IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

Julia L. Bienias, David M. Lassman, Scott A. Scheleur, and Howard Hogan, U.S. Bureau of the Census
Julia L. Bienias, U. S. Bureau of the Census, SRD FOB 3000-4, Washington, DC 20233

Editing is an important step in producing estimates from survey data. In many settings, trained analysts examine the data to find unusual or unexpected values, which may be the result of errors made by the respondent or in the data-capture processes. Accurately identifying the values most likely to be in error is an essential part of efficient editing.

Graphical methods have been used to improve the efficiency and accuracy of the editing process (e.g., Esposito, Fox, Lin, & Tidemann, in press; Granquist, 1990; Houston & Bruce, 1992; Hughes, McDermid, & Linacre, 1990). We describe the application of graphical methods from exploratory data analysis to identifying potentially incorrect data points.

2. Two Surveys

The Annual Survey of Communication Services

The Annual Survey of Communication Services (ASCS) is a mail survey (with telephone follow-) covering all employer firms that are primarily engaged in providing point-to-point communication services (e.g., telephone, television, radio), as defined in Major Group 48 of the 1987 edition of the *Standard Industrial Classification Manual*. The ASCS provides detailed revenue and expense statistics from a sample of approximately 2,000. (See U.S. Bureau of the Census, 1992.)

The Monthly Wholesale Trade Survey

The scope of the Monthly Wholesale Trade Survey (MWTS) is all employer firms engaged in wholesale trade, as defined by Major Groups 50 and 51 of the 1987 edition of the *Standard Industrial Classification Manual*. It is a mail survey (with telephone follow-up) of approximately 7,000 firms, of which 3,500 receive forms in a given month (U.S. Bureau of the Census, 1994.)

3. Current Procedures: Issues

Analysts review cases that have failed completeness, internal consistency, or tolerance edits through an interactive correction system or a paper listing. They also use a database query system to try to find problem cases that have not already been identified.

Unfortunately, examining one case at a time does not permit the analyst to obtain a broad view of the behavior of the industry as a whole. Such a view can be of great benefit in determining the impact of an individual unit on the aggregate estimate. In addition, it undoubtedly leads analysts to examine more cases than necessary. Finally, for a few of the ASCS tolerance edits, constant parameter levels derived from previous surveys have been hard-coded into the programs. This implicitly assumes the relations among the variables are static over time, which may not be the case.

4. Exploratory Data Analysis

4.1 Background

Exploratory data analysis (EDA) can be described as "a set of tools for finding what we might have otherwise missed" in data (see Tukey, 1977). These tools fit well in the survey processing environment.

EDA emphasizes displaying and fitting data using methods that are relatively insensitive to the presence of outliers in the data ("resistant" methods). This is particularly valuable during editing, when we expect "wild" observations. In addition, they allow for efficient examination of large amounts of information, an aspect that is particularly valuable in the time- and resource-constrained survey production environment.

We found the following particularly useful, and they are illustrated here: univariate boxplots, bivariate scatter plots, bivariate fitting, and transformations.

This paper reports general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

We thank the other members of our working group, without whom the work described here would not have been possible: Thomas Bell, Willard Caldwell, Vicki Garrett, Imelda Hall, Donald Hundertmark, Jennifer Juzwiak, William Knowlton, David Stachurski, and Georgeann Wright. We also acknowledge the many other members of Business Division and of Economic Statistical Methods and Programming Division who have supported this effort and who continue to support it.

4.2 Boxplots

Boxplots allow quick visual analysis of the location, spread, and shape of a distribution. We defined potential outliers as being beyond the whiskers, which were based on a length 1.5 times the interquartile range (see Tukey, 1977; Hoaglin, Mosteller, and Tukey, 1983).

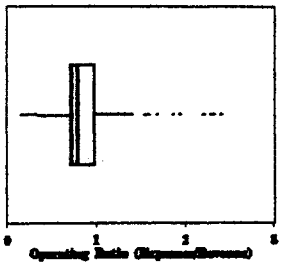

Figure 1

Figure 1 demonstrates the use of the boxplot for operating ratio (expenses/revenue) data from the ASCS.[1] The median operating ratio is .7978 and fifty percent of the points lie between .7269 and .9811. The left and right whisker values are .3760 and 1.3401. The cases flagged by the use of the boxplot are different (and fewer in number) from the cases that would have been flagged by the current hard-coded edit parameters, .9 and 1.1. Those parameters fail to help us isolate the "true" outlier cases, as they result in too many cases being flagged. Note the whisker-based bounds are not symmetric around one (consistent with the findings of Granquist, 1990).

