HOW EXPLORATORY DATA ANALYSIS IS IMPROVING THE WAY WE COLLECT BUSINESS STATISTICS

Howard Hogan, Census Bureau
Howard Hogan, Bureau of the Census, Services Division, Washington, DC 20233*

Key Words: Resistant Regression, Re-expression, Editing, Sample Selection

1. Introduction

It has been twenty five years since John Tukey published the preliminary limited edition of *Exploratory Data Analysis.* All the standard texts have been out for ten or more years: Tukey (1977), Mosteller and Tukey (1977), Mosteller, Hoaglin, and Tukey (1983, 1985). So how is it that these important statistical ideas are only now being used to improve the way we collect business statistics, and why has it taken so long?

Let me give some background. I have been interested in Exploratory Data Analysis (EDA) for twenty years. I have taught an EDA course for more than ten years at the U.S. Department of Agriculture Evening School and have given occasional lectures at the U.S. Census Bureau and elsewhere, including Statistics Sweden. Typically, the reaction from statisticians at government statistical offices has been unfavorable. EDA, they felt, was unsuitable for the production of official government statistics. EDA was too informal, too intuitive, too subjective for the needs of government. End-user acceptance was posed as a hurdle. "Are we going to publish boxplots?" was typically the reaction.

The results from EDA may sometimes be unsuitable for publication as official statistics. However, as this paper will explain, EDA has proven effective in improving the production of those "official statistics."

Two years ago, when I joined the economic statistics area, there was almost no use of explicit exploratory data analysis techniques in the Census Bureau's establishment surveys. By EDA techniques, I mean techniques that explicitly use the "4 R's," resistance, re-expression, residuals, and graphical revelation. Rather, much of the work was non-graphical. It relied on "exceptions lists" of the top 20 cases or 5 percent. It relied on ordinary least squares, sometimes refit after eliminating outliers. Even systematic analysis of residuals was not routine.

At the time, we identified three obstacles to the use of EDA methods by those producing trade and transportation statistics.

* Lack of high-resolution graphics hardware.
* Lack of EDA software.
* Lack of training.

We are overcoming these obstacles in a number of ways. Those who are interested are beginning to have access to the needed equipment, software, and tools. We have successfully introduced the method in a number of situations. On the other hand, we have yet to make EDA a routine part of the corporate culture. The work of many staffs that collect economic data has been little changed by our efforts.

I will discuss how we introduced EDA into four areas:

i) Annual Survey of Communication Services
ii) Commodity Flow Survey
iii) Business Sample Revision
iv) Construction Price Index

The techniques we introduced were certainly not new, and can be considered standard. Thus my focus will not be on the methodology. For that the reader is directed to the references given in paragraph one. Nor am I interested in the substantive results of the application of EDA to a particular problem. Instead, my interest is in the organizational barrier to introducing EDA, the cost effectiveness of these techniques, and the reaction of survey managers and subject matter analysts.

2. Annual Survey of Communication Services

EDA found its most enthusiastic welcome in the Annual Survey of Communication Services. As the title suggests, this is an annual survey. This means

that there is time to analyze the current results as well as an opportunity to learn from previous experience. Further, the data sets are not overly large. The survey collects approximately 30 variables on 2000 firms. It covers telephone, radio, television broadcasting, cable television, and other related activities.

Previously, analysts reviewed cases that failed simple computer edits such as range checks. They also reviewed the top 20 cases, as defined for example, by absolute size or absolute change from last year. This approach seemed to catch most of the big errors.

The goal of introducing EDA was to make the process more efficient, i.e., to save time and free the staff time from data editing. The techniques were simple log transformations followed by either box plots or scatter plots. Most of the work was done within SAS and SAS/INSIGHT.

Figure 1 shows the kind of displays the analysts now use. This is the kind of graphs that the analysts work with, it is not altered to achieve publication quality. It presents boxplots in log terms for the operating expense ratio for tax exempt TV and radio broadcasting. Within the interactive tools, we are able to click on the outliers and quickly identify the firm. These outliers are typically reviewed by the analysts. Analysts also look at plots log-revenue against log expenses. Often, analysts find non-linear relations due to the inclusion of new firms or tax exempt firms in the data set. Removing the linear relation lets us look at the residuals, and identify outliers.

