DEVELOPMENT OF EDIT PARAMETERS FOR 1992 ECONOMIC CENSUS ENTERPRISE REPORTS

Sungsoo Oh, Dave Paletz, Jay Kim, and Eddie Salyers
Bureau of the Census, Washington, D.C. 20233
Contact: Eddie J. Salyers, Agriculture and Financial Statistics Division, Bureau of the Census, Washington, D.C. 20233, Phone 301-763-7234

KEY WORDS: edit, imputation, SPEER, outliers, dendrogram

1. INTRODUCTION

The Enterprise Statistics program is part of the U.S. Bureau of the Census's Economic Census conducted every five years and includes the Large Company Report and Auxiliary Establishment Report. Both surveys collect financial statistics. The Large Company report is based on responses to questionnaires sent to companies with 500 or more employees. The Auxiliary Establishment report is based on questionnaires mailed to auxiliary units, namely establishments which support activities of other components of the company. Examples are research and development centers, warehouses, and administrative offices. The data from these enterprise statistics have been editing using SPEER (Structured Programs for Economic Editing and Referrals) system since 1982, see Greenberg and Petkunas (1990). SPEER makes use of upper and lower bounds of ratio values for items for detecting suspicious values (outliers) which may be erroneous reported data. Reported data outside the bounds are either "flagged" for analyst review or replaced with estimates by an automated system.

In previous Censuses these bounds were based on prior Census data from five years earlier with adjustments to account for inflation and the analyst's expert knowledge.

In the 1992 Economic Census, the Enterprise Programs increased coverage to include companies engaged primarily in finance; insurance; real estate; communications, electric, gas, and sanitary services; and selected transportation industries for the first time. The lack of prior data for several industries and our desire to develop a system that could be used to generate bounds from current unedited data led to the development of the methodology discussed in this paper.

2. BRIEF OVERVIEW OF SPEER

SPEER is designed to edit and impute continuous data. SPEER can greatly facilitate editing because it can determine the logical consistency of a set of edits prior to receipt of data. Also, for each record failing one or more edits it determines the minimum number of fields needing change.

SPEER checks the ratio of two data items (e.g., annual payroll to number of employees) against pre-determined bounds. If the ratio falls outside these bounds SPEER will impute one or both of the data items so that the ratio using the new value(s) falls within the accepted bounds. SPEER checks all predetermined ratios against respective bounds. Items involved in bounds failures are one by one marked for imputation based on a predetermined order of unreliability. The marking continues until no ratios formed from the remaining unmarked items fail edits. Marked items are then imputed in an agreed order and so that no impute causes an edit failure.

SPEER requires that parameters for edit bounds and imputation be determined ahead of time. Analysts who have expertise in the data to be edited are best qualified to do this. However, the analysts who design a survey sometimes have only limited expertise in this area or there may be new survey data for which experience does not exist as was the case here. Even if the analysts are well acquainted with the data they may wish to thoroughly examine it. This can be automated and systemized by applying a statistical or mathematical method to compute the bounds. Once computed the analysts can review these parameters and replace any or all of them own using any specialized knowledge they have. Analysts often will want to follow up the

automated application of the method with their own expertise. This knowledge cannot be fully programmed into any software.

The relatively large number of parameters required for the Enterprise edit provide an additional reason for development of an automated methodology for generating parameters. There were actually two types of data to be edited. The Large Company Report required us to compute 50 ranges and/or central values. We needed separate sets of parameters for each of 65 different industrial classifications. This brought the total parameters needed for Large Company data to 3250. We had to compute 29 ranges and/or central values for the Auxiliary data. There were 4 functional classifications amongst the Auxiliary data. So we needed 116 ranges for this data. Between the two types we had to calculate a grand total of 3366 ranges and/or central values. It was clear we needed automated methods to assist in determining edit bounds.

3. PARAMETER DEVELOPMENT

The methods used to automate the setting of edit parameters and improving the quality of these parameters include the use of cluster analysis and dendrograms, correlation studies, outlier tests, and difference criteria.

3.1 Correlation and Cluster Analysis for Establishing SPEER Parameters.

The first step in the development of the editing parameters is to determine closely related items whose ratios will be evaluated by the edit system. Values of these ratios can be compared to predetermined parameters to identify suspicious data.

We determined meaningful ratios using 1987 data. To do this we used correlation coefficients to test which data items could be reliably used to form ratios for the edit.

3.2 Cluster Analysis and Dendrograms

Once correlation analysis is performed on all possible data pairs, cluster analysis using dendrograms are used to test core items and satellite items. These were predetermined in previous surveys based on analyst's expertise. Core items consist of key variables that are interrelated and thus are edited by using all possible pairs among the items. Satellite items which are closely related with a core item are edited by using the core item. To test the close relationships among items, cluster analysis was performed.

Cluster analysis is used to discover grouping of data objects. Groupings can be made based on a measure of similarities between objects. The measure of similarities can be distances between objects or coefficients for pairs of variables. When the measure is correlation coefficients between the variables, cluster analysis can be employed using the correlations as a distance between objects.