4.3 Scatter Plots

A scatter plot of two variables is a simple and particularly useful technique. When the data are appropriately transformed, one can use a variety of methods to remove a linear relation from the scatter and then examine the residuals from the linear fit. Looking at the residuals from a fit allows us to examine the data on a finer scale (see Section 4.5).

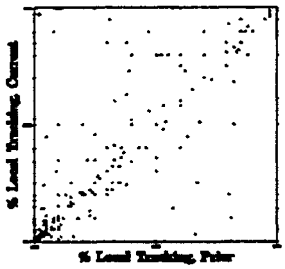

Figure 2

As a vivid illustration of the kinds of problems encountered in editing data, we used another survey for which we had raw responses to a particularly problematic question. One item in the Motor Freight Transportation and Warehousing Survey is the percent of revenue derived from local trucking, a question believed to be confusing to respondents may define "local" in different ways.

Figure 2, a scatter plot of these unedited data for the current versus prior period, shows a weak linear relation. Cases along the 45° line are companies whose year-to-year reports are consistent; reports are more inconsistent the further they are from the 45° line. Some of the cases along the vertical axis are "births" to the survey (cases selected during the current period to reflect new firms). Births should be analyzed separately, because they have only current-year data.

4.4 Transformations

Transforming the data so patterns can be more easily discerned is important to all graphical and data-fitting methods. It is used to obtain symmetry in the data, to promote linearity, and to equalize spreads between data sets. These properties are assumed, implicitly or explicitly, by many of the techniques we use to analyze data (e.g., boxplots). Economic data are typically skewed, and we want to spend our time investigating those data points that are particularly unusual, given that we expect many points far from the bulk of the data. For these data, transformations that lead to the expansion of lower data values and to shrinking the spread of larger data values are particularly useful (Hoaglin, Mosteller, & Tukey, 1983.)

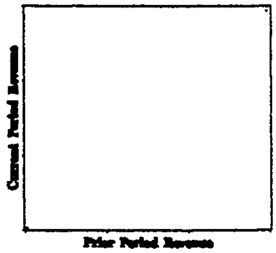

Figure 3a

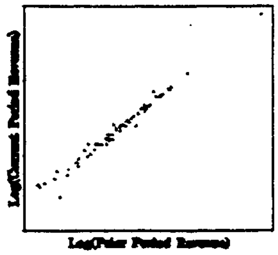

Figure 3b

Figure 3 is a an example of the use of transformations for the ASCS. The scatter plot of untransformed revenue data (Fig. 3a) reveals little, as one case is many times larger than the other cases. Hiding the large case was unsuccessful, as the next largest case was still many times larger than the remaining cases. Instead, taking logs of the data showed a useful scatter plot (Fig. 3b) and a strong linear relation, which is what we expect for a plot of current and prior data. Cases that do not appear to be following this linear relation would thus be considered unusual. We now see the case that appeared to be an outlier in Figure 3a is in line with the rest of the data.

In a different example (not shown), we found the fourth root to be the best transformation for a scatter plot of current MWTS inventory data against current sales data. A log transformation, in contrast, created a negative skew in the data ("over"-transforming them).

4.5 Fitting

We describe two approaches to fitting, ordinary bivariate linear regression and resistant regression. The goal of both methods is to remove a linear relation in the data so the analyst can concentrate on other aspects (Hoaglin, Mosteller, & Tukey, 1983; Velleman & Hoaglin, 1981).

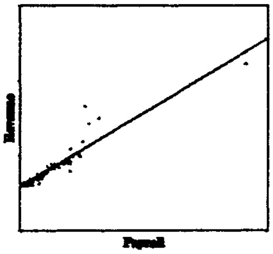

Figure 4a

Figure 4a shows the ordinary least squares regression (OLS) of ASCS revenue on payroll; many points are clustered near the origin and two cases are in the upper right corner. Removing the two large cases led to new outlying points. Such an iterative approach has the disadvantage of being subjective and of essentially requiring analysts to identify outliers first.