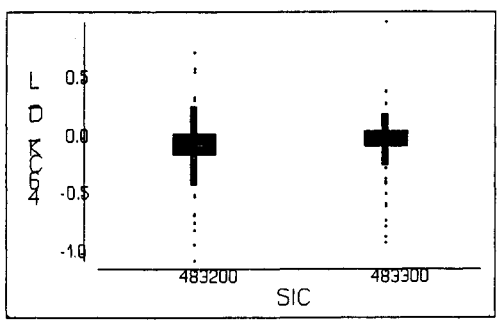

Figure 1. Boxplot of log expenses minus log revenue for tax exempt TV (483300) and RADIO (483200)

Several advantages of the new approach are already evident. Analysts like to see the "big picture." It lets them see how important the largest differences are relative to other differences. A few, but not all, analysts have given up reviewing the "top 20," and now review only statistical outliers. One dramatic improvement occurred when the Branch plotted the data for one survey relative to the edit cutoffs. It turned out that the median fell outside the cutoffs! The edits were quickly reset with immediate gains in efficiency. Of course, the staff might have discovered the problem without EDA. However, the fact is that the edits had been in place for two years without the problem being noticed.

Other benefits of the new approach are noticeable. There has been a modest improvement in data quality as a few more erroneous values have been identified. However, a real benefit is the ability to review less good data and concentrate more on actual problems. The 1993 annual survey finished one month sooner than it did the previous year. This was accomplished with fewer staff (3 vs 4). The analysts are now analyzing more data and spending less time editing forms. Because of this they can now refocus on being experts in these industries. The number of questionnaires selected for more detailed analyst review was reduced from near 2000 to around 500. Given this success, the EDA approach will now be applied to the Service Annual Survey, and the Transportation Annual Survey.

Key to the success has been a supportive branch chief and one very enthusiastic analyst (David Lassman) who knew SAS and learned EDA. The branch chief arranged for the whole branch to take a special 12 classroom-hour short course on EDA and graphical methods. This ensured a common basis and vocabulary for the staff. The SAS resource person built a simple menu within SAS to call up the data and perform a few simple analyses. The staff also has a high percentage of recent college graduates who have majors or minors in mathematics and training in computers. However, personal outlook seemed more important than math background.

The staff used SAS because most of the data were on computing platforms readily accessible via SAS. We acquired used X-terminals to utilize the SAS/INSIGHT package. Analysts particularly like the ability to "click" upon individual cases to retrieve the identifier. The drawback of INSIGHT

(Version 6.08) is its inability to do more than the most simple transformations. Even taking absolute value requires a trick. Re-transforming the predicted results to the original scale requires leaving INSIGHT. Furthermore, we had to write our resistant regression program within SAS but outside the SAS/INSIGHT module--thus requiring more exiting and entering.

The success of EDA in the annual surveys can be contrasted with our experience with the monthly surveys. Each month the Census bureau publishes monthly survey on retail trade, wholesale trade and inventories. We have had great difficulty introducing either EDA techniques or EDA approaches into these processes.

A major barrier seems to be that the press of the monthly cycles leaves little time for investment into new approaches. Currently, the data must be retrieved across VAX clusters from a DEC/Rdb Database. Pulling from multiple tables across clusters or even to another platform is very time consuming. It takes approximately 1/2 day to read the data, with only 2 or 3 days to edit the data prior to the next cycle, it is not a feasible option at this time. Annuals, by comparison, have 3-4 weeks between data cycles and thus are much more flexible. On monthly surveys, the time simply is not available. We have various plans to develop a menu-driven system, but they remain plans, since we are always pressed to complete our mandated surveys on time.

3. Commodity Flow Survey

The Commodity Flow Survey (CFS) attempts to measure the volume and direction of shipments with the U.S. The sampling unit is the establishment. The unit of analysis is the shipment, which can be anything from a 100-car train shipment of coal to one pair of socks air-mailed from a mail order catalog. Nearly 200,000 establishments were in the sample. They reported on each of nearly 17 million sample shipments.

The 1993 CFS was a new survey. This gave us a chance to use EDA techniques without any "this is the way we have always done it." Our first task was to edit the data files. We had no previous data to base edits on. Further, we had a huge data set. We wanted to perform edits on each of 1200 Standard Transportation Commodity Code groups (STCC's). We needed something that was quick, simple, and robust.

Some of our early work was graphical. (See Dembroski 1994.) However, graphical analysis, STCC by STCC was too time consuming. Instead, we chose to look at:

$$\log_{10} (\text{Weight/Value}),$$

Where

Weight was the reported weight of each shipment, supposedly in pounds.
Value was the reported value, in dollars.

We used the robust outlier cutoffs:

$$F_u + c\ \text{IFR}$$
$$F_l - c\ \text{IFR},$$

where

F_u: Upper Fourth (Quantile)
F_l: Lower Fourth
IFR: Inner Fourth Range: $F_u - F_l$
c: scaling constant set at either 2 or 3.

This robust edit proved quick and useful in weeding out mis-reported units of weight. The distribution of many of the STCC's proved to be bi- or even multi-modal, as many shipments were reported properly in pounds, but other shipments were reported in tons, gallons, etc. One STCC had over 3000 outliers out of 25,000 shipments, with one half the outliers being beyond 3 IFR. Another STCC had 5,800 outliers out of 75,000 records, with 4,000 records being beyond 3 IFR. Recall that in the Gaussian distribution, one would expect fewer than one case in a thousand to lie beyond 3 IFR of the fourths.