We used SAS software to form clusters and develop dendrograms. Dendrograms illustrate cluster relationship among the items in a tree structure. The items in the closer branch of the tree diagram are more closely correlated. The items that are remotely linked represent items that are poorly correlated.

Once the core items and satellite items have been determined, the next step in developing edit parameters is the setting of lower bounds, upper bounds, and central values in order to determine outliers.

3.3 Data Preparation

To set the bounds from the current unedited data we had to remove obvious outlier cases that could be identified easily using standard robust statistical techniques. After outliers are removed from the data set (as described in this section), then we proceeded to use the remaining data points to obtain edit parameters as described in Section 4.

In previous surveys bounds were set (based on prior census data) at points that would "cut" the lowest and highest two percent of a set of ratio values. A study on the distributions of 1987 data showed skewness to the right (having a long right tail) for the majority of the item ratios. Cutting 2 percent of the right tail as in the 1987 approach tends to leave in outliers and resulted in setting the upper bounds wider than optimal bounds. Some data sets used in the edit were skewed to the left and others were fairly symmetrical. We employed a

methodology that determines the type of distribution of the data and then selects the appropriate outlier test.

Operationally, skewed distributions were assumed to be of Gamma type and non-skewed distributions were assumed to be of Normal type. An outlier test is done on each extreme observation(s) until the last extreme case(s) fits under the curve.

The following procedures explain how we removed outlier(s) from a given data set. The data are in ratio form for SPEER parameters.

A. First the elements of a given data set are listed in ascending order,

$$X_{[1]} \le X_{[2]} \le \ldots \le X_{[n]}$$

B. Then the data set is tested for skewness using the following test.

$$\text{let } m_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \quad \text{and} \quad m_3 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{n}$$

compute the value of skewness, $S_k = \frac{m_3}{m_2 \sqrt{m_2}}$

C. Statistical test of skewness

If S_k is significantly different from 0, the data are considered to be skewed and not to be best represented by a symmetrical distribution (normal, etc.). The test is described below.

Under the null hypothesis H_0: that the absolute value of

$S_k = 0$, the value of S_k is tested at the .5% level of significance ($\alpha = 0.005$). The level α was chosen at 0.005 due to the sensitivity of the skewness test on the ratio data sets.

If $\left| S_k \right|$ exceeds the .5% cutoff point with sample size n, then the data is skewed.

If $S_k > 0$, then it is **skewed to the right.** Otherwise, it is **skewed to the left.**

Otherwise, it is **not skewed (4c).**

The following section gives explains testing for outlier(s) in the three types of data distributions.

D. Testing outliers

Once the type of distribution is determined the appropriate outlier test is applied, see Barnett, V. and Lewis, T. (1978), pages 76-94.

i.) Skewed to the right (outlier test 1)

The test used is for a single upper outlier assuming a Gamma distribution and the orgin not equal to 0.

Let

$$\frac{x_n - a}{\sum_{i=1}^{n} (X_i - na)} \quad \text{where} \quad a = \min(x_1, \ldots,)$$

$$\text{If } \left| n \left(1 + \frac{t}{1-t}\right)^{-(n-1)} \right| \le 0.05$$

then X_n is an outlier and delete it. After deleting X_n, the measure of skewness S_k is re-computed and appropriate test for outlier(s) is applied and the outliers deleted. This process is continued until all the outliers are deleted.

Otherwise, the X_n is not considered as an outlier based on the assumed distribution the procedure is stopped.

ii.) Skewed to the left (outlier test 2)

The test used is for a single lower outlier in a Gamma distribution.

Let $t = \frac{X_1}{\sum_{i=1}^{n} X_i}$

If $\left| 1-n\,(1+\frac{t}{1-t})^{-(n-1)} \right| \leq 0.05$

then delete X_1. After deleting X_1, the measure of skewness S_k is re-computed and the appropriate outlier test selected. Continue this process until all the outliers are deleted. Otherwise, the X_1 is not considered as an outlier based on the assumed distribution and the procedure is stopped.

iii.) Not skewed (outlier test 3)

If the data are not skewed, a normal distribution is assumed and both the upper and lower values limits of the data must be tested to eliminate outliers.

a.) Test for Upper outlier X_n

Let $t = \frac{X_n - \overline{X}}{s}$ where $s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$

If t > 5 % level of significance $\alpha = 0.05$ (5 % point), then delete X_n, recompute skewness S_k, and follow the testing procedure. Otherwise, X_n is not considered as an outlier based on the assumed distribution and the procedure is stopped.

b.) Test for Lower outlier X_1

Let $t = \frac{X_1 - \overline{X}}{s}$

If $|t|$ > 5 % point level of significance, then delete X_1, recompute skewness S_k, and follow the testing procedure.

Otherwise, X_1 is not considered as an outlier based on the assumed distribution and the procedure is stopped.

Note: When the desired sample size n is not listed on the given table, linear interpolation can be applied to determine the critical value for a statistical testing.

