “parent” government. In the edit process, supplemental data will be edited with the special district data as they are received. A parent government will be edited after the main government has been aggregated with all of its dependent parts. The dependent parts will have been edited, but the main government will not be edited until it is edited as a complete parent government. Consequently, the research was done in parts. We researched parameters that could be used to edit special districts and supplements together. We then researched parameters for the aggregated parent governments. The questionnaires for the 48 largest local governments (referred to as jacket units) are not edited using this system, so they were left out of the research files.

There are some functions like various utility expenditures where we do not have enough cases to examine edit parameters separately for each type of government, even on a national level. It was decided that cities and townships would always be collapsed together. When there are not enough responding counties, they will be collapsed with the cities and townships. In this paper, we will examine some of the results from the national by type of government collapsing.

Note that reported unedited data are not available for California, so this state has not been included in the study. Separate studies will be needed later to determine what is best for California. Until that time, national by type of government edit parameters will have to be used to edit California data. Also, in preparing our files for our research, we did not include any records where either the numerator or denominator variable was zero. Such cases will be edited using a logical edit when the change in the data exceeds a predetermined amount.

In looking at special districts, we examined the data by function code, i.e., a code which defines the special district as an airport authority, a utility, a sewer district, etc. For those units where the data were sparse, a predetermined collapsing mechanism was used to gain enough units for our research.

As previously mentioned, we started our research with an examination of a new ratio edit that compares the Finance Survey's annual salaries data to the Employment Survey's payroll. Other ratios include Intergovernmental Revenue to Nontrust Revenue, Total Salaries to Current Operating Expense, General Revenue from Own Sources to Nontrust Revenue, and six other similar ratios of aggregates. We also examined 13 ratios of less aggregated variables, such as air transportation revenue to air transportation expenditure, education revenue to education expenditure, etc.

3. Fences Methods

Although resistant methods are designed to deal with different distributions, the resistant fences rules implicitly assume symmetry. For some ASGF ratios, approximate symmetry is a reasonable assumption. For most, it is not. In an effort to develop meaningful ratio tolerances for those distributions where the assumption fails, we applied two variations of the resistant fences method to the ASGF data: resistant fences rules on symmetrized data and asymmetric fences.

3.1 Resistant Fences (RF) Method

For some researched ratios it was possible to form an approximately symmetric distribution using a log transformation of the data. For these ratios the resistant fences rules were applied to the symmetrized data.

Thompson (2001) describes the resistant fences method for detecting outliers. Given an ordered distribution of ratios, let q_{25} be the first quartile, q_{75} be the third quartile, and H be the interquartile range (q_{75}-q_{25}). [H should be greater than 0 in order for this method to be effective.] Ratios less than $q_{25} – k*H$ or greater than $q_{75} + k*H$ will be flagged. The value substituted for k defines the fences rule. Generally accepted values of k, and those used in our study, are 1.5, 2, and 3 corresponding to the inner, middle, and outer fences rules respectively.

3.2 Asymmetric Fences (AF) Method

When the distribution of a ratio was moderately skewed, it was possible to apply the AF method. The AF method takes the skewness of the data into account and elongates the bounds in the direction of the longer tail of the distribution.

The AF method, as described by Thompson (2001), defines outliers to be ratios that are less than $q_{25} – k*(m – q_{25})$ or greater than $q_{75} + k*(q_{75} – m)$, where m is the sample median. For this method, the traditional values of k, and those used in our study, are 3, 4, and 6, corresponding to the inner, middle, and outer fences rules respectively.

3.3 Application

Following guidelines set forth by Thompson (2005, 2001), cell size, degree of skewness (s_k), and correlation of ratio items were used to determine the initial method and fences rule to apply to each ratio.

For small cell sizes, $20 \leq n < 80$, the AF method was applied. For medium ($80 \leq n < 3{,}000$) and large ($n \geq 3{,}000$) cell sizes, degree of skewness was used to help determine the initial method used. A skewness value of 0 indicates the distribution is perfectly symmetric. For large cell sizes, highly skewed ($s_k > 10$) ratios were transformed using the natural logarithm, and the RF method was applied. If the transformation did not achieve approximate symmetry, then the AF method was used. For large cell sizes that were only moderately skewed ($s_k \leq 10$) the AF method was applied. A literature review revealed no conclusive findings have been reported for medium sized cells. Consequently, for this cell size, both the RF and AF methods were used in our initial stages of parameter development.