The CFS illustrates an important aspect of what would hinder the use of exploratory data analysis in a production environment. Although we successfully implemented EDA techniques, we had a harder time encouraging a true EDA approach. The "traditional" data editing approach had been "Do a, b, and c, and you are done." With true exploratory data analysis, the approach is "Do a and decide whether to do b, c, or d, which might lead you to think about what is really going on." The way you proceed depends on what you find. Many of the people involved with CFS were far more comfortable with the traditional linear processing

approach than with the less structured world of EDA.

Because the CFS was not an ongoing survey, we were not able to develop a structured step-by-step analysis that anyone could then follow. Instead, the true EDA work fell to a few, with the rest of the staff reverting to approaches with which they were comfortable.

4. Business Sample Revision: BSR-97

Sampling may seem an odd field for EDA, but in fact, it has been one of our most successful applications. Every five years, the Census Bureau draws a new sample for its surveys of wholesale trade, retail trade, and service industries. Large firms are selected with certainty (probability one) based on their sales and inventory, as measured by the latest economic census, in this case the 1992 Census. For other firms, we draw a stratified sample with probability based on their measured size. Rather than using the latest census, the size measures were updated from administrative records from 1994. These records give us payroll rather than the desired sales and inventory.

To convert payroll to sales or inventory, we fit a no-intercept linear regression using 1992 Census data. We then used this fit to predict 1994 sales and inventory.

In previous sample revisions, we have used ordinary least squares (OLS). Outliers were defined as points with residuals of more than three standard deviations. These were excluded and the regression was run again. The new coefficient was used if it differed from the old by more than ten percent.

Resistant (biweight) regression has both theoretical and practical importance. First, it allows for partially down-weighting large values. Before, residuals just smaller than three standard deviations were fully accepted, those greater were completely rejected. With the biweight, the discontinuity is avoided. Second, several iterations are run automatically to fit the model. The earlier method only allowed for two passes.

The practical gains were also important. The survey designers have more confidence in the results. This meant that they had to spend less time checking over the fits. This is important because we had to fit 125 models for wholesale, 250 for retail and 500 for services.

Subject matter specialists felt that the new coefficients were more reasonable, especially for wholesale where the relation between payroll and sales is looser than in the other two trade areas. In general, staff felt that the new approach represented a great improvement.

One reason that this application was such a success is that all work was done by staff trained in mathematical statistics, as opposed to those with a more subject matter oriented background. Their training allowed them to see the weakness in the previous method and adapt quickly to the new approach.

5. Price Index of New Single-Family Houses Under Construction

Most of our applications of EDA techniques are for internal use only. However, in at least one case, we have been able to incorporate results directly in our publications.

The Price Index of New Single-Family Houses Under Construction is based on data from the Housing Starts, Sales, and Completions Survey. It is computed monthly and published with the Value of New Construction Put in Place series. The index uses a "hedonic" or regression methodology.

The price of a "fixed market basket" of housing characteristics is calculated for the current period and a base year. The Laspeyres Price Index uses a market basket developed from base year housing characteristics. The Paasche Index uses the current market basket. The price index is a measure of the change in price of the market basket from the base year to the current period.

To develop the index, a linear regression is fit. The response variable is the logarithm of value of construction, that is, total price minus value of the lot. The explanatory variables are a list of housing characteristics, such as logarithm of floor area, as well as dummy variables for number of bedrooms, number of bathrooms, etc. The model is fit for the base year and current year. Thus, the Laspeyres index is compiled as:

$$L_t = \frac{\text{antilog}\{\sum_i b_i(t)\ Q_i(\emptyset)\}}{\text{antilog}\{\sum_i b_i(\emptyset)\ Q_i(\emptyset)\}} \times 100$$

the Paasche index is:

$$P_t = \frac{\text{antilog}\{\sum_i b_i(t)\ Q_i(t)\}}{\text{antilog}\{\sum_i b_i(\emptyset)\ Q_i(t)\}} \times 100$$

where

$b_i(t)$ are the regression coefficients for the current period,
$b_i(\emptyset)$ are the coefficients for the base period,
$Q_i(t)$ are the explanatory variables for the current period,
$Q_i(\emptyset)$ are the explanatory variables for the base period.

Thus, the regression model is used to answer the question of how much housing construction of a fixed quality would cost. (See Luery, 1990.)

Even in log-terms, the regressions proved sensitive to large, expensive, and in some sense, unusual houses. The practice had been to exclude houses based on high construction value before doing the final fit.