4. Difference Test

After removing outliers we begin a methodology to define edit bounds in order to detect "suspicious" data values that may remain. To define bounds we use a "difference" approach. This approach is an attempt to mathematically replicate the process often used by analyst of examining the data distributions for "unnatural" breaks in the data.

We experimented with additive and proportional differences (intervals) between adjacent observations. We also looked at the additive difference of the moving average of items *i* through *j* and items *i+1* through *j+1*. After producing sample distributions we decided the proportional difference would work best. To illustrate the proportional difference, if the smallest two observations (say, $x_{[1]}$, $x_{[2]}$) were 0.5 and 0.8 respectively, their proportional interval would be 0.8 / 0.5 = 1.6.

For this application it was necessary to determine a **cutoff** which all intervals larger than can be identified as a possible location for an outlier boundary. It made sense to base the cutoff on some characteristic of all the distribution intervals. That would make the cutoff more indicative of outlier boundary locations. We looked at both the mean and median proportional differences (intervals) and the median appeared to be the better choice. A **cutoff factor** can be chosen to multiply by the median interval to obtain the interval size cutoff. Ideally, the cutoff factor will help us target those intervals we wish to examine. The cutoff level depends on the level of stringency desired in identifying outliers.

Using several distribution samples, we found that most (about 75% to 95%) of the intervals were smaller than 1.2 times the median difference. We felt using this level would result in the desired bounds for the edit. So we used 1.2 as our cutoff factor. However, after review the cutoff factor was adjusted for some data items.

We wanted to limit the selection of outlier boundaries to the tails of distributions. We agreed that for a given sample size the outlier tier could exceed not a given percentile. This controls the maximum number of cases that are identified by the edit for review. The percentile are set to limit the number of cases to be reviewed.

If more than one outlier boundary could be drawn the innermost one was used. The lower and upper limits for the ratio were set to the lowest and highest observations inside the boundaries. If no intervals exceeded the cutoff, the limit was set to the most extreme observation. The limits were then padded. We multiplied the upper limit by the median interval and divided the lower limit by the median to widen the acceptance bounds by a small amount to allow for leniency.

5. Mean vs Median

Finally, SPEER requires central values of ratios in the edit imputation process. In prior surveys, the mean was used. However, considering the number of cells with cases with fewer than 25 observations and the skewness of data, a question arose as to whether using the mean indiscriminately as a central value is appropriate. It was determined that median was a better estimate of what we considered the central tendency, since the ratio distributions were heavily skewed.

6. Concluding Remarks

The methodology presented in this paper proved to be a reliable, flexible, and efficient system for setting lower and upper bounds for editing the 1992 Economic Census Enterprise Data. The methodology is applicable for determining edit bounds and subsequently "suspicious" values in data irrespective of the editing system. The two methods for identifying outliers and suspicious data values can be used either together as was the case here or separately depending on the application. The use of a flexible cutoff factor for the difference test permits the data analyst to adjust the cutoffs as needed. The methodology is not dependent on distribution assumptions in the data. Furthermore, these methods should be applicable to the analysis of other continuous data sets for identification of suspicious data values.

Acknowledgement

The authors wish to Phil Thompson, Donna Hambric, and Barbara Boney of Economic Census and Surveys Division for their valuable input regarding Enterprise Statistics processing and Bill Winkler, Tom Petkunas of Statistical Research Division for their valuable comments in the process of developing these methods, and Bob Ormsby of Computer Sciences Corporation, Professional Services Group for his in-depth assistance in SAS programming. We also thank Brian Richards of Economic Statistical Methods Division and Brian Greenberg of Industry Division for their review and comments.

References

1. Barnett, V and Lewis, T., (1978), Outliers in Statistical Data, John Wiley & Sons, New York

2. Draper, L., Petkunas T., and Greenberg, B. (1990) "On-Line Capabilities in SPEER", Statistics Canada Symposium 90

3. Greenberg B. and Petkunas T. (1990) Overview of the SPEER System (Structured Programs for Economic Editing and Referrals). Bureau of the Census, Statistical Research Division Report Series

4. Greenberg, B. and Surdi, R (1990) "SPEER(Structured Programs for Economic Editing and Referrals)." Proceedings of Section on Survey Research Methods, American Statistical Association, 421-426

5. Grubbs, F and Beck, G., (1972), Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations, Vol. 14, No. 6, Technometrics

6. Johnson, R., (1992), Applied Multivariate Statistical Analysis, Prentice Hall, Englewood, NJ

7. Paletz, D. (1993),Documentation of Distance Measurement Algorithm for Selection of Outliers (D-

MASO), Internal Documentation, Bureau of the Census

8. Paletz, D. and Winkler, B. (1994), Generalized D-MASO Software Documentation, Internal Documentation, Bureau of the Census

9. SAS Institute, (1990), SAS/STAT User's Guide

10.Snedecor and Cochran, (1967), Statistical Methods, Iowa State University Press, Ames, Iowa

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessary reflect those of the Census Bureau. Please address correspondence to: Eddie Salyers, Economic Census and Surveys Divison, FOB 2553-3, Washington, D.C. 20233