Once the initial method was determined, the fences rule was chosen based on the correlation between ratio items. According to Thompson (2005, 2001), if ratio items are highly correlated, $\rho \geq 0.70$, both inner and middle values of k are appropriate. For ratio items that are not highly correlated, $\rho < 0.70$, the outer fences rule is most appropriate.

3.4 Problems

When working with the ratio of Personnel Expenditures from ASGF to a similar payroll variable from the Survey of Public Employment and Payroll, we encountered a problem with certain special district function codes. As mentioned above, in order to effectively apply the RF method, the value of interquartile range must be greater than 0. For several function codes the interquartile range was 0. This was a direct result of having most ratios equal to 1.

After consulting with the analysts we learned that for some function codes it is not surprising to have all ratios equal to 1. For these function codes we did not develop tolerances. For the others, because a value of 1 is desired and expected for this ratio, ratios equal to 1 would pass the edit. The solution then, was to remove all ratios of 1 for these function codes (assuming they would pass the edit) and then develop bounds using the remaining data.

3.5 Findings

In general, cell sizes for general purpose units were almost always classified as medium and the correlations classified as high. As a result, we used the inner and middle fences rules for both the AF and RF methods in our initial development of tolerances for all general purpose types. After analysts viewed the total number of failures and individual IDs that failed each method they decided the bounds created using the inner fences rule were too narrow. This was a direct result of too many acceptable ratios being flagged. Consequently, our second attempt at bounds development involved running both the AF and RF methods with the middle and outer fences rules for all general purpose types.

Using the asymmetric fences method often resulted in the development of a negative lower bound. Having a negative lower bound is undesirable because all units with a value in the denominator greater than that in the numerator automatically pass the edit. To resolve this issue, we ran the asymmetric fences method twice, once for the original ratio and once for the inverse ratio. This allows no values to be overlooked.

4. The Hidiroglou-Berthelot Method

The Hidiroglou-Berthelot (HB) method has a number of promising features for editing the ASGF. One such feature is that it does not require that we develop edit parameters prior to data collection. With the right statistical software, analysts can simply run submitted data through a computer program that outputs flagged records (McLaughlin 2005). Another promising feature of the HB method is that it focuses on questionable values that largely affect population totals – an important requirement for editing ASGF data.

4.1 Methodology

Hidiroglou and Berthelot (1986) offer this method to edit periodic data from a sample of *n* units. For unit *i*, given variable *x* at times *t* and

t+1, we can edit the ratio $r_i = x_i(t+1)/x_i(t)$, where i=1,…,n, by doing a series of transformations. The first transformation on r_i is:

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

In this equation, r_m is the median ratio over all i. The next transformation is:

$$e_i = s_i \{ \max(x_i(t), x_i(t+1)) \}^u ,$$

where 0 ≤ u ≤ 1. Hidiroglou and Berthelot recommend using u to allow "a control on the magnitude of the data." In other words, the larger the value of u, the more importance we place on values with large influences on population totals.

The final step before calculating the confidence interval is to find the quartile deviations d_{q1} and d_{q3}:

$$d_{q1} = \max(e_m - e_{q1}, |ae_m|),$$
$$d_{q3} = \max(e_{q3} - e_m, |ae_m|).$$

In these equations, e_{q1}, e_m, and e_{q3} are the 1st quartile, median and 3rd quartile, respectively. Hidiroglou and Berthelot recommend a value of 0.05 for a. We also found in our research that this is a reasonable value. An outlier is an e_i that falls outside the interval (e_m-cd_{q1}, e_m+cd_{q3}), where c is a constant that controls the width of the confidence interval, and hence, the number of records that are flagged. Values of c vary from survey to survey and ratio to ratio and are largely dependent on the resources available for editing.

For this research it was important to extend the HB method for use with a current year ratio, $r_{curr,i} = x_i(t)/y_i(t)$. Sigman (2005) shows that the difference lies in how we treat e_i. For the current ratio:

$$e_i = s_i \{ \max(x_i(t), r_{curr,m} * y_i(t)) \}^u ,$$

where $r_{curr,m}$ is the median current year ratio over all i. Similarly, we made the necessary adjustments to our computer code.