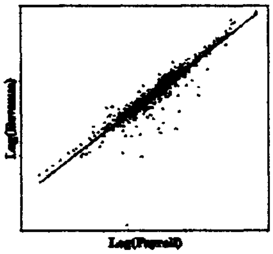

Figure 4b

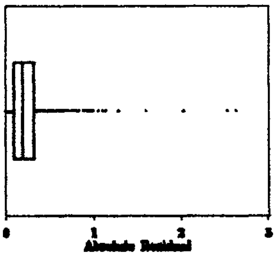

Figure 4c

Alternatively, we used OLS on log-transformed data (Fig. 4b), and the result is strongly linear. An obvious outlier can be seen near the bottom center. A pattern seen in the residuals revealed a pattern was due to the inclusion of tax-exempt cases. Tax-exempt cases should be examined separately from taxable cases, because our revenue item only includes taxable receipts. Removing both those cases and a single outlier near the bottom of the graph and refitting the data led to the distribution of absolute residuals shown in Figure 4c. This plot can be used to detect outliers, as with a cutoff level $C = K * $ (median absolute residual). (We found $K=4$ identified "true" outliers.)

Unfortunately, outliers may have great influence in an OLS fit, and thus it may not always work as well as in our example. As an alternative, we investigated resistant fitting using the biweight function developed by Tukey (Mosteller & Tukey 1977; McNeil, 1977). This widely-tested iterative weighted-least-squares fitting procedure is efficient; we reached our prespecified tolerance in fewer than 5 iterations.

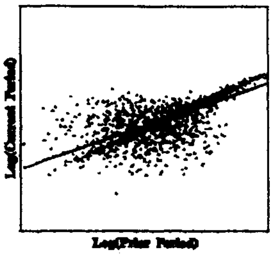

Figure 5

We applied resistant regression to the MWTS, fitting current logged inventory data to logged inventory data from the prior year (using $c=4$ and a tolerance of .01; see McNeil, 1977). Figure 5 shows the data and the line from the OLS fit. It is easy to see the OLS fit missed the central tendency of the point cloud, and resistant regression (Fig. 6) more effectively removed the linearity from the data.

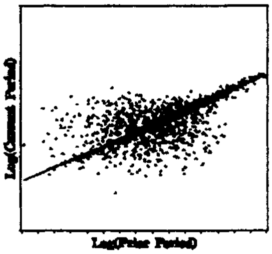

Figure 6

5. Summary and Extensions

Principles and methods from EDA can be used to improve the efficiency and accuracy of editing, by helping analysts see patterns in the data and use that information to choose cases for follow-up. Building a successful editing system using this approach is more than just selecting the correct statistical tools. The system must be acceptable to the people who will use it. Creating such acceptance requires training the analysts in the methods described here, as well as incorporating the tools into the current production environment and existing computer systems. To date, we have been successful in getting many people to try the methods on several surveys.

These techniques can be applied to raw or weighted data; applying them to the latter allows the analyst to ascertain the effect of a given case on the estimate. Also, they can be used as the basis for batch-based edit parameters, if necessary.

Note

[1] To protect the confidentiality of our data, we have not provided details about the particular subset of data analyzed in each plot, nor have we labeled axes when such information could be revealing.

References

Esposito, R., Fox, J. K., Lin, D., & Tidemann, K. (in press). ARIES: A visual path in the investigation of statistical data. *Computational and Graphical Statistics*.

Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, pp. 225-34. Ottawa: Statistics Canada.

Hoaglin, D.C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*. NY: Wiley.

Houston, G., & Bruce, A. G. (1992, February). Graphical editing for business and economic surveys. Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch.

Hughes, P.J., McDermid, I., & Linacre, S. J. (1990). The use of graphical methods in editing (with discussion). *Proceedings of the 1990 Bureau of the Census Annual Research Conference*, pp. 538-54. Washington, DC: U.S. Department of Commerce.

Lee, H. (in press). Outliers in survey sampling. In B. Cox et al. (Eds.), *Survey Methods for Business, Farms, and Institutions*. NY: Wiley.

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*. Reading, MA: Addison Wesley.

McNeil, D. R. (1977). *Interactive Data Analysis*. NY: Wiley.

Office of Management and Budget. (1987). *Standard Industrial Classification Manual*. Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012).

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

U.S. Bureau of the Census. (1992). *Annual Survey of Communication Services: 1992* . Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BC/92).

U.S. Bureau of the Census. (1994, April). *Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993*. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BW/93-RV).

Velleman, P.F., & Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.