Recently, we have switched to resistant biweight regression. Those doing the work have been pleased. It is less arbitrary and agrees with their judgment. They do not need to be constantly changing the cutoffs. More important, the chief customer, Bureau of Economic Analysis, is pleased. In their report back to us, they said:

"The resistant regression technique appears to be an improvement in procedures because it allows for the appropriate inclusion of extreme observations that were formerly excluded from index calculations." (Donahoe, 1995)

This final example disproves the last of the objections cited at the beginning--the belief that EDA techniques are never appropriate for preparing official statistics. The right technique for the right application can be quite effective.

7. Conclusions

It seems clear that the introduction of exploratory data analysis has improved the way we collect, edit, and publish business statistics. EDA methods have helped us detect outliers more efficiently. The result is better data, more quickly and with less cost. Simple re-expression by logs or occasionally roots have made the work easier. Similarly, resistant regression has proved effective in a number of applications. The resulting fits have helped us draw samples. Finally, at least in one instance, the results have been incorporated directly in publication, with apparent customer satisfaction.

To achieve this success, several barriers had to be overcome. We had to acquire new hardware for the staffs that manage the surveys. We also had to acquire and develop new software. Finally, we had to provide training and select appropriate EDA techniques for our work. The result of the effort so far has shown great promise.

However, barriers remain. A major barrier is simply the difficulty and delay of getting the data onto the same computer platform as the software. In the case of monthly surveys, this problem has stopped almost all progress. There is no doubt that this obstacle can, and eventually, will be removed. However, it will require programming resources, currently in short supply.

Similarly, our success so far has come mainly with people who are comfortable in a less structured environment. Not all methodologists have quickly adopted the new tools. Primarily, relatively recent college graduates, comfortable with new software, have accepted the challenge. Deadlines are a key factor. Those working on annual or five year projects have had the time to learn and see that learning applied. Monthly surveys, with tight deadlines, have proved harder.

EDA, by its nature, cannot be fully structured. Each analysis should raise new questions calling for new analyses. It is unrealistic to believe that many production- oriented people will ever operate in this environment.

Rather, I would see us moving to a multi-tiered system. There would be a few analysts conducting true exploratory data analysis. They would re-express, fit models, and examine residuals. More importantly, over the long term they would develop standard models and methods for "second level" analysts to use.

At another level, some analysts would be trained on a limited number of EDA methods. They would use boxplots and residual plots. They might use re-expression, but only to logs. They might use a computerized menu system with limited options. Since to do this, they would need graphics terminals and links to the data, and these are currently limited, these analysts would not be able to do all the editing.

The third level would be analysts and clerks who would use the results of EDA. Cases for review might be based on a system that, say, re-expressed the data in logarithms, fit a resistant model, and selected all residuals greater than three midspreads from zero. However, the clerks would only need to see a listing of cases for review. It is clear that this will take programming resources to develop. We are now beginning to incorporate EDA techniques into a generalized edit system. This process is at the very beginning. The lessons we have learned so far can help guide this process.

In all, we have made good progress. We have achieved some modest success. However, there is much to do to truly explore what EDA can do to improve our methods.

References

Bienias, Julia; Lassman, David; Scheleur, Scott, and Hogan, Howard; 1994, "Improving Outlier Detection in Two Established Surveys," *Proceedings of the Survey Research Methods Section*, American Statistical Association. Also published in *Proceedings of the Seminar on New Directions in Statistical Methodology.*

Caldwell, Carol V., 1994, "Regression Plans for BSR-97," Memorandum for the Record, Bureau of the Census, Washington, DC.

Dembroski, Bruce A., 1994, "SAS/Edit Analysis: Suggested Approach," Note to John L. Fowler, Bureau of the Census, Washington, DC.

Donohoe, Jerry, 1995, "Memorandum for Donald Luery: Construction Index Research at the Census Bureau," Bureau of Economic Analysis, Washington, DC.

Hoaglin, David C.; Mosteller, Frederick; and Tukey, John W.; 1983, *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York.

Hoaglin, David C.: Mosteller, Frederick; and Tukey, John W., 1985, *Exploring Data Tables, Trends, and Shapes*, John Wiley, New York.

Luery, Donald, 1990, "Price Indexes of Single-Family Houses Under Construction," Bureau of the Census, Washington, DC.

Tukey, John W., 1970, *Exploratory Data Analysis*, (Limited Preliminary Edition), Addison-Wesley, Reading, MA.

Tukey, John W., 1977, *EDA: Exploratory Data Analysis*, Addison-Wesley, MA.

* This article reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The research on which this paper reports is being conducted by David Lassman, Scott Scheleur, James Burton, Carol King, Carol Caldwell, Bruce Dembroski, Michael Kornbau, and others. The author is indebted to all these people for help in preparing this paper.