4.2 Application

When the HB was tested on the ratio Finance Payroll Expenditures to Employment Payroll Expenditures, the parameters were initially set as follows: a=0.05, u=0.5, and c=10. After reviewing the output, it was noted that: 1) smaller values than expected are flagged, 2) more values are flagged than analysts expect to review, and 3) too many "reasonable" values are flagged. To address 1), for payroll expenditures and all other ratios, the value of u was set to 1. To address 2) and 3), the confidence interval was widened. Specifically, analysts were given output for different values of c - 15, 20, and 25 - to compare.

5. Combined Conclusions

As was mentioned in section 2, edit parameters were developed for editing supplements and special districts together. When these units were initially combined for edit research, there was concern that the distributions of supplements and special districts were different, because the supplements' weights are pulled from the parent records, and the special districts have their own weights. To ensure that we were making valid comparisons, we tested the difference in the means of the ratio Total Revenues to Total Expenditures for supplements and special districts. We conducted the significance tests at the national and the national by function code levels, for weighted and unweighted data.

At the national by function code level, the differences were not significant (α=0.05). At the national level using unweighted data, the difference was not significant (α=0.05). However, at the national level using weighted data, the difference was significant (α=0.05). Given that there are no function codes for some supplements, this is what we would expect. Therefore, whenever editing supplements and special districts together, it must be done nationally by function code.

For each ratio, analysts received lists of IDs that failed the following methods: AF middle and outer, RF middle and outer, and HB conservative (c=60) and moderate (c=30). After reviewing all the general purpose output in depth, analysts decided that the AF middle and HB methods were better than the RF methods for editing ASGF data at the national by type level.

The analysts then examined the individual units that each method (AF middle and HB) flagged for closer review. As expected, the HB method did an excellent job of flagging units that would largely impact national level estimates. However, we determined that state level estimates are currently our primary concern. The AF middle method did a good job of flagging units that would largely impact state level estimates. Keeping this in mind, the analysts chose the AF middle method for bound development and implementation in 2006.

We believe that the HB method would be the method of choice if the edit cells were formed at the state by type of government level versus a national by type of government level. Because the edit cells were national by type the bounds created by HB were heavily influenced by large states like New York. Consequently, units impacting smaller state level estimates were not flagged for review. Due to time constraints we are unable to research this now. When there is time, this will be researched and if the HB method fares well at the state by type level it is possible it will be implemented in the future.

Analysts completed a similar review for special district governments and the conclusions were the same. For 2006, edit bounds will be created using the AF middle method for all ratios. The HB method will be researched at the state by function code level when time permits.

6. Topics for Future Research

After we have two years worth of data available using the new questionnaire format, we will examine edit parameters for historical ratio edits (current year to prior year ratios).

As we started our research, we had one overriding question to answer: Which method is the best for determining edit parameters that will fail true errors, while reducing analyst workloads for each type of edit? Over the next few years, we want to answer several other questions: Should state by type of government parameters be used, or will national, regional, or census division by type of government be sufficient? How often should the parameters be recalculated? Do the methods hold up over time? Are the best methods for determining edit parameters for the surveys also the best methods for the census? Over the next few survey years, we hope to answer these questions with future research.

7. References

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

McLaughlin, J. F., Craig, T. L., and O'Shea, P. (2005), "Improving the Edit Process for the Public Libraries Survey While Migrating to a Web-based Format," Proceedings of the Joint Statistical Meetings, 7-11 August, Minneapolis.

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. (2005), "Applications of Resistant Fences Techniques in the Economic Directorate," internal presentation to U.S. Census Bureau employees, November 22, 2005, Washington D.C., U.S. Census Bureau

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.

Thomson, K. J. and Hostetter S. L. (2000), "Investigation of Selective Editing Procedures for the U.S. Bureau of the Census Economic Programs," Proceedings of the Second International Conference on Establishment Surveys, 17-21 June, Buffalo, NY

Sigman, R. (2005), "Statistical Methods Used to Detect Cell-Level and Respondent-Level Outliers in the 2002 Economic Census of the Services Sector," Proceedings of the Joint Statistical Meetings, 7-11 August, Minneapolis.

ACKNOWLEDGEMENTS

The authors would like to thank Terri L. Craig, Loretta A. McKenzie, Aaron Gray, and Stan Batton for their help on this project. The authors would also like to thank Rita Petroni and Maria Garcia for their helpful comments on this